IBM® Information Management Software

IBM Information Governance Solutions

- New approaches for developing long lasting and trusted information assets
- A strong foundation of data policies and practices
- Scenarios to help guide your development

Chuck Ballard
John Baldwin
Alex Baryudin
Gary Brunell
Christopher Giardina
Marc Haber
Erik A O'neill
Sandeep Shah

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM Information Governance Solutions

April 2014

SG24-8164-00

Note: Before using this information and the product it supports, read the information in “Notices” on page vii.

First Edition (April 2014)

This edition applies to Version 9.1.04 of InfoSphere Optim, Version 11 of InfoSphere Master Data Management and Version 9.1.2 of InfoSphere Information Server, V9 of InfoSphere Guardium Data Activity Monitor, v2.5 of InfoSphere Guardium Data Redaction, v1.1 of InfoSphere Guardium Data Encryption Expert, v4.6 (with FixPak 4.6.2) of InfoSphere Discovery, and V6.0 of Altas eDiscovery and IT eDiscovery process Management Solutions.

Contents

Notices vii
Trademarks viii

Preface ix
Authors x
Now you can become a published author, too! xiv
Comments welcome xiv
Stay connected to IBM Redbooks xv

Chapter 1. Information Governance: foundations and solutions 1
1.1 Common themes and focus of this book 3
1.2 Thought leadership and innovation 3
1.3 Implementing information governance 4
1.3.1 Some final foundation governance thoughts 6
1.3.2 Understanding why implementing IG can be hard 7
1.4 The Information Governance Component Model 9
1.4.1 Developing your approach 11
1.4.2 Breaking down the IG Component Model value 13

Chapter 2. Information Governance organizational structures 19
2.1 Introduction to organizational structures 20
2.2 Introduction to information governance roles 20
2.3 Governance layers 22
2.3.1 Governing functions and roles 22
2.3.2 Governed functions and roles 24
2.4 Organizational structures and the communications model 27
2.4.1 Understanding the SCM 29
2.5 Application of data stewardship 37
2.5.1 Creating and managing data stewards 38
2.5.2 Applying data stewards to information assets 39

Chapter 3. Business definitions and policies in IBM InfoSphere Information Server 41
3.1 Introduction to business definitions 43
3.2 Introduction to business policies 43
3.3 InfoSphere Information Server 44
3.3.1 IBM InfoSphere Business Glossary 45
3.3.2 Business Glossary subject material and council 48
3.3.3 Managing Business Glossary categories and terms 49

3.3.4 Managing Business Glossary Policies and Rules 51
3.3.5 Managing Business Labels 54
3.4 Benefit and value 55

Chapter 4. Workflows and business specifications 57
4.1 Introduction to Workflows 59
4.2 Introduction to business specifications 60
4.3 InfoSphere Information Server 61
4.3.1 InfoSphere Blueprint Director 62
4.3.2 InfoSphere FastTrack 69
4.4 Benefit and value 72

Chapter 5. Metrics and measurements 75
5.1 Metrics and Measurements community 76
5.2 Required infrastructure 77
5.3 Capitalizing on your metrics and measurements 79
5.4 Scorecards and dashboards 84

Chapter 6. Business drivers for information governance 91
6.1 A vision of trust 92
6.2 Classic business drivers 92
6.3 Business reasons for information governance initiatives 93
6.3.1 Integration quality scenario 94
6.3.2 Information Lifecycle Management scenarios 95
6.3.3 Master Data Management scenario 97
6.3.4 Data Security scenario 98

Chapter 7. Data Quality 99
7.1 What is Data Quality? 100
7.2 Data quality scenario 102
7.2.1 Assess existing application and associated processes 102
7.2.2 Identify data owners: Data stewards 104
7.2.3 Perform data quality assessment 105
7.2.4 Work with data stewards to establish rules for data validation 107
7.2.5 Work with data stewards to establish match criteria 108
7.2.6 Perform lifecycle of application development, user, and acceptance testing 118
7.2.7 Design new and improve existing processes for managing de-duplication 119
7.2.8 Organization structures, roles, and responsibilities 120

Chapter 8. Information Lifecycle Management 121
8.1 What is ILM 122
8.2 ILM: Decommissioning 124

8.2.1 The high-level steps to enterprise decommissioning 125
8.2.2 Understand the information: Assess the application landscape . . . 126
8.2.3 Organization structures, roles, and responsibilities. 134
8.2.4 New decommissioning policies and processes. 145
8.2.5 Refine program and communicate 149
8.2.6 Implement technologies to support the new IG capability. 160
8.2.7 Recruit and enable resources 167
8.2.8 Decommission applications. 169
8.2.9 Build into an operational capability 177
8.2.10 Monitor, measure, and report 179

Chapter 9. Test data management 181
9.1 Provisioning. 184
9.2 Privatization. 185
9.3 Approach 186
9.4 Assessment. 187
9.5 Gap Analysis: Aligning capability and maturity 189
9.6 Organizational alignment. 190
9.7 TDM definitions, policies, and processes 192
9.8 Alignment of technology capabilities 198
9.9 Operationalizing your TDM/DP approach 201
9.10 Value-based measurements and monitoring. 202

Chapter 10. Master Data Management 205
10.1 What is Master Data Management 206
10.1.1 MDM functionality that can support information governance 208
10.2 MDM scenario - outsourcing and acquisition 213
10.2.1 Analysis. 215
10.2.2 Design. 217
10.2.3 Summary of information governance support 218

Chapter 11. Data protection and security scenario 221
11.1 Scope, objectives, and goal 222
11.2 Approach. 222
11.2.1 Discover your data 223
11.2.2 Identify and classify sensitive data 224
11.2.3 Secure and protect 224
11.2.4 Monitor and audit 234

Appendix A. Additional material 239
Locating the Web material 239
Using the Web material 240
Downloading and extracting the Web material 240

Related publications 241
IBM Redbooks 241
Other publications 241
Online resources 241
Help from IBM 242

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:
This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Redbooks (logo) ®	FileNet®	Optim™
AIX®	Global Business Services®	QualityStage®
CGOC™	Guardium®	Rational Team Concert™
Component Business Model™	IBM®	Rational®
DataStage®	InfoSphere®	Redbooks®

The following terms are trademarks of other companies:

Performance View, and Kenexa device are trademarks or registered trademarks of Kenexa, an IBM Company.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.

Preface

Managing information within the enterprise has always been a vital and important task to support the day-to-day business operations and to enable analysis of that data for decision making to better manage and grow the business for improved profitability. To do all that, clearly the data must be accurate and organized so it is accessible and understandable to all who need it.

That task has grown in importance as the volume of enterprise data has been growing significantly (analyst estimates of 40 - 50% growth per year are not uncommon) over the years. However, most of that data has been what we call “structured” data, which is the type that can fit neatly into rows and columns and be more easily analyzed. Now we are in the era of “big data.” This significantly increases the volume of data available, but it is in a form called “unstructured” data. That is, data from sources that are not as easily organized, such as data from emails, spreadsheets, sensors, video, audio, and social media sites. There is valuable information in all that data but it calls for new processes to enable it to be analyzed.

All this has brought with it a renewed and critical need to manage and organize that data with clarity of meaning, understandability, and interoperability. That is, you must be able to integrate this data when it is from within an enterprise but also importantly when it is from many different external sources.

What is described here has been and is being done to varying extents. It is called “*information governance*.” Governing this information however has proven to be challenging. But without governance, much of the data can be less useful and perhaps even used incorrectly, significantly impacting enterprise decision making. So we must also respect the needs for information security, consistency, and validity or else suffer the potential economic and legal consequences. Implementing sound governance practices needs to be an integral part of the information control in our organizations.

This IBM® Redbooks® publication focuses on the building blocks of a solid governance program. It examines some familiar governance initiative scenarios, identifying how they underpin key governance initiatives, such as Master Data Management, Quality Management, Security and Privacy, and Information Lifecycle Management. IBM Information Management and Governance solutions provide a comprehensive suite to help organizations better understand and build their governance solutions. The book also identifies new and innovative approaches that are developed by IBM practice leaders that can help as you implement the foundation capabilities in your organizations.

Authors

This book was produced by a team of specialists from around the world working with the IBM International Technical Support Organization, in San Jose, CA.

Chuck Ballard is a Project Manager at the International Technical Support Organization, in San Jose, California. He has over 35 years experience, holding positions in the areas of Product Engineering, Sales, Marketing, Technical Support, and Management. His expertise is in the areas of database, data management, data warehousing, business intelligence, and process re-engineering. He has written extensively on these subjects, taught classes, and presented at conferences and seminars worldwide. Chuck has both a Bachelor's degree and a Master's degree in Industrial Engineering from Purdue University.

John Baldwin has extensive experience in data modeling and database design. Since 2007, he has worked for IBM Software Group in the worldwide MDM Center of Excellence. As a Solution Architect, John has implemented many Master Data Management and Reference Data Management solutions, principally in the banking and insurance sectors. Before joining IBM, he worked in the Transportation and Logistics sector designing and managing real-time messaging and tracking systems; in the Broadcast Media sector building systems for capturing program metadata; and in the Public sector designing and building screen-based examination systems and implementing accounting and payroll packages.

Alex Baryudin is currently serving as a Technical Architect in Information Management Big Data COE. Alex is an IBM Senior Certified Professional and currently has over 18 years of experience in designing and implementing Data Quality projects that have included data warehousing, SOA, integration of IBM QualityStage® and IBM MDM Server, as well as other enterprise applications. Alex has created and contributed to best practices pertaining to data quality components to support IBM InfoSphere® suite of products.

Gary Brunell is a Data Security, Compliance, and Optimization leader for the IBM North America Information Governance Practice. In this role, he is responsible for: Architecture design and implementation of enterprise data and information lifecycle management, as well as related data privacy and governance solutions. Gary has over 30 years' experience in the design, development, and implementation of complex application and data management technologies. He has been involved with a number of leading-edge technology companies where he held various leadership positions in Research and Development, Sales, Marketing, Technical Support, Professional Services, and Management. Gary is an IBM Certified "Thought Leader" and was certified by the Open Group as a "Distinguished IT Specialist". He has authored many papers, holds a patent in automated software testing, and is a frequent speaker at industry events and executive briefings. Gary earned his engineering credentials from the University of Illinois.

Christopher Giardina is an IBM Worldwide Center of Excellence Data Security, Compliance, Optimization leader for Information Lifecycle Management (ILM) and Governance, with a focus on IBM InfoSphere Optim™ solutions. Chris is based in NE Pennsylvania and works with client and IBM/partner project teams implementing ILM around the world, everything from helping to scope and shape implementations, create, promote and teach best practice methods and enablement, and stepping in to troubleshoot or mitigate project or technical escalations. He has over 25 years of experience in business process and Information Management including database and Information Management systems design, development and implementation, solution implementation, and advanced technical account support. Chris is also an IBM Level 3 IT Certified Data Management Specialist, very active in the IBM IT Certification Specialist program. Chris is a certified Project Management Professional since 2003, and holds a Bachelor's degree in Management Science and an MBA degree in Production and Operations Information Management.

Marc Haber is the Product Architect for IBM InfoSphere Information Server, specializing in Business Glossary and Metadata Workbench. He joined IBM as part of the acquisition of Unicorn Software in 2006 and has worked in development, consultation, and product management roles. As Product Architect, Marc is responsible for working with customers to advance their understanding of the InfoSphere products and assist them in the planning, enablement, and rollout of metadata initiatives and strategies. He has authored and delivered previous IBM Redbooks publications, numerous presentations and Engineering Guides, as well as delivered training at conferences and individual customer locations.

Erik A O'neill is a Senior Product Manager on the IBM Master Data Management (MDM) team. He has held senior product management roles at IBM for over 10 years, with a primary focus on Information Management and MDM. Erik is currently leading the development of the IBM Reference Data Management Hub: An MDM domain and stewardship application specifically designed for management of enterprise reference data, and is an evangelist on the merits of an MDM approach for the governance of enterprise reference data. Before IBM, Erik worked in sales and product management roles for both start-ups and established enterprise software companies in the US, Europe, and SE Asia. Erik received his BA Degree in Experimental Physics and MA from Trinity College Dublin, Ireland and now lives in Northern California.

Sandeep Shah is an Accelerated Value Specialist in the IBM Software Group, in Houston, Texas. He has 10 years of IT consulting experience in customer relationship management and enablement, and designing and implementing robust database security solutions to improve business efficiency and functionality of Fortune 500 customers. His expertise is in the areas of database security, data management, data warehousing, and business intelligence. Sandeep holds a Bachelor's degree in Computer Engineering from Gujarat University in India and is currently pursuing his MBA at UCLA Anderson School of Management.

Other contributors

We want to thank others who contributed to this book, in the form of written content, subject expertise, and support.

A special thanks for significant contributions to the following

- Thomas Jesionowski: Sr. Managing Consultant - GBS Strategy and Analytics CoC, IBM Global Business Services®, Seattle, WA, US; Designer, Author, and Co-US Patent Holder - Structured Communications Model

- Scott Kimbleton: Sr. Managing Consultant - IBM Global Business Services Lead, Information Governance North America Lead, Global Information Governance Center of Excellence, FL, US; Designer and Author - Information Governance Component Model

From IBM Locations Worldwide

- Joe DiPietro: COE Leader for IBM Guardium®, IBM Software Group, Information Management, Littleton, MA, US
- Peter T Nguyen: Information Technology Consultant, Global Business Services, Southbank, Victoria, AU
- Yosef Rozenblit: WW Services Lead - InfoSphere Guardium, IBM Software Group, Information Management, Littleton, MA, US
- Rakesh Shah: Senior Product Manager, Information Management, IBM Software Group, Atlanta, GA US
- Kathryn Zeidenstein: InfoSphere Guardium Evangelist, IBM Software Group, Information Management, San Jose, CA, US

From the International Technical Support Organization

- Mary Comianos: Publications Management, San Jose, CA
- Ann Lund: Residency Administration, Poughkeepsie, NY

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Information Governance: foundations and solutions

Information governance (IG) refers to the disciplines, technologies, and solutions that are used to manage information within an enterprise in support of business and legal requirements. Information governance encompasses a broad set of subjects covering information quality, information protection, and information lifecycle management.

IBM has been a thought leader in information governance for many years. In 2004, IBM formed the IBM Data Governance Council as a leadership forum for organizations concerned with data governance issues. The IBM Data Governance Council has grown to over 55 member organizations that span the complete industry spectrum including companies, universities, and IBM Business Partners. The IBM Data Governance Council has established benchmarks and best practices for successful Data Governance. A key milestone in 2006 was the development of the Data Governance Maturity Model. In 2010, IBM changed the name of the Data Governance Council to IBM Information Governance Council, highlighting that information governance is a business concern, and not just a data concern.

IBM has also been instrumental in founding the Compliance, Governance, and Oversight Council IBM CGOC™ in 2004, which brings together over 1900 professionals with an interest in discovery, retention, privacy, and governance. In addition IBM has both public and internal Information Governance communities

that focus on collaboration, discussion, and development of Information Governance best practices. IBM Global Business Services (GBS) has an Information Governance Center of Excellence (COE) consisting of over 250 professionals supporting clients in multiple industries around the world. IBM GBS has deep expertise in the design, development, and deployment of information governance initiatives.

1.1 Common themes and focus of this book

While the specific focus and goals of the different information governance forums and communities may vary, there is a strong alignment on the basic principles and best practices. The Data Governance Maturity Model has become the leading model for framing the components of a governance program. Nevertheless, the broad scope and scale of information governance, and depth of complexity in each individual domain can make it difficult for an organization to bridge between the model and real world implementations.

This IBM Redbooks publication explores practical approaches to implementing information governance and shows how Enabling and Supporting Domains of the IBM Data Governance Council Maturity Model complement the Core Disciplines in real world scenarios:

- We place a stronger emphasis on the distinction between the Enabling and Supporting Domains versus the Core Disciplines
- We provide an in-depth analysis into the foundational capabilities that are associated with the Enabling and Supporting Domains
- We provide a series of “Core Scenarios” that demonstrate the dependencies of Core Disciplines upon these foundational capabilities

1.2 Thought leadership and innovation

Information governance is also a constantly evolving and maturing science in and of itself. This Redbooks publication introduces some thought-leading new concepts and approaches to information governance that have been developed by IBM to help bridge the gap between theory and implementation.

The IBM Information Governance Component Model (IGCM) provides an integrated model of people, process, and technology and breaks down the Information Governance Council Maturity Model disciplines into a matrix of executable non-overlapping implementable components. These components can be combined selectively to achieve a specific implementation and governance goal and the model offers techniques to select the right components to meet a particular business need.

The IBM Data Governance Structured Communication Model (SCM) is a patented model developed by IBM Global Business Services that aligns organizational structure and technology infrastructure. It describes a common set of functional roles required for data governance and provides a way to model the communication channels between these roles as repeatable processes. In turn,

the repeatable process can be used as the foundation to optimize the use of software tools that are used to manage data across the information supply chain.

A word on automation

In order to reach the next levels of maturity, organizations need better ways of automating the development and application of data governance solutions. This Redbooks publication describes how a number of IBM Solutions and software tools can be combined with the processes and models to provide a higher degree of automation. This includes the role of centralized governance repositories, how we can measure and monitor the success of our data governance activities using dashboards and reporting, and how software tools can help to standardize governance process flows.

Information governance is a developing science and as the nature of information used and managed within the enterprise changes, so the science and discipline of information governance must adapt in step. In particular, *big data* brings new requirements and concerns for information governance and a new set of technologies and processes. In view of its unique concerns, IBM is writing a companion Redbooks publication that focuses on the specific topic of information governance in the context of big data.

1.3 Implementing information governance

As described in the previous section, an important objective we sought to achieve with this book was to acknowledge that there is much discussion today around information governance and the many concepts and components that organizations must consider, enable, and deploy to make it work. At the same time, we acknowledge the sometimes overwhelming nature of the content that is associated with this subject.

In the balance of this book, we attempt to “break it down” a little more, beyond just the labels of Enabling and Supporting and Core disciplines, and add some detailed explanations and practical examples. In addition, we share some new, evolving work in IBM that is focused on helping organizations with planning their approach to shaping and building their information governance practices.

Key in this description is understanding and embracing the important distinction between the focus and efforts that are needed to build a governance program, versus the execution or operationalizing of governance type activity.

In the subsequent chapters of this book, we dive a little deeper into what we will be referring to as the *Foundation* type disciplines. In the Data Governance Council Maturity Model (DGCMM) (see Figure 1-1), though all disciplines are expected to be “matured”, in the early stages of any governance-focused initiative these foundation discipline areas are key to establishing accountability and baseline information asset definitions. Metadata management and centralized glossaries of all information asset definitions and policies are the typical manifestations of those foundation disciplines, but the work and efforts must extend beyond that. In the well documented Data Governance Council Maturity Model these are captured in the bands *other than* the Core disciplines, and they make up the building blocks or unifying governance principles and practices that will support any specific governance-related initiatives. These are very important for the “establishing a Governance Program” objective.

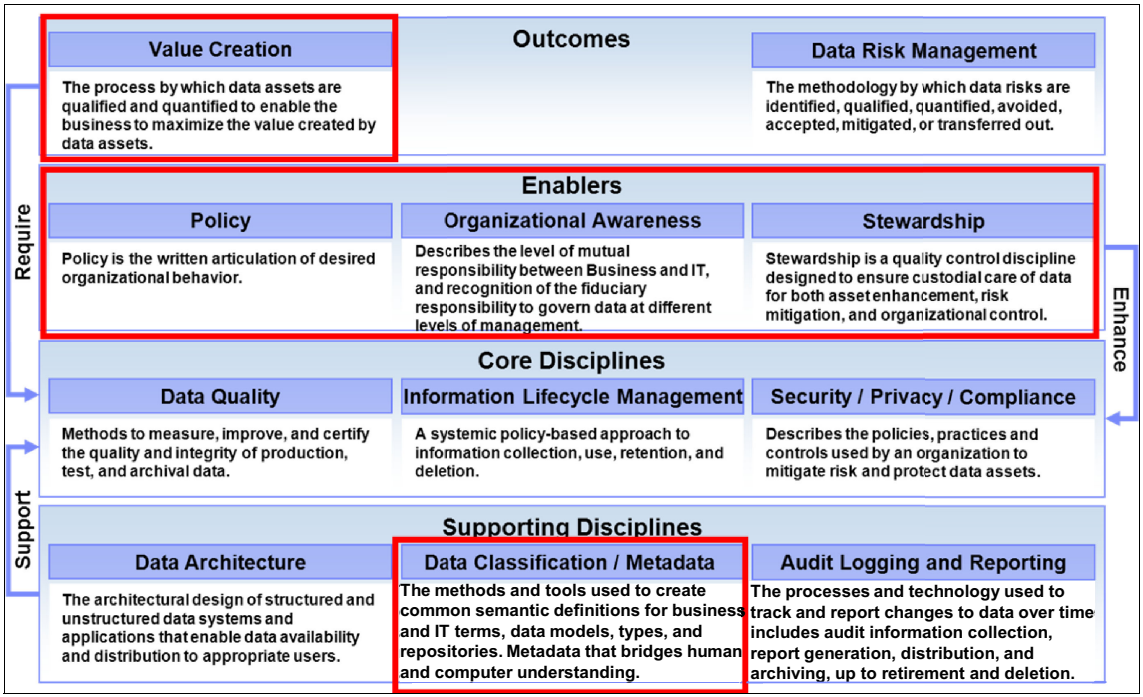


Figure 1-1 DGCMM with foundation governance discipline areas highlighted

In Chapter 2, “Information Governance organizational structures” on page 19, we examine Organization Structures and Roles and examine some new concepts (such as the Structured Communications Model) around how to “get organized” and approach getting started with implementing this governance foundation cornerstone.

In Chapter 3, “Business definitions and policies in IBM InfoSphere Information Server” on page 41, we delve into the IG definitions and policies from which most other IG foundation and functional capabilities are created.

In Chapter 4, “Workflows and business specifications” on page 57, we describe translating and enacting policies via the establishing of procedures and workflows that lead and guide governance workers, as well as the specifications and related collateral that support this work.

And in Chapter 5, “Metrics and measurements” on page 75, we look at Milestone and Measurement, a key but often under-emphasized foundation governance component that must be put in place to both officially establish the aligned organization objectives and to monitor initial and ongoing IG acceptance, adherence and impacts.

1.3.1 Some final foundation governance thoughts

When approaching the complex work of establishing foundation governance capabilities, another way to look at the elements of governance to be created and matured is to group the produced outputs:

- One collection of *foundation* definitions, policies, processes, and people that are more part of the umbrella framework for *all of* the candidate functional (or core) governance capabilities you seek to support. These are commonly the business of the Strategic and Control responsible and accountable parties in the organization.
- Supplemental sets of *functional* governance capability component parts that take shape and develop while getting each functional governance capability to an operational state. These will often be very specific processes and collateral, guided and created from the governance strategy and control outputs, but more operational and practice oriented.

The umbrella capabilities are often the more common or shared type, applicable to multiple areas, such as the cross organization definition of the business object “customer”, or the baseline policy requiring the full describing of all new data elements within the central metadata repository. The more specific, functional elements might be updates or added attributes to base definitions that identify elements as sensitive data or cross-system identifiers, or policies that require the specifying of what specific content in the “customer” business object must be retained in archived data or available for legal discovery.

Figure 1-2 on page 7 shows the Foundation and Functional governance components.

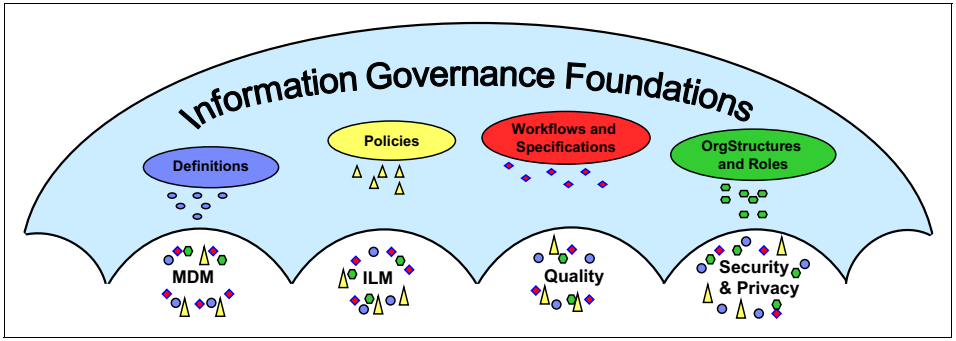


Figure 1-2 Foundation and functional governance components

This interwoven framework of elements of capability is what then support the building of good governance and ultimately confidence in the organization's information and the decisions upon which they are based.

Finally, any discussion on establishing of the foundation governance components is not complete without noting the need to establish a formal *Information (or Data) Governance Office (IGO).* This should be a dedicated group that is in effect the embodiment of the IG efforts. All governance efforts include an element of change and iteration and adaptability, and without a body established that actually owns the "minding" of this, the veracity and credibility of the governance effort will be questioned if not scuttled. Establishing an IGO is also a positive sign that the organization is legitimizing the efforts, and intends to employ monitoring and enforcement.

More is described on IGOs in Chapter 2, "Information Governance organizational structures" on page 19.

1.3.2 Understanding why implementing IG can be hard

In the balance of this chapter, we spend a little more time on this topic of "How do I get started on this information governance journey", building the right amount of the foundations and actually getting on to executing practical governance activities.

There is no perfect meter or measuring stick that can tell you when to begin, or why or in what degrees of sophistication. In many cases, getting started is more "push than pull"; responding to a company specific or industry-driven compliance concern, or a desire to capitalize on anticipated efficiencies to be gained from more control and knowledge. And in most well established organizations there will be aspects or components of governance in place.

The DGCMM (http://www.infogovcommunity.com) provides ways to assess where you are from a maturity perspective and validly prescribes “gauges” for documenting evidential levels of maturity in your governance practices. The IBM Data Governance Unified Process also outlines a very explicit roadmap to follow, milestones, and critical checkpoints and transitions to achieve to commence on the governance journey. Practically though, an often expressed or observed reality is that the physical world of governing information is much more involved. It has many levels and layers of organization and political and technological brokering. It requires juggling and sequencing of activity, much more complexity than two-dimensional pages in a book or a planning guide can describe or instruct.

A good goal for any governance initiative is to separate or compartmentalize:

- The activities that will be initiated to build the overall “governance program management’ from.
- More custodial level activities necessary on a day-to-day basis that will need to eventually be integrated into the operations.

This is accomplished in part by the appropriate levels of up-front study and effort to design and implement the policies and processes that will make the daily integration of good governance behaviors possible. In early phases, it is not necessary that everyone across the entire organization have the understanding of and responsibility for shaping and defining all of the strategic aspects of every discipline. And in many cases there will be some existing “controls” already well established and socialized that can be leveraged. In the planning and building stages of any governance initiative, a reasonable role-to-responsibility assignment is key to being successful. Ultimately all levels should understand and respect their contribution and be able to execute the same with minimal impacts on their usual work day. In fact, this should be a high-level objective for every policy and process that will go into effect to better ensure adoption and acceptance.

However, changes and improvements will be necessary that will add value. In Chapter 2, “Information Governance organizational structures” on page 19, we examine an IBM patented approach for modeling a very logical and effective governance-friendly role and responsibility, communications-oriented framework. In the balance of this chapter, we explain more about another practical and value-adding approach for addressing the “How do I do this governance stuff?” question, called the *Information Governance Component Model*.

1.4 The Information Governance Component Model

Organizations that are successful in starting and implementing information governance programs often seem to be able to “persevere” in the face of a very challenging objective in even the most demanding of environments. A certain “persistence” and “selective battle fighting” talent seems a common characteristic of these winners. Probably more accurately this boils down to an ability to control the players and activities of those players to guide their focus on those things that are most suited for each group’s specific roles. So cross organization looking, governance program management type players can be asked and expected to “look across all of the disciplines” of governance maturity for systematic and integrated capability development. And those components of the governance team that must contribute a narrower, more specialized set of input are depended upon to contribute what they can best provide.

A complete “Stop everything we are doing and let us build information governance across the entire organization” approach is not usually realistic. So intelligent and practical methods need to be employed. A general method commonly used for these types of complex problem-solving efforts is to “break the problem down” into consumable chunks of activity, but in a way that makes clear how the pieces will eventually fit together.

The Information Governance Component Model (IGCM) approach attempts to do that, as depicted in Figure 1-3 on page 10, across the 11 disciplines of governance (promoted via the DGCMM) there are varying levels of Strategic, Controlling, and Operational level capabilities that need to be implemented and matured. Breaking down the effort into the discrete units of “accomplishment”, each composed of elements of people, process, and technology, enables you to describe the levels of accountability associated with each with greater precision. The model also makes it clear as to what discipline and unit blocks make up the foundation or “prerequisite” capabilities (upper left highlighted) of information governance, and the vertical breakdown along the bands of Strategic, Controlling and Operational detail make it clear how the first two foundation capabilities must be in place to support any long-term operational governance activity.

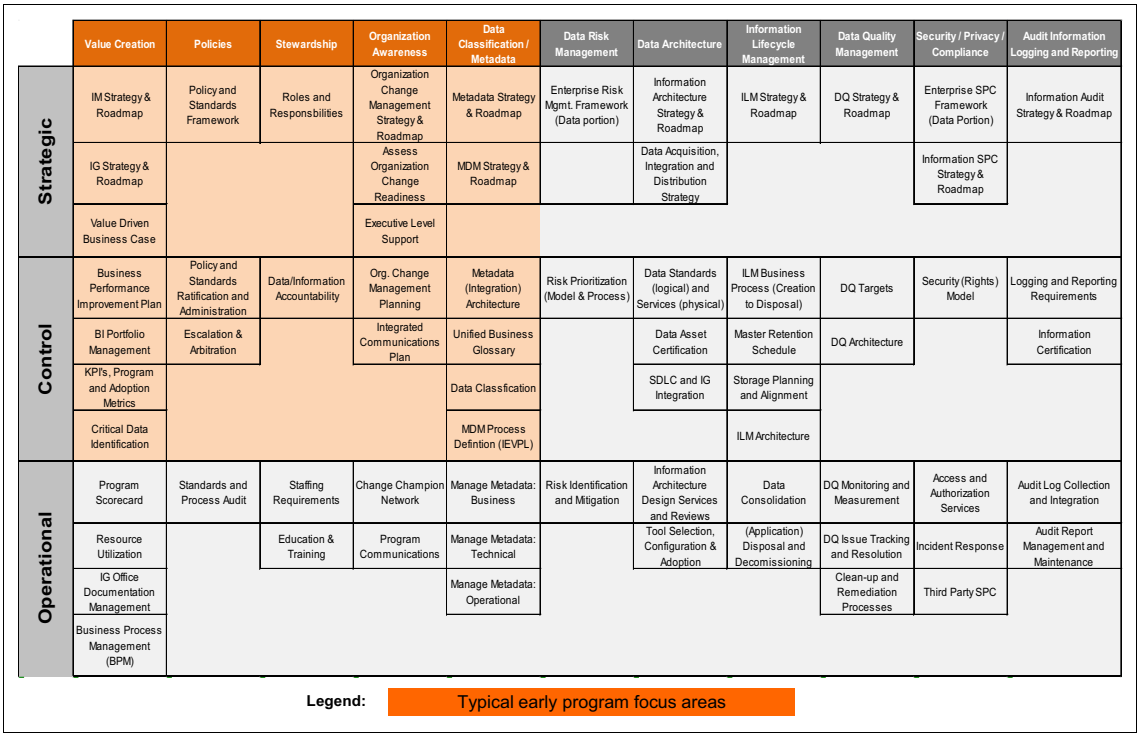


Figure 1-3 Information Governance Component Model example

To summarize, the IGCM is a breakdown like the 11 primary disciplines of information governance, into components of discrete executable nature, that can be prioritized and aligned with each client's expected (functional) governance capability objective. And these should also feather into any core capability, solution set type methodologies, for initial and ongoing implementation:

- The IG component model is based on a long standing IBM Component Model built with many years of business process implementation and change management experience from our IBM Strategy and Transformation practice.
- Components are exclusive and independent - to reduce overlap, each defined in terms of that component's specific people, process, and technology dependencies. (atomic)
- Components are unique but flexible, can be prioritized and reorganized based on specific client needs. Included processes can be moved between them to facilitate greater control or industry nuances or constraints.
- Components are aligned along the IG accountability lines of Strategic, Control, and Operational levels, and this "plays nicely" with classic

IG-prescribed organizational structures and process flows (including the Structured Communications Model).

- Flexible implementation is probable, and will lead to a natural progression of accomplishing operational control, as strategic components are realized and accepted.

1.4.1 Developing your approach

Once your organization makes the concerted decision to implement information governance, many will approach the initiative with the classic steps of:

1. Build the foundations: Efforts and investments in researching, defining, instantiating information definitions, policies, procedures, and organizational structures
2. Focus on implementing the core objective: Quality, MDM, privacy, and masking ... injecting and embedding the governance foundations into the operations as an ongoing concern

Though this approach is generally valid, it often lacks the focus and accountability types of mapping to the organization that using an approach such as the IGCM can help to make more evident and tangible. Using a modeling approach such as the IGCM allows the governance leadership to look across all of the 11 disciplines of information governance to identify the distinct blocks of capabilities that must be put into play to produce a coordinated and managed governance outcome.

To kick off this process, one approach used by IBM Information Governance practice leaders to help clients develop a better sense of focus, is to start with a simplified or compressed view of the IGCM (see Figure 1-4 on page 12). This “governance wheel” diagram acknowledges that a big step in this journey is getting started, the need to have an aligned vision of where you are starting and why. The wheel diagram presents this in such a way as to make it clear how even in your initial efforts you should be aware of the possible reuse and redeployment of any built-out governance capabilities eventually across all information and implementation domains. For example, in all cases, establishing measures of success (objectives), policies, standards, accountability, and control mechanisms will be essential.

Starting from the simple wheel and working toward the right, a simple “Let us take a step back and get focused” set of discussions and decisions would follow the “arrow inward” path:

1. What information domains are you looking to better govern? What functional area or areas of information is the driver behind the need for governance?

2. Where in the set of applications or implementation domains do you want to establish better governance capabilities and processes?
3. What specific kinds of governance capabilities do you need to apply in these implementation domains?
4. How will these governance capabilities be achieved?

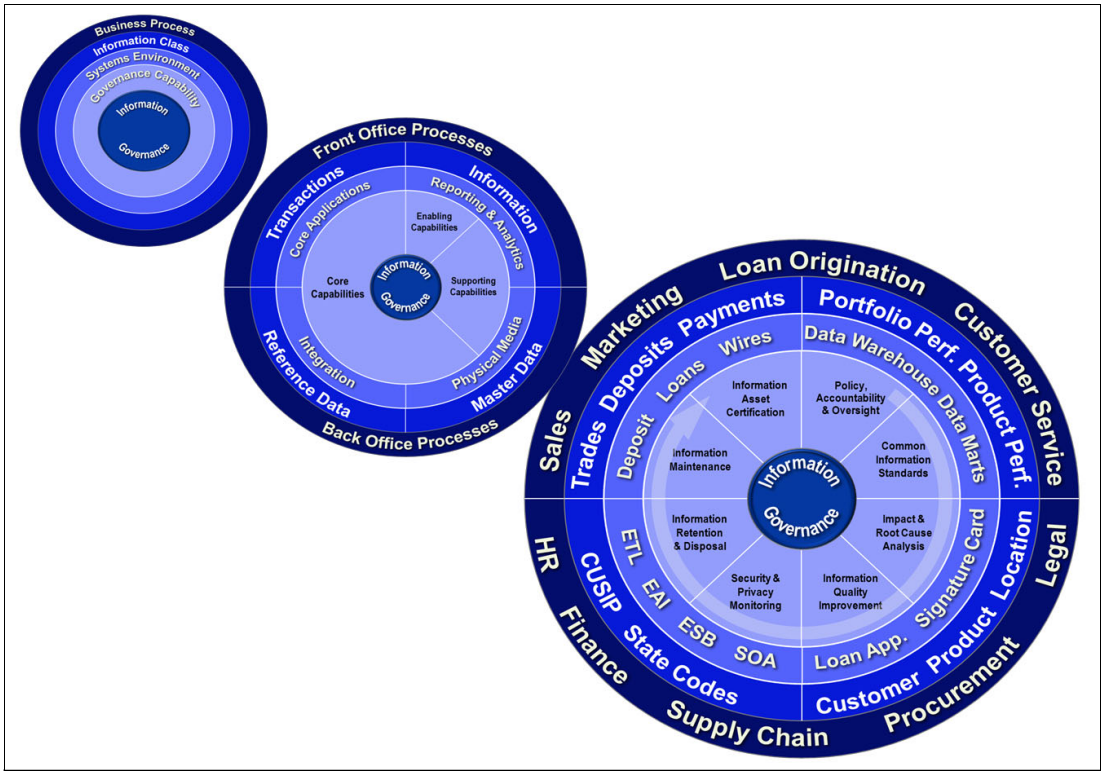


Figure 1-4 Capability framework wheel model

The framework wheel can be used to help provide a *dimensional vision* of where you are going with your governance initiatives and help you move to the next steps of breaking it down to the next levels of decomposition.

The next steps in the process would be to leverage something like the IGCM grid or heat map approach (Figure 1-3 on page 10). This approach helps with zeroing in on the requisite and core governance objective focused disciplines (listed horizontally across the top) that you will be focused on, combined with the type of activity breakdown (strategic, controlling, or operational) that will drive the kinds of work and the types of people that will be necessary to enlist to accomplish the same.

In almost all cases, as Figure 1-3 on page 10 highlights, there will be ‘the upper left” set of activity-resource units of work, which are commonly considered the preliminary or prerequisite types of capabilities needed no matter what the core objectives.

Then, as you work across and down the model, you can begin to both plot and validate existing capabilities along with required governance development (specific processes and specifications) that will be necessary to accomplish your targeted core governance objectives.

The simple value of this component model approach is that it allows governance program management and specific governance steward-type specialists alike to each focus on what is most important for them to accomplish. These mutually exclusive components of accomplishment collectively work together to provide the governance framework.

1.4.2 Breaking down the IG Component Model value

The IGCM approach is based in many ways on the Data Governance Maturity Model premises (11 key disciplines), and adds more drill-down, more component breakdown across a responsibility thread to help organizations focus on work that is required to achieve their targeted governance goals. So a good first step in understanding and appreciating the value that an IGCM approach brings would be to walk through a simple, core governance type objective, such as Security and Privacy Monitoring, first using the DGMM and then with an example IGCM model mapping.

In Figure 1-5 on page 14, we see the standard DGCMM, with one slight modification; the Enablers band has been brought to the lowest tier (just to better present the more incremental “capability layer building” metaphor). Our core governance objective of security and privacy, at the top layer is probably best represented as a risk management type outcome - the appropriate and compliance guided treatment and protection of all sensitive data across the source, integration, and analytics data environments within the organization.

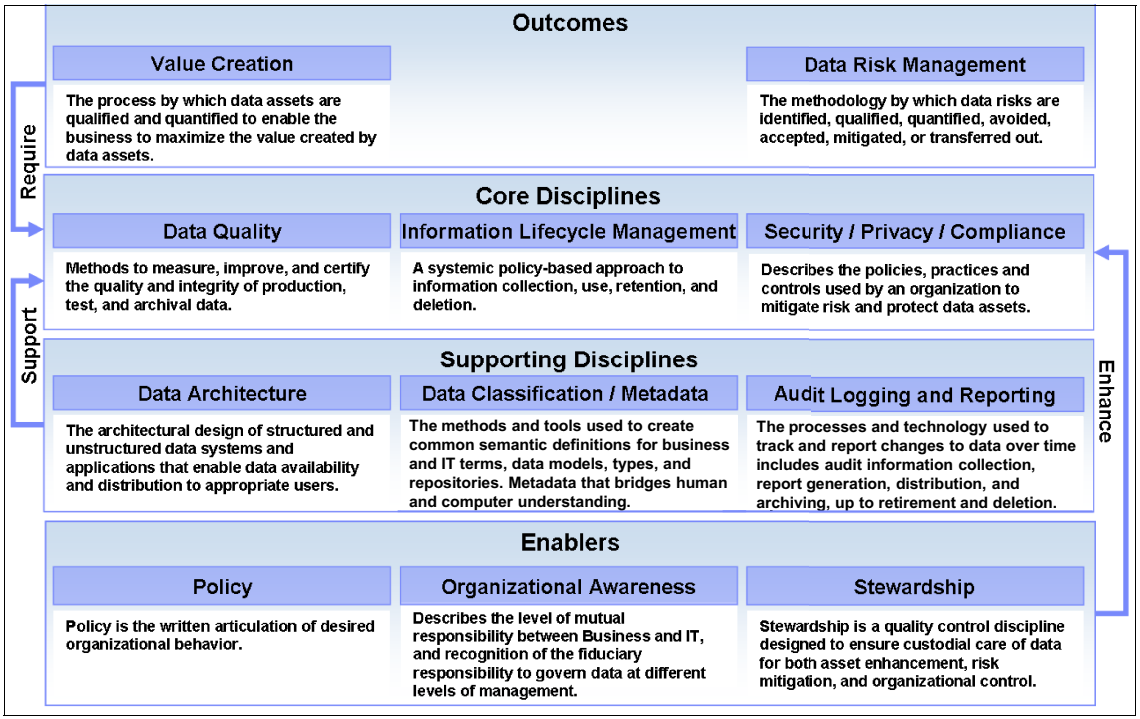


Figure 1-5 Modified DCMM: Eleven disciplines of information governance

Working in Figure 1-5 from the bottom up, the Enablers layer prescribes a need for some degree of organization awareness and stewardship as well as some policy that would articulate the correct organization behaviors, in this case, regarding security and privacy.

Next, the Supporting Disciplines. This band identifies a need for specific data architectures and classification capabilities. And with a security and privacy goal, these could imply the need for some form of definitive organization data classification tools and repositories that would both provide the ability to establish and control sensitive data classification. Auditing and reporting are also included at this layer, and again we could infer that this would include the need to monitor and measure the amount or percentages of sensitive data classified and controlled, versus the entire landscape, as well as any out of compliance or neglected privatization anomalies.

And finally we hit the core discipline level, that would suggest, assuming you are supported by the layers below, you would now be able to design and build the core discipline-specific policies and procedures and tooling for execution of the appropriate levels of securing and privatizing all data assets.

To be clear, these are all very valuable points of guidance. And the additional detail found within the DGCMM assessment materials, at http://www.infogovcommunity.com, would also provide some evidential examples of where and how your efforts might represent a low or high level of maturity. But this is where the value of using a component model approach can help you move to the next levels of a more focused or targeted work breakdown, that at the same time provides the benefits of allowing you to recognize and plot the existing components of capability already operational. Using this modeling approach can help you move to the next steps with greater focus.

So, looking at Figure 1-6 on page 16, an IGCM approach, for a Security and Privacy core objective (and assuming that Figure 1-3 on page 10 still applies regarding the understanding that OVERALL Governance Program Management will need to ensure that the “orange highlighted” capabilities exist in some degree or capacity), we can begin to construct, by discipline and responsibility-type of effort, a typical next breakdown of work components to guide the organization’s governance development efforts.

Focusing first on the “upper left” foundation areas, though to some extent there would need to be initiatives in all component boxes, we can highlight (yellow) those that are clear prerequisites. For example, a core initiative of this type is not possible without Critical Data Identification and assigning of specific Information Accountability. These would need to be defined and governed (or enforced) within the “control layer” of your governance framework. Other related Roles and Responsibilities must also be established and socialized.

Looking across disciplines, organization Data Standards and Classification capabilities, the core initiative is effectively not even viable without these capabilities in place. And operationally, Change Champions and Governance Program Communications channels must be in place to support execution and adoption.

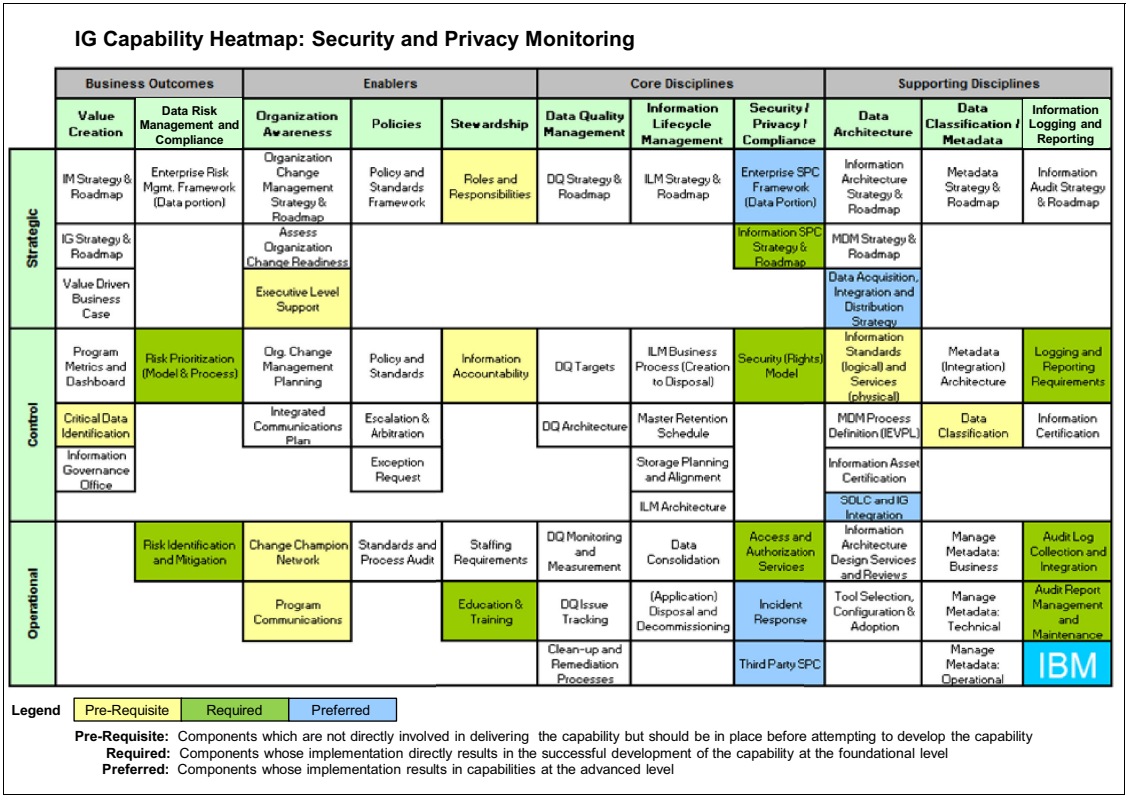


Figure 1-6 Information Governance Capability heatmap

Green colored components then highlight the additional required work effort areas that would need accomplishing. And blue areas provide recommended additions, often aligned with or as indicators of advanced levels of governance maturity.

Viewing the governance capability initiative in this shape helps governance program management visualize the “inventory” of governance areas that must be managed, and makes clearer how the responsibilities will be spread across organization roles of Control-through-Operations. And though additional breakdown efforts will need to be defined, the model provides the first-level picture of the “run state” of this particular core governance capability initiative.

Some final thoughts on the Information Governance Component Model:

- The Model provides us with a useful tool to work with information governance stakeholders to get from the initial phases of identification of need and high-level objectives to a more componentized mapping of the foundation and core discipline pieces that can guide successive efforts of implementation.

- The model makes clearer the domain of responsibility for the Data (Information) Governance Office: the “orange” or upper left components.
- The model follows the same theory and reasoning of existing IG frameworks (DGCMM and Unified Process roadmaps) but adds “next level decomposition of effort” outlining, a form of MECE (mutually exclusive, completely exhaustive) and aligned with the IBM existing Component Business Model™ approaches to implementing governance and related information technologies.
- The model provides the ability to define IG efforts in terms of non-overlapping components defined in terms of the People, Process, and Technology formula, which is at the core of any change initiative.
- The model components are organized into Strategic, Control, or Operational accountability levels, that present a more explicit intersection between activity to accomplish with responsible level to execute.
- The model helps make clearer the need to have in place or implement the Strategic and Control components before any operational efforts are undertaken.
- Depending on the organization maturity with strategic and control levels, core initiatives can be tailored to adjust for gaps, that is, *all* components do not have to be *perfectly* in place to build some initial success, as long as the organization understands the gaps in the model and the ramifications of same.
- The approach can be used for one division, to allow early adopters to succeed, while highlighting potential gaps and deficiencies critical to expanding the success to more divisions across the organization.

A final added benefit of the IGCM breakdown is how the decomposition can assist in more clearly identifying where automation of processes can be applied. This includes how technology might connect initiatives across the 11 governance discipline-based framework. As new or enhanced governance-oriented process definitions are architected, the “technology” ingredient of the People, Process, and Technology formula (at the core of all governance initiatives) can potentially be identified earlier in the design effort.

The total IG CM ends up comprising some 70-plus discrete components of implementation focus that would fulfill the full palette of areas necessary to operate a comprehensive IG program. The choices for any individual organization’s first IG initiative would depend upon the primary area or use case that the organization is attempting to deploy. Once the final set of components to address are aligned and agreed to, the final steps in developing a concerted approach would then examine correct services and software required to create a specific and tangible, use case-based implementation program.

2

Information Governance organizational structures

A thorough and considered approach to organizational structures is a necessary part of establishing foundation governance principles. Typically, existing organizational structures and roles will not be adequate and must be expanded to include the necessary responsibilities and ownership to support effective governance practices.

Organizational structures include those who define and own the information governance (IG) requirements, those who interpret and manage the processes for the implementation of the requirements, and those who develop and participate in the implementation.

This chapter outlines the benefits, definitions, and representative structure for an effective organization structure within an IG program and then takes a deeper look at some concepts for implementing this capability (briefly mentioned in Chapter 1, “Information Governance: foundations and solutions” on page 1, developed by IBM in the form of the Structured Communications Model (SCM). This chapter examines key communications paths between a definitive set of organizational structures and roles and how a managed approach to such can help guide organizations to a better understanding and grasp on implementing traceable and repeatable IG capabilities.

2.1 Introduction to organizational structures

Organizational structures provide the framework for management and decision making within an enterprise and usually align with functional or departmental substructures.

To implement and manage the disciplines of information governance, an organization must ensure that the framework for execution is well-defined and integrated into daily operations. This is partly accomplished via establishing or adjusting of organizational structures, and in particular the creation of an Information Governance Council or Steering Board. Driven by executive-level sponsorship, any new organizational structures and roles are led by the council. The Information Governance Council can then foster and build the expertise and efficiency necessary for the application, review, and results monitoring of all the information governance standards.

Those already in the “business” of information governance know that establishing and supporting governance focused structures is a primary condition for building success and maintaining a healthy business environment. The existence of specific and dedicated roles and structures is evidence of higher levels of information governance maturity. They are the underpinnings of compliant and trusted information, leading to a stronger, and more confident ability to compete successfully.

Organizational structures and roles supporting information governance is a committed partnership between all aspects of the enterprise, from the business owners and shareholders to the developers and information technology support personnel. Everyone contributes in time and effort and has input and visibility to the process and the outcomes.

2.2 Introduction to information governance roles

Data or information governance (within this book, these terms are used interchangeably) should be applied to all business information and data assets in a clear and transparent manner. Supporting this initiative requires a clearly defined set of roles and responsibilities, in addition to aligned definitions and well-established processes.

The governance community members, collectively, are responsible for publishing the approved and controlled set of definitions, policies, rules, and benchmarks that must be applied to and control all information. Through the prescription of consistent use and processing policy including appropriate traceability, the

business benefits of improved confidence and transparency in information and the decisions which that information guides, can be achieved.

There are probably an infinite number of ways to approach assigning personnel within an organization to support the governance community. The tasks often fall to “IG passionate” proponents who might or might not be the best or most empowered set of players to make the initiative effective. A simple but effective model to help with understanding a best “dimensional” approach is depicted in Figure 2-1.

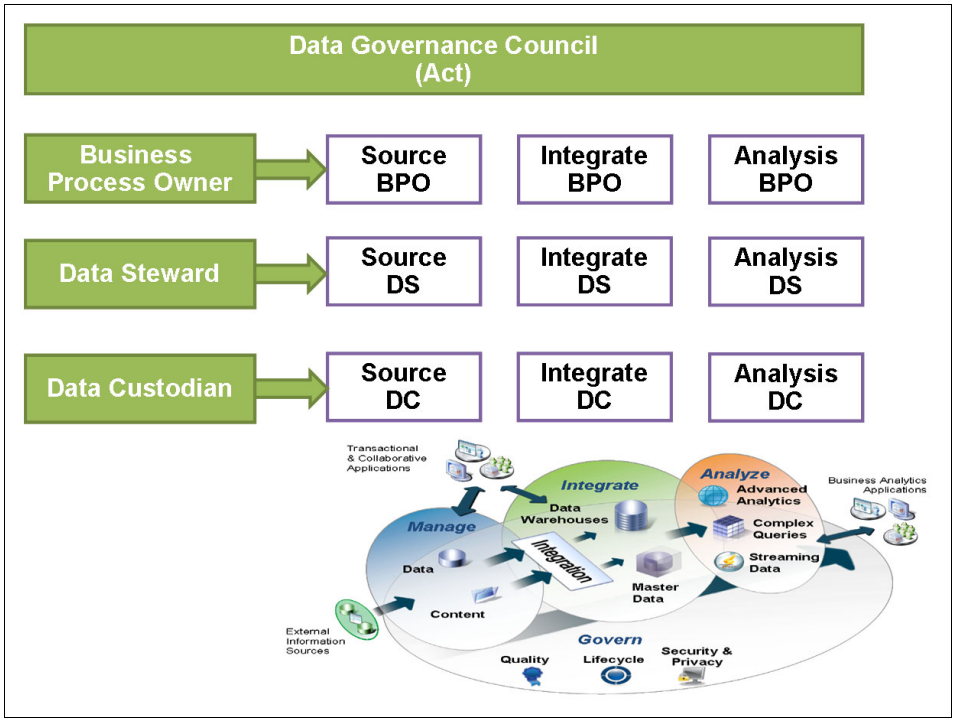


Figure 2-1 Suggested organization governance role model, supporting functional processes

This model overlays the information supply chain, information governance requirements, and organizational structures. The information supply chain depicts the flow of information through an organization, from sources, through its integration and aggregation and analysis, while the organizational structures depict the owners, stewards, and custodians charged with implementing and guarding the information assets. This simple framework forms a *communication network map* or model to support the execution and management of the governance processes. When these roles of “governed” resources are combined

with the “governing” bodies and roles (Figure 2-2) you can be more confident that you have established a complete set of organizational structures and roles. This “Structured Communications Model” or SCM will be discussed in more detail in the 2.4, “Organizational structures and the communications model” on page 27 section later in this chapter.

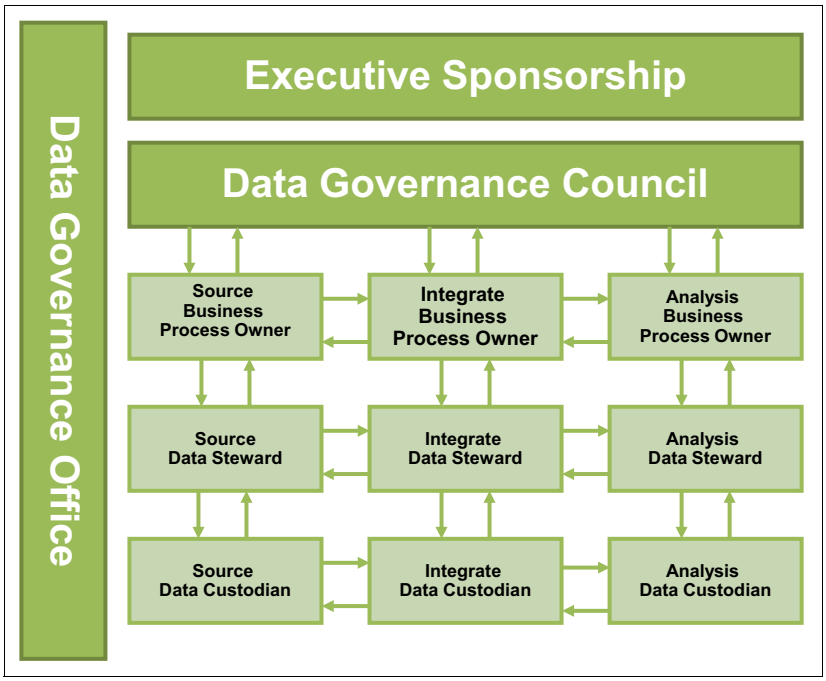


Figure 2-2 Complete governing and governed bodies and roles of IG

2.3 Governance layers

Information governance is a strategic initiative involving multiple functions across the enterprise. An information governance program (and overall “committee” of oversight) should include a governing body with an agreed upon common set of procedures, and a plan to communicate and execute those procedures. Thereby, the definition of an organizational structure is typically a layered structure consisting of varying levels of management and responsibility

2.3.1 Governing functions and roles

In this section, we describe the governing functions and roles in the governance layers.

Executive Sponsor

Sponsors are commonly the individual or department that drives the requirement, funding, oversight, and guidance of the information governance initiative. Regardless, if the motivation is initially a reactive one, for example the compliance with external regulation, or adherence to internal corporate policies, any such initiative requires a sponsor who can ensure that the organization structure is properly staffed and authorized to complete their mission. The sponsor additionally clarifies and defines the specific scope of the governance initiative and helps establish milestones and goals for completion.

Data Governance Council

This group consists of senior representatives from across the business, which would primarily be senior level representation across technical and business departments. The team articulates the vision and plan for the implementation of the information governance initiatives as defined by the sponsor. The council additionally sets the strategic direction, and provides review and oversight on its execution. This team is tasked with ensuring compliance with all laws and regulations regarding data or information use and storage for the purposes of conducting business. With its executive power, the team can identify and allocate the resources needed for the execution of the program. This group will also set the higher level, organization-aligned measurable objectives and goals in cooperation with the executive sponsor. These are then broken down and used by individual governance guided programs and projects to shape their specific goals and objectives.

Data Governance Office (DGO)

This group of both dedicated and virtual staff will also typically include a specifically nominated program manager or champion to lead the office and ensure the development and implementation of the initiatives. The DGO may often be referred to as the *Governance Center of Excellence* (COE), and may in fact be composed of representatives from more granular COE organized teams from across the enterprise, such as Information Architects, or Data Quality or ILM specialists. This office provides the logistics and organizational support for the operational teams and further guides the separate operational teams in their development of the processes and procedures for the various governance activities upon which they are focused. Their operational oversight and methodology type guidance ensures that each's initiative is based on policies, principles, and guidelines that are provided by the Information Governance Council and Executive Sponsor. The team further oversees the compliance and operational performance of the operational teams.

Together, this "governing" team is responsible for providing and setting the data content, definition, context and associated definitions, rules, policies, and specifications for all information/data content. The "guidance", which this team

establishes (and enforces) is then utilized as the instructions and template for the governed roles and functions to adhere to. For example, the activity of capturing and recording *metadata*, is a requirement to properly identify information within some formally identified (guide) Information Catalog to be able to aid all users of this content (policy/guide) in properly understanding the meaning and structure of information across the enterprise.

An enterprise typically has a number of Business Process Owners, Data Stewards, and Data Custodians aligning to the different subject or operational areas, such as Reporting and Analysis, Data Integration, Data Quality, Data Warehousing, Master Data, and Life Cycle Management. Without these guides and boundaries, the team is void from benefiting from a controlled vocabulary and clearly defined and published set of governance policies, rules, and business guidelines.

2.3.2 Governed functions and roles

In this section, we describe the governed functions and roles in the Governance Layers.

Business Process Owner

The *Business Process Owner* is responsible for the use, integration, and processing of information, as depicted and specified within the Information Flow Diagram (in Figure 2-1 on page 21). The owner further is accountable for ensuring the application of data quality routines and compliance regulations when integrating with or otherwise using such information.

Data Steward

The *Data Steward* is the caretaker of information and is responsible for managing and maintaining its content. Stewards, themselves, are not as typically responsible for the physical integration or usage of information, rather ensuring the content is well-defined, structured, accessible, and compliant. The steward is an expert in a particular area or topic and possesses an intimate knowledge of the structure, definition, and intent of the information.

Data Custodian

The *Data Custodian* manages the physical ownership of information and the implementation of the processes for data integration and data quality. The custodian is the person responsible (directly or via coordination of other technical roles) for the usage, custody, transport, and storage of information and specifications.

In addition to the defined set of responsibilities of the Business Process Owners, Data Stewards, and Data Custodians, governance focus can also be divided into functional responsibility bands as seen in the Information Flow diagram (see Figure 2-1 on page 21).

Sources and acquisition

The data sources and information assets are depicted within the Information Flow Diagram (in Figure 2-1 on page 21). All sources must include: A *Business Owner* who is responsible for the care of information, a *Steward* who governs the content of information and ensures information is well-defined and confident, and a *Custodian* who implements the business rules and routines readying information for integration.

Integration

The extract, transform, and load (ETL) processes and systems that are depicted within the Information Flow Diagram move information. All processes and systems must include the following:

- A Business Owner who is responsible for the architecture, specification of the systems and processes.
- A Steward who governs the structure of information and advises developers and custodian on usage of information.
- A Custodian who implements the systems and processes that are based on the defined specification and definitions.

Analysis and reporting

The analysis reports and user applications that are built upon the processed sources and systems, as depicted within the Information Flow Diagram. All reports and applications must include the following:

- A Business Owner who is responsible for the definition, processes, and distribution of the report or application.
- A Steward who governs the data quality and compliance of the sources included within the reports and applications.
- A Custodian who consolidates the sources and makes the reports and applications available.

Combining the (vertical) organization roles with the (horizontal) architecture functions yields a simple but clear three-by-three roles and responsibility model that can be used by governance councils and offices when examining and measuring up their current and required organizational structures to support their initiatives. Additionally, these can often be one-to-many or many-to-many types of relationships, depending upon the particular IG project or application initiative.

As pictured in Figure 2-3, this model does not necessarily prescribe an army of BP Owner, or Steward, or Custodian atomic role teams, but has the flexibility to adapt to the situation. For example, in a smaller organization a single resource may fill multiple roles, perhaps even across functions. And in a larger, more complex project, a single role may require the segmentation of work into multiple resources. The key is for the appropriate team to ensure that the governance role is filled.

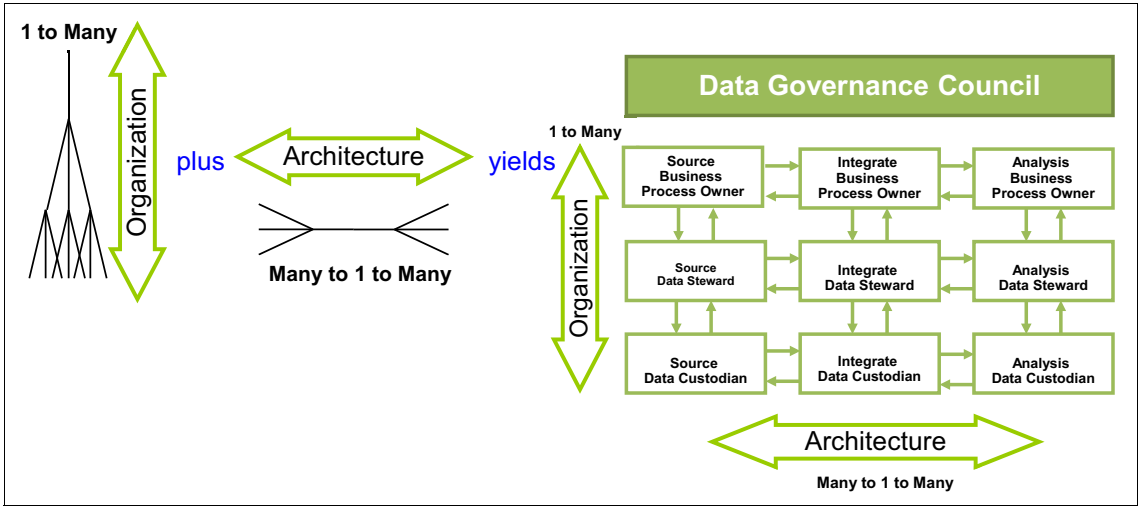


Figure 2-3 Roles and responsibility by organization and architecture

The type of efforts and tasks these roles engage in and execute follow the typical lifecycle of information processing streams. The governance-related aspect would include activities such as:

- Identify and name a specific set of data elements
- Develop and provide an approved and explicit business definition for all data elements
- Develop and provide an approved and explicit set of business rules and policies for the handling, storage, and usage for all data elements
- Define and provide the structure for all data elements in the form of a model diagram or otherwise acceptable format
- Define and provide the master data requirements for all data elements
- Define and provide the quality standards, data definitions, metrics, and standardization rules for all data elements
- Define the traceability and integration process and requirements for all data elements

The expectation is that these efforts and activities are guided and controlled by policies and processes for good information governance that is established by the governing bodies (Council) and administered via the Governance Office.

The complete organization structure picture might look like something that is shown in Figure 2-4.

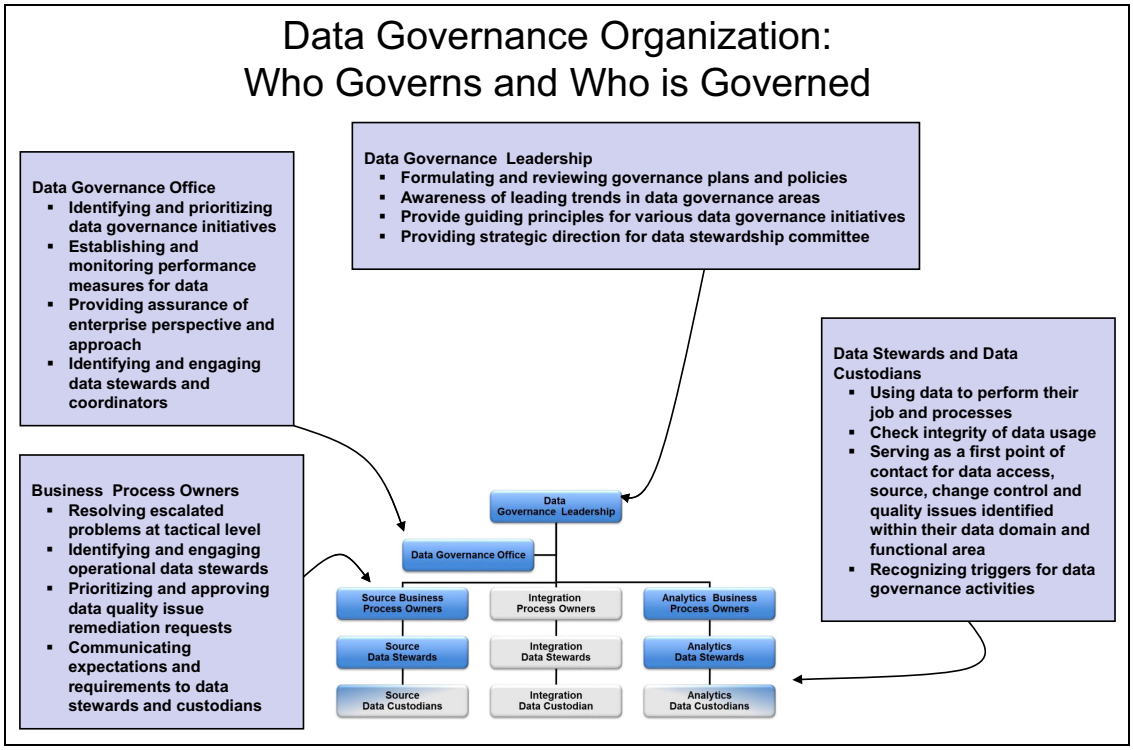


Figure 2-4 The governing and the governed

2.4 Organizational structures and the communications model

Thus far in this chapter we have described a hierarchical breakdown of organization structure and supporting roles that most any information governance or change management proponent should have little problem understanding and supporting. The establishing of new or better defined and elaborated duties of information governance often require expanded departments and formal responsibilities.

When we discuss organizational structures, we are implicitly also describing how people and departments will be grouped or aligned with common objectives. As important as the “what” the structure will look like, is the “how” the pieces will interact. Interacting includes communicating, verbally and procedurally, to efficiently and effectively reach the stated objectives.

Work from IBM on the Structured Communications Model (Figure 2-5) seeks to emphasize this latter “how” dimension and promotes the premise that in order to ensure that our new governance structures are successful, we must go a bit further than just defining the shape and flows of information. We must actively structure and manage these information flows and communications as we would manage any other “process flow” in the organization that is integral to the organization’s success.

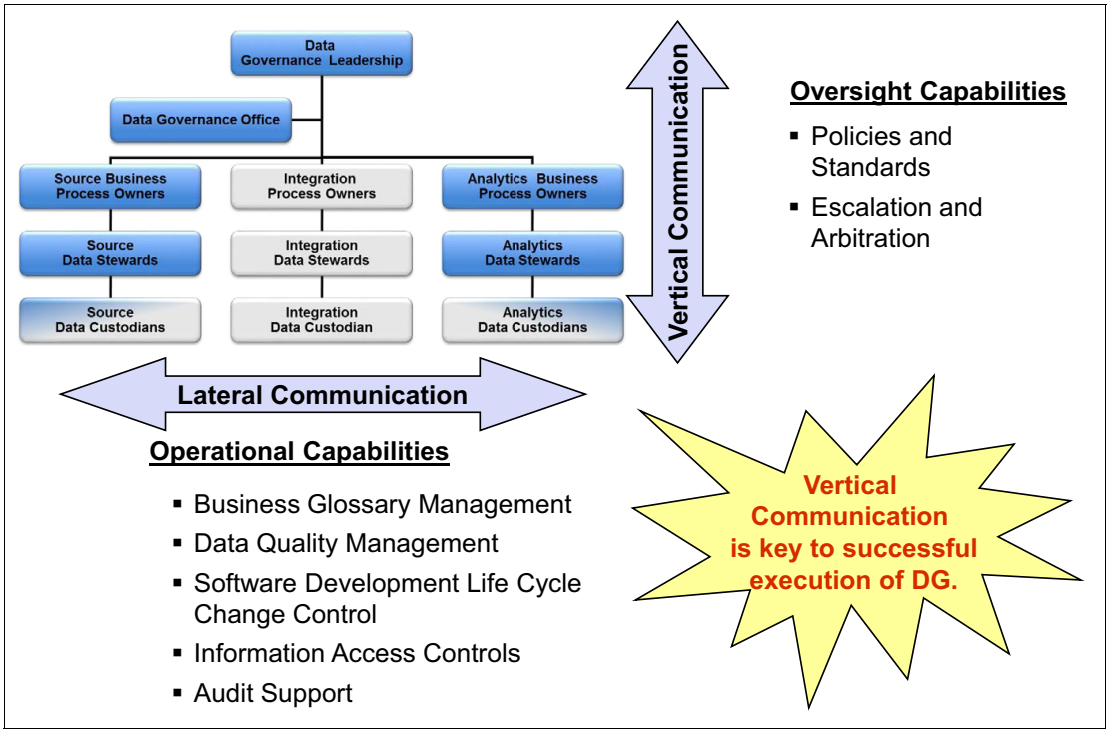


Figure 2-5 IBM Information Governance Structured Communications Model

The SCM provides the structure that is needed to unify the interoperability of a wide range of data management and governance-oriented tools.

2.4.1 Understanding the SCM

The basic premise with SCM is the following:

Communication paths among the nodes of an IG aligned organizational framework form the backbone of an Information Governance Structured - Communication Model. Once recognized as a “structured model” it provides the following benefits:

- You can define a specific set of responsible stakeholders
- It presents a definitive set of communication flows
- The direction and alignment of flows reflect the classic movement of information through the organization that drives the need for governance type activity
- The model lends itself to automation (both the automating of the core processes of applying governance behaviors as well as promoting and growing governance assets and collateral)

To better understand the nodes and structures, we can start with a commonly understood “Information Supply Chain Flow” model applicable to any given enterprise or industry. Refer to Figure 2-6 on page 30.

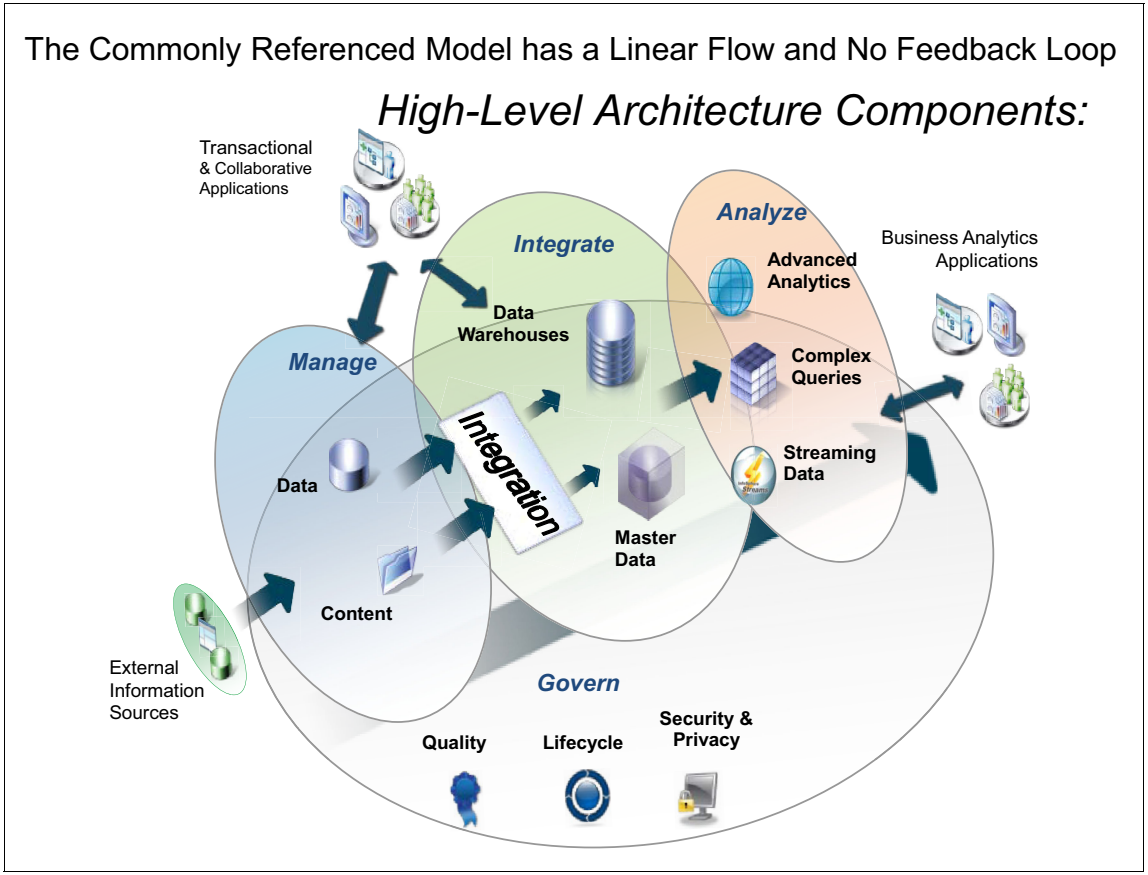


Figure 2-6 Basic information supply chain

We know that we need to govern the information “flowing” through this “supply chain”, and if not, it will exemplify the root of our “lack of lineage understanding” (classic business driver), commonly behind many Information Governance initiatives (That is, how and where did this piece of information come from, or, what are my obligations for keeping or managing this “information nugget”?). We know that as this information flows through the organization, it will typically go through numerous transformations and processing, from sources-through-usage states, and eventually into more elaborate analytics driven shapes and sizes. Our desire is to create control and predictability in these processes, but the natural characteristics of these flows are that they can be disjointed or elongated and non-uniform, which impedes and obscures our understanding of state, transition and origin.

The SCM simply states that a declarative mapping of People and Communications Flows, along this already naturally occurring information supply

chain flow, is what can be *the medium through which* IG operational policies and procedures can be instantiated and persisted. This becomes the linkage between the Acquire → Absorb → Analyze stages that can be shaped and guided with a governance compass.

To be effective, this model of communications must be officially recognized and managed within the organization, to both enforce the initially identified governance objectives, and to also expose and accentuate where more or new explicit types of governance details (information asset-related characteristics such as definitions, policies and processes) might be in order to improve or mature the overall governance activity (the feedback loop, back to the executive or sponsoring council and control, noted in Figure 2-7).

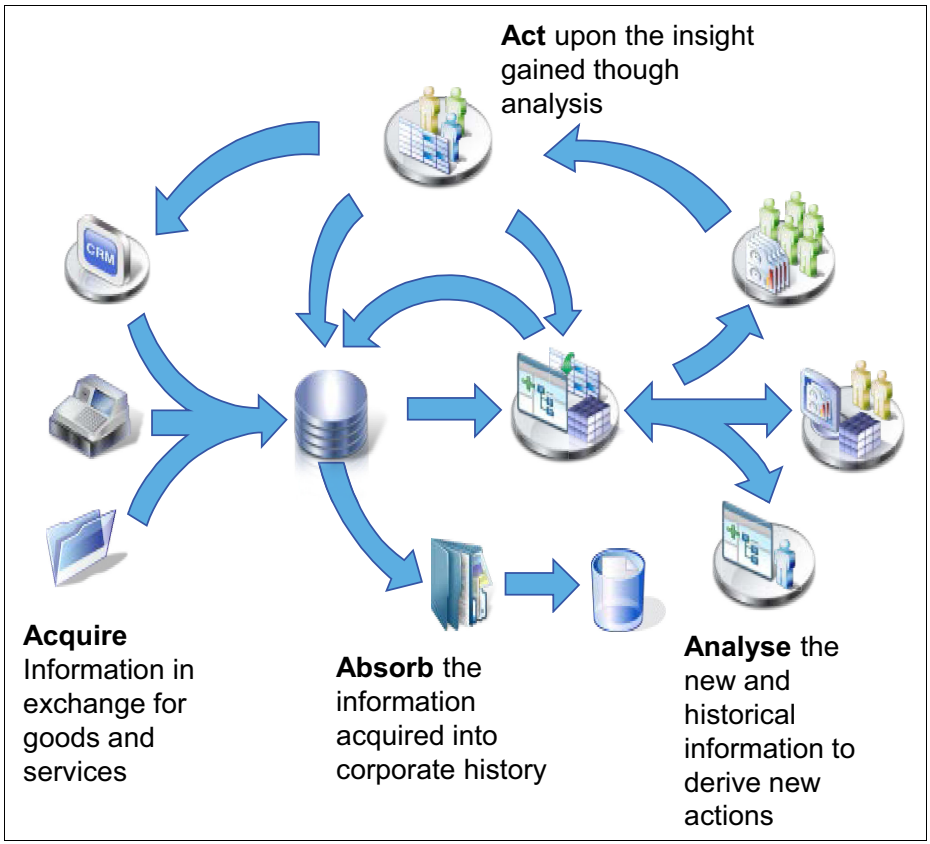


Figure 2-7 Information supply chain with feedback loop

Compare the Information Supply Chain shown in Figure 2-7 on page 31 with the connecting roles shown in Figure 2-8. The organization roles within IG are established horizontally as well as vertically. This is effectively the SCM: managed roles and communications at the convergence of hierarchies and technical architectures. This pattern, with little variance from naturally occurring information flows, enables:

- Higher-level organizational structure roles (executive sponsor, council) being corporate-business objectives oriented
- Process Owners and Stewards to some degree “straddling” both business and technical understanding
- Custodians being accountable for the technology/implementation levels

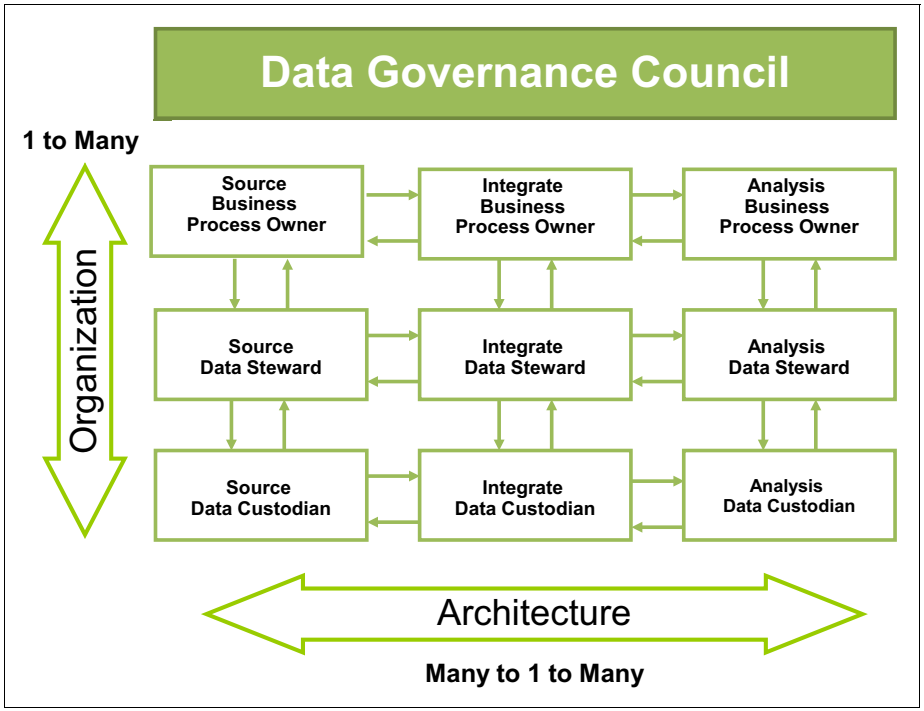


Figure 2-8 Structured Communications Model: Communications flows

At all levels, existing and new governance definitions, policies, and processes may be defined for or related to specific process flow paths and roles, with varying degrees of decomposition. So for example, an entire policy may be Business Process (owner) associated, but individual rules of this policy might be defined to guide stewardship and custodial behaviors.

In 2013, IBM secured a patent of the SCM to promote a deeper understanding and adoption of SCM for governance organization framework modeling and implementation. In summary, the SCM brings to your governance initiative a tangible approach to implementing an often very (business) political and theoretical capability implementation exercise, by providing a clearer, simpler implementation approach and direction:

- Structure the communications between business and IT stakeholders of business critical data, ensuring streamlined, complete, and efficient responses to information governance requests.
- Maintain a viable scope: Recognize that there are a limited or finite number of functional roles to be governed, and therefore a limited number of communication channels (as described in the diagrams).
- Model and express the communication channels as a repeatable, well-defined process.
- Define the specific functional roles (Business Process Owner, Data Steward, and Data Custodian) with specific activities for each, to govern the information across the various enterprise systems.
- Automate: Once defined, this repeatable and structured (foundation set of) process can then be automated. That is, more confidently leveraged and optimized via the use of software tools applied to manage data throughout the information supply chain.

A couple of examples that might demonstrate how you would leverage the investment in managing this mapping of communications flows to drive information governance.

In Conducting an Impact Analysis - (how a new application or perhaps just an additional set of new attributes would affect the Integration and Analysis process flows) - you would use the communications pathways as a map to trace or interrogate the responsible roles and the governance processes and systems they use to track governance-related detail, from Source-to-Integrate-to Analyze, along and through the vertical as well as horizontal pathways of communication (see numbered mapping in the information flow in Figure 2-9 on page 34).

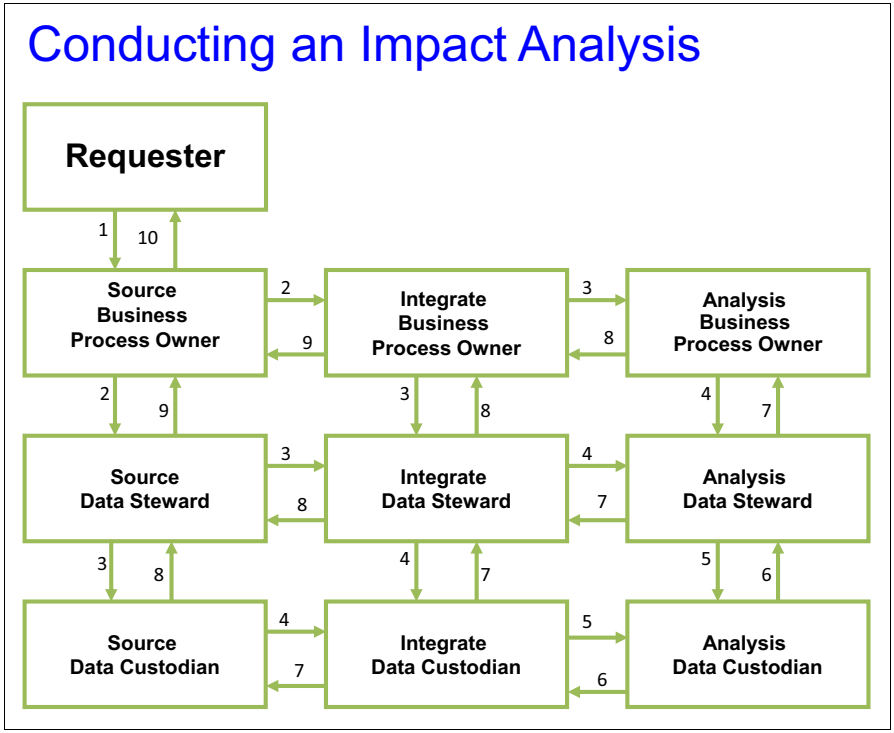


Figure 2-9 Conducting an Impact Analysis: Governance communications flow map

Performing a Root Cause Analysis - (how and where a specific set of information, downstream in the Analysis phase, was originated and transformed) - you trace backwards through the communications model, again interrogating the responsible roles and governance processes/systems used, to document and validate lineage (see numbered mapping in the information flow in Figure 2-10 on page 35).

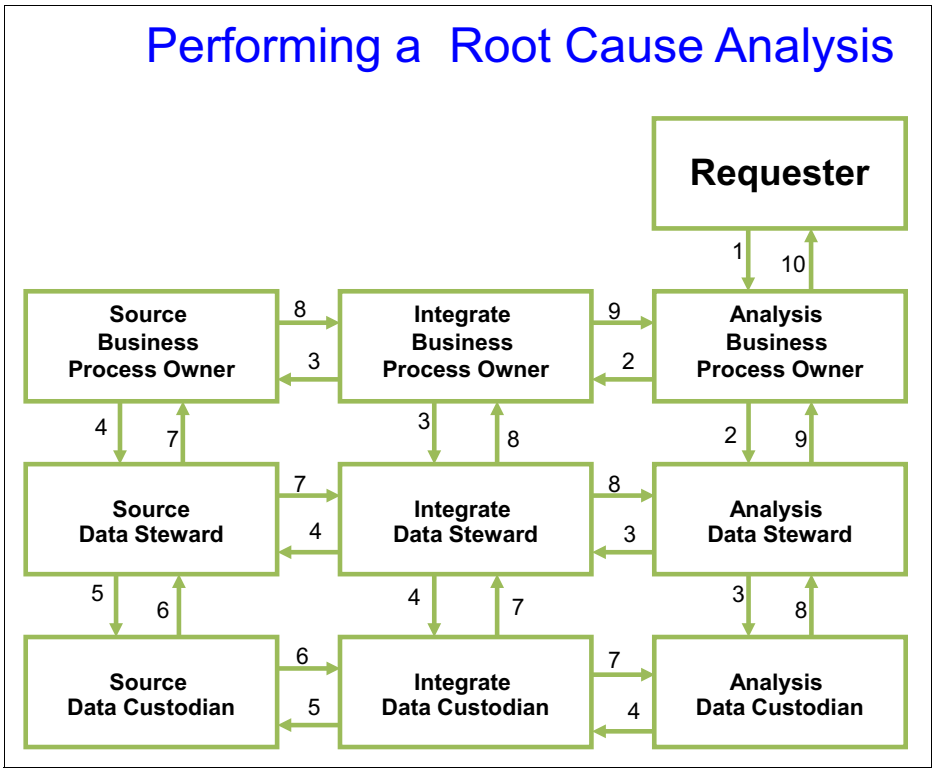


Figure 2-10 Root Cause Analysis: Governance communications flow map

One final note, related to the last bullet in the bulleted list on page 33 on software automation, each of the connecting arrows in this communication map can be linked to or automated with some aspect of IBM InfoSphere solutions software. In Figure 2-11 on page 36 and Figure 2-12 on page 37, a more explicit breakdown is provided to demonstrate an example of how technology (InfoSphere solutions) can be mapped more definitively and confidently across the enterprise to support these governance activities.

In Figure 2-11 on page 36, the logical governance duties and functions are described, and in Figure 2-12 on page 37, a sample of specific IBM software solutions are mapped to the grid. This perspective is useful for governance leadership to examine and evaluate automation alternatives, as well as to establish requirements for candidacy qualification.

This is significant because a growing base in the governance community is concerned with the lack of automated solutions to support both the delivery of governance as well as the monitoring and measuring of compliance and impacts.

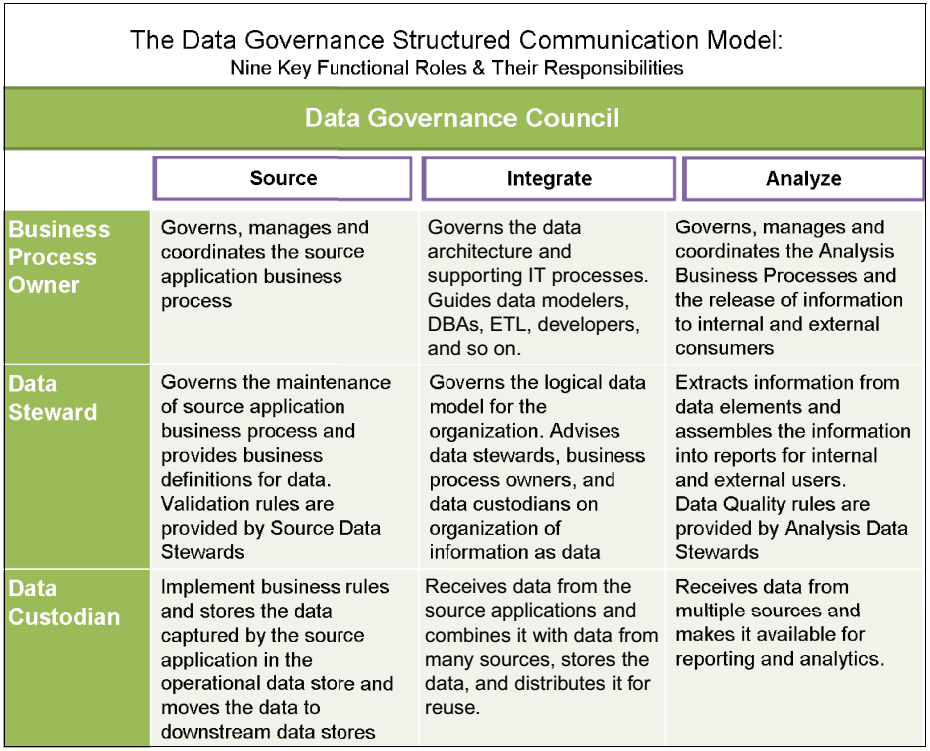


Figure 2-11 IBM patented: SCM with role descriptions

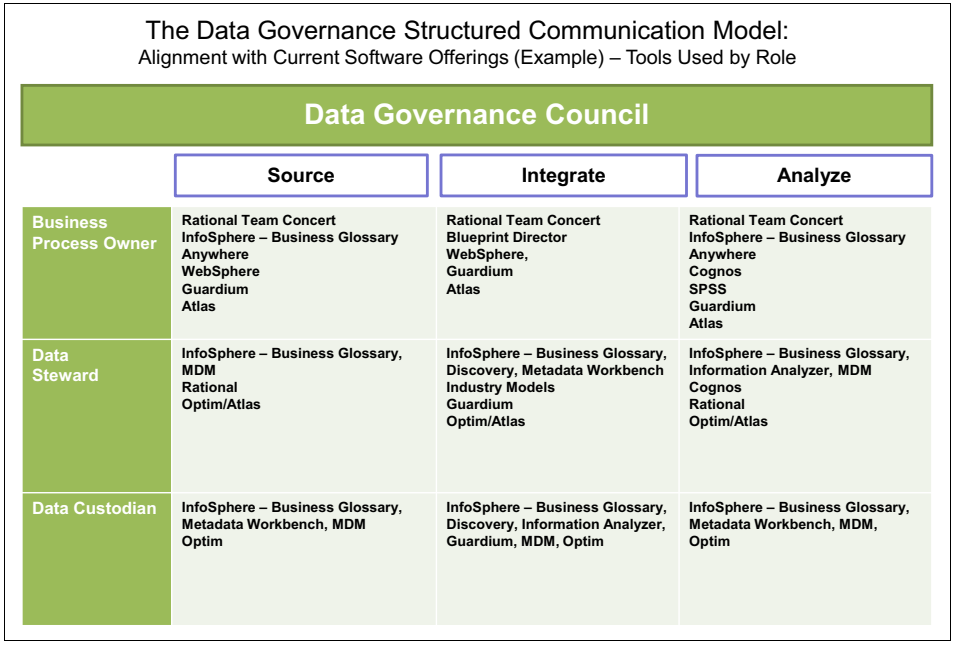


Figure 2-12 IBM Patented: Structured Communications Model for Information Governance with Tool Mapping

2.5 Application of data stewardship

The Data Governance Committee and the Operational Team members must be able to identify and ensure clear understanding of the corporate objectives via regular discussion and interaction among the committee:

- Which stakeholders should be part of an Information Governance Council?
- What type of commitment is required of them?
- What will they be expected to do in their meetings? Between meetings?
- What support will they get, and from whom?

Once a specific governance capability or iterative improvement to the same is completed and delivered, there must be an initial assessment of objective attainment as well as continual review to ensure ongoing compliance and any changes in requirements. Thus, a best practice and requirement of information governance is the creation, assignment, and deploying of on-going data stewards and data custodians, in addition to business unit owners or subject matter experts.

InfoSphere Information Server Business Glossary allows for the management of data stewards and their association to the specific information assets for which the steward is responsible. Data stewardship assignments from Business Glossary are aligned with its basic definition, the owner, and person most familiar with its content, structure, and intended usage. Additionally, the steward is also involved and responsible for the application of data quality rules and routines for the information asset.

2.5.1 Creating and managing data stewards

Stewards reflect a defined user who is responsible for a specific information asset. The steward is often the subject matter expert or owner of this information, and will understand the meaning, structure, intent, and usage of the information. The steward may also be further responsible for first level monitoring and analyzing the information, to ensure its compliance and adherence to the set of policies and rules defined, as well as enriching the meaning and understanding of the information asset for all users.

Data stewards are easily created and managed within InfoSphere Information Server Business Glossary, where they may also be associated with specific information assets. Users can therefore easily identify with the owner and person most familiar with the information for purposes of questions or otherwise. This is depicted in Figure 2-13 on page 39.

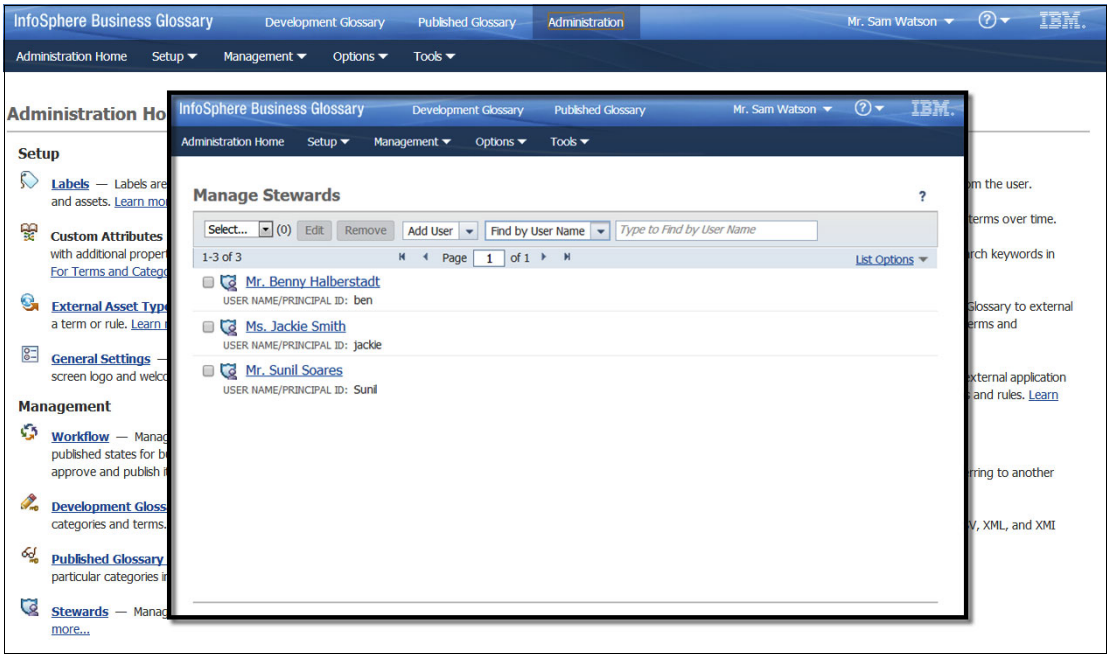


Figure 2-13 Display of Manage Stewards from InfoSphere Information Server Business Glossary

2.5.2 Applying data stewards to information assets

Information assets may easily be explored and managed within InfoSphere Information Server Business Glossary. Information assets reflect the inventory or catalog of key system information, data assets, and information governance procedures of a given enterprise. Users leverage the catalog to understand, explore, and analyze such information to better understand their defined meaning, related specifications and requirements, and structural context and usage. Associating a data steward to an asset adds an additional element defining the owner of the asset who has (or should have) a complete understanding of this information. This scenario is depicted in Figure 2-14 on page 40.

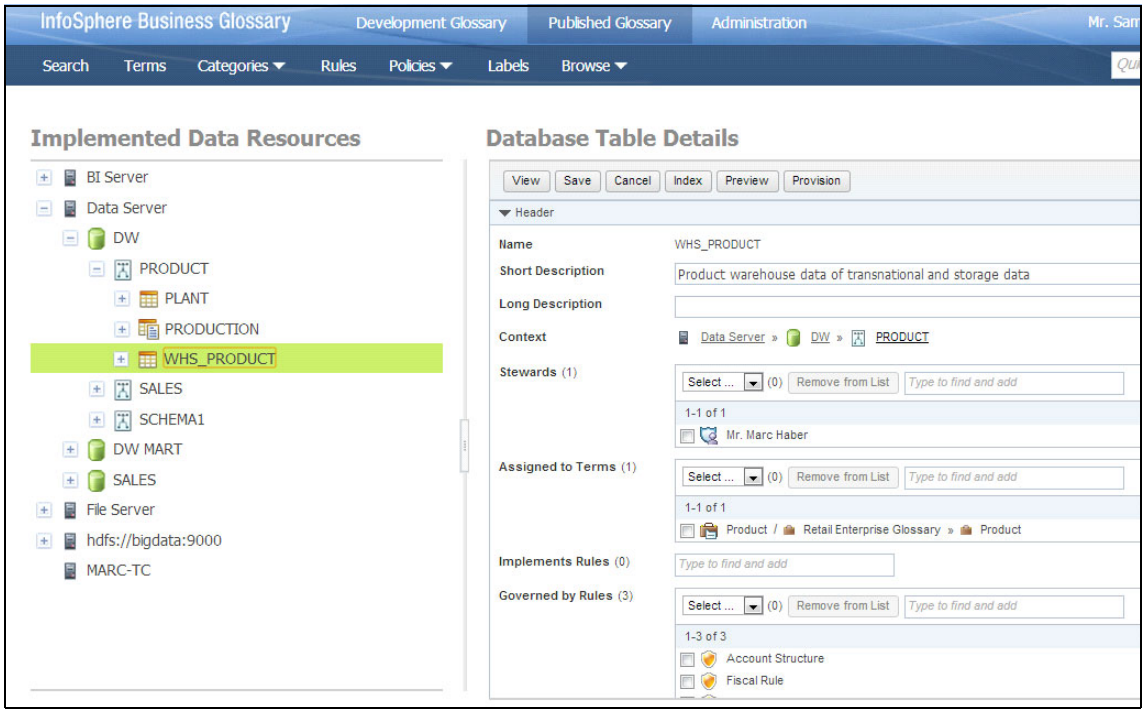


Figure 2-14 Display of Manage Stewards from InfoSphere Information Server Business Glossary

3

Business definitions and policies in IBM InfoSphere Information Server

The building blocks for any integration, quality, or security development effort require having well-defined and understood definitions as well as a clear set of policies, which govern and dictate its implementation and application.

Definitions apply meaning, identification, and context to the processes, routines, and assets. Definitions further provide an easy entry point for searching, cataloging, and understanding information.

Policies help define and supply a broad range of requirements for these processes, routines, and assets. Policies further allow for the declaration of the business requirements for security, privacy, and other such regulations in a clear and easy to understand manner.

This chapter describes the benefits, requirements, and implementation in the development of business definitions and a set of business policies, and the real benefit that they provide within any organization.

The core set of information governance foundation principles are depicted in Figure 3-1.

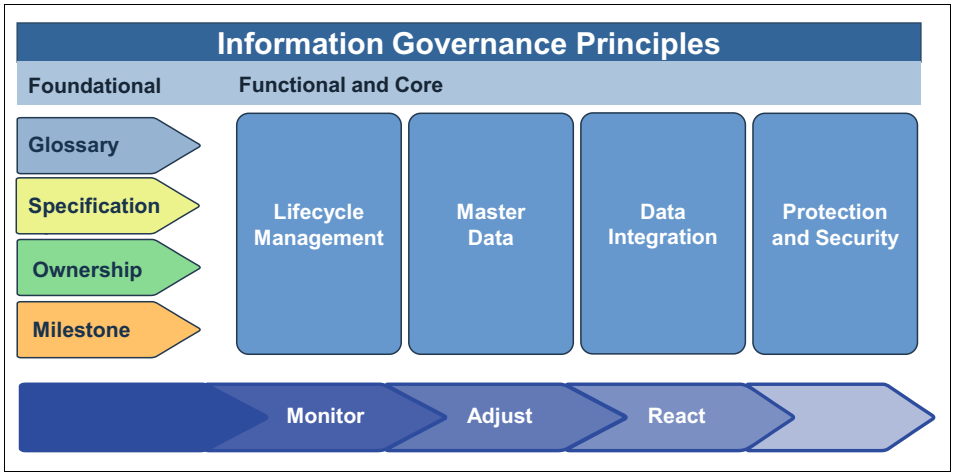


Figure 3-1 Core principles of Data Governance

3.1 Introduction to business definitions

Organizations today are realizing that the establishment of a common set of definitions is a requirement to achieve efficiency, promote communication, streamline, and ensure standardization and compliance. These definitions remove ambiguity within the organization and allow leveraging and sharing the correct understanding and meaning of information.

Business definitions impart an extended understanding and meaning of information and provide greater knowledge and comprehension about the data generated, processes, data stores, and usages to support operations. Through a well-defined taxonomy (hierarchy) and a rich set of formatted business definitions, users clearly benefit in their ability to easily search, browse, and navigate such definitions, their meaning, usage, and applied information assets.

Consider the following scenarios:

- I am being tasked with implementing a data integration effort to identify “High Value Customers”. What is the definition of a “High Value Customer”? Are there any specific instructions for identifying such customers? Additionally, how can I identify other processes, assets or owners which leverage “High Value Customers”?
- I am inspecting a new regulatory requirement and would like to understand the business interpretation and meaning of key terms within this requirement.
- I am reviewing a new set of quality processes and want to confirm the meaning and usage requirements of these processes and inspect the details of those processes previously defined and implemented.

A business definition will offer and promote a singular authoritative and accepted source for the meaning of information and further can provide a rich set of added dimensions. Definitions also help identity and locate and identify the assets and processes found within the organization.

3.2 Introduction to business policies

Business regulations and standards for the continuous operational integration and storage of information are many and transforming. Organizations today are tasked with reacting and adapting to these regulations and standards in the realm of data privacy, master data, standardization, and truth-in-reporting. For organizations to remain agile in their ability to implement these regulations and standards, they must be able to deliver a set of well defined and easily understood requirements to all members of the organization.

Business owners and analysts must posses the ability to publish a catalog of such requirements, in the form of business policies and rules. Business policies and rules, similar to business definitions, are predicated on expressing the precise requirements, compulsory regulations and standards or compliance definitions in simple and easy to understand language. Business policies and rules additionally convey the source of such regulations or standards and the additional details that complete them.

Developers tasked with implementing will benefit from understanding these requirements in greater detail, including their intended purpose and wanted results. Analysts and managers will benefit from a comprehensive view of how such requirements have been implemented and general compliance of information assets.

Consider the following scenarios:

- I am a Data Steward investigating rejected contract reimpressments due to invalid or missing contract numbers. I need to understand the precise requirements for contract numbers, and which processes and assets source such information.
- I have been asked to prepare operational data for testing purposes for a new data integration process at an off-site facility. I need to ensure the privacy and compliance rules are respected, and I could benefit from better understanding the meaning of such rules and their origin.

A set of business policies and rules will provide for a comprehensive and detailed set of regulations and standards that can be shared and benefited across the organization. Such policies and rules can identify the processes implementing them or the assets that they govern. Such policies and rules can also identify the compliance of the organization upon these set of requirements, and ultimately help determine any unnecessary risk or suspect data, and express confidence in its information.

3.3 InfoSphere Information Server

Information Server provides a single unified platform that enables organizations to understand, define, cleanse, transform, and deliver trustworthy and context-rich information. Information Server is a multi-tiered platform, which includes a suite of applications focused on all aspects of the information supply chain or data integration domain. The platform further aligns all objectives of the business and is optimized for connectivity to diverse data sources. Refer to Figure 3-2 on page 45.

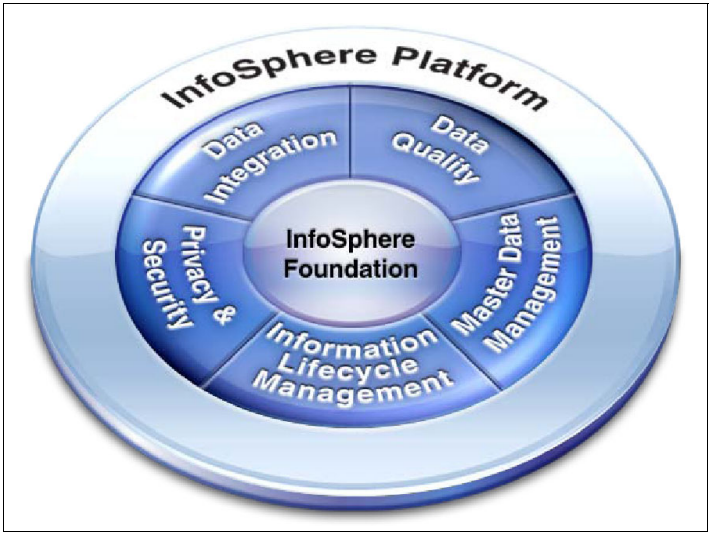


Figure 3-2 InfoSphere Information Server Foundation

3.3.1 IBM InfoSphere Business Glossary

Business Glossary provides for the ability to deliver a solid foundation for Data Governance and integration to the user through Business Categories and Terms, Business Policies and Rules, and Business Labels.

Organizations will enjoy the simplicity of the Business Glossary in its ability to deliver a rich catalog of business information and robust features within a web application. Users may search for a specific definition or policy, understanding their implied meaning and usage. Users may further browse and view the structure of information assets, in addition to benefiting from their implied business definition, required business policy, data owner, and usage within transformation processes or applications. Refer to Figure 3-3 on page 46.

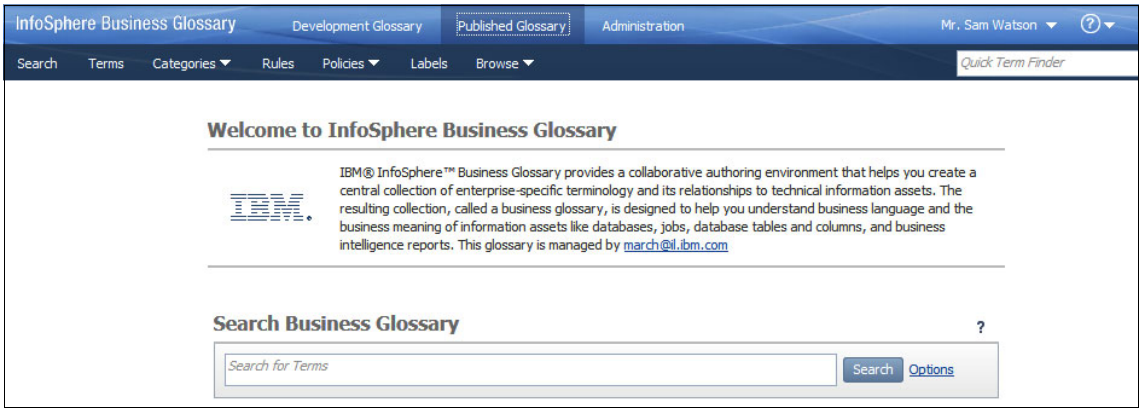


Figure 3-3 IBM InfoSphere Business Glossary Version 9.1.2

Business Terms reflect the definitions and core subject area of the organization and are contained within Business Categories. Categories often lend an added semantic definition to a term, and aide in the grouping together of terms for a particular subject. For example, a category titled *Product Classifiers*, will imply the terms contained within that category define the classification types for a product.

Business Terms additionally benefit from their relationship to Information Assets, such as Database Tables, Business Intelligence Reports, or Logical Data Models. As well, Business Terms also benefit from their relationship to other terms. For example, a term may define generically the concept of an account, however *account* may reference and define more precisely *Commercial Account* and *Individual Account*. Refer to Figure 3-4 on page 47.

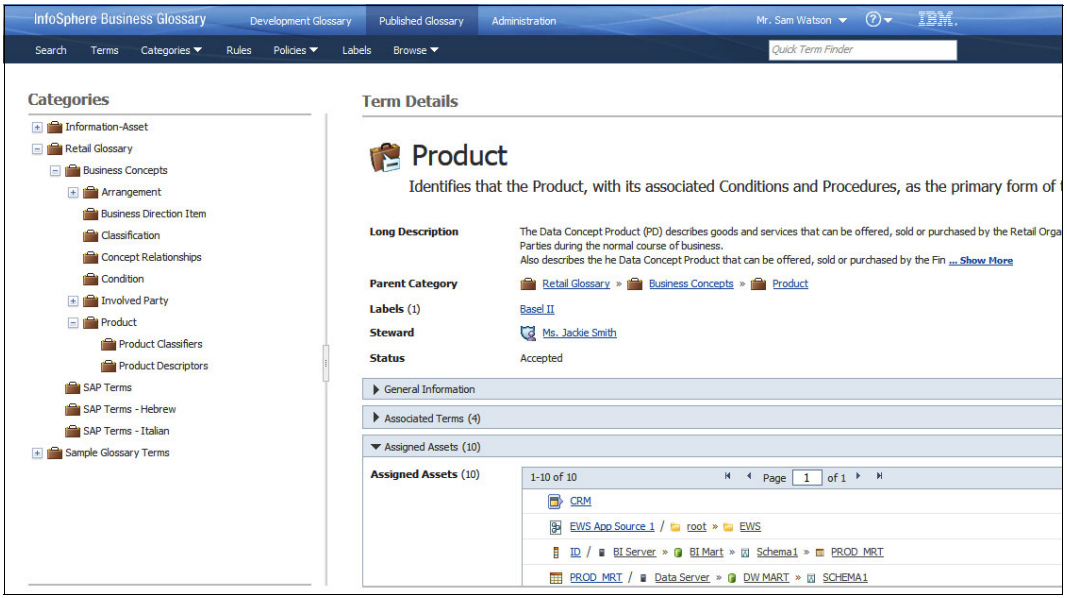


Figure 3-4 Display of business categories and a business term

Business Policies formulate the abstract set of requirements and standards which define the scope and set the guidelines for data integration and data quality routines. Business Policies further reference Business Rules, which reflect and define the specific details and action of the policy.

Business Policies and Business Rules must be specific and exacting, discounting any uncertainty in their definition. They must also be unambiguous, comprehensive, and uniform so that policies and rules may easily be incorporated or applied. For example, a Business Policy may define a set of data quality requirements for customer data, and include Business Rules that will implement such requirements.

Business Rules benefit from their ability to link with the Information Assets on which the rules must be applied, such as Database Columns or Application Systems. Business Rules also benefit from their ability to link with the processes or runtime procedures, which implement the Business Rules. Refer to Figure 3-5 on page 48.

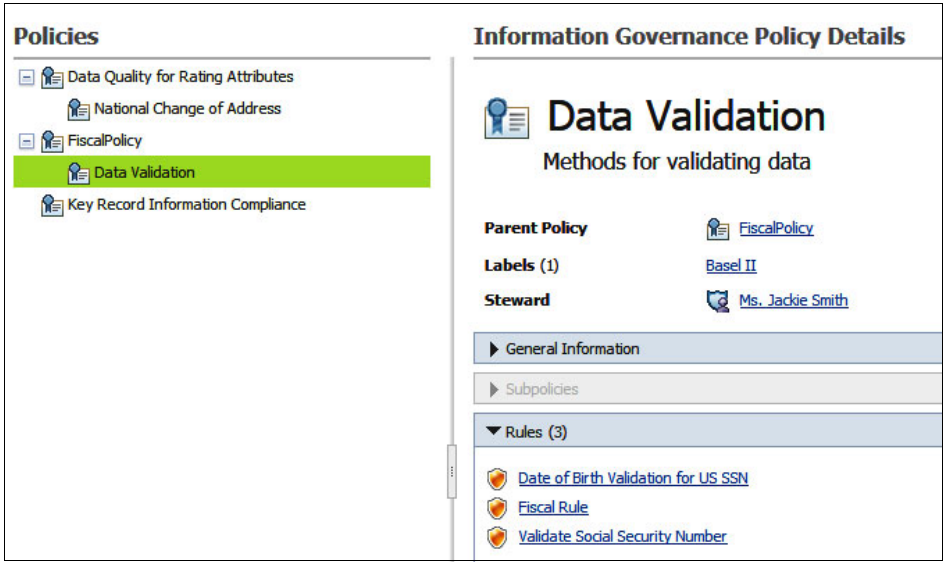


Figure 3-5 Display of business policies and a business rule

In this chapter, you learn the process for creating and managing Business Categories, Business Terms, Business Policies, Business Rules, and Business Labels within IBM InfoSphere Business Glossary. As well, you will understand the process for best developing such information and the persons involved in such a process and how to make use of and incorporate existing data dictionaries or data standards.

3.3.2 Business Glossary subject material and council

Business Glossary is designed to provide a common and consumable vocabulary to facilitate communication and understanding throughout the organization. Business Glossary is also a hub for the implementation and application of data governance principles and their requirements.

As such, Business Glossary offers a collaborative environment for the drafting and maintaining of information. Formal governance initiatives require a process for the authoring and management of Business Categories, Business Terms, Business Policies, and Business Rules. This process allows for the review and approval of information before its ultimate publication and visibility within the organization.

Domain or subject-matter experts will often possess the knowledge required in defining of terms and rules. These individuals further are considered the owners

of such information, whose understanding of these concepts and their definitions are critical in capturing and publishing their meaning and usage requirements.

Before drafting and publishing information within Business Glossary, the formulation of a comprehensive Data Governance process, as that depicted below, is a critical first step. This process, in particular, defines and outlines the tasks, owners, and expected results for such information. Refer to Figure 3-6.

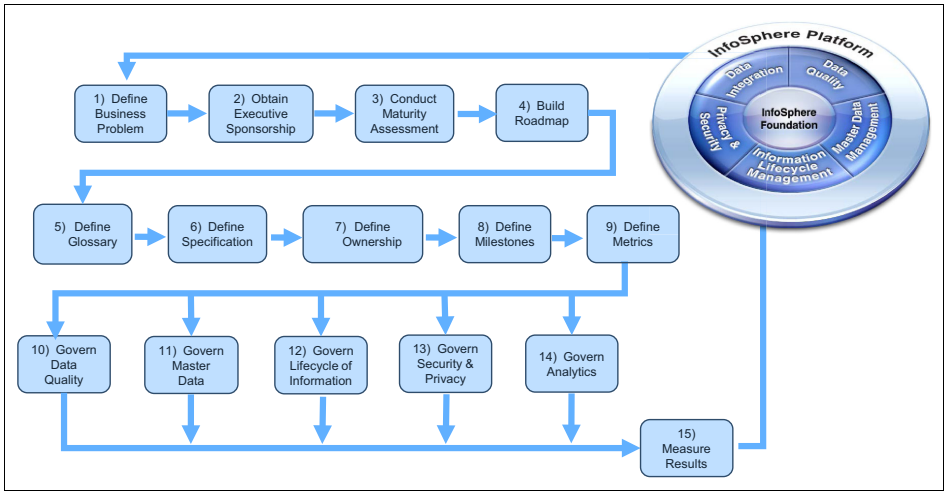


Figure 3-6 Data Governance process defining business definitions and policies

3.3.3 Managing Business Glossary categories and terms

Knowing where to begin is often the most difficult and monumental challenge. Organizations are typically as diverse as they are complex. The individual segments and divisions within the organization each will want to participate and benefit from a Business Glossary initiative. After all, it is the Business Glossary that will deliver to the benefit of the entire organization a clearly defined and well understood set of Business Categories and Business Terms in a single tool.

General guidelines dictate initially concentrating upon a single high-value information area or scope within a data integration implementation. For example, a new Data Warehouse initiative or a heavily regulated area within a data store with changing requirements and lax governance insight.

Set the focus on the semantic business definitions, and borrow from the language of the business, in the form of Logical or Business Intelligence models. Leverage existing lexicons or industry standards. Poll and gather real-time

understanding for how such concepts and definitions are currently used or applied.

Establish benefit and garner interest within a selected and focused area of the organization. Demonstrate the richness of a Business Term in that it includes relationships to Information Assets, links to external specifications or documentation, and details specific to the application and usage of such a term. Adoption across the organization may not be immediate, however it will follow as users realize the benefits of a singular portal for defined information.

Finally, develop milestones for the establishment and publication of Business Categories and Business Terms, including the correct set of authors, approvers, and publishers. Authors define the domain or subject-matter expert who will draft information and the details of such information. Approvers review, comment, and accept or reject the drafted information. Finally, the publishers will make available the approved information for everyone within Business Glossary.

To support such a glossary approval process, Business Glossary is seen as having a Development Glossary area for the drafting, approval, and publication of information, and a Published Glossary area for the read-only display of information. Refer to Figure 3-7.

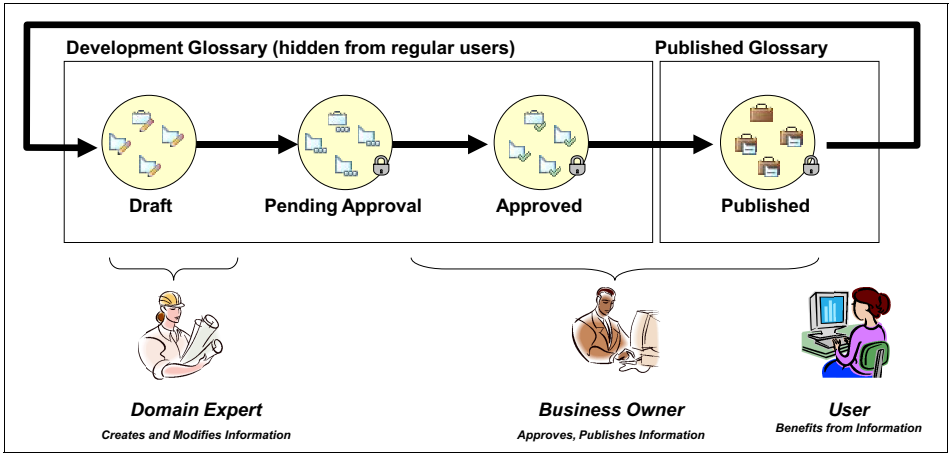


Figure 3-7 Business Glossary workflow process for the drafting, approval, and publishing of information

Creating a new Business Category or Business Term is a simple process. It is equally as simple to modify and perfect information based on feedback or changing requirements. Therefore, the process is iterative and the definitions of terms are continually reviewed and adapted.

First, begin with creating a set of hierarchal Business Categories to reflect the containers of the terms to be created. Each category should be unique and reflect a distinction in its definition.

Secondly, create a set of terms that must appear within a Business Category. It is good practice to additionally include a description, status, and example details. With the maturity of the Business Glossary, the term can be modified to include relationships to Information Assets, other Terms, Business Labels, or a Data Steward.

When the draft of categories and terms are complete, initiate the approval and publication process so that the information is available to all users. Finally, institute a further review process with key business owners and stakeholders, gathering any feedback and understanding the key benefits that they may have received. Refer to Figure 3-8.

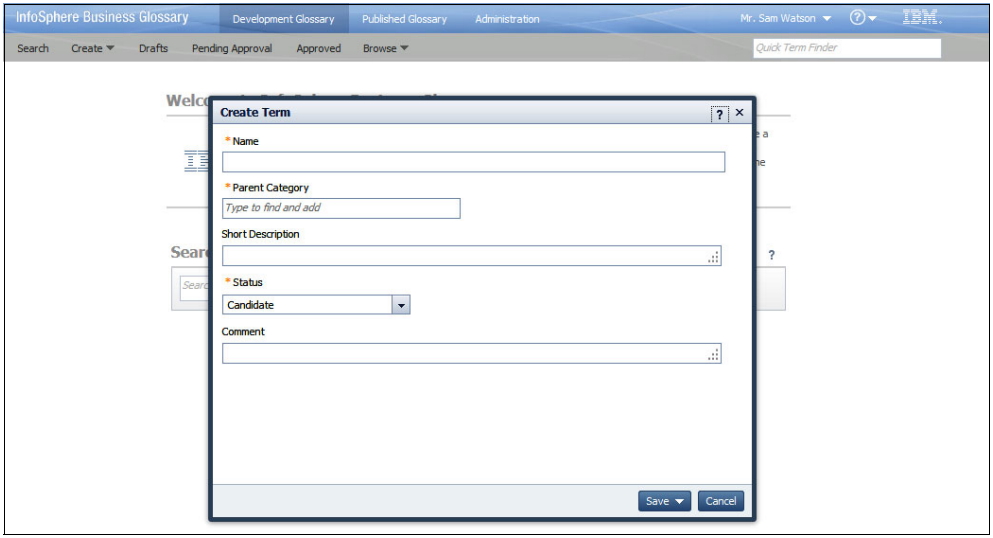


Figure 3-8 Initial creation process for a Business Glossary term

3.3.4 Managing Business Glossary Policies and Rules

Business Policies reflect the standards and requirements that an organization must adhere to, the functions and practices that apply those standards, and the responsibility to comply with them as well. Policies define and broaden the understanding and knowledge of the standards as they relate to the assets and process of the Information Supply Chain. For example, the data privacy, data cleansing, or data masking requirements that must be applied.

Business Policies further allow the business owners and compliance officers to declare the intended behavior and usage of information. Business Policies leverage existing Business Terms in helping define the precise scope and definition, and leverage Information Assets to communicate their precise governance lifecycle.

Business Policies contain and reference the Business Rules which implement the requirements and standards defined by the policy. Therefore, a hierarchy of Business Policies may be created to express the diverse and multitude of standards and requirements, where each policy imparts additional identification information and a set of contained rules.

Business Rules assert the definitions and constraints for the structure of information and, which are applied to assets of the Information Supply Chain. Data Governance requires that Business Rules are clearly written, defined, and approved, as well as have a process for applying such rules to Information Assets, reviewing the rule results, and reacting accordingly.

Business Policies and Business Rules further convey a broad set of standards and requirements that affect organization objectives, operations, and direction. These policies additionally lay down a response for given situations, circumstances, and rule failures. Policies also determine the formulation and implementation of the strategy, which directs the plans, decisions, and actions to achieve a defined set of objectives.

Creating a new Business Policy or Business Rule is a simple process. It is equally as simple to modify and update information based on changing and evolving requirements. Therefore, the process is iterative and policies and rules must continually be reviewed and adapted.

Start with creating a set of hierarchal Business Policies reflecting the domains and subject of the standards and requirements. Each policy should be unique and reflect a distinction in its definition, usage, and set of contained or referenced rules. Refer to Figure 3-9 on page 53.

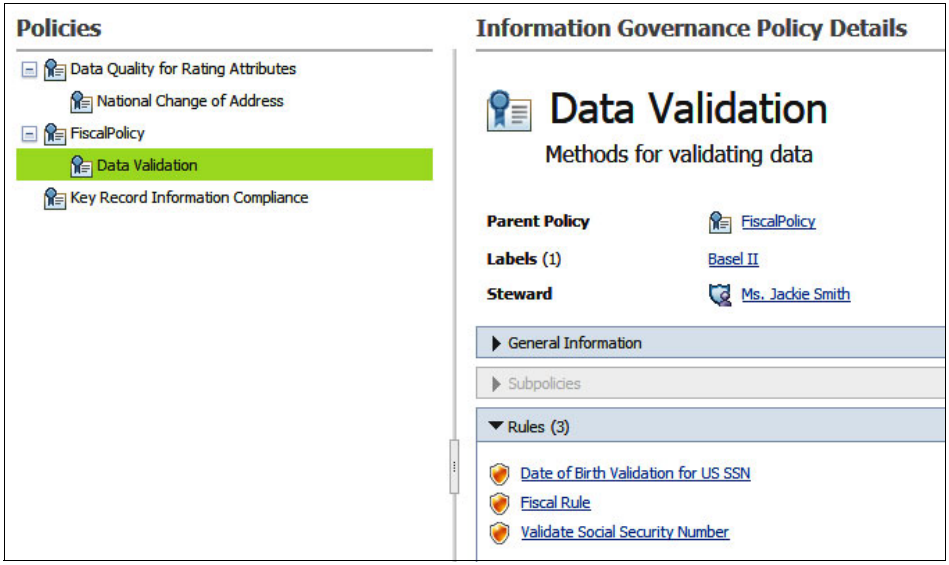


Figure 3-9 Display of business policies and a business rule

Next, create a set of Business Rules which should appear within a Business Policy. It is good practice to additionally include a description, owner, and link to the process and application used to implement the rule, or the Information Assets on which the rule is applied.

When the process of creating new or modifying policies and rules is complete, initiate the approval and publication process so that the information is available to all users. Finally, institute a further review process with key users, business owners, and stakeholders, gathering any feedback and understanding the key benefits received. Refer to Figure 3-10 on page 54.

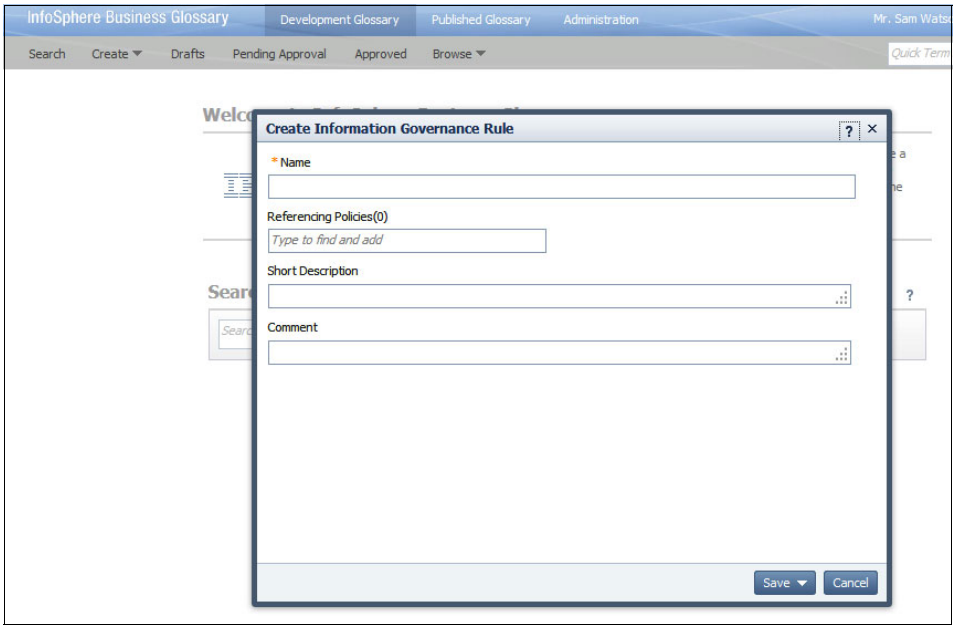


Figure 3-10 Initial creation process for a Business Glossary policy

3.3.5 Managing Business Labels

Without an ability to easily identity and locate information, the definitions and policies developed and made available by any organization will fail in their adoption and general usage. Ultimately, if such definitions and policies do not gain acceptance and are not fully developed nor made an integral part of any process, the efforts to develop and maintain such definitions and policies will cease.

Therefore, the enablement of users to gain access to information is as critical in the Data Governance plan as are the publication of definitions and rules.

Business Labels satisfy these requirements in their ability to tag and associate subject areas for information assets or process. Labels can denote geographies, divisions of an organization, integration projects, or applied business standards. Information assets may include multiple labels, allowing for flexibility in their application and the extended understanding of such assets. For example, a set of tables within an operational-data-store may be shared across multiple projects, and include a reference to a business standard.

Business Labels may easily be created and managed, in addition to the process of associating such labels with information assets. Users may leverage such added descriptive information in searching and identifying specific assets.

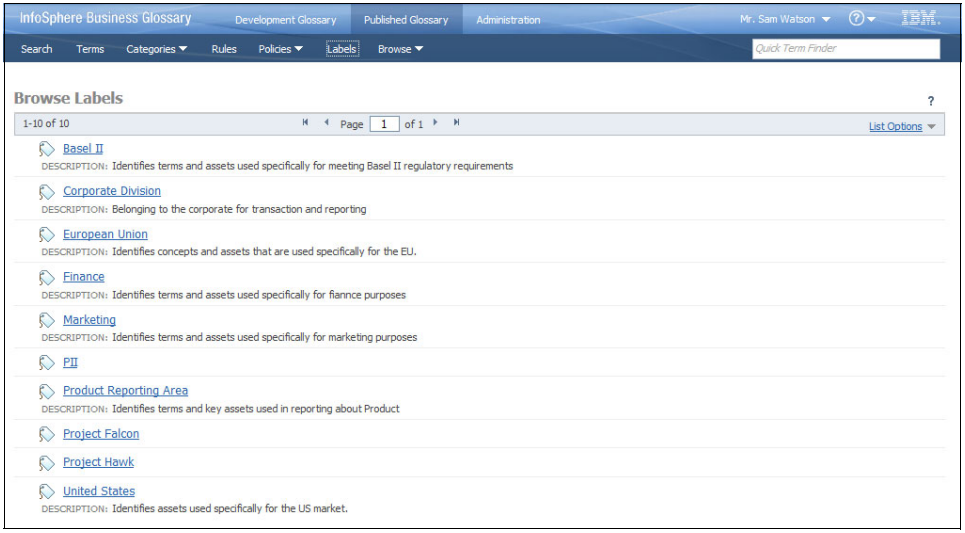


Figure 3-11 Display of Business Labels

3.4 Benefit and value

A comprehensive set of definitions and policies benefit an entire organization that needs to reference, understand, and analyze the information assets and processes, which that organization generates and contains. Users may consult and use such definitions and policies for understanding the structure and meaning of information in real-word terms, in addition to exploring the requirements for the usage and integration of such information.

IBM InfoSphere Information Server as a platform, and IBM InfoSphere Business Glossary allows for the easy navigation of Information Assets which are categorized by Business Categories and Terms and structured by Business Policies and Business Rules in addition to being tagged with Business Labels.

Developers will understand the intent of the requirement and definition of information, leading to agility and streamlined integration processes.

Business owners will gain confidence in their information and will be satisfied in their ability to publish and make known the data requirements and project standards.

A core principle of Data Governance includes defining key concepts and terms, and publishing known requirements. These principles lead to more informed business decisions, reduced operational inefficiencies, and adoption of the business standards throughout the organization.

4

Workflows and business specifications

The world is experiencing unprecedented growth in the sheer volume of data that is continuously generated. Managing this growth and ensuing competency in the data translates into a comprehensive need for the administration of the processes that create, store, use, and distribute information. Without such processes and planning in the form of a specification document or integration workflow, erroneous data and errors in analytics could lead to incorrect strategies or business decisions.

Workflows depict an abstract and broad overview that annotates an integration process or implementation project, highlighting their primary components and systems. Specifications, however, detail the individual components and subcomponents in precise detail. Specifications and workflows both must reference the absolute set of requirements and standards, and include data quality metrics and reference to master data strategies.

This chapter describes the benefits, requirements, and implementation in the development of Workflows and Business Specifications and the real benefit they provide within any organization.

Workflows and Specifications are included in the core set of foundation principles. They describe the acquisition, transformation, and movement of data

within the Information Supply Chain. An example of such a process is depicted in Figure 4-1.

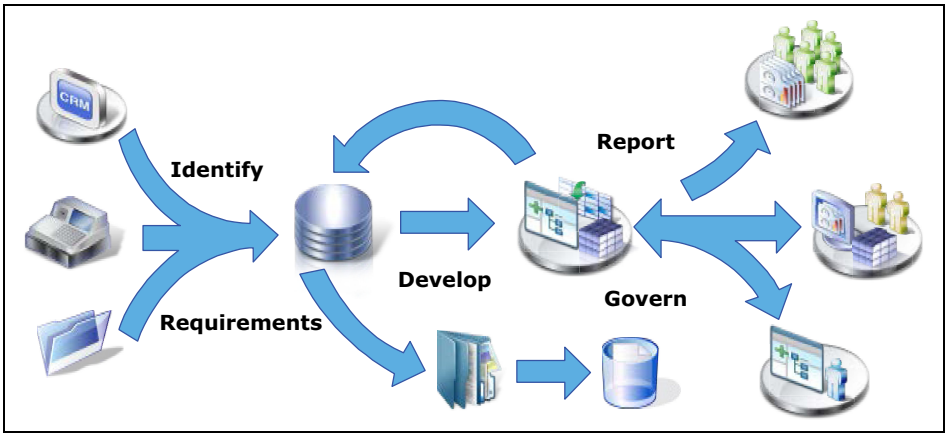


Figure 4-1 Example workflow

4.1 Introduction to Workflows

In the context here, workflow means an abstract data flow, depicting the extraction and load of data between separate source and storage systems. Workflows further identify where the data quality standardization, data privacy concerns, master data, or other elements of the Data Governance set of requirements must be applied. Workflows, therefore, are seen as a visual design plan and a point of reference by which to measure the implementation of the plan.

Additionally, workflows include a reference to the methodologies, tasks, owners, and milestones associated with the components found within them. Workflows further must be complete and include all key components. For any process or flow, there must a corresponding description of the elements involved and their subsequent operation or usage. A sample workflow is depicted in Figure 4-2.

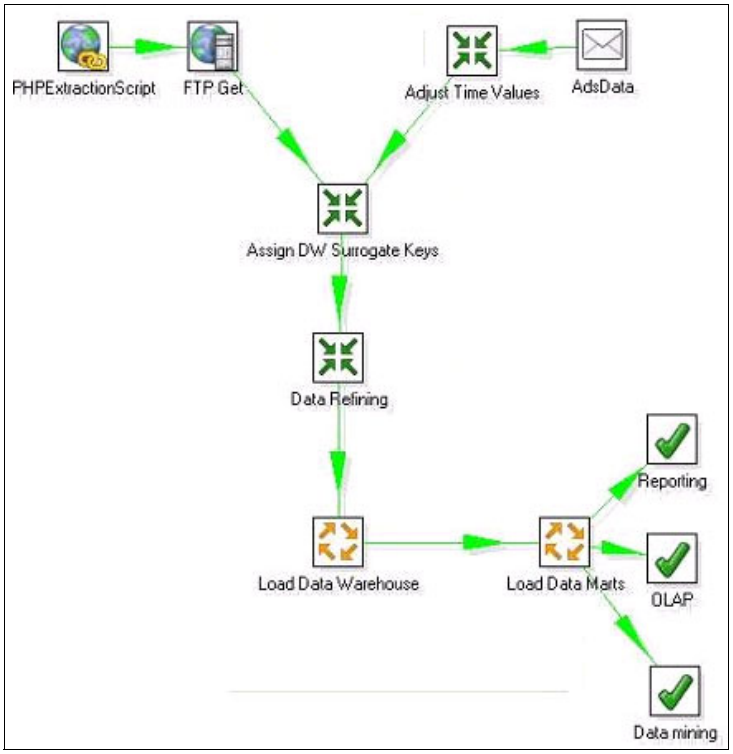


Figure 4-2 Sample workflow depicting an information flow

In this manner, workflows are able to fully support the data quality and regulatory requirements, as well as fully document and make known the data systems and applications where information is created or generated and the services and applications that consume such data. The documentation of these systems allows for the clear application of Business Definitions and Business Policies and recognition of the data lineage and analysis, reflecting confidence in the data.

Organizations further develop workflows to provide business users with information for their decision-making process and planning, and developers for implementing such processes. These solutions will be modified by the business as they continue to grow and gather additional requirements, and easily adapted by the developers because the workflow helps to clearly identify and communicate these changes.

Workflows are said to describe the business need and their objectives, and this chapter provides you with an understanding of how to develop and distribute such workflows.

4.2 Introduction to business specifications

The specification includes the exacting requirements to build and deliver an integrated solution that incorporates the foundation principles of the Data Governance solution. The process for drafting such specifications must be inclusive, allowing input from the business owners and compliance officers, and must be shared, giving insight to everyone within the organization.

Specifications detail the precise and explicit information about the requirements for the design and implementation of a definitive process or data item, such as how shall the system behave or how shall data be transformed. They also must additionally express the intended usage, dimension, and quality factors accurately describing the conceptual implementation of the requirement. A solid design is absolute to ensure compliance with the stated objective, requirement, and standards. Specifications must precede any development, as void of such it would not be possible to deliver a clear understanding of the scope and requirements and ensure their compliance.

The requirements included by the specification are a high-level description of the functional design, planned implementation and expected results. Organizations

realize in order to gain a better grip on how information is managed, created, used, and distributed throughout, specifications detailing this functional design, and the people and processes included within, are a prerequisite. Therefore, specifications can also be said to address the quality and reliability of the information that is used throughout an organization,

Additionally, it is required that specifications reference a common and shared understanding of the business definitions and policies, using a common and well-defined business language. Specifications must further directly relate to a shared set of technical assets, which will be developed upon, analyzed, and profiled. In this manner, specifications deliver a clear set of instruction, requirement, and the business intent.

4.3 InfoSphere Information Server

IBM InfoSphere Information Server supports the data governance initiatives and requirements as a solid platform for the documentation of requirements and standards, implementation of the supporting processes and insight and visibility of the continual integration and development efforts. This is depicted in Figure 4-3 on page 62.

InfoSphere Blueprint Director is a component of Information Server that allows for the declaration of the business and technical vision, requirements, and integration, in the form of workflows.

InfoSphere FastTrack is a component of Information Server and defines and captures the precise design requirements and transformation logic in the form of specifications.

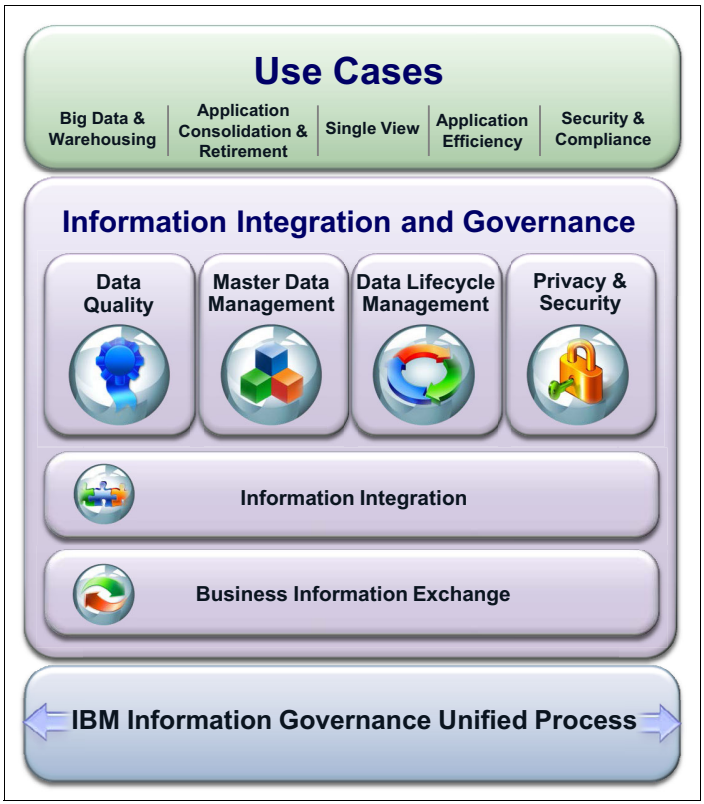


Figure 4-3 IBM Information Server information integration

4.3.1 InfoSphere Blueprint Director

InfoSphere Blueprint Director is a graphical design tool that allows for the creating and sharing of abstract information flows or implementation plans, including reference of the data governance, data integration, and data quality requirements and initiatives. These flows or plans are collectively known as *blueprints*, similar to those more familiar blueprints used when designing a building and detailing the building-blocks of the construction process.

Blueprints help deliver the reference architecture and a visual information landscape. Blueprints are unique in their capability to express and share the exacting requirements and associated methods, tasks, and processes. Blueprints must remain a living document, where they are continually revised to incorporate changes in the requirements or implementation.

The blueprint captures and delivers the following important factors:

- The consumers of information
- The core information assets, those applications and analytics consuming the data, and those source data systems
- The compliance and regulatory requirements for information assets and processes

The blueprint ultimately delivers and conveys the precise requirements of the information landscape to the organization. An example of the InfoSphere Blueprint Director is depicted in Figure 4-4.

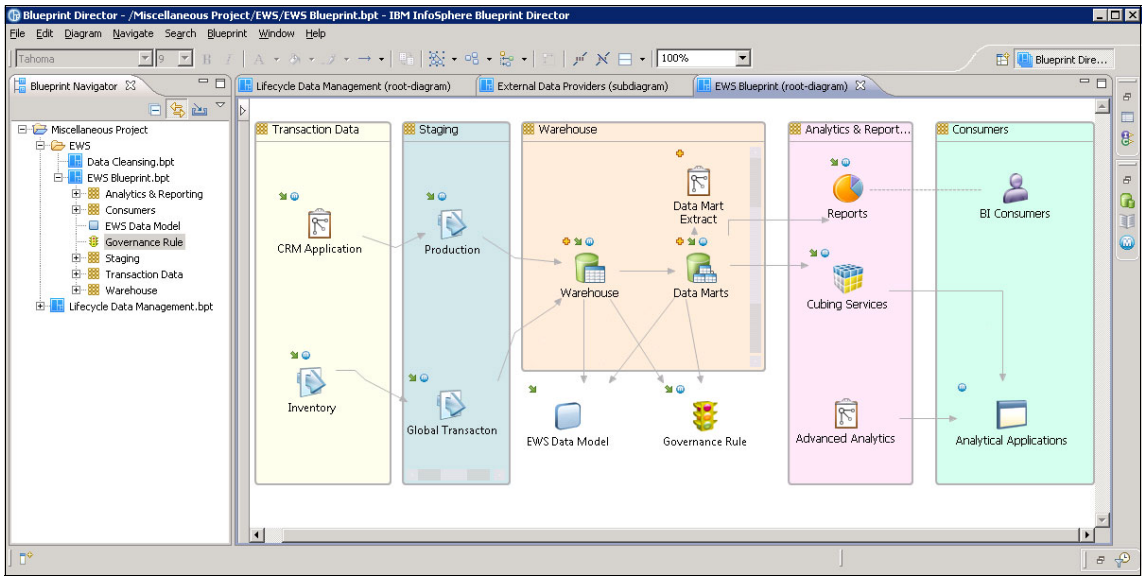


Figure 4-4 InfoSphere Blueprint Director design canvas

InfoSphere Blueprint Director is a component of InfoSphere Information Server and can be installed as a stand-alone desktop application and configured to share information through InfoSphere Business Glossary and InfoSphere Metadata Workbench.

Creating and managing blueprints

Selecting an existing template is the easiest and simplest way to begin the process of documenting and sharing the information flow. Templates save time in the planning and execution while including reference to the main components that are to be documented.

Templates are available for the following:

- Information Lifecycle Management
- Business Driven Development
- Managed Data Cleansing
- Delivering Trusted Master Data

Blueprint Director includes several ready-to-use content templates that can be easily customized to fit any project requirements. Alternatively, a new blueprint may be created to more precisely match the information project requirements. Additionally, Blueprint Director comprises a unique graphical design canvas onto which the components representing the processes, tasks, data sources, or other objects may be depicted, annotated, and joined. Objects within a design canvas include a label indicating their purpose or usage as well as a link to a method definition, development task, project milestone, or physical asset represented within the InfoSphere Information Server. The Blueprint Director design palette and canvas is depicted in Figure 4-5 on page 65.

Furthermore, blueprints support a hierarchical view, allowing nested levels of processes and subprocesses depicting greater detail of their requirement. Once a template has been selected, it is displayed and ready for use within the design canvas. Using the design palette a user may select from a list of components to modify and enrich the blueprint. Each component has a specific purpose and reflects a data source, application, or process. Components can be dragged and dropped onto the canvas, illustrating the objects or hops within the information flow diagram. Connection components are used to depict the flow of data, whereas group components allow for documenting greater detail and information within subdiagrams and referenced blueprints.

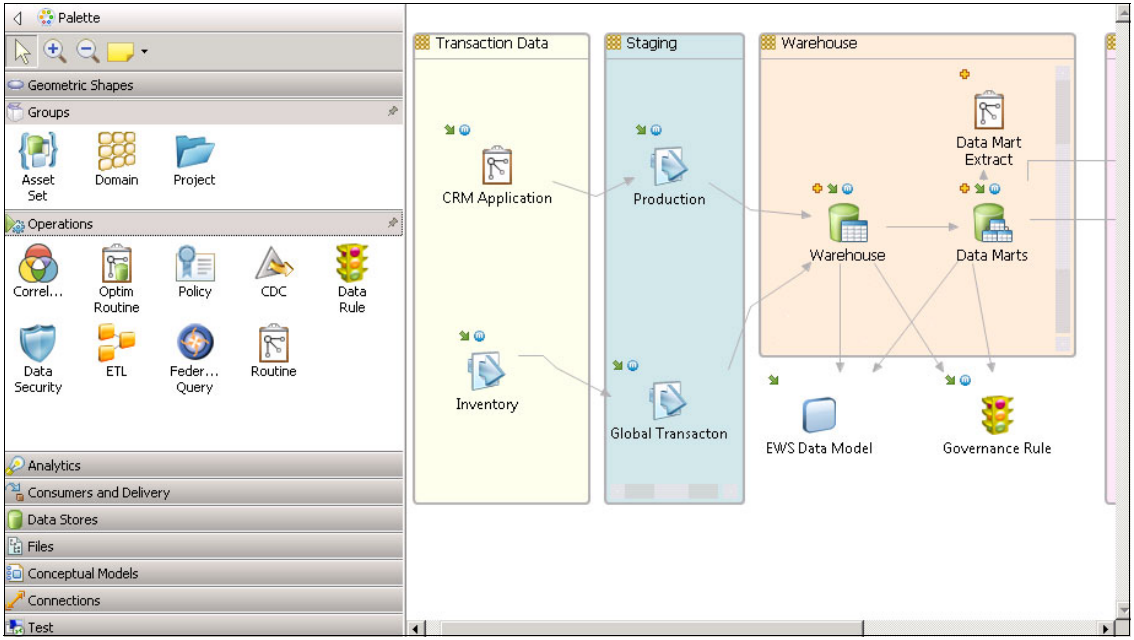


Figure 4-5 Blueprint Director design palette and canvas editor

To ensure effective communication and clarity between the different persons using a Blueprint, is important to ensure that the correct terminology is used within the project initiative. Therefore, when specifying the key data elements and processes within a blueprint, reference the shared Business Definitions and Business Policies that have been previously created linking those definitions and policies to the objects within the blueprint. As well, link to the Shared Information Assets previously created or imported into IBM InfoSphere Information Server.

Drag a Business Glossary Term from the Glossary Browser onto an object within the design canvas, creating an association between the business definition and the implied meaning and semantic of the component. Similarly drag an Asset from the metadata browser onto an object as well. This is depicted in Figure 4-6 on page 66.

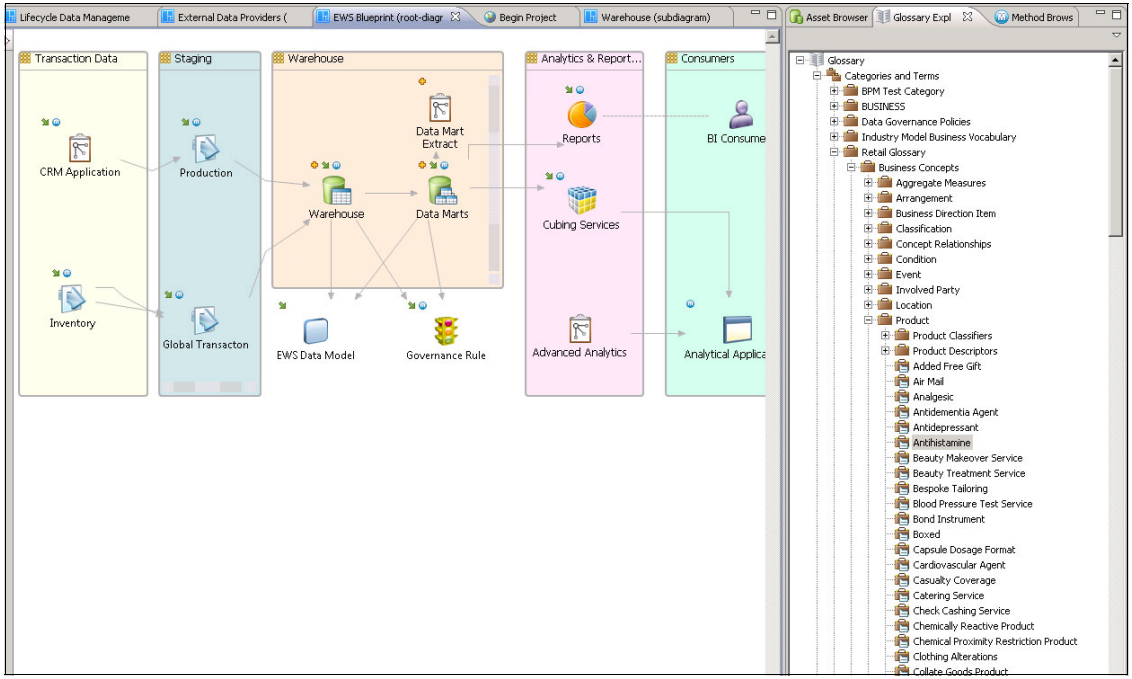


Figure 4-6 Blueprint Director glossary term selection and viewer

Blueprint Methodologies and Task Management

Methods describe the work process and condition for the usage or deployment of a system, application, or data source. Methods further characterize the roles, tasks, and artifacts that are required for the implementation of the given method, requirement, or standard.

Methods expose key phases, capability patterns, and activities for a selected project. This view provides a high-level overview and guidance for the required steps of a particular project or implementation. Users have access to the corresponding methods in a hierarchical view for high-level phases and activities, plus detailed descriptions. This is depicted in Figure 4-7 on page 67.

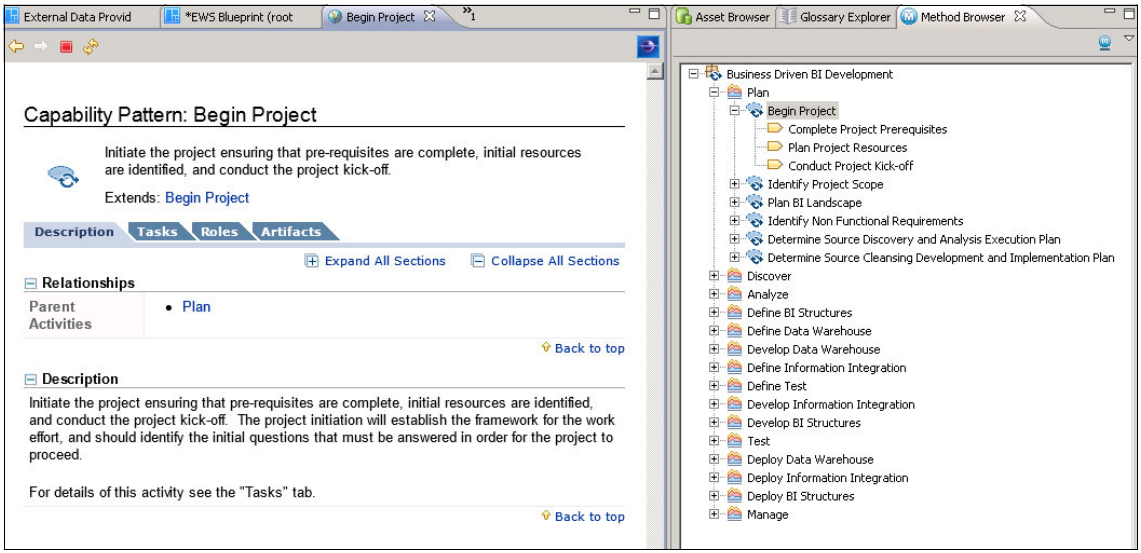


Figure 4-7 Blueprint Director method selection and display

Methods further include a brief description of the tasks that are required to implement and satisfy the methodology. Tasks reflect the individual steps or activities that must be accomplished, define the ownership and the relative ordering of such task, defining the intended workflow of the method. Tasks can further be managed and assigned to their owners within IBM Rational® Team Concert™, which can be integrated with InfoSphere Information Server Blueprint Director for this purpose. This is depicted in Figure 4-8.

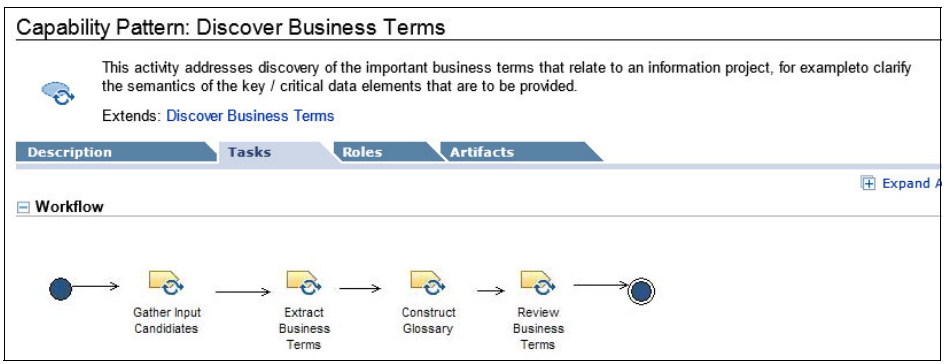


Figure 4-8 Blueprint Director task definition and workflow

The Method Browser visualizes a methodology in the context of a blueprint. Drag a selected method from the Method Browser onto the design canvas to associate the method with an object of the blueprint.

Publishing and sharing blueprints

When a blueprint has been completed and reviewed, share the blueprint across the organization, conveying a high-level overview and guidance for the required steps of a particular project or implementation.

Published blueprints are shared to the Information Server and may be reviewed within IBM InfoSphere Business Glossary or Metadata Workbench. The summary view includes an overview image from the blueprint as well as the name, description, and version included when publishing the blueprint. Administrators may further enrich and annotate blueprints in linking them to Business Labels, Data Stewards, Glossary Terms, and Governance Rules. This added information imparts additional insight for the blueprint.

Users can further view in detail the blueprint, and the specific information that is associated with its components. This information contains the descriptions, methodologies, milestones, and links to the physical assets that they represent. This is depicted in Figure 4-9.

Figure 4-9 Interactive view of a blueprint seen from IBM InfoSphere Metadata Workbench

Ultimately, through the publications of blueprints, developers will now benefit from their ability to review, understand, and adopt the methods and requirements that are contained within.

4.3.2 InfoSphere FastTrack

IBM InfoSphere FastTrack increases the ease and efficiency by which one creates and distributes specifications thus accelerating the development cycle. From such specifications it is possible to generate IBM InfoSphere DataStage® extract, transform, and load (ETL) jobs.

FastTrack further delivers a collaborative development environment for documenting and tracking business requirements, optimizing data integration development, increasing productivity, and maximizing business value.

Within FastTrack, you can create and manage mapping projects for the declaration of mapping specifications, which contain the business logic and function for the extraction, lookup, and transformation of data. These mapping projects strengthen and support the Data Governance initiatives through the deliverance of the business objective implementation requirements. An example is depicted in Figure 4-10 on page 70.

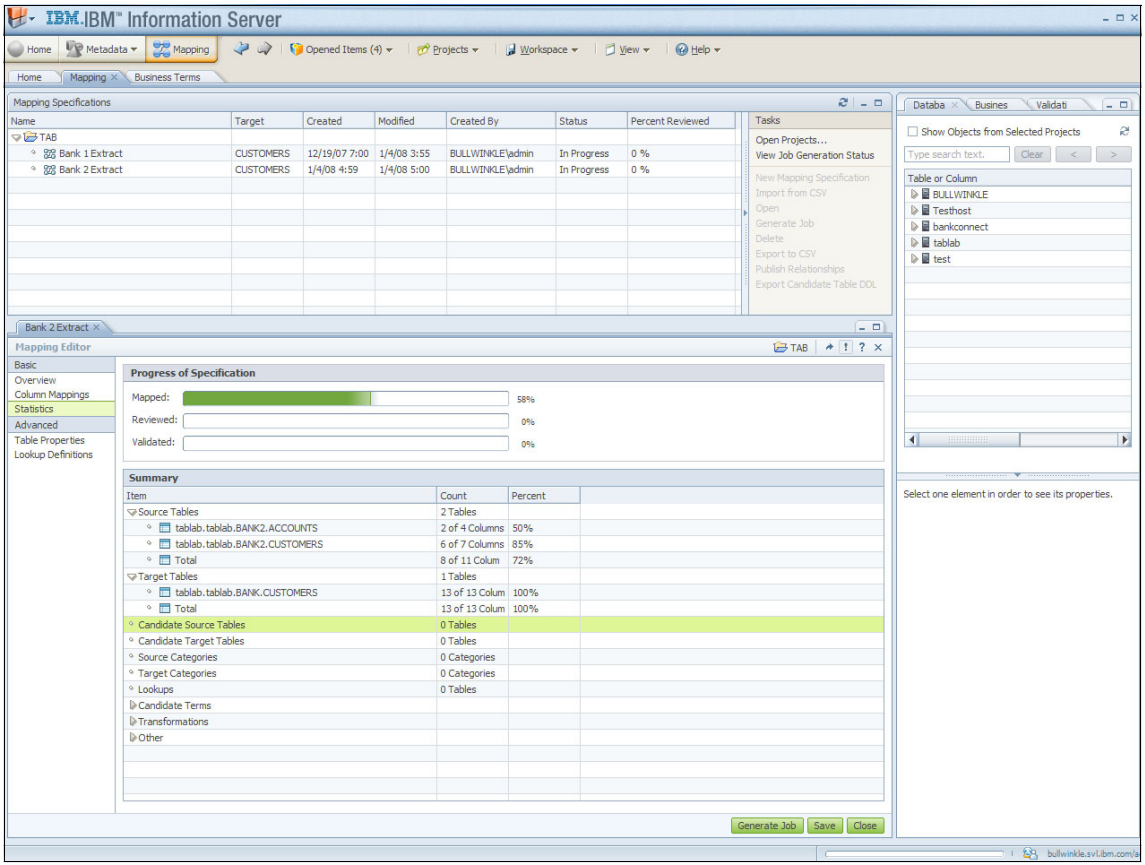


Figure 4-10 IBM InfoSphere FastTrack mapping specification

Creating and managing mapping specifications

Creating a mapping specification is a simple process and documents the source-to-target flow, leveraging existing shared information assets, glossary terms, and system information. Mapping specifications further document and annotate the business requirements and data standards.

Every mapping specification must reference either a source or target asset, or both, in the form of database columns. A mapping may declare the flow of multiple source columns to a single target, for example the concatenation of given name and surname into full name, or the generation of a target column, for example a system date or lookup value.

Source and target assets can be searched within the Metadata Asset Browser and dragged onto the mapping canvas. The Metadata Asset Browser reflects those Information Assets available for display within the Information Server.

Alternately, searching upon Glossary Terms allow for the declaration of candidate source or target assets, or discover source or target columns that are based on term definitions and usage patterns.

Mappings can be further refined by declaring transformation rules, expressions, status, and definitions. These annotations reflect the complete set of requirements and standards for the inclusion and usage of information within the data integration processes. Rules reflect the general declaration of the requirement, standard, or process, in language that is understood by the developer. Whereas expressions reflect a more precise formula for the transformation or calculation of information. This is depicted in Figure 4-11 on page 72.

When mapping specifications are complete, they may be viewed within IBM InfoSphere Metadata Workbench thus more quickly delivering such specifications and requirements to the developers and others within the organization.

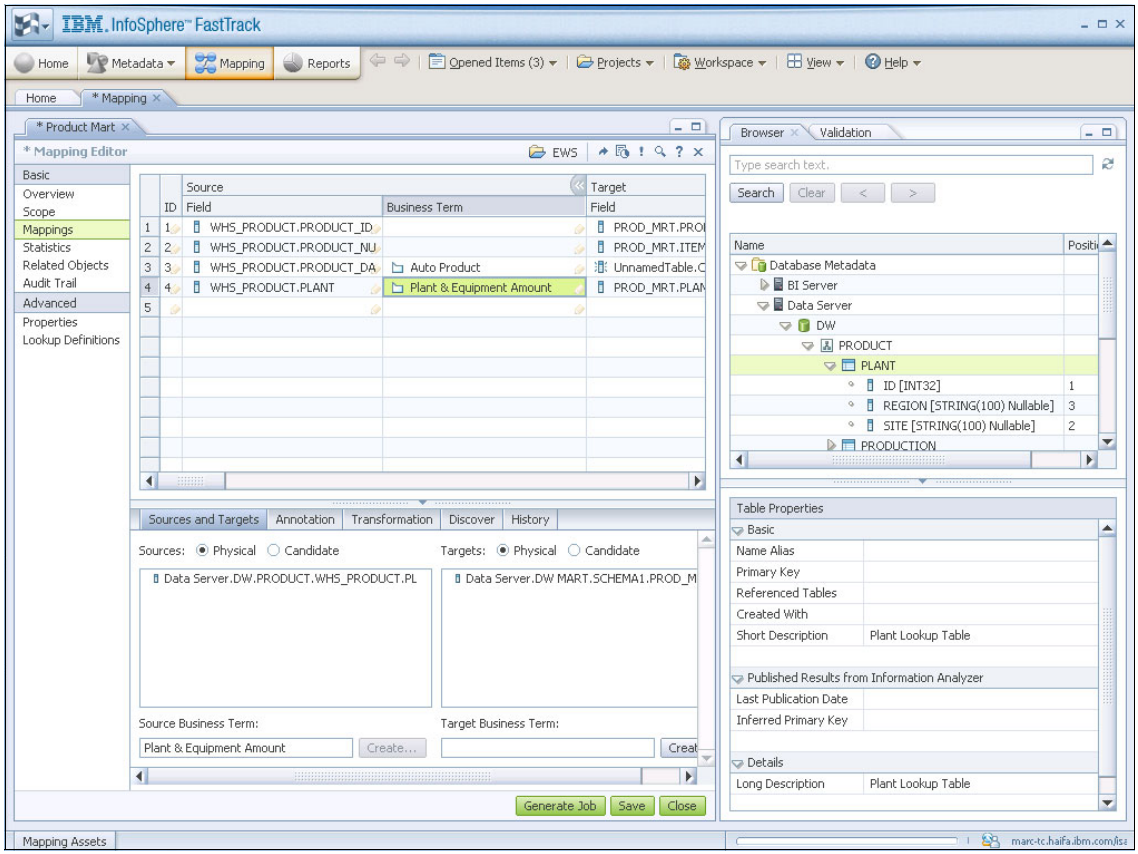


Figure 4-11 FastTrack mapping specification design

4.4 Benefit and value

A comprehensive set of workflows and specifications benefit an entire organization, which needs to deliver and implement integration projects for the collection, standardization, and delivery of information. Business owners can fully describe and make known the intent, requirements, and expectations of data, while developers can easily search, understand, and incorporate such requirements and standards into the data integration processes and ETL jobs. In this way, the workflows and specifications deliver transparency of the data requirements and faster time to value.

The core principles of Data Governance include drafting workflows and specification for the handling of data, defining the compliance and standardization rules as well as the strategies for master data. These principles

lead to competence in the data sources, trust in the analytical reports and applications and better regulatory compliance.

Ultimately, workflows and specifications are among core building blocks and foundation principles of any Data Governance initiative.

5

Metrics and measurements

When it comes to metrics and measurements one size does not fit all. Information Governance Owners, Stewards, and Custodians have very different profiles, interests, and perspectives. Some are policy-centric, some are rule-centric, some need summaries, and others need details. Even the volume, frequency, granularity visualizations, and interpretations may differ. Owners and Custodians are going to see the same data and react very differently. An enforcement rating of 90% might be a cause for celebration while others may be very concerned with an exception rate of 10%.

Information Governance Metrics need to be business policy centric. That is, the stated objectives and goals of a governance program should in some way be represented as or connected to existing or new policies against which measurable criteria and results of achievement can be plotted. This is a key awareness that warrants repeating. If I have high-level objectives within my organization-wide governance program that cannot in some way be mapped to specific policies and rules (and owners or actors responsible for following same), producing any valuable measurements of success will be challenged.

The anatomy of a business policy consists of layers of policies and rules. A single business policy might consist of multiple subpolicies. Each policy consists of one or more business and technical rules. Some typical policy-based questions would include:

- Does a policy exist? Who owns it? Who uses it?
- Is a steward assigned to this policy?

- What is the definition of the policy, and terms used to define the policies, and rules?
- What is the composition of the policy, that is, what business and technical rules are included in this policy?
- What level of frequency are the policies and rules applied?

Getting valued policy and rule-based measurements and metrics requires a level of business and policy integration to support:

- Data Quality
- Master Data Management
- Information Life Cycle Management
- Privacy and Security

5.1 Metrics and Measurements community

Since defining metrics and measurements for data governance requires the ability to support different roles and varying degrees of granular and holistic views of policies and their enforcement, a sound initial approach is to use a combination top down and bottom strategy. The goal is to develop the right amount of objective setting and subsequent measurement supporting the operational, administrative, and strategic viewpoints. A good way to start with this approach is to use the perspectives of the typical hierarchy of roles involved in delivering good governance. Refer to Chapter 2, “Information Governance organizational structures” on page 19. Some information is common to all roles but for others it may be optional or require further levels of detail or consolidation. For example, stewards and custodians typically require more details about the application of specific business and technical rules than on overall policies. On the other hand, Business Owners and Governance Council members will need more detail about overall policy adherence and application. All roles however need to know what level of compliance they are at, and who is responsible. Take a closer look at each role:

- *Business owners* are focused on seeing and understanding the process used to support a policy, and its overall level of enforcement. In general the measurement and metrics they use do not require large volumes or near real-time updates. In general, performance is something that is of interest but not from the technical paradigm of speeds and feeds. The Business Owner perspective is performance over time. For example, they might ask the questions of “How are we doing today?” or “How well did we do last year at this time?” Another interest in performance would be to gather measurements and metrics necessary to establish and monitor system level agreements.

Other questions they might want to know relate to responsibility such as “Is there anyone formally responsible for supporting a particular policy or initiative?”

- *Information Governance Stewards* sit in the middle between Business Owners and the Information Governance Custodians and are essential to an information governance strategy. They require the most and comprehensive amount of information necessary to champion the alignment between the business and technology contingencies. This role requires good people, process, technology, and business skills that can quickly react and function using top down and bottom up strategy. This requires the ability to do both high-level summaries, trends by policy and rules, and the ability to access detail information related to rule exceptions by person, rule, policy, and data source. They are the largest consumer of data and metadata and hence are a good place to start when assessing the best amounts of metric setting and measurement to undertake.
- *Information Governance Custodians* are focused on following and adhering to a predefined process. They are process and rule centric. Though not usually directly responsible for the development or detailed articulation of a policy or rule, they are responsible for ensuring that the elements of a policy are in tact and executing. They view things differently than the business owner and governance council member who are very policy-centric. One of their objectives is to achieve predictable results in a quantifiable and predictable time frame.

Supporting these three roles requires connectivity and access to multiple different data sources with different levels of detail and summarization. Each profile has different representations and demands on the data composing the measurements, such as volume and currency.

A common requirement across these varying viewpoints will be a need for intuitive types of visualization: Incorporating trends, summaries, charts, and iconic gauges. Ideally, this visualization needs to include the capability to drill through the high-level summaries and gain access to more detailed information. Finally, on capturing, correlating, and rendering IG measurements and metrics, any measurement and reporting capability should include considerations for appropriate infrastructure to support these needs. We discuss this further in the next section.

5.2 Required infrastructure

Satisfying a broad audience from executive to the front line requires access to potentially large volumes of data from several different data and repository

sources, processing outputs, and tools. The on-going exercise of generating governance metrics will be the timely collection and correlation of this data and metadata. Breaking it down further, Information Governance Measurements and Metrics depend on the ability to:

- Identify and organize an access approach to the volumes of data and metadata related to governance. Measurements and metrics require connectivity and access to large volumes of data from potentially many systems' operational data stores and repositories that both store business policy-based information as well as instances of governance activity executions. This is often one of the most challenging tasks when initiating a governance program as quite often the use of a centralized definitions and policy-based repository may also be in the beginning stages of maturation. Many of the governance execution activities and processes and tooling will often use different techniques to produce and expose the information. Though challenging, this should be one of the primary focal areas for definition and control by the Governance Council, establishing some standards. For example, some integration may be done through a physical file level, such as using a common file format like comma-separated values (CSV) file, or Extensible Markup Language (XML). Other systems may offer database-like mechanisms that expose the source as a relational data source using ODBC or JDBC that supports SQL access. Others may offer programmatic mechanisms and have application programming interfaces (APIs) used to access and collect information from disparate sources. And finally, an additional approach gaining momentum is to look at landing zone type data stores such as Hadoop, which can provide both lower-cost storage with more highly efficient (parallel) processing. This approach also lends itself well to the use of analytical processing tools. Regardless of the technique used, these volumes of data need to be brought together with some standardized approaches.
- Amalgamate the data and metadata retrieved from disparate data sources using federation and integration techniques that render this as a unified data source available for processing. Standards established for defining governance metadata will greatly facilitate the "matching up" the definitions and policies with the empirical outputs from governance activity executions. For example, establishing a minimum set of attributes that must be collected to instantiate a new customer could be leveraged by both existing and new application processes that collect and manage data. In this way, gaps can be detected before any integration or quality driven initiatives would run into these as a reactive discovery process. Or, centrally identifying all sensitive data elements, ideally by application, would allow for automated comparisons of any data masking activity against the required baseline, simplifying and hardening any compliance analysis and reporting requirements. Creating a single view of information governance data and metadata is essential for processing this information with simple and sophisticated analytical and

reporting tools. These views are used to satisfy the organizational roles such as:

  - Information Governance Program Office
  - Business Process Owner
  - Information Governance Steward
  - Information Governance Custodian

- Connect analytical tools and scorecards to the amalgamated data and metadata sources. These correlations and metrics should be clear in how they are related to governance objectives as represented in the established policies, rules, and rule enforcement. And though these may initially start out as simple spreadsheets of facts of accomplishment, the presentation of the governance objectives and results are typically better served with multi-dimensional analysis that can provide the strategic, tactical, operational, and administrative views into information governance policies and rules that we have thus far been discussing. For example, the ability to visualize policies, rules, measurements, and metrics using iconic indicators or widgets that have an ability to link related items and drill through into more details is a desired capability that will support lineage-type inquiries.

5.3 Capitalizing on your metrics and measurements

Turning collected governance information into meaningful governance metrics that program leadership can begin to use to manage the governance initiatives starts with the ability to view actual results as compared to stated objectives, as well as some historical result sets that will present trend (are we getting better or worse at this?). As discussed in section 5.2, “Required infrastructure” on page 77, at a highest level, leadership will be looking for summary information that will help them gauge the progress of the overall program and provide indicators to help with the difficult job of “steering the ocean liner”: the small rudder corrections that can have very large impacts. So for example, overall compliance reports for establishing clear data lineage for all downstream content may see initial spikes in improvement but then taper or level off (stay static). This could be because the additional policies required to be established are not making progress or it could be due to a lack of adherence and enforcement of those policies that have been established. This additional detail needs to come through in the reporting so that at a high level, the governance sponsors and leadership can take the correct actions (for example, we need more or bigger rudders or we need to apply more pressure). And as discussed in section 5.2, “Required infrastructure” on page 77, this will likely require the drill-down or blending of this detail from the summary visualizations into more understanding. Drill downs into the policy and ownership or responsibility levels of the governance framework should include details of the policy definition, assigned

stewards and custodians with responsibilities, composition (specific rules), and usage (what systems or content to which are applicable). The steward levels will usually be the most fluent in both the details of the rules and applications in relation to the summary projections, and will most usually be the best resources to help with both defining the best visualization techniques that can support the high-level summaries with the right amount of detailed drill-down and well as explaining and presenting the visualizations to the governance leadership.

Some examples include an overall view of governance performance for Data Quality, depicted in Figure 5-1. This information summary lets users see an aggregation of information. Each icon supports the ability to drill through into the details of a governance initiative.

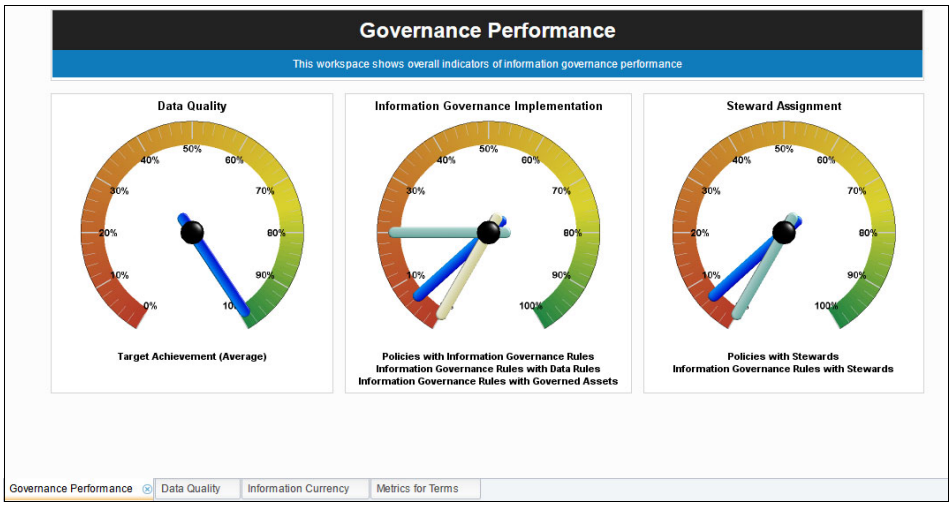


Figure 5-1 Information Governance performance view

A policy-centric view, such as that depicted in Figure 5-2 on page 81, might show expected (100%) achievements and then current performance against this goal. Ideally, however to provide more meaningful information to governance leadership, indications (or drill down) as to the overall numbers of policies that might be essential to achieving higher (targeted) levels of overall governance program compliance should also be visible. So for example, in *English*, this might translate to: We have achieved a 90% policy adoption and use rate, for the policies we have established, but overall *compliance wise* we are still only at 75% of the required target. This might indicate to senior leadership a need to drive harder on the creation and instantiation of the balance of the required policies.

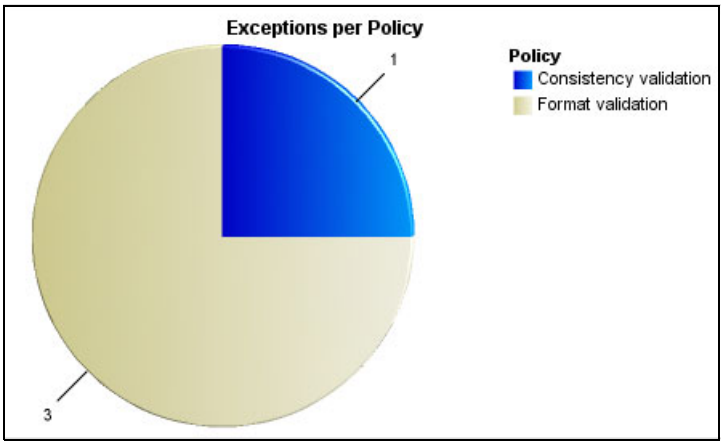


Figure 5-2 Policy-centric view

If however overall governance compliance attainment is in line or on target, further drill downs into rule type feedback might be warranted to identify holdouts or mitigating obstacles. In some cases (assuming the exceptions are valid), these might even support an adjustment to the overall compliance and adoption goals.

Some additional visualization perspectives are described here that should provide guidance for your governance metric and reporting design efforts.

Rule Centric Measurements will illustrate if the rule meets a specific information governance policy. Linkage between business rule and policy are very helpful techniques to promote adoption of the information governance rule. Since there is a one-to-many relationship between policies and rules, it is important to see the composition and hierarchy of each policy, business, and technical rule. Figure 5-3 on page 82 breaks things into lower-level details.

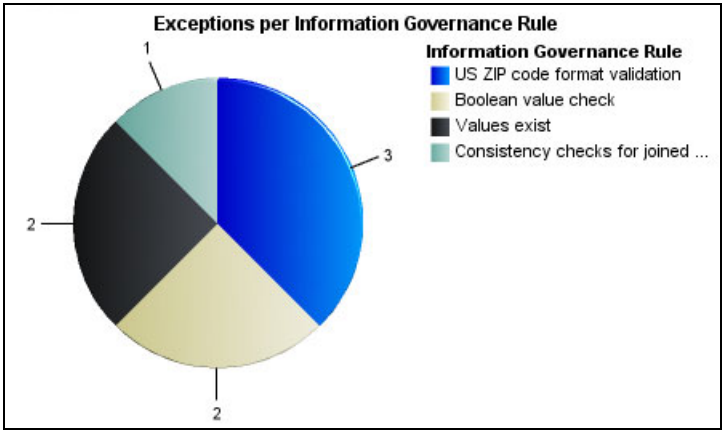


Figure 5-3 Information governance rule-centric view

A Rule Exception Centric Measurement provides identification of rules and quantifies the rules and the enforcement exceptions. These types of metrics help channel efforts into the reduction or elimination of exceptions depicted in Figure 5-4.

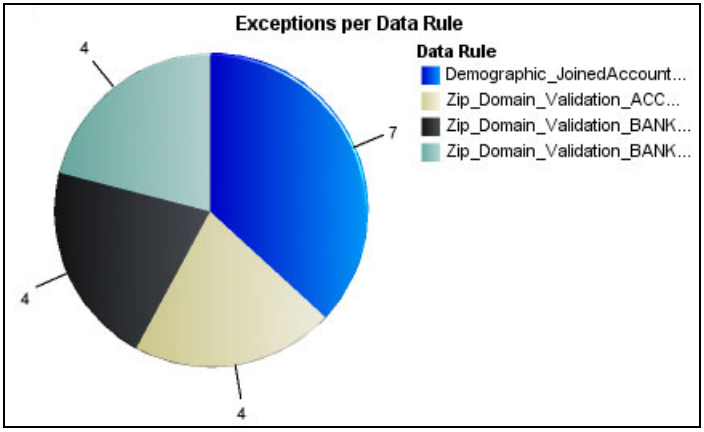


Figure 5-4 Technical rule exception-centric view

A steward-centric view is shown in Figure 5-5 on page 83.

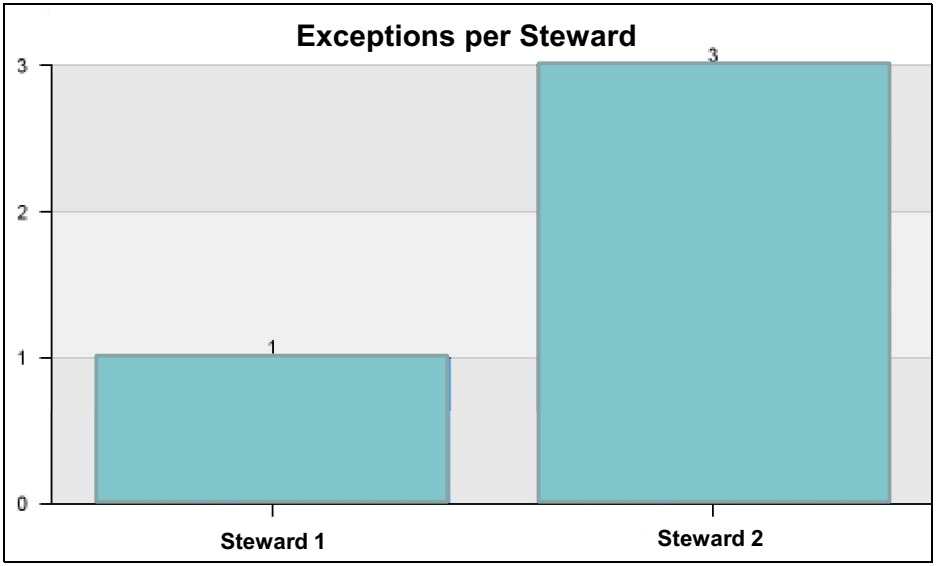


Figure 5-5 Steward-centric view

- Heath Check Summaries provide an operational summary of rule enforcement trends, and statistics that are necessary to answer to the following commonly asked questions:
  - What rule ran when?
  - What data was processed?
  - How much data was processed?
  - How well did it go?
  - Were there any issues?
  - When was this done last?
  - What policy is enforced by each particular rule?

An example of a health check summary is depicted in Figure 5-6.

Health Summary by Data Rules

Data Rule	Trend	Target	Percent Passed	Total Records	No. of Passed	Percent Not Met	Last Run Time	Policy
Zip_Domain_Validation_BANK_CUSTOMERS		92%	88%	3,008	2,650	12%	Dec 2, 2013 12:20:15 PM	Format validation
Zip_Domain_Validation_BANK_CLIENTS		92%	89%	5,112	4,528	11%	Dec 2, 2013 12:13:44 PM	Format validation
Zip_Domain_Validation_ACCOUNT_HOLDERS		92%	89%	2,941	2,604	11%	Dec 2, 2013 12:14:40 PM	Format validation
Name_Domain_Validation_BANK_CUSTOMERS		92%	100%	3,008	3,007	0%	Dec 2, 2013 12:21:10 PM	Completeness validation
Name_Domain_Validation_BANK_CLIENTS		92%	100%	5,112	5,112	0%	Dec 2, 2013 12:18:07 PM	Completeness validation

Figure 5-6 Health check summary

5.4 Scorecards and dashboards

The most common and desirable approach to view your governance measurements and metrics is using a formalized approach known as *score cards*. Policy-based scorecards have evolved from simple spreadsheet calculations to digital dashboards.

Operational Dashboards are used to empower managers or customer-facing workers who need to take immediate actions that will improve daily business performance. Operational Dashboards place more emphasis on monitoring than on taking action. For example, an operational dashboard might have a series of stoplights to reflect an application testing policy for de-identifying sensitive data. A widget might be a good visualization that all of the application privacy policies had been applied, enabling the provisioning of privatized data for application testing. Typically these results need to be current, or near real time. These metrics and measurements allow each IG Owner, Steward, and Custodian to view important issues, assess the impact, and determine the next step. Ergonomics is probably one of the most important requirements of an Operational Dashboard.

Tactical Dashboards track projects and processes measurements. They are more devoted to analysis with a focus on completeness more than run time. Tactical Dashboards do not require immediate response or action. They are not as time critical as an Operational Dashboard. They do require access to large amounts of detailed data for use by business intelligence and analytical tools.

Project Dashboards are used to track time and task specifics. They are used mostly to determine overall project visibility and progress towards milestones or performance. These dashboards do not have the same volume requirements as an Operational or Tactical Dashboard; however, they do need to track activities at the speed of operations. For example, projects typically have a ramp-up time, which proceeds at a pretty regular pace. Towards the end of the projects the frequency of updates increases.

Strategic Dashboards are used by business executives to view or monitor cross-functional progress of a business strategy. Typically these results Dashboards leverage the use of several visualizations, charts, summaries, trends, and icons. The information used to support these dashboards is not updated in near real time. Typically, the data that the Strategic Dashboard uses is updated on a monthly, quarterly, or yearly basis.

Public and External Dashboards are used when public or external entities require access to the data. These dashboards are similar to others except they are more tightly scoped and have a limited access to information with a limited

amount of drill-down capability. Typically, these results updated on an as-needed basis using historical data.

Rule-level Widgets are used to provide metrics and details that are related to the rules and records. A common use of a rule dashboard is to highlight the percentage of records that fail a particular rule. If and when there are exceptions, the dashboard requires access to details related to the definition of the rule and its exceptions. This information is often accessed in near real time and offers the users the ability to navigate or toggle back and forth between rules, records, and policies as necessary.

Record-level Widgets are used to relate particular records to one or more rules. This is very helpful if policies and rules fail or are broken. If a particular record fails because of a particular rule, some investigation is needed. It is important to recognize why this happened and determine if this is due to process or technology, and ascertain if there are other upstream or downstream affects on other policies and rules. The ability to identify records with the highest number of rule violations is a great way to review the business and technical policy and rule definition.

Data source Widgets are focused on the data and the percentage of records that break or fail at least one rule. It shows the average number of rules that are broken and identifies whether or not trends are developing, such as whether degradation is occurring between baseline measurements and the current results.

Based on the different types of dashboards and measurements, one should be able to see that the dashboards are used to render metrics and measurements that support the business and infrastructure teams information governance efforts by role and focus. Earlier in this section, we discussed infrastructure requirements that addressed data integration, analytics, and reporting. While this might not seem like anything new, the approach that IBM has taken is to migrate from a product-centric view into a solution-centric view as a means to simplify and standardized an approach for a governance framework.

We have already begun this effort with IBM InfoSphere Information Server products that provide seamless integration with components to record the definitions of terms, policies, and assets and tied those to physical implementation rules. Imagine a Chief Governance Office wants to understand the current status of a governance program, they will want answers to the following questions:

- Do policies have stewards assigned?
- Do policies have information governance rules assigned?
- Do information governance rules have assets assigned?
- Are information governance rules implemented?

Typically, Chief Governance Officers want to reacquaint themselves with specifics in a define policy hierarchy before they narrow their attention into specific details. Figure 5-6 on page 83 illustrates a policy hierarchy. While the navigation has occurred through a dashboard, it is actually drilling into and accessing information contained in the IBM InfoSphere Business Glossary. Having a common vocabulary and understanding of the policy is as meaningful as the current status of that policy.

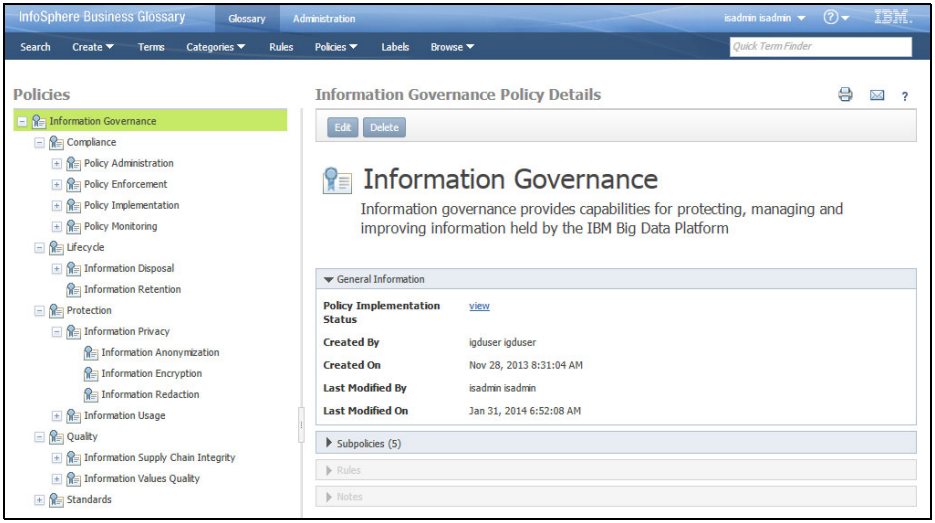


Figure 5-6 Policy centric view: Definition and status

Figure 5-7 on page 87 shows an example of IG Policy implementation and Rule status.

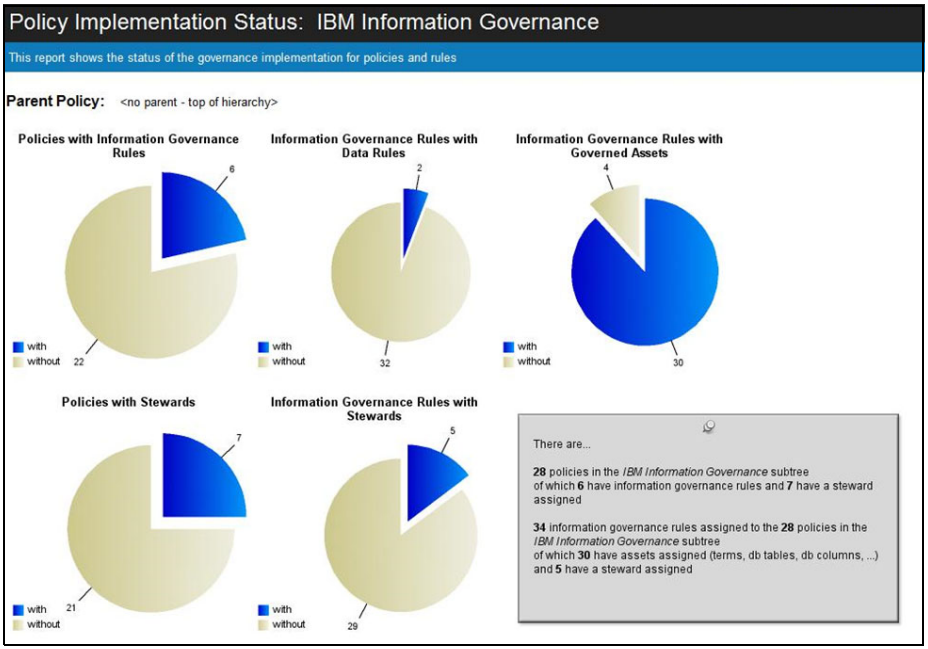


Figure 5-7 Governance policy implementation and rule status

Another solution to consider is when a Business Data Steward wants to understand if there are quality issues on the systems that are being governed. This scenario is depicted in Figure 5-8 on page 88. The steward will want the following questions answered:

- What is the overall rating of data quality across the governed systems?
- Which policies have the highest number of exceptions?
- Which data rules have caused these exceptions?
- When have these rules been run and how many records were processed?

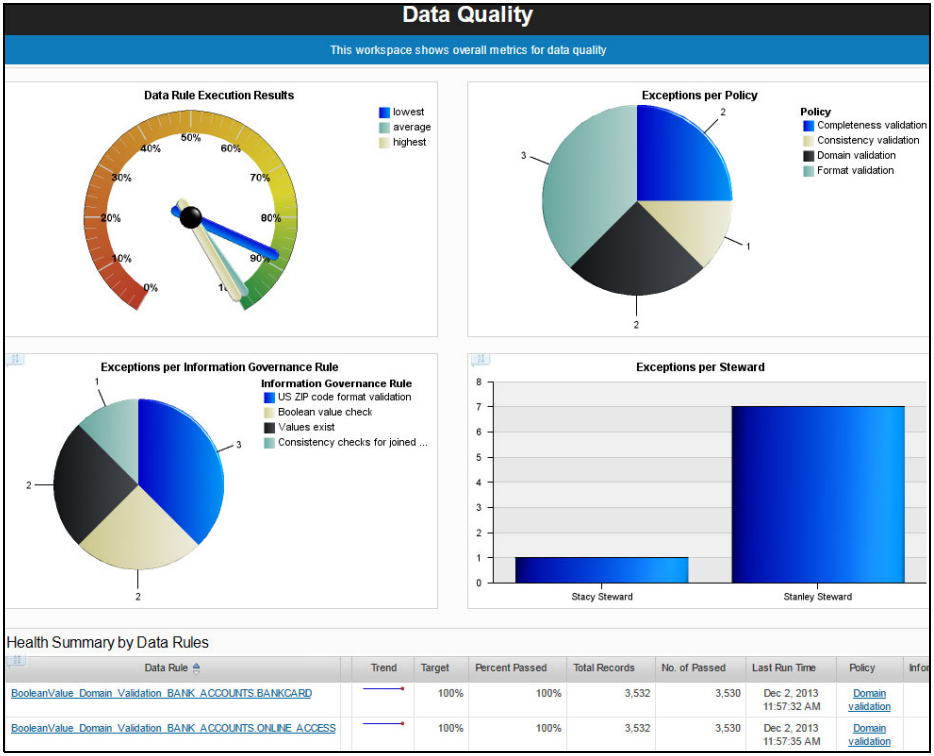


Figure 5-8 Steward-centric data quality metrics

Figure 5-9 on page 89 shows you a summary of the technical rules metrics.

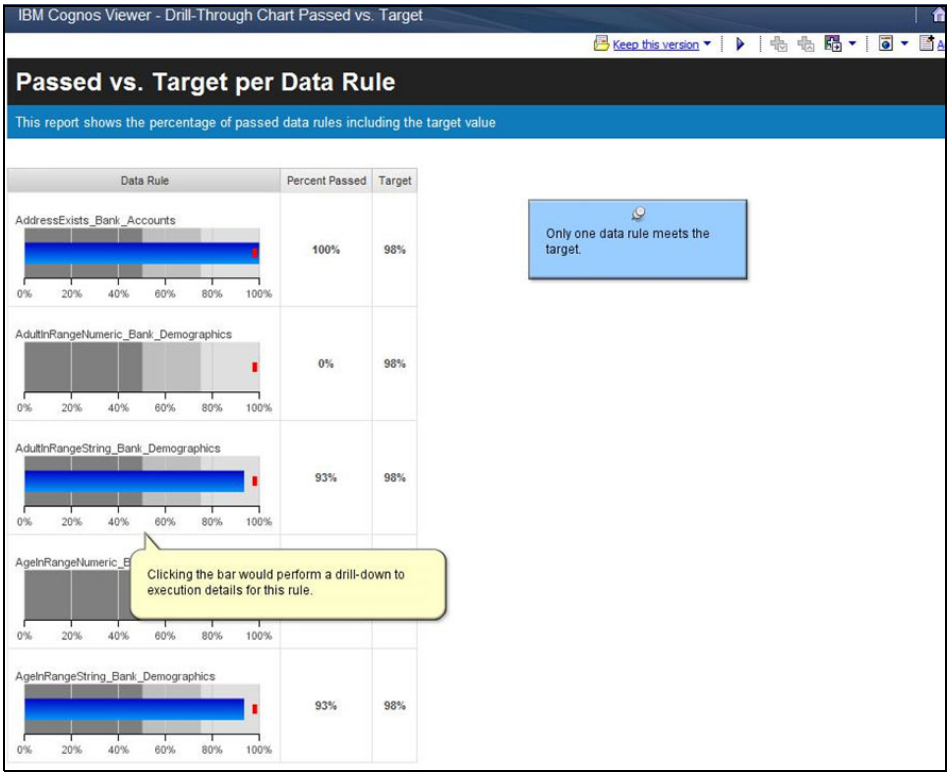


Figure 5-9 Technical rule summary metrics

Throughout this chapter and the book, we have discussed terminology and the application of policies, and rules for information governance. One thing can be fairly certain, which is that nothing stays the same over time. For this reason, we must be able to examine definitions and usage of policies, terms, and assets. We need to be able to see the currency of the information both as an aggregation and the discrete elements. The information currency is depicted in Figure 5-10 on page 90.

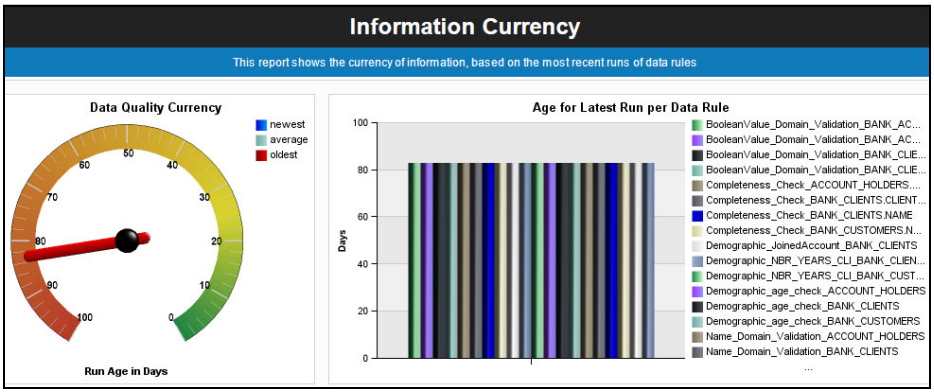


Figure 5-10 Information Currency view

In this chapter, we discussed measurement and metrics by examining who and how they are defined, what is included, what is required, and finished with some examples using the IBM Information Governance Dashboard. The Information Governance Dashboard was released late 2013 with an initial focus on Information Governance for Data Quality. Our plans are to continue to extend metrics and measurements to include additional Information Governance products. Our stated direction is to deliver high-quality Information Governance solutions while reducing the level of effort required to implement Information Governance solutions.

6

Business drivers for information governance

In this chapter, we review a number of the often discussed and documented drivers underlying better information governance maturity and introduce some of the more common capability scenarios that are undertaken to achieve this goal. These scenarios (aligned to IBM InfoSphere solutions) represent some of the typical starting or launching points that our clients undertake every day, and when approached and executed with a solid footing of foundation capabilities, can propel organizations down the information governance maturation path.

Each scenario is driven into more thoroughly in the chapters that follow, providing more details about both the logical flow and approach to implementation as well as the key intersections with the principles of foundation governance that will be instrumental to their success.

6.1 A vision of trust

There are many definitions and descriptions of the drivers behind why organizations, private and public, seek to climb up the levels of information governance maturity. In essence, these come down to achieving a level of completeness, confidence and trust with the data and the information it supports.

Years before the “information at the speed of light” days, there was greater leniency, greater forgiveness for less than perfect information in making key business and policy decisions. Lags in time, gaps in detail were the status quo, and business prowess was often measured in part by the insight and foresight that the business leaders possessed.

Today, the powerful data access and analysis capabilities make it almost unacceptable or at least disappointing to make a less than informed decision. And in business of course, “disappointing” can carry much heavier penalties and ramifications.

Yet still we often struggle to make sense of, or to feel confident with, or to trust, that the intelligence gathered for every decision or interaction involving sets of information is accurate and current, properly secured, and ubiquitously recorded and documented. You want to know that it represents all that you can or need to know. So that *whatever* the action or decision that you are about to make related to it, such as the following, is the best decision that can be made:

- Buying a new car
- Signing off on a new parts vendor
- Allocating risk-oriented buffer management funding, such as reserves
- Approving a new offshore application testing outsourcing contract
- Executing a comprehensive legal hold or discovery process
- Validating IT storage and management budgets

This need for trust and confidence is what underlies the more tangible manifestations of these concerns, and is further discussed in the following sections.

6.2 Classic business drivers

There are a number of common categories and subcategories of business drivers that motivate organizations to inquire about and ultimately engage in projects and programs to achieve better information governance. You could argue that all of the drivers ultimately come down to a calculation of maintaining or controlling some value, such as revenues, profits, expenses, and stock price,

so these might better be called *business reasons* behind the drivers for controlling value. In the end, they are the measure of a solid set of governance capabilities.

6.3 Business reasons for information governance initiatives

There is typically never just one incentive or motivation for organizations to embark on a governance-related initiative. Many factors will usually come into play. Though there can be more immediate drivers, such as an audit or infraction of a compliance-related issue, or a substantial financial impact or penalty resulting from improper information handling or bad decision making, more often the driver will be actions to avoid the impacts of these misgoverned situations and the more positive desire for the resulting or potential benefits.

Below is a list of some of these common drivers, followed by short descriptions of some core or functional governance-related capability initiatives followed by our clients to address these drivers.

And though each scenario may start with one motivation, each or many of these motivations can be factored into the justifications behind the common scenario initiatives in the following ways:

1. Cost Reduction
   a. Systems, licensing, and storage/storage management
   b. Information handling, maintenance, and management
   c. Content research and discovery
2. Risk and Compliance
   a. Eliminate improper or non-compliant processing or exposure
   b. Increase the quality of decision making
   c. Reduce the numbers of “blind spots”
3. Opportunity and Distinctive Competency
   a. Increased transaction value
   b. Increased consumer value
   c. Unique and distinctive offerings
4. Efficiency
   a. Unified understanding of information
   b. Non-redundant systems and data
   c. Streamlined processing
   d. Reduced defects and errors

The following information governance scenarios are the focus of the balance of this book. These classic examples are based on interactions and programs that we face daily in support of our IBM client base. Each scenario maps directly to a

core (or more comprehensive IBM solution-based) discipline highlighted in the Data Governance Maturity Model and the implementation of each will expose and highlight the key dependencies and interactions with foundation governance disciplines. It is not uncommon to work with clients on the building of these core capabilities coincidentally with the building and maturing of their foundation governance principles, that in all scenarios will evolve gradually over a scale of less to more maturity.

It is our expectation that the reader is able to relate to and find affinity with many of these scenarios regarding their own experiences and is able to use both the implementation approach outlines and the foundation governance interaction discussion to assist in the development and maturing their own information governance capability endeavors.

6.3.1 Integration quality scenario

The first integration quality scenario involves a classic data warehouse load from a set of enterprise applications and the resulting challenges and distrust of the results, specifically around the stated goal of having a “single view of the customer.”

The business drivers

The primary driver is the desire for the efficient-single view of information goal, in this case customer. Additional drivers associated to this could be streamlined processing and decision making, which can result in both higher “value” transactions for both company and customer, as well as reduced errors and any related processing costs.

The scope of the initiative

The scope is to implement an enterprise data warehouse (DW) solution by creating and maintaining unique customer records in the DW based on two data sources (as examples, a core customer banking application and credit card system).

The objectives

Based on the two sources of information, identify critical elements that make up the complete definition of a customer for the DW and then formally define the specific data elements required to instantiate and accept a customer record (set of records). In this process, identify and define the data elements to be used as match criteria for initial and ongoing DW loads.

The challenges

Classic DW implementations invariably expose the complexities of merging cross organizationally managed content, and highlight the core quality governance capabilities that are essential to increase the validity, value, and trust in the resulting merged content.

6.3.2 Information Lifecycle Management scenarios

The next scenarios look at two of the major disciplines included in the larger Information Governance capability of Information Lifecycle Management (ILM): Enterprise decommissioning and enterprise test data management. Each will share a number of foundation capability needs around understanding of applications and associated process landscapes, as well as supporting policies and processes that lead to successful deployments.

ILM: Enterprise Decommissioning scenario

In this section, we discuss the Enterprise Decommissioning scenario.

The business drivers

Typical business drivers include the impacts to costs and management of information and associated systems footprints due to increased trends in certain industries of acquisitions and divestitures, forcing often hectic exercises of application rationalization and centralization. These often manifest as footprint (and cost) reduction objectives. As equally motivating however is the growing compliance and regulation ramifications surrounding the inappropriate retention and overall handling of aged and retiring information. So risk reduction and audit drivers can be as or even more pertinent.

Decommissioning is often a first endeavor into the ILM governance arena because it can focus on applications and sets of information that are usually less critical to the day-to-day operations. This allows organizations to build their governance and technical ILM capabilities before then moving on to the next logical steps of overall application (including live/production) data growth (archiving) management.

The scope of the initiative

The scope is to implement an enterprise-wide application (structured and unstructured content) decommissioning capability for applications that have reached end-of-life or are being rationalized into centralized systems.

The objectives

Reduce the overall enterprise application data footprint via a process of assessing, targeting, and executing an appropriate content (electronic) archiving

technique, adhering to company and industry records management policies and regulations, and ensuring appropriate and timely access to archived content as determined required. This should reduce the direct and indirect costs that are associated to the retiring of physical assets and support reduced risks in the records and legal management processes responsible for this content.

The challenges

The challenges are a centralized understanding of application landscapes and applicable records and retention policies, data and application ownership, and personnel, policies, and processes for appropriate execution and ongoing (defensible) administration.

ILM: Enterprise test data management scenario:

In this section, we discuss the test data management scenario.

The business drivers

Cost reduction and improved management of all shapes of testing environments and related processing is at the core of this capability. Supplemental drivers include the impacts of these increased efficiencies, such as faster development -to-production turnarounds of key functionality that can translate into faster time to market, distinctive capabilities supporting premier market offerings, and placement.

The scope of the initiative

Implement an enterprise application test environment management capability used to perform (and eventually automate) individual and group functional, system, regression, performance, and user acceptance testing across multiple platforms, technologies, and systems.

The objectives

Reduce the size, effort, time, and cost to provision and maintain relational application databases that are used for domestic and offshore testing with real but fictitious data without compromising the integrity and consistency of the data that is required to meet or exceed regulatory and compliance initiatives during planned and unplanned testing cycles.

The challenges

The challenges are a centralized understanding of application landscapes and federation of data impacts to comprehensive testing environment needs, along with adequate application testing process, and related business object content knowledge (including detailed understanding of sensitive data items). Implementing this capability must also often grapple with non-uniform and low levels of foundation governance principles applied to testing activity policies, processes, and personnel.

6.3.3 Master Data Management scenario

In this section, we discuss the Master Data Management (MDM) scenario.

The business drivers

The drivers include cross category elements such as efficiency of information management and handling (costs reduction), and single source of decision-making detail. But the core business objective is the truly mastered set of content, the one trustworthy, centralized truth, whether it be about customers, products, orders, or any master worthy set of content. Even with just the goal of a single view of customer, costs of marketing and selling can be reduced, improved transactional value (for client and customer) are achievable, and a deeper, trusted relationship with customers can lead to distinctive competencies difficult for competitors to replicate. Risk reduction drivers can also play a supporting role, where single views can lead to centrally definable and managed policies for privacy and protection.

The scope of the initiative

Implement an enterprise master data management capability such that all relevant applications and business functions enterprise-wide will use this single customer profile.

The objectives

Produce an integrated and consistent view of customer data across all applications and business functions.

Success will be measured by the ability to market, cross-sell, or up-sell and facilitate fraud detection and risk exposure. Comply with data privacy policies and align with regulatory and compliance standards, such as, the Foreign Account Tax Compliance Act (FATCA) and the Health Insurance Portability and Accountability Act (HIPAA).

The challenges

The recognition and acknowledgement for the need for master data management is indicative of a level of organization maturity and complexity that requires higher degrees of information governance. It is often hard to separate what aspects of implementation are more MDM versus foundation governance oriented, such as the defining of core data and information definitions (for example, what content actually makes up the complete record of information defining a customer). These types of interdependencies, it could be argued, support the need for the foundation information governance principles more than any other core information governance initiative.

6.3.4 Data Security scenario

In this section, we discuss the Data Security scenario.

The business drivers

Risk reduction, compliance, and overall data security is the primary business driver, which in turn can reduce cost and penalties due to non-compliant or negligent behaviors. Establishing foundation and core information governance data security policies and processes reduces the “blind spots” vulnerabilities that can lead to embarrassing exposures. Indirectly, better and more timely decisions and actions can result when higher levels of confidence prevail around the security of information content.

The scope of the initiative

Implement an enterprise sensitive data security capability for both structured and unstructured information content in enterprise database systems, data warehouses and file shares.

The objectives

Protect data via real-time monitoring of all data traffic activities, including actions by privileged users. Also, protect sensitive data repositories against new threats or other malicious activity and continually monitor for weakness.

Measurable goals include safeguarding sensitive data by detecting unauthorized or suspicious activity and alerting key personnel, while meeting and exceeding standards and regulations for data protection, such as the Sarbanes-Oxley Act (SOX), the payment card industry (PCI), and HIPAA. Long-term objectives and measures include reducing operational costs via automation of cross organization (individual database, management, and related systems level implemented) security, and protection policies and audit repositories.

The challenges

Securing data and the related information it composes will depend heavily on foundation principles established around what and who requires protecting and monitoring. These metadata and composite information level content definitions and policies need centrally managed control and coordination, and will be a key focal point for foundation governance efforts. Automation via intelligent tooling across diverse landscapes can also pose technological challenges that can create artificial hurdles and roadblocks.

You are encouraged to read on in this Redbooks publication, examining the detail involved in each of these information governance capability initiatives to learn more about how they may impact your comprehensive information governance maturity.

7

Data Quality

In this chapter, we introduce you to the information governance core discipline of Data Quality (DQ) and then take you through one of the very common implementation scenarios that will better demonstrate the value of data quality. We describe in detail standard practice steps that are usually followed to build and enable data quality practice in the enterprise.

7.1 What is Data Quality?

Data Quality, which is depicted in Figure 7-1, is one of the Information Governance disciplines.

Data Quality is not just a technology. It also includes roles and organizational structures, processes for monitoring, measuring and remediating data quality issues, and links to broader information governance activities via data-quality-specific policies.

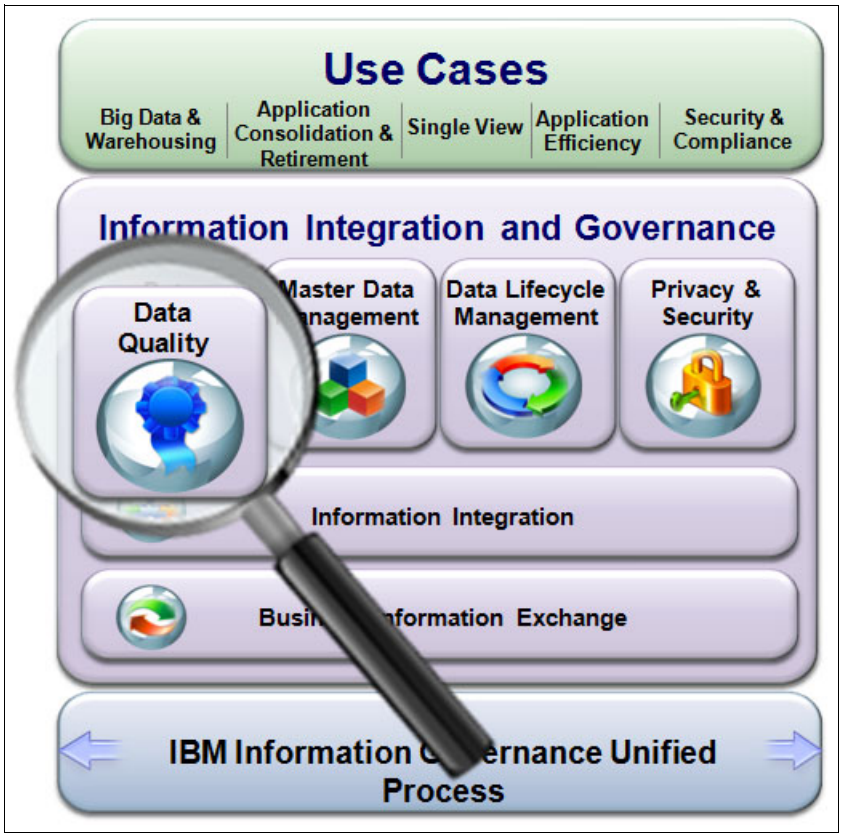


Figure 7-1 Information governance: Data Quality focus

There are many definitions of Data Quality in the information governance world. A very widely accepted definition of Data Quality defines it as data fitness to serve its purpose in a given context. This means that the same data could be absolutely perfect for one particular system because it fulfills all the needs that this system has and for another system it could be absolutely disastrous. When

talking about Data Quality, there are multiple aspects of it that need to be considered:

- Accuracy
- Completeness
- Relevance
- Reliability
- Consistency across data sources
- Accessibility

Data quality is a cornerstone to the business operation in the organization, and is affected by the way data is entered, stored, and managed.

Most enterprises are running distinct sales, services, marketing, manufacturing, and financial applications, each with its own master reference data that satisfies requirements of each individual system. There is often no single system that is the universally agreed-to system of record. However, in data integration efforts, old data must be repurposed for new systems. A complete and accurate integrated view is greatly dependent on the quality of the raw data.

Data quality is not necessarily a data entry problem, but an issue of data integration, standardization, harmonization, and reconciliation. When you look at the lack of confidence in information and examine why that is, you might realize that it is because information is pervasive across the organization. On one side, you are dealing with fragmented silos of data that were accumulated through many years that lack data quality, and there might have been difficulties in organizing and consolidating the data in a way that makes sense to the business. Add this to the sheer growth in volume, velocity, and variety of data that is being collected every day to create this complex mass of information.

On the other side, the business is looking to receive the information they need and not more. They want to know how they can receive it in a structured way and not the way it was captured and how they can get it all in a timely manner for decision making. That is where organizations must tailor the information to the diverse needs of users on the demand side. These users want information that is relevant to their role and have that information accessible to them wherever, whenever, and however it is needed. The information also must be usable, with the level of transparency, accuracy, and usability that is relevant to their role. This can range from at-a-glance views for the average business user to monitor the business, to a detailed business analyst who has total freedom to explore and conduct different what-if business scenarios.

7.2 Data quality scenario

A typical banking industry example is referred to through this chapter. In this scenario, a company needs to implement a data warehouse solution by creating and maintaining unique customer records in a data warehouse that is based on two data sources (core banking and credit card).

Based on two sources they must identify critical data elements that need to be loaded into the data warehouse, and define which data elements are required to accept a customer record. And they must define which data elements are used as match criteria for the initial load and the delta process.

There has been an implementation of the data warehouse already, however the results that the process produces are not satisfactory. That is, the customer view that is produced by this process is not trusted. From the initial discussion with the project team, elements from the data flow, such as the initial load process, delta load process, data standardization, and matching, seem to be in place. So the project team needs to step back and determine what are the main causes for the failed implementation.

7.2.1 Assess existing application and associated processes

In order to determine what is causing problems in the existing process, it is necessary to look at all aspects of the implementation. As the existing application contains a significant amount of DataStage and QualityStage processes, you can start by looking at the problems that were reported by the client. That allows you to narrow down the problem area. The following are some common problems reported:

1. Some customer records that should match, do not get matched and there is no clear reason why.
2. Some customer records that should not match, do match.

A recent sample evaluation of the customer data loaded into data warehouse showed that there are about 30% duplicate customers.

Several reasons could be behind these problems. The implemented process consisted of a batch initial load process where the first source (core banking system) went through a validation, standardization, and de-duplication match process and was loaded into data warehouse. After that, the second source (credit card) was taken through a delta process and loaded. After both sources are loaded, change records from both sources are processed daily to keep the data warehouse consistent with the changes.

The flows for the initial and delta load processes are presented in Figure 7-2 and Figure 7-3 on page 104.

In the initial load process, the following steps are performed:

1. Data from the first source is read by the DataStage process
2. Data gets validated
3. Data gets standardized by QualityStage standardization rules
4. Data is deduplicated by a QualityStage match
5. Data survivorship is performed for the records that need to be collapsed
6. Data is loaded into data warehouse

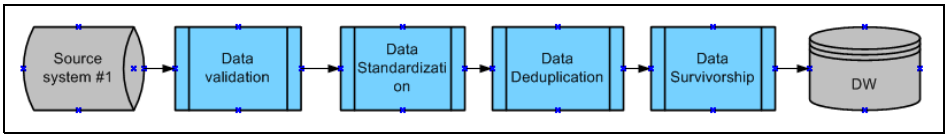


Figure 7-2 Initial load process

In the delta process, most of the steps are reused from initial load, but there are some differences. The biggest difference is that the data coming into the delta file needs to be compared and matched to the records already loaded into the data warehouse. After a comparison is done, the records in the delta file that did not match to any of the records in the data warehouse need to go through a de-duplication match to ensure that no duplicates are loaded.

QualityStage match specifications are set up in a way that the results produced could be divided into three groups:

1. Records that matched with high confidence and could be collapsed
2. Records that match with low confidence and need to be reviewed by a subject matter expert (match suspects)
3. Records that did not match (residuals)

Records that fall into the second category are stored in a data warehouse as separate customers under the assumption that the Data Steward will be constantly reviewing these suspect matches and make a decision to collapse matched records or break them apart.

Figure 7-3 on page 104 shows the delta load process.

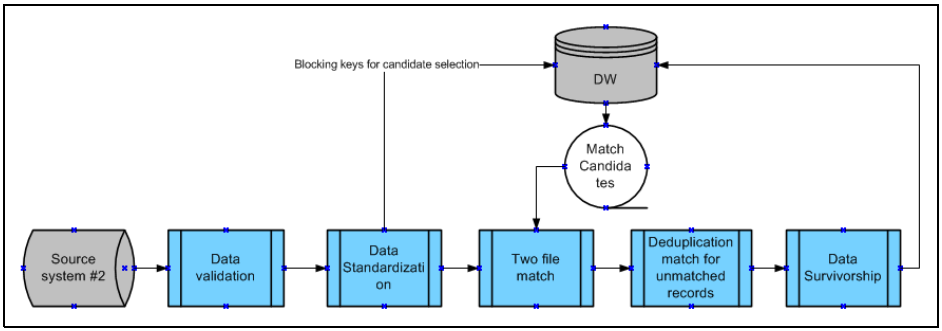


Figure 7-3 Delta load process

The following list contains descriptors of the potential issues that were discovered during initial process analysis:

- Reviewing the process, it was identified that there is only minimal data cleansing being done. No policies or requirements were defined across sources to conduct consistent data cleansing.
- Reviewing match processes, it was noticed that match specifications in the de-duplication match and two file match in the delta processes are not aligned because slightly different match criteria was used. No policies or requirements were defined around customer match rules.
- The candidate selection process that brings match candidates from the data warehouse into two file match should be based on the blocking criteria of the match, and in this implementation it is out of sync.
- Although subject matter experts are supposed to review and resolve match suspect pairs, that was not happening consistently and hugely affects ongoing delta processes as incoming delta records are matched to a number of suspects rather than records representing a consolidated customer view. An appropriate organizational structure was not put in place to accommodate the issue resolution.
- There is no process in place to make sure that new delta feeds do not introduce new data anomalies. That is, no policies or requirements were defined around data quality monitoring and no organizational structure was put in place to conduct ongoing data quality monitoring.

7.2.2 Identify data owners: Data stewards

Data integration implementation always requires heavy involvement of data owners. It is not possible for IT personnel to make an accurate decision on data validity, data completeness, and match rules. In our example scenario, data is consolidated from two lines of business inside the bank. During the assessment

of the implementation, it was found that at the time of the initial implementation, the subject matter experts were involved just for clarification of the data fields from their specific line of business and were not greatly involved in the decisions on cleansing and matching for the target data warehouse. Even if information about the specific sources was identified and captured in some centralized glossary, it is not sufficient for the target data warehouse. These decisions were made by IT personnel. The example banking company realized the problem of not having a central information governance body and had to establish such an organization after the initial implementation. The Data Governance body will be playing role of data owner by reconciling issues between two sources and finding solutions that will fit both sources.

7.2.3 Perform data quality assessment

As noted in 7.2.1, “Assess existing application and associated processes” on page 102, initial assessment of the process revealed that there is very little data cleansing being done during the load process. There could be several reasons for that:

Good reasons (but very rarely seen):

- Data is cleansed in the source
- Company policies and systems are very strict and do not allow bad data to be entered
- Source data quality fits perfectly for the new target source

Bad reasons (and very common):

- The existing implementation ignored bad data issues, assuming that it will not affect the load processes
- Cleanse data in the source systems and let it trickle down into the data warehouse as the changes were made.

In our specific case, a decision was made to cleanse the source systems manually with the idea that cleansing involves not just identification of bad values but additional research that will allow enrichment of the data. That sounds good. However, nobody realized at the time of the decision that it takes years to go through records manually.

To move forward with the tuning of this application, the first technical step that always needs to happen is to run a data assessment. It is vital to know what is in your data so you can build appropriate validation and cleansing rules. Running a quick data assessment on the fields that are used in the record match criteria and already in the data warehouse revealed the bad condition of the data.

Some results of the assessment are presented in Figure 7-4 and Figure 7-5.

BIRTH_DT	FREQ
12/31/9999	150217
1/1/2001	29211
1/1/1970	24653
1/1/1991	17804
1/1/1996	17155
1/1/1995	16385
1/1/1980	15772
1/1/1994	13059
12/31/1960	12787

Passport #	FREQ
123456789	21
999999	21
123456	20
abcd	19
XX112245	19
YYB447930	19
70	19
K1923	19
99	19

Figure 7-4 Sample Data Quality assessments for date of birth and passport

ADDR_LINE_ONE	FREQ
MISSING	738
RETIRED	400
PO BOX	310
LONDON	304
ENGINEERING DEPARTMENT	277
MATH FACULTY	269
WILSON UNIVERSITY	202
MISING	183
MAIN STREET	177

National ID	FREQ
01	6748
0	942
-	732
1	711
*	690
000000000001	648
123456789	603
345723747	601
MISSING	598

Figure 7-5 Sample Data Quality assessments for address and national ID

As you can see from the frequency distribution of the data values, those coming in are not good quality. So what are the obvious problems that can cause?

- If bad values are not identified, you do not really know if you have complete information about your customer.
- If using these fields for matching (which the example banking company did), the match results will be distorted. For example, National ID carries high importance in the match. But just by looking at the first line of the frequency distribution report, you can see that 6748 records have value of “01”. That will definitely affect how records are matched together generating a false positive. On the other hand, if there is another record with a valid National ID and we try to match to record with a bad National ID, the match penalizes different values and we might end up with false negatives.

- Blocking criteria in the QualityStage match is a mechanism to select records from file that could possibly match together. It is done based on equality of some portion of the record. For example, you can block on National ID and then compare and score whatever other data is part of the match criteria (such as name, address, and phone). The size of the candidate pool that is created by blocking is a question of a balance. Make blocking criteria too conservative (small blocks) and you will miss some matches. Make blocking criteria too relaxed (large blocks - thousands of records in a block) and it will increase execution time. Having a high frequency of invalid and default data in the fields that are used as blocking criteria (such as National ID or Date of Birth) will result in extended execution time.
- As part of the delta process, an SQL query is executed against the data warehouse to select all possible match candidates and bring them into a QualityStage match. The SQL query is done based on the match blocking criteria. Again, if an incoming record is populated with bad data in the fields that are used for blocking, the SQL query will try to bring all records with matching values from data warehouse. For example, if an incoming delta record contains Date of Birth field “12/31/9999”, the SQL query brings 150217 records into QualityStage match, which is another cause of bad performance.

7.2.4 Work with data stewards to establish rules for data validation

As you can see from data warehouse data quality assessment results, there is work that needs to be done to clean up incoming data. Looking at the result of the assessment, IT personnel probably can have an educated guess for what values are valid and what values are not for some fields. But it is obvious that help of a subject matter expert (SME) is imperative here because only an SME would know for sure what could be treated as invalid or default values. In our case, where multiple systems are involved in loading data warehouse, involvement of a data steward who can actually coordinate resolution of different anomalies coming from two different systems is essential. As previously described, there are multiple areas where bad data could have an impact. So you have to be aware of the impact when deciding on the ways of mitigating the issue. First, of course, it is necessary to identify values that need to be cleaned up. There are multiple reasons why fields could carry bad data. Some of them are listed here:

- Default values were populated in some fields by batch program as fields must carry value and cannot be blank or null. Data stewards will know or will have to investigate which values were populated as default (Example: date of birth of 12/31/9999).
- Usage of the fields according to metadata defined is not enforced. For example, you have multiple lines of address where you need to enter the address in a free form, but the data entry personnel chose to use it to enter comments.

- Field contains multiple values because it was easier for the source system to reuse the existing field for multi-purpose rather than extend the source data model.

So running a data quality assessment will help the data steward to identify these problems. The next logical step is to decide what should be done with these anomalies.

There are a couple of ways to deal with the anomalies. And it all depends on whether or not you need to propagate all values from the source systems to the target data warehouse. In our experience, most of the time if the piece of bad data is identified, there is no value of storing it in the data warehouse. In that case, validation routines need to be updated to make sure that invalid values are identified either by the data format or reference table lookups and cleaned up. If for some reason there is a necessity to load all the values into data warehouse, the invalid values need to be identified and cleaned up before those records are used in a match process so that bad data does not affect the results of the matching.

Obviously, all the decisions around data cleansing need to be documented and data cleansing policies need to be created so they can enforce processes around data quality, monitoring, and cleansing.

7.2.5 Work with data stewards to establish match criteria

One of the main tasks that must be performed during data integration implementation is to understand the requirements that the client has for the QualityStage process, and for the QualityStage match or matches in particular. A statement such as “We want and need to identify a unique customer” or “We want and need to identify a unique household” is not sufficient to translate into technical specification for the match. And as soon as you start asking questions, often, different departments have something different in mind when they talk about a customer. The role of the data steward is really important here because that data steward needs to resolve the differences in customer definitions between departments and drive customer match logic in QualityStage and as a result, define policies around the data warehouse customer definitions, matching rules, and match suspects handling.

To determine the rules for matching, more questions need to be asked. For example:

- Are there individual and business customers?
- Is a unique customer determined by the name at the same address or it could be across addresses? Is it different between an individual and a business?

- What address should be used in matching if there are many (such as Mail-To and Bill-To)?
- What name should be used in matching if there are many (such as nickname, database administrator (DBA), or also known as (AKA) used?
- What other fields are maintained in the source system to help with the match? As examples, date of birth, phone numbers, email address, Social Security Number (SSN), and Taxpayer Identification Number (TIN).

As a result of these question and answer sessions (yes, in most cases there will be multiple), you get an initial common understanding of what (in business terms) a business considers a match.

To get even more clarity and firm up a common understanding between the technical and business side on the expected results, it is necessary to review examples of different scenarios which that match can present. Ideally, it needs to be done with the customer data to which the business can relate. These examples become a way for both sides to understand and document expected match behavior. The following are a few sample scenarios of the match.

The record pairings in Figure 7-6 through Figure 7-12 on page 111 contain data in which the Address and Postal Code are the same, however variations exist in the Individual Name.

In Figure 7-6, should these records be considered a match, a non-match, or require manual review? Why? How do transposed words in Individual Name impact the decision?

1-A						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	DAVID SOMEFIRSTNAME	47 N ANYSTREET LANE	MIAMI	FL	33178	US
2	SOMEFIRSTNAME DAVID	47 N ANYSTREET LANE	MIAMI	FL	33178	US

Figure 7-6 Match example 1-A

In Figure 7-7 on page 110, should these records be considered a match, a non-match, or require manual review? Why? How do additional words in Individual Name impact the decision.

1-B						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT NOTREALNAME	700 N ANYSTREET DR	SACRAMENTO	CA	31244	US
2	HUBERT NOTREALNAME JR	700 N ANYSTREET DR	SACRAMENTO	CA	31244	US

Figure 7-7 Match example 1-B

In Figure 7-8, should these records be considered a match, a non-match, or require manual review? Why? How does the usage of a nickname in Individual Name impact the decision?

1-C						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	BILL NOTREALNAME	3480 ANYSTREET RD	GREENBOROUGH	NC	95658	US
2	WILLIAM NOTREALNAME	3480 ANYSTREET RD	GREENBOROUGH	NC	95658	US

Figure 7-8 Match example 1-C

In Figure 7-9, should these records be considered a match, a non-match, or require manual review? Why? How does the presence of a maiden name in Individual Name impact the decision?

1-D						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	KATHY NOTREALNAME	41 S MYSTREET ST	WESTWOOD	MA	02090	US
2	KATHY NOTREALNAME-MAIDENAME	41 S MYSTREET ST	WESTWOOD	MA	02090	US

Figure 7-9 Match example 1-D

In Figure 7-10, should these records be considered a match, a non-match, or require manual review? Why? How does a variation in the middle name of the Individual Name field impact the decision?

1-E						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT F NOTREALNAME	333 ANYSTREET ST	FAIRFIELD	CT	01234	US
2	HUBERT X NOTREALNAME	333 ANYSTREET ST	FAIRFIELD	CT	01234	US

Figure 7-10 Match example 1-E

In Figure 7-11, should these records be considered a match, a non-match, or require manual review? Why? How does a variation in suffix in Individual Name affect your decision?

1-F						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT NOTREALNAME PHD	25 ANYSTREET	WASHINGTON	DC	20001	US
2	HUBERT NOTREALNAME	25 ANYSTREET PL	WASHINGTON	DC	20001	US

Figure 7-11 Match example 1-F

In Figure 7-12, should these records be considered a match, a non-match, or require manual review? Why? How does a variation in the last name in Individual Name affect your decision?

1-G						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT NOTREALNAME	200 ANYSTREET RD STE 102	ALEXANDREA	VA	22314	US
2	HUBERT NOTREALNAMES	200 ANYSTREET RD STE 102	ALEXANDREA	VA	22314	US

Figure 7-12 Match example 1-G

The record pairings in Figure 7-13 through Figure 7-16 on page 112 contain data in which the Individual Name is the same, however variations exist in the Address and Postal Code data.

In Figure 7-13, should these records be considered a match, a non-match, or require manual review? Why? Does the absence of a Postal Code in record 1 affect your decision?

2-A						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT NOTREALNAME	450 ANYSTREET	CHARLESTON	SC		US
2	HUBERT NOTREALNAME	450 ANYSTREET TRL	CHARLESTON	SC	29824	US

Figure 7-13 Match example 2-A

In Figure 7-14 on page 112, should these records be considered a match, a non-match, or require manual review? Why? Does the difference in the house number affect your decision?

2-B						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT NOTREALNAME	6201 ANYSTREET PIKE	FALLS CHURCH	VA	22044	US
2	HUBERT NOTREALNAME	6200 ANYSTREET PIKE	FALLS CHURCH	VA	22044	US

Figure 7-14 Match example 2-B

In Figure 7-15, should these records be considered a match, a non-match, or require manual review? Why? Does the additional suite level data included in the record change the decision you made in the previous example?

2-C						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT NOTREALNAME	6201 ANYSTREET PIKE STE 405	FALLS CHURCH	VA	22044	US
2	HUBERT NOTREALNAME	6200 ANYSTREET PIKE STE 405	FALLS CHURCH	VA	22044	US

Figure 7-15 Match example 2-C

In Figure 7-16, should these records be considered a match, a non-match, or require manual review? Why? Does the difference in the house number affect your decision? Does the absence of a unit value in Record 1 affect your decision? What if the Postal Code of Record 2 was “03814”?

2-D						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT NOTREALNAME	140a ANYSTREET RD	PORTSMOUTH	NH	03814	US
2	HUBERT NOTREALNAME	140 ANYSTREET RD UNIT B2B	PORTSMOUTH	NH	03815	US

Figure 7-16 Match example 2-D

The record pairings in Figure 7-17 on page 113 through Figure 7-19 on page 113 contain data in which slight variations exist in the Individual Name and in the Address and Postal Code data.

In Figure 7-17 on page 113, should these records be considered a match, a non-match, or require manual review? Why? Does the difference in Individual Name and Address (“HUBERT NOTREALNAME versus HUBERT NOTREALNAME III”) impact your decision?

3-A						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	HUBERT NOTREALNAME	800 WHATSTREET AVE	BOSTON	MA	02914	US
2	HUBERT NOTREALNAME III	800 W HATSTREET AVE	BOSTON	MA	02914	US

Figure 7-17 Match example 3-A

In Figure 7-18, should these records be considered a match, a non-match, or require manual review? Why? Does the difference in the Individual Name and Address affect your decision? Does absence of a Postal Code affect your decision?

3-B						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	DAVID SOMEFIRSTNAME	720 SOMELOW CRESCENT	LITTLE ROCK	MD		US
2	SOMEFIRSTNAME DAVID	SOMELOUGH CRESENT	LITTLE ROCK	MD		US

Figure 7-18 Match example 3-B

In Figure 7-19, should these records be considered a match, a non-match, or require manual review? Why? Does the difference in Individual Name, Address, and Postal Code affect your decision?

3-C						
Rec Id	Individual Name	Address	City	State	Postal Code	Country
1	MR HUBERT NOTREALNAME	5151 ANYSTREET AVE	FAYETTVILLE	NC	27612	US
2	HUBERT NOTREALNAME PHD	5151 W ANYSTREET AVE	FAYETTVILLE	NC	27621	US

Figure 7-19 Match example 3-C

As you can see, there is much complexity that is introduced during match reviews, and numerous decisions that need to be made. It is imperative that these decisions are documented and match policies created so there is consistency during match suspect resolution and matching during incorporating additional sources into data warehouse.

Define match success criteria

It is very important to understand and define what results that the match process needs to produce.

Typical thinking is as follows:

“The more matches I find the better” or “I want to find at least 20% duplicates”.
“The less clerical records I have the better” or “Can we reduce the number of clericals to 100 per day?”.

However, that type of thinking is fundamentally wrong.

Match is all about quality. The goal for a “good” match is to correctly identify records as either matches, non-matches, or pairs requiring further review.

Figure 7-20 shows typical match results distribution.

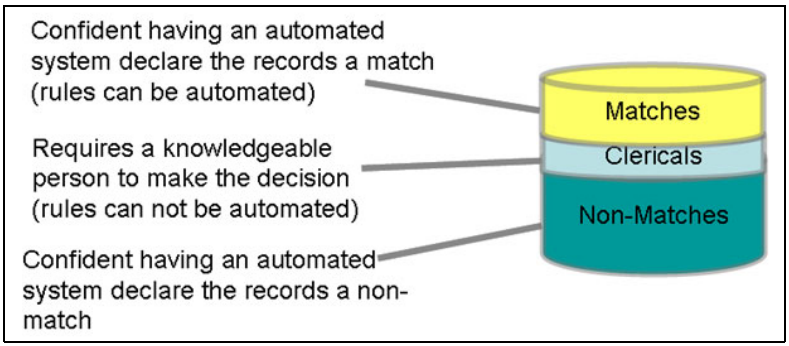


Figure 7-20 Typical match record distribution

Keep in mind that whatever you do, in most cases it is not possible to be 100% accurate in identifying these groups. Lack of data in the records and erroneous data in the record will create false positive matches and false negative matches. The process of match tuning will have to make sure that the number of false positive and false negative records are reduced.

Figure 7-21 presents the typical approach for match tuning and optimization to achieve the best possible match results.

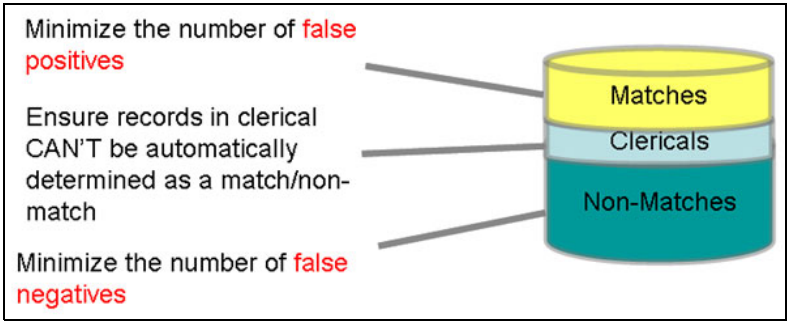


Figure 7-21 Typical approach to match record distribution optimization

If you visualize a portion of the match histogram shown in Figure 7-22, you will see that one of the ultimate goals for the match tuning should be minimizing the size of the shaded (gray) area of the histogram.

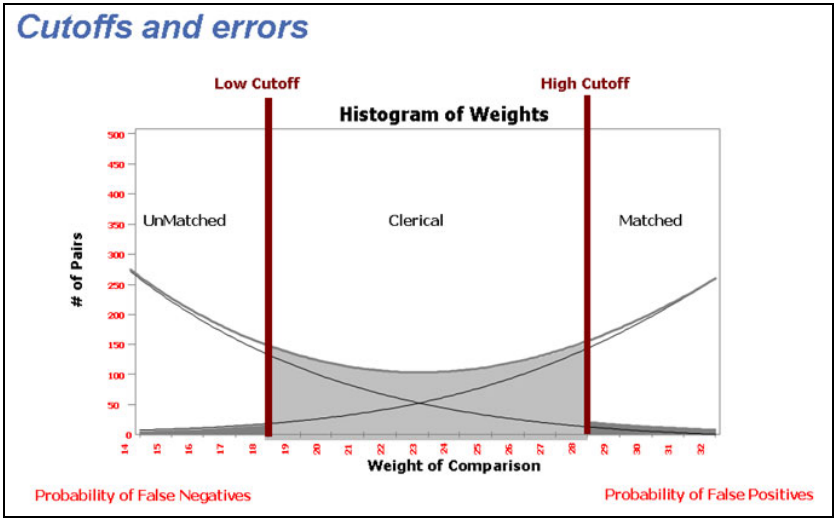


Figure 7-22 Section of match histogram around clerical review records

Fields to use in the match

After describing and determining the purpose of the match with data stewards and what information domain should be used to accomplish the task, the question of what fields from these domains should be used in the actual match specifications arises.

Generally, raw information in the source systems is not acceptable to match on because it is either stored in a free form format or, even if it is parsed into individual columns, content does not always correspond to the metadata. It is very difficult to achieve high accuracy match results when using data in these formats.

As a rule, business intelligence fields and match fields that are generated by Domain Specific Standardization should be used for match specifications.

In addition, single domain fields, such as SSN or TIN (that might require some validation) should be reviewed as possible additional fields to help in determining the match. Initial investigation results could be used to determine the accuracy and field population.

Initial blocking strategy

Within a file with many records, it is very difficult to find matches when trying to compare every single pair in the file, although that seems a logical approach, except for the large files job, which takes a very long time to run because of the number of comparisons that need to be made. (n*(n-1)/2).

Blocking is a scheme that reduces the number of pairs of records that needs to be examined. In blocking, files are partitioned into mutually exclusive and exhaustive blocks designed to increase the number of matches while decreasing the number of pairs to compare. Consequently, blocking is generally implemented by partitioning files based on the values of one or more fields.

The following guidelines should be used when selecting fields for blocking:

- Select fields that are reliable, well populated, and have a good frequency distribution of values
- Use multiple fields in a block
- Go from most specific in the first pass – to more lenient in future passes. Your first pass should result in the most obvious and “best” matches
- Multiple Pass Strategy: Use multiple passes for variation and to “widen the net”
- Use match fields generated from standardization (NYSIIS, Soundex) rather than the business intelligence fields when applicable (First Name, Primary Name, Street Name, City Name)

Of course, if a pair of records blocking fields are not the same in any of the passes of the match specification, that pair does not have a chance to be linked. So defining blocking fields and multiple passes is a balance between scope and accuracy to compare a reasonable amount of “like” records.

Match blocking and candidate selection for presented scenario

Reviewing blocking criteria for the match in the example banking company revealed that the match was set up with five passes with the following blocking criteria:

- Pass 1: Person National ID
- Pass 2: Last Name + Date of Birth
- Pass 3: Last Name
- Pass 4: Date of Birth
- Pass 5: Postal Code

Candidate selection in the delta process is done by four SQL queries against the data warehouse and is based on the following fields in the input records:

- Query 1: Person National ID
- Query 2: Last Name + Date of Birth
- Query 3: Date of Birth
- Query 4: Address Line One

Given the guidelines for the blocking fields specification, the following are problems in the existing blocking and candidate selection:

1. Each of the passes 3, 4, and 5 could possibly bring a very large number of candidates into the match. For example, if incoming records have a Date of Birth of 12/31/9999, pass number 4 will have 150217 candidates to match against (based on our data quality assessment results).
2. As you can see, match blocking criteria and candidate selection criteria are not aligned and that could generally result in missing matches.
3. Query 2 and Query 3 are actually overlapping, and Query 2 results contained within Query 3. It means that running Query 2 is absolutely redundant.

Obviously, selection of the appropriate blocking criteria and candidate selection is a task that needs to be resolved by a QualityStage practitioner. However, data steward involvement is crucial in understanding and rationalizing data quality assessment results and setting up and tuning the match.

Select fields for match comparison

Extended reviews of the match with data stewards and different scenarios of records coming together presented in section “7.2.5, “Work with data stewards to establish match criteria” on page 108” will reveal what fields should be used to do records comparison.

Taking these recommendations and implementing them is a task for a QualityStage practitioner. These are guidelines on the match field selection:

- Do not use Match Fields generated from Standardization. Use Business Intelligence Fields (such as First Name, Primary Name, Street Name, and City Name) and apply a fuzzy comparison:
  - UNCERT (Street Name, City Name)
  - Name UNCERT (Middle Name, First Name Maybe)
  - MultUNCERT (Individual Name)
- Use Exact Comparison methods on shorter fields (as examples, House Number, Codes, Gender, and Floor Value) as well as on fields well standardized (such as Street Type, Directions, State Code, Country Code, and Floor Type)
- Use default match parameters in the initial run
- Use default match cutoff values of 0 in the initial run

7.2.6 Perform lifecycle of application development, user, and acceptance testing

Data Quality applications, as any other software applications, generally require that you must go through a full cycle of application development. As previously described, initial development is an iterative process that includes development tasks by a QualityStage practitioner and review sessions with a data steward to verify the behavior of the validation/standardization and match processes. In most cases, customer counts are in millions and if it is generally acceptable to do initial testing of validation and standardization routines with full volumes, it is not efficient to do match development using the full volume of data. As part of the initial testing, one of the tasks that needs to be resolved by the QualityStage practitioner and data steward is the selection of one, or a few, sample sets. These sample sets need to be representative as much as possible. For example, if during discussion of the match strategy it was determined that Last Name must be the same (allowing some misspellings) in the match pair, the input match sample could be selected by first letter of the Last Name or first letter of phonetic representation of Last Name.

A few iterations of validation and standardization and match testing will ensure that these processes are producing the results that are required. However, the full process of a data warehouse load is much more complex, involving multiple jobs, fields formatting, running SQL queries, and so forth. So it is imperative that user acceptance testing creates different scenarios, runs them through and verifies the results. Again, the role of Data Steward is crucial here to collect, document, and reconcile examples from different sources.

Examples of scenarios for the user acceptance testing could be as follows:

1. A record for “John Smith” was loaded from a core banking system in the initial load. A record for “John S Smith” with the same Address, SSN, and Date of Birth is loaded from the credit card system in delta process. Expectations are that the records are matched and merged.
2. The record for “John Smith” was loaded from a core banking system in the initial load. The record for “John S Smith” with the different address, slightly different SSN, and the same Date of Birth is loaded from the credit card system in the delta process. Expectations are that it will result in a new record created in the data warehouse. As a subsequent update to the credit card system, the record for “John S Smith” is updated with the new SSN and Address that matches the SSN and Address from the core banking system. Expectations are that the result of loading the updated record from the credit card system will cause the merge of records for John Smith in the data warehouse.

7.2.7 Design new and improve existing processes for managing de-duplication

As previously described in section “Define match success criteria” on page 113, there is a choice to set up a match in a way that some of the records where we are not sure in the match outcome to be a set of records that need human intervention. Normally, these records are loaded into target data warehouse as separate customers, and Data Stewards review the pairs and make a decision on whether or not to merge the records. However, it is important to make sure during match development that the size of the bucket for the clerical records is as little as possible (only the records that must be there, are there) because otherwise:

1. It requires enormous manual effort.
2. A subsequent delta match will be performed against unresolved match pairs and that will possibly cause the bucket of clerical records to grow exponentially.

In the data warehouse implementation in our scenario, the bucket size of the clerical records was very large with only a single Data Steward assigned for suspect resolution. The number of suspect pairs that could be resolved by the Data Steward was significantly lower than the number of clerical pairs that the delta process was producing on a daily basis.

That is a problem, and its resolution is twofold:

1. In rewriting the match specification, the example banking company needs to make sure that the match is set up in a way that match pairs end up in the clerical bucket only when it is absolutely necessary.
2. Business processes are set up so enough Data Stewards are reviewing and resolving clerical matches to avoid exponential growth of clerical pairs.

7.2.8 Organization structures, roles, and responsibilities

In the chapter scenario, the absence of a central data governance body during initial implementation of the data warehouse drastically affected the implementation outcome. Creation of a data governance body was an essential piece to overlook, identify, and mitigate data quality issues. Data Stewards play a crucial role in success of the implementation. As project development continues, Data Stewards take on different tasks, at first helping to define requirements, mitigating data issues, and resolving discrepancies between data sources. After implementation is in place, Data Stewards take on a role of resolving generated clerical matches, and monitoring data quality of the new records coming in from the sources. As time goes by and the size of clerical match records hopefully start going down, changes might be necessary to the organizational structures to reassign Data Stewards to different tasks. So the roles and responsibilities of Data Stewards are evolving as time and project implementation phases are completed.

8

Information Lifecycle Management

In this chapter, we introduce you to the Information Governance (IG) core discipline of Information Lifecycle Management (ILM), identify its major subdisciplines, and then focus on one of the subdiscipline areas to better demonstrate the governance-related activity and rigor that is involved in successfully deploying such initiatives. We describe in detail commonly used steps that are followed to build and enable these ILM capabilities and demonstrate their intersections with foundation IG principles and IBM InfoSphere solutions.

8.1 What is ILM

Data or Information Lifecycle Management, Figure 8-1, is the Information Governance core capability that focuses on the roles, processes, policies, and technology that is established to both understand and then appropriately manage the variety of information existing across the enterprise.

With ILM, the characteristic of the information that guides and prescribes the appropriate approaches for handing and managing the respective content is the information's current value and “state or age” in life. This “state” evaluation includes currency and relevance of the information to the organization that must be factored into key decision areas such as proper preservation and protection. This applies to both structured and unstructured types of content and these decisions must include cross functional organization inputs and support.

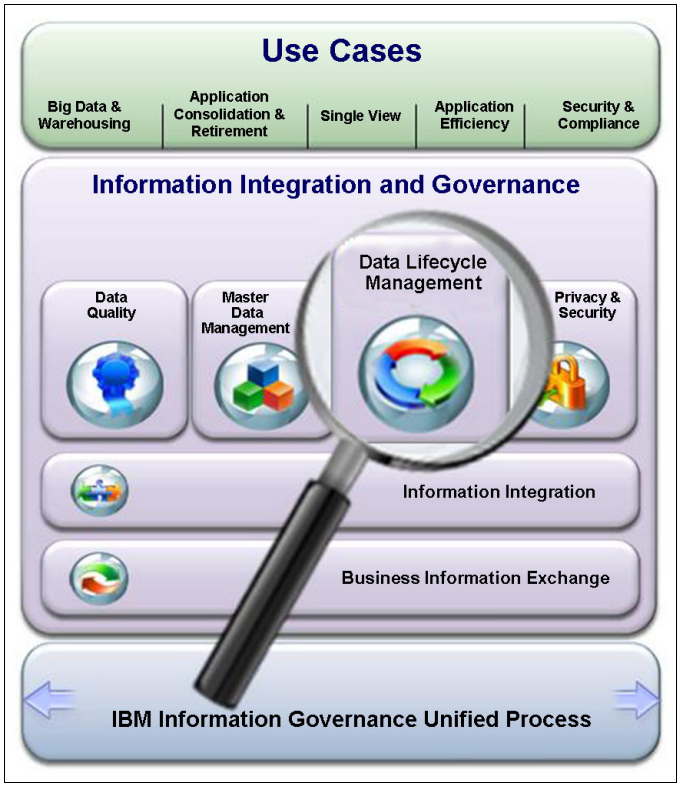


Figure 8-1 IG image: ILM focus

Initiatives such as Data Archiving, Application Retirement (also known as *Decommissioning*), test data management, and Data Masking, are all ILM

disciplines that organizations must apply to their information based on its age and usefulness throughout the organization.

In this and the next chapter, we do a deep dive into two of these ILM capabilities:

- ILM: Enterprise application decommissioning
- ILM: Enterprise test data management and masking

We follow an IG-oriented perspective to explain implementing these ILM capabilities, and focus on the use of the IBM InfoSphere family of solutions as the primary implementation tools. We identify key approaches for successful implementation and identify many of the intersections and dependencies with IG foundation principles.

Some ILM basics

A common foundation governance issue facing companies across industries seeking to implement ILM capabilities, is the need to better understand their information solutions (or applications. We use these terms interchangeably throughout this chapter). Better understanding leads to making more informed decisions. A thorough understanding would include details of the usefulness state of this information, specific application use cases and the technology, systems, and storage that are supporting it.

A most obvious and common *state of being* for information is the “alive and active” state: Current, dynamic, operations relevant content that is typically associated to running the enterprise on a day-to-day basis. It is not difficult to validate the investment in the ILM techniques of governance in this context because its value is usually readily apparent. Proper handling and security and protection are easily justified. The content is at the center of major application implementations such as Enterprise Resource Planning (ERP) or Customer Relationship Management (CRM), and any variety of Business Intelligence or Analytics systems. And it can be appropriately acted upon with the ILM practices such as test data management and data masking.

The explosive growth of big data initiatives is a testament to this “lifecycle relevance” of information. The influx of big data content into many enterprises is highlighting the accelerated rates not only at which information content must be consumed and processed, but also managed within these new and often accelerated time lines of relevant life.

But because data is growing, iterating, and aging all the time, other key ILM best practice techniques must be applied to this information. No matter the actual *length* of the “information life” time line, sound information management practices require constant evaluation and adjustments to ensure that the proper handing and treatment of information is accounting for the trade offs between content age, value, and responsibility. This may require new or enhanced approaches to

assessing and evaluating information, as well as the applications and systems that support them. It requires people and empowered processes to make the often tough, but cost-benefit oriented decisions around the most appropriate handing of these systems and their content.

The discussions that follow on ILM: Decommissioning and ILM: Test data management help you better understand some of the appropriate IG-oriented techniques for treatment of information content from the ILM point of view. The two implementation scenarios that are outlined exemplify successful approaches that are used to managing applications and their associated information from an ILM perspective. In each scenario, efforts are made to call out the dependencies and intersections with foundation IG principles as well as other core IG disciplines (such as security and privacy) supporting overall business value, risk minimization, and compliance assurance objectives.

8.2 ILM: Decommissioning

Data archiving, whether from active applications or retired/retiring systems, is the process of removing older and usually less valuable data content from its current systems and storage medium (typically high cost, online transaction processing-based systems), and off to a usually less costly, non-transaction oriented environment. Done correctly, the benefits include lower cost and risk for the management of this content, as well as the positive impacts of lightening the load on the active application and storage processing resources. In the case of application retirement (or decommissioning), the net impacts will also include the benefits reaped from the retirement of the supporting infrastructure, licenses, and any associated management costs.

Retirement projects focus on applications that have reached or exceeded the “quiescence” stage of life. For various reasons (functionality and data have migrated to other applications, divestiture of business, and so on), the application and data are no longer as relevant transactionally (or operationally). This usually implies a lower business value component to business data retention and a higher focus on record retention and any current or predicted legal oriented compliance or inquiry needs.

And though some content may be disposed of immediately (has already completely satisfied its retention requirements), any other archiving of the data (and any related supporting content) is not the “final deleting” of the data from all locations. It is more a “segmenting out” of the older and usually less valuable data to a more appropriately matched type of storage and accessibility. This latter concept is one of the more complicated concepts and decision areas

organizations grapple with when it comes to data archiving, especially when there is little or no information governance guiding or leading these initiatives.

Most owners or direct users of applications have a hard time defining the “value” of their old versus most current content, other than for the most obvious use cases. This is to be expected because, historically, in traditional data management environments, for structured and unstructured content, most of us have all been “insulated” from ever having to wrangle with any of these types of decisions. All of the data is being provided via the same high end, production-capable support systems and subsystems. There has been no need to make any distinction of what is most valuable versus what is of lesser value. Whether for older content cluttering up the screens and views of the live application, or “stacks” of older applications that are used by maybe one or two agents rarely, it has been easier to “not make” any decisions surrounding proper treatment (or governance) than to run the risk of making a wrong decision. This issue is at the crux of the Information Governance relevance for ILM-Data Archiving and Decommissioning scenarios, and is the most pertinent of arguments as to why IG disciplines are essential in these use cases.

An even more challenging set of characteristics with the decommissioning use case of ILM is the typical lack of clear ownership, understanding, and responsibility that prevails for many of the viable candidates for decommissioning. In some cases, this may even be due to these applications being allowed to linger for years and years, while their respective owners have moved on or retired. Whatever the causes, this is further evidence of the need for the application of governance foundation principle basics, ideally much further upstream of any organized decommissioning initiative, but at a minimum as part of the initial phases of activity to implement such a program.

8.2.1 The high-level steps to enterprise decommissioning

To be successful with an enterprise decommissioning initiative, the organization and the leaders should follow these steps:

1. Assess and understand the landscape of applications and associated content that is targeted for retirement and decommissioning.
2. Build new or adjust existing organization structures and roles to support new and varied governance-related activity that is required in enterprise decommissioning.
3. Build definitions, policies, and processes to support and guide these new roles for directing and controlling the new capability such as:
   – Determining how decommissioning decisions are to be made.

- Evaluate the techniques and technologies to be employed for decommissioning.
- Determine and control how and who will be involved in the decision and archiving project workflows.
- Direct the who, how, and where for creating and processing archived application content that eventually will be recorded and integrated into the *operationalized* solution decommissioning capability.

4. Communicate the new program and framework changes to the enterprise and build the necessary collateral and tools to support.
5. Select technologies, implement, and enable resources to staff and support the various capability process executions.
6. Establish metrics, execute projects, and perfect it all into an operational capability.
7. Apply regular monitoring and reporting on objectives/achievements.

In the balance of this discussion, we take a deeper dive into this commonly observed set of steps that can guide the building of an enterprise decommissioning capability. This approach assumes the use of InfoSphere Optim Data Archiving solutions for the physical archiving /decommissioning of structured data. We also identify related IBM InfoSphere governance-oriented solutions that can facilitate the associated tasks involved in application decommissioning project executions, and others that support the overall ILM capability maturity goals. We look at where intersection points exist with IG foundation principles (for key definitions, policies, roles, and responsibilities and processes) and provide some insight into how foundation governance disciplines support higher value outcomes.

8.2.2 Understand the information: Assess the application landscape

One of the most challenging aspects of establishing an enterprise decommissioning capability lies in the first step of assessing and fully understanding the “age-to-value” characteristics of information. A very large part of improving the overall ILM capabilities within an enterprise comes back to fundamental information governance principles of understanding the data.

This understanding and value assessment examines and records characteristics such as age and value, but especially with aging content in live applications, age alone may not be the only determining factor. The key is to decipher and document the current useful operational value from the historic “origins” types of value, removing all non-logical or emotional attachments to the content. The oldest data in an application could be as relevant as the newest if there is evidence that it supports some critical, transactionally relevant (meaning must be

accessible with the highest or most extreme service level agreements (SLAs)) use cases. Similarly, the most current piece of content could be a candidate for archiving hours after it is created (which may be the case with some forms of streaming big data, for example).

With enterprise decommissioning initiatives, it is usually assumed that the candidate applications and associated information is already at a state where the content is either now replicated in some surviving application or intrinsically no longer as relevant operationally to the organization.

This relevance or importance rating must be examined and assessed via a set of key groups or “stakeholders”. Figure 8-2, further describes these primary players:

- The business: The owners and users of the applications
- Records management: Official coordinators of records of information policy, preservation treatment, and disposition of data
- Legal: Legal or audit department teams that know of any current or future potential for information “holds” or legal discovery requirements
- IT: All technology team members that must manage applications, including any special or complicated technical details regarding applications or data

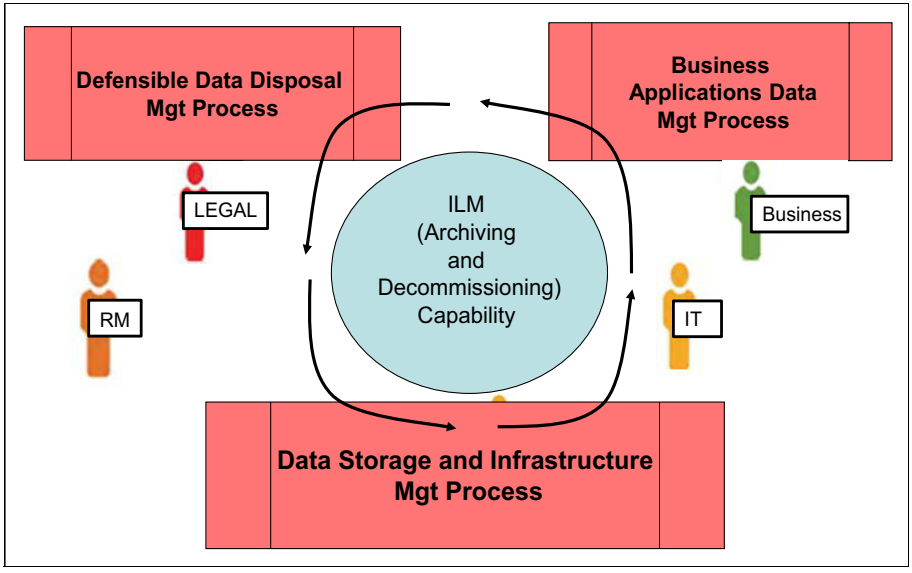


Figure 8-2 ILM stakeholders

Each stakeholder community brings a different perspective and set of requirements to the shaping of the application decommissioning capability:

The *Business* is concerned with competing requirements of retaining access to any necessary information to run the business successfully while not incurring the costs of maintaining old or outdated systems and storage (that also clutters up and slows down their transactional application processes).

The *Records Management* department is focused on the “types” of information the applications store, what “classes” of data this represents and the appropriate handling and retention implications of these assigned classes.

The *Legal* stakeholders represent the organizations that are concerned with a need now or in the future for accessing or providing access to information content as either part of a litigation-related action or imposed information discovery request.

And *Information Technology* is looking for direction from all of these stakeholders, as well as organization leadership, to balance an ever decreasing budget for infrastructure, systems and technology, with ever increasing needs for more information integrity, security, privacy, and availability.

These perspectives must be captured, organized, and evaluated for each candidate application or source of *decommissionable* content in order to both assist the capability leads with making the correct decisions around size of effort and technique for archiving the content as well as the ongoing management, access and eventual disposition. These decisions require enabled and authorized bodies, possibly requiring organization changes (see 8.2.3, “Organization structures, roles, and responsibilities” on page 134 as well as policies and procedures for both collecting and organizing the information and making the decisions. (See 8.2.4, “New decommissioning policies and processes” on page 145).

Figure 8-3 on page 129 identifies some of these enabling and supporting (foundation) elements using a familiar Information Governance model.

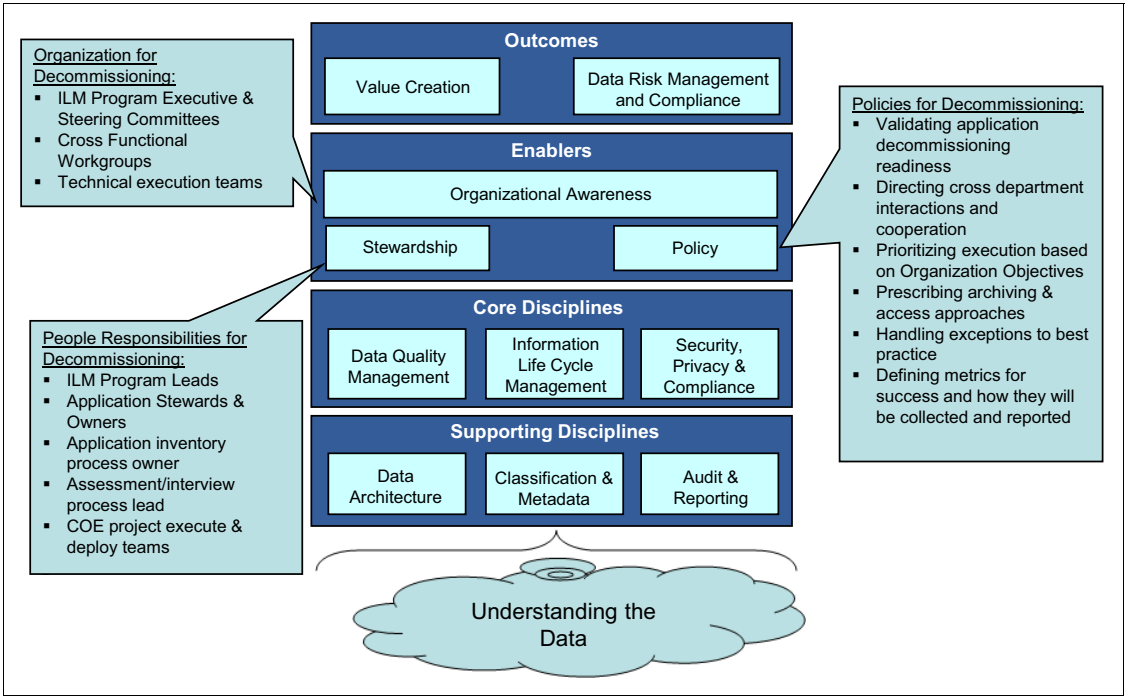


Figure 8-3 IBM Data Governance Council Maturity Model: Foundation Details

Table 8-1 on page 130 identifies a sampling of typical application characteristics that should be available for the decision-making processes involved in decommissioning capability planning purposes. In a perfect world, much of this information would be available as a product of ongoing efforts at implementing the foundation IG capabilities, such as information definitions, policies, and workflows. But the reality is that even if some of this foundation capability is in place, efforts may need to be conducted to enhance or supplement (unify) varying stakeholder perspectives. With new IG capability efforts, new procedures and ideally prescriptive organization policies will need to be established and leveraged to handle any gaps in knowledge, conflicting opinions, or exception cases. All of these should be driven by IG-oriented, organization-wide success criteria and measurement capabilities.

In the second column of the table, for a number of the characteristics, some technology suggestions for tools that can be applied to assist with capturing this content are also identified.

Table 8-1 Understanding the decommissioning candidates

Characteristics	Details (examples)	Facilitating tools and technology
Candidate for Decommissioning:		
Is the solution actually DECOMM ready	Has the appropriate evaluation been done to decide the solution is a fit candidate for decommissioning?	IBM InfoSphere Guardium tools can be employed to monitor DBMS level activity to help with assessing and validating low/no activity systems/databases.
Is there any current legal hold or legal matters pending?	Legal department status regarding legal holds on any content or existence of any open legal matters.	IBM Atlas and ILG family of solutions can be utilized to both record, track and manage, record management policy details such as retention and legal hold.
Are all current retention requirements already satisfied?	Can the Records Management team sign off on any/all retention requirements as already being complete?	IBM Atlas and ILG family of solutions can be utilized to both record, track and manage, record management policy details such as retention and legal hold.
Solution Migration/ Rationalization	Is the solution in the process of undergoing migration or rationalization to a surviving solution, and are there any urgent or impending hardware/license EOL requirements?	The family of InfoSphere Information Server solutions, including Business Glossary, can be a source for defining cross systems integration activity as well as information lineage.
What candidate forms of archiving approach are applicable and available?	Are there organizations aligned and approved - candidate decommissioning format approaches? (examples): -Paper/Digital Reports -Export content/application to a File System -Short term retain or virtualized environment -Express (one chunk) or time period sliced (Optim relational) Archives	This is where ILM Program and Center of Excellence leadership and guidance can be leveraged to focus on the shaping of prescriptive policies. Among other traditional forms of archiving content (such as printed reports) InfoSphere Optim archiving solutions will be a key component to support intelligent and query capable, electronically archived content. InfoSphere Optim Search may also be utilized to support cross application or enterprise data search type flexibility.

Characteristics	Details (examples)	Facilitating tools and technology
Information solution inventory:		
Vendor-based solution or in-house built?	Is the information solution a packaged or vendor-based system, or was it built in-house? If a vendor system, will access to knowledgeable resources be possible?	
Core data characteristics	The DBMS or system storing the data, all code pages that the content depends upon, OS requirements, environment characteristics, data sizes, row or items counts, the array of data types (including any BLOB, CLOB, and so on), storage types and medium, storage reclamation capability.	IBM InfoSphere Discovery & Metadata Workbench can be useful in analyzing and classifying database content. IBM InfoSphere Business Glossary can also be leveraged as a centralized knowledge repository for this detail.
Core information characteristics	Is the information single or multi-geography/country based? Is the information content inseparably dependent upon the application (for example, a cube type analytics system or proprietary algorithmic processing).	IBM InfoSphere Discovery & Metadata Workbench can be useful in analyzing and classifying database content. IBM InfoSphere Business Glossary can also be leveraged as a centralized knowledge repository for this detail.
Ownership-rights to decide	Who owns or has decision rights over the application and its content. Have the four (Business, Legal, Records Management, IT) primary stakeholders provided their key perspectives and requirements?	IBM InfoSphere Business Glossary can also be leveraged as a centralized knowledge repository for this detail. IBM Atlas can document legal and records management responsibilities.
Data/Business Objects	What are the specific Business and or Legal/Records Management sets of subject areas or groups of content that represent a complete business object or legal presentation of information, (for example- a Customer, an Order, a Financial Statement, a Clinical Trial), and how many of them are there in the solution?	IBM InfoSphere Discovery can be used to analyze and assess/present potential Data Objects (related entities) within structured data sources, which many times can be used directly in Optim DG and TDM solutions to support the building of business object (or Access Definition) configurations.

Characteristics	Details (examples)	Facilitating tools and technology
Source of Record	Of the business objects identified, are all or part of them the definitive “source of record” for this information, or is this replicated or shared with another system?	IBM InfoSphere Discovery Transformation analysis capabilities can identify cross data source mappings and overlaps and share this metadata across the InfoSphere family of tools.
Federated Content	Is information shared with other application solutions? Is any of the application content a participant in an up stream or down stream system (for example, data warehouse)? Does the complete business object of information consist of a union of this content? (In other words, is the object incomplete without the federated pieces?)	same as above
Data Sensitivity/ Classifications	Is any of the information solution content considered sensitive or private and is the handling (masking or encryption) well understood - especially when stored outside of the original application?	IBM InfoSphere Discovery Sensitive Data analysis capabilities can identify sensitive data types and share this metadata across the InfoSphere family of tools.
Record Classes (mapped to content)	What specific Records Management classes apply to the information solution content and has this been mapped down to the business object level?	
Legal/Business Discovery or Reporting	When and if any archived content viewing or reporting is required, are these requirements and SLAs clear and well defined? For all stakeholders?	IBM InfoSphere Optim and Optim Connect solutions provide access to archives via any ODBC/JDBC compliant tools. Additionally, InfoSphere Optim Search can be deployed to create web search like flexibility for searching across multiple enterprise archives.
When did the solution start collecting data? What is oldest data in system?	Key data content facts that provide statistics on data age and access.	IBM InfoSphere Discovery & Metadata Workbench can be useful in analyzing and classifying database content. IBM InfoSphere Business Information Exchange can also be leveraged as a centralized knowledge repository for this detail.

These (and other) detailed characteristics are key pieces of information for two primary reasons. The first is to facilitate a more accurate and cost effective coordination and management effort for the evaluation of candidate applications. An organization lacking any current decommissioning operations must build their operational capability to not only process the actual decommissioning project executions but to also perfect the process that evaluates and determines appropriate decommissioning approach and processing decisions. Understanding the landscape is critical to making informed and comprehensive decisions. And, in the larger scope of IG, this information is part of the foundation understanding of the enterprise's applications or solutions.

The second reason is this detailed understanding allows the organization to assemble a more complete level of effort assessment (overall application complexity levels/groupings of candidate solutions), facilitating more precise and predictive, resource and technical capability planning, as well as an opportunity to consider the design and enforcement of some prescriptive ILM techniques and approaches, that can result in higher returns on investment for the overall program.

For example, if a common/consolidated technique for providing access to archived content can be identified for many or the majority of different decommission candidate applications, that can satisfy each's requirements to at least some minimum SLA, this can be a "prescribed" solution component and save resource investment in attempts to build many one-off type solutions.

The collecting of the solution content characteristics is usually accomplished via some initial (and then ongoing) application surveying and inventory process (an IG foundation capability - Data Definitions and Metadata Understanding). As equally as important will be the organization structure, policy, process and monitor/measurement capabilities that will need to be leveraged or matured to even get adequately through this initial step of "understanding your data". Without these, the ability to successfully collect and organize accurate, aligned, and corporate evaluated characteristics is at risk, and hence puts at risk the ability to create a suitable and responsible ILM-decommissioning capability.

Finding and assigning owners of applications, establishing processes for identifying and interrogating applications, building inventory collection tools and collateral - all of these become key factors of success. There is a spreadsheet, "Understand Data with Foundation Principle Area" available that reinforces this interdependency with foundation IG principles and capabilities by providing some examples of specific governance foundation dependencies for many of the "understand data" characteristics we identified in this section. For instructions about how to access that spreadsheet, refer to Appendix A, "Additional material" on page 239.

8.2.3 Organization structures, roles, and responsibilities

Most organizations faced with implementing new or expanded information governance initiatives accept that there will need to be changes or additions to their existing organizational structures and personnel or roles. The more challenging decisions are determining to what extent these changes need to occur and at what pace. Most organizations will shy away from “over the top” type efforts, building the perfect new monolith of committees and resources and execution handlers, before they understand how deep and involved their governance initiative will be, especially in economic times where everyday leaders are being asked to do more with less. Still, realistic IG champions will recognize that new structures and roles will be a key factor of success, and finding the correct pace of development and player assignment to support the initiative will be essential.

All of this applies to establishing an enterprise decommissioning capability, and a common approach to organizing departments and people often looks something like the model that is pictured in Figure 8-4 on page 135.

If you compare this to the information in Chapter 3, “Business definitions and policies in IBM InfoSphere Information Server” on page 41 and the Structured Communications Model discussions, you should be able to pick out the key components of:

- Executive Sponsorship (Executive Committee)
- Data Governance Council (Cross Functional Steering Committee)
- Data Governance Office (depicted as the Working Group below, who might be new members or existing members of the DGO, with specific to ILM focus)
- Business Process Owners, Stewards, Custodians (Operational and Execution teams)

As noted in Chapter 2, “Information Governance organizational structures” on page 19, precisely what you name these organization components is much less important than the assignment of clear responsibilities and the management of the same.

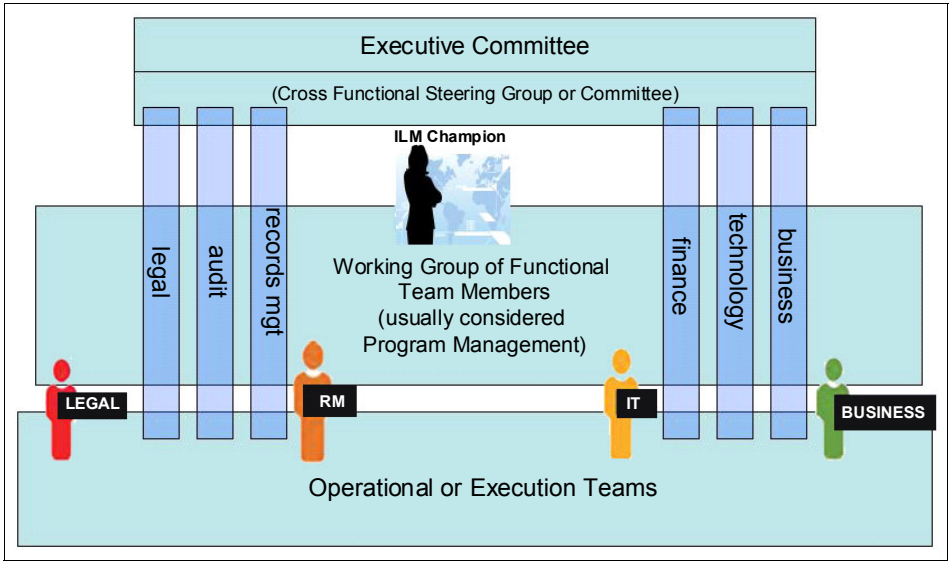


Figure 8-4 ILM organization structure

1. An Executive level Steering Group or Committee: A cross functional committee that represents the senior, C level management vision and objectives of the organization that in theory should be supporting and driving the ILM (in this case Solution Decommissioning) initiative.
2. A Working Group of cross functional leadership that will be charged with driving and shaping and building the framework for success. This includes not only the core capability (such as decommissioning in this case) but any and all supporting (foundation) IG capabilities.
3. Operational or Execution teams: For whatever is determined are the comprehensive set of improved or newly created operational capabilities. For example, in our enterprise decommissioning scenario the key (new) execution capability and team will be the Optim project archiving team. But enhancements to existing or new roles may also be required for the current records management, legal and business teams to support new ILM processes and decision making.
4. IG/ILM Champion: An officially recognized and empowered, multi-faceted governance champion. This role might be an expanded duty of an existing IT or Business team member but is becoming more recognized as a distinct role, especially in the early stages of the core governance capability development. Eventually, as the capability gets instantiated and operationalized, the champion role may move on to the next IG core initiative or back to more foundation governance initiatives (there is never usually a shortage of new initiatives in most organizations where the skills of this champion can be put

to good use). The skills that are required from this role include a cross department navigation and negotiation expertise, combined with a blend of the right amount of business process and technology and architecture understanding. The champion is a person who can converse with and manage both C level and project team level personnel, and possesses a passion and conviction for accomplishing the targeted IG core capability implementation and operationalization goals.

5. Functionally Aligned Information Stewards and Custodians: These are a commonly understood necessity in any IG initiative and should become some of the leading players in the operational execution as the capability matures.

A caution is warranted here regarding value and effectiveness of establishing focused organizational structures and roles. It is not uncommon to see a reasonable framework of new or modified people and roles mapped out, and not embrace the empowerment and monitoring of the same. One without the other will ultimately put a strain on the overall effectiveness and success of the initiative. This is a key discussion of focus in Chapter 2, “Information Governance organizational structures” on page 19.

To demonstrate the potential dilemma, in most organizations, setting up IG supporting structures and roles will either:

- Start with existing resources and add new roles and responsibilities that support the IG initiative.
- Add new resources and associated roles and responsibilities supporting the initiative.
- Some combination of both.

But setting these up, assigning people or creating new “jobs”, is really the easy part. The gaps get exposed when the new policies and processes and other underlying foundation capabilities are not established in support of these roles. This dependency needs to be understood and developed as part of these organization changes. All are critical to support the success of the initiative. Responsibilities will very often be too loosely defined, ownership without empowerment, and not enough attention on management and monitoring.

Without this empowerment, a common result is a very top heavy and bottom heavy set of resources, that present the facade of governance, but cannot be effective because the middle layers of connective roles and responsibilities (stewardship) are neither clearly enough defined or adequately authorized to support the building of an effective model for timely information governance execution. In Chapter 3, “Business definitions and policies in IBM InfoSphere Information Server” on page 41, we discuss these organization and role concerns, and introduce the IBM Structured Communications Model as an approach for better understanding the dynamics of this area, balancing the

correct set of roles with the appropriate number of structured interactions, communications, and outcome flows.

Table 8-2 outlines a typical set of roles that are required in an ILM and Decommissioning or Archiving COE initiative from the Working Group and Operational team perspective. A “Steward” indicator is tagged to some of the roles to identify where the role typically has broader, ongoing IG role significance. The description column includes some solution decommissioning project participation examples, roles that they might play in each decommissioning execution.

IBM InfoSphere Optim Solutions are assumed to be the technical solutions that are leveraged to deliver the ILM capability.

Table 8-2 ILM COE and supporting organizational roles

Roles	Description
ILM COE: definition	The ILM COE is usually a virtual group, made up of key stakeholders and empowered resources, led by an assigned person or set of persons who will be responsible for identifying, designing, and building the roles, definitions, and policies and processes that are required to create and then operationalize the ILM capability. The term Optim COE is also used - but this can also sometimes refer more to the technical, archive execution team. More commonly it will be a group rooted in the Data Governance Office but with ILM-specific focus and know-how.

Roles	Description
ILM Program Lead/Architect	► Chairs/leads (champion) oversight or steering committee and participates in planning, solution evaluations, and review and approvals of new COE projects. ► Leads program level, business process improvement, integration, and change management efforts. ► Assist with design and implementation of COE operational workflows and policies. ► Leads ILM/COE reporting and monitoring, report out to senior committees. ► Leads solution inventory efforts: meetings, decision reviews, schedule, and capacity planning. ► Assist in complex Optim technical designs and project initiation processes.
COE/ILM Project Manager (PM)	► Responsible for leading the operation decommissioning/archiving project team in delivering a solution to the LOB customer using a governance body approved, archiving solution implementation methodology. ► Responsible for coordinating and scheduling of all (core COE and non-core) project team members to support timely and efficient project execution. ► Responsible for organizing and collecting appropriate business measurements and terms and conditions for the project according to the project team aligned and validated scope, objectives, and approach. ► Leads (with Solution Implementation Manager) project kick-off and key project meetings and milestone transitions. ► Leads formal delivery, regular reporting and sign-off of all solution deliverable documentation.

Roles	Description
Optim Solution Implementation Manager (SIM)	► Assist program leads in design and build of new ILM processes and policies. ► Support program leads in solution inventory efforts; meetings, decision reviews, schedule, and capacity planning. ► Assist and guide project manager and program lead in individual initiation and planning tasks - new Optim projects. ► Lead efforts in defining and maintaining the scope of individual decommissioning or archiving projects. ► Assist project manager in plan articulation, project kick-off, and key methodology meetings/milestone transitions. ► Provide expert guidance in solution design, configure, and deploy phases. ► Coordinates any solution architecture framework modifications and change control with ILM program architects.
Optim Developer/Delivery Consultants	► Execute all core solution methodology (Analyze, Design, Configure, and Test) tasks. ► Assist project kick-off with PM and SIM, as well as all other key project methodology meetings and milestone transitions. ► Installation of any required solution component software. ► Leads preparation for deploy tasks, migration of tested solutions to application and systems teams, and supports delivery of all project deliverables.

Roles	Description
Line of Business:definition	Line-of-business type resources will include all business functional types, those closest to specific solutions (applications) being considered for archiving or decommissioning as well as more day-to-day oriented departments such as business operations and finance. These subject matter experts (SMEs) bring critical requirements insight to the ultimate solution design and eventual execution.
Business Solution Subject Matter Experts (SME) *Business Process Owner/Steward*	From the application team, business process group, finance, and operations management. Primary Performer in: ► Project kick-off and follow on meetings/discussions ► Requirements workshops ► Solution requirements and design validation ► Functional and performance testing Contributing Performer in: ► Data Model/Data analysis ► Business Object discovery ► Production readiness planning/validation
Business Solution Data Specialists *Steward*	Typically, most knowledgeable in the specific data usage and designs - a classic data steward type role. Primary Performer in: ► Project kick-off and follow on meetings/discussions ► Requirements workshops ► Data Model/Data analysis ► Business Object discovery ► Collect volumes and statistics ► Solution requirements and design validation ► Functional and performance testing Contributing Performer in: ► Pre-project interviews

Roles	Description
Business Solution System Architect/Security Administrators *Business Process Owner/Steward*	Personnel responsible for the day to day and ongoing systems operations and management. Critical for understanding current and future organization architecture and technology designs/plans. Primary Performer in: ► Pre-project interviews and prerequisites ► Preliminary and ongoing Solution Architecture design ► Solution installations ► Requirements workshops ► Solution requirements and design validation ► Systems and performance testing ► Production readiness planning/validation ► Solution deployment/execution Contributing Performer in: ► Project kick-off and follow on meetings/discussions
Records Management:definition	Critical players who understand record class type assignments to application content. In many larger ILM and elated governance initiatives these teams are often also undergoing process maturity assessment and improvement efforts.

Roles	Description
Records Management Coordinator (RMC) *Business Process Owner/Steward*	RMCs are the stakeholders who can help the ILM team understand the appropriate record classes for each business object (or subject area) of content in each application and the respective record retention processing requirements. Primary Performer in: ► Lead guidance on records retention schedules ► Assists with mapping data sources to record classes/retention schedules ► Project kick-off and follow on meetings/discussions ► Requirements workshops Contributing Performer in: ► Production readiness planning/execution ► Solution requirements and design validation ► Functional (UAT) and performance testing
Records Management Process Owner	Bring to the discussion any current or planned process improvement characteristics and interfaces that core ILM team should be considering. For example, to perfect more automated and integrated processing and management capability. Primary Performer: ► Project kick-off follow-on meetings/discussions ► Preliminary and ongoing Solution Architecture design ► Requirements workshops ► Solution requirements and design validation Contributing Performer: ► Functional and performance testing

Roles	Description
Legal/Audit:definition	The legal/audit team is a key ILM team component and will be responsible for establishing and defining any content handing and management from a legal or compliance (audit) perspective. Legal event management, eDiscovery, and reporting all play a part in the solution decommissioning design and execution.
Legal Hold & Discovery Coordinator *Steward*	Primary Performer: ► Pre-project interviews and prerequisites ► Requirements workshops ► Solution requirements and design validation Contributing Performer: ► Project kick-off and follow on meetings/discussions ► Production readiness planning/execution
Legal Counsel	(same as above)
Compliance Specialists	(same as above)
Information Technology:definition	The typical array of IT specialists needs to be part of the ILM team. Their roles are usually more prominent in the early planning/assessment and then ongoing/operationalization.

Roles	Description
Solution Database Administrators (DBAs)	Primary Performer: ► Pre-project interviews and prerequisites ► Preliminary and ongoing Solution Architecture design ► Solution installations ► Solution requirements and design validation ► Systems and performance testing Contributing Performer: ► Project kick-off and follow on meetings/discussions ► Data Model/Data analysis ► Business Object discovery ► Collect volumes and statistics data ► Requirements workshops ► Production readiness planning/execution
Solution and Systems (Infrastructure) Administrators	Primary Performer: ► Pre-project interviews and prerequisites ► Preliminary and ongoing Solution Architecture design ► Solution installations ► All functional, systems, and performance testing ► Production readiness planning/execution ► Post deploy monitoring Contributing Performer: ► Project kick-off and follow on meetings/discussions ► Requirements workshops
Storage Systems Administrators	(same as above)

A good way to evaluate the policies and processes and underlying definitions that are essential to support your new organization structures within your decommissioning program is to build a cross matrix between these new roles and the IG foundation principles areas.

There is a spreadsheet, “Roles and Responsibilities with Foundation Principle Area” available that provides an example mapping between many of the roles

outlined in Table 8-2 on page 137 and the IG Foundation capabilities, with some details about ownership and authorization regarding policies and processes (or workflows). For instructions on how to access that spreadsheet, refer to Appendix A, “Additional material” on page 239. This information assumes an InfoSphere Optim-based archiving and decommissioning approach.

More is discussed on these critical governance components in the following 8.2.4, “New decommissioning policies and processes” on page 145 section.

The final take away for you should be an appreciation of how just creating new roles and committees and working groups, is not enough, and that these people and structures must be empowered to create or leverage the IG program established policies and processes to affect the changes that are necessary to implement any core governance capability initiative.

8.2.4 New decommissioning policies and processes

Identifying and building the correct number of ILM decommissioning policies and processes (or workflows - these terms will be used interchangeably) are key to achieving a successful enterprise decommissioning capability. In many ways, the final set of policy-supported procedures are what in effect make the operationalization of the capability possible.

There are a number of IG foundation capability type policies that should be articulated, implemented, and enforced as part of overall foundation IG work. Some of these will be very atomic, lower-level prescriptions for behavior and duty with regards to data stewardship. For example, a policy for the correct way to add new data elements to a given data model and the amount of rigor to exert to assess the sensitivity and record classification for the new content to ensure these attributes are not owned or mastered in some other system.

In the discussion here, we focus on procedure type policies that are very much the outline or rules for the related processes they empower. They will usually contain some stated purpose for the policy, a set of users or stakeholders to which it applies, and some objectives for measuring the policy and process effectiveness. The purposes should always be in alignment with the organization's high-level vision and objectives for the ILM initiative.

A simple way to look at policies and processes: The policy describes what and why some IG action needs to be accomplished or some control needs to be put in place and the related process describes who and how it should be done. They are effectively ways to define the reasons and approaches for the behaviors necessary to make the necessary IG decisions to be successful.

In building an enterprise decommissioning capability, we see some recurring policy - process sets that both enable the initiative to get off the ground and running and that lead to a repeatable and measurable operational model.

Table 8-3 identifies some examples of common processes (or workflows) and the related policies that are essential to support them in an enterprise decommissioning implementation.

Table 8-3 Common ILM: Decommissioning process and policy

ILM/Decommissioning process	Related policy (example)	Policy details (example)
Evaluate current application landscape of candidates for decommissioning.	Application ownership and responsibility rules	Guidance and rules for determining who has authority for making decisions that are related to solutions and the definition of the comprehensive scope of these decisions.
	Application Decommissioning Participation and IG-related obligations	Guidance to Business, IT, Records/Legal departments on why and how decommissioning fits into organization-wide IG objectives. Should also include specific department participation requirements and objectives as target achievement goals. Examples would include when and how departments should participate in solution inventory exercises and “fit for decommissioning” evaluation meetings.
Evaluate each application technical complexity and incorporate into planning and scheduling steps for each decommissioning project.	Technical IG obligations as related to ILM and decommissioning	Various department team members are required to contribute to technical meetings and analysis efforts. Policies in this area identify required participation or IG-related metadata management obligations that each stakeholder is responsible for maintaining on an ongoing basis (as part of regular IG data management) or for specific ILM efforts.
	ILM Program level: Guidance and prescriptions for decommissioning (archiving of content) approaches and post archive access.	At an organization level, ILM leadership should establish the guidelines for the approved set of archiving and archive data access approaches to include any prescriptions for aligned designs across the organization to increase value/lower total cost of ownership.

ILM/Decommissioning process	Related policy (example)	Policy details (example)
Execute solution decommissioning and manage archives.	Project implementation methods	Standard and repeatable decommissioning project implementation steps and standards, including supporting collateral.
	Archive retention management policies	Organization-wide rules and standards for applying retention management to existing archives.
	Inter-departmental funding and charge back models	Rules and models for assessing and assigning costs for any decommissioning to the appropriate stakeholder departments.
Collect and evaluate decommissioning outcomes and success criteria.	Decommissioning auditing/reporting objectives and success measures	Standard ILM program direction and guidance on reporting outputs and accomplishments of individual decommissioning efforts and for overall ILM/Decommissioning COE objectives and success measures.
Execute storage and systems reclamation.	Storage management and reclamation policies	Specific policies that support viable and cost effective procedures for assessing current solution storage allocations, utilization rates, and reclamation techniques and strategies that must be put in place to support reaping the benefits that application archiving and decommissioning can provide.

ILM/Decommissioning process	Related policy (example)	Policy details (example)
Manage Record Classes and Retention/Disposition and Legal processing requirements for all information solutions - in alignment with IT and Privacy and Audit processes/policies.	Policies supporting procedures for everything from collecting and recording all records management required intelligence, including legal inquiry and discovery, to acceptable methods for initiating and executing discovery, holds and retention/disposition of archived content. These must be supported by policy driven processes for managing ownership, stewardship, provisioning, and auditing of all information systems	The Records and Legal Management stakeholder area includes a number of processes and supporting policies that are candidates for evaluation and improvement (maturation). The IBM Information Lifecycle Governance (ILG) solutions team in alignment with the non-profit Compliance and Governance and Oversight Council have established a set of (currently) 16 processes that organizations should both assess and mature to achieve higher levels of ILM-related information governance and control. The first nine of these focus on Legal and Records Management and can be read about in more detail in the Information Lifecycle Governance Leader Reference Guide at the following URL: https://www.cgoc.com/files/CGOC_ILG_LeaderReferenceGuide.pdf See also the following URL: https://www.cgoc.com

With enterprise decommissioning initiatives that develop in concert with the perfecting of the supporting IG policies and procedures, it is crucial to make sure these are established with at least a rudimentary formality. As we have discussed in the 8.2.3, “Organization structures, roles, and responsibilities” on page 134, well-defined organization structures and processes can be ineffective if there are no policies and legitimate authority to support the fulfillment of their duties.

These policies often need to be devised and enacted with some urgency to support the core capability implementation at hand. They require more than the usual amount of scrutiny and adjustment in their initial deployment. But they ultimately contribute to long-term ILM program archive and decommissioning management success.

The following high-level workflow diagrams further depict some of the key processes or workflows that need to be established and matured in any enterprise decommissioning effort. These would also apply to an ongoing, live application archiving capability. The workflows emphasize the cross functional participation that is required to be successful. For instructions on how to download the “Work Flow.zip” file from the Additional Materials Directory on the ITSO web site folder, refer to Appendix A, “Additional material” on page 239.

8.2.5 Refine program and communicate

As the organization builds out the people, policies, and processes there will be a need to create and perfect additional specifications and supporting collateral - everything from ILM program marketing plans to project execution deliverable materials. These are necessary to support the socialization and acceptance of the new IG initiative and to standardize inputs and outputs of process flows (making it more of a standard operation).

IBM Optim solution methodologies for implementing IG-oriented solutions are one type of this collateral and are an excellent source for the individual project execution level guidance. These are used in all IBM services supported engagements and provide an excellent starting point to developing sets of decommissioning and archiving factory artifacts. These are usually integrated into a client's project management office type materials to ensure their acceptance and use. In the case of Optim solutions, the IBM Information Management Unified Methods (or IMUM) group oversees and manages a set of solution implementation materials accessible to all IBM Software Group (SWG) and Global Business Services (GBS) consultants.

However, decommissioning project execution materials will only make up part of the required collateral. There are numerous other materials that need to be created and matured to support this IG initiative.

In the example workflows that are identified in 8.2.4, "New decommissioning policies and processes" on page 145, a "Supporting Collateral and Specifications" band at the bottom of each highlights some of the categories of materials that will be key to IG maturity. We now examine some of these categories and describe some examples.

Communications plan materials

- ILM program marketing materials
- Meeting and process guidance

This category of materials is focused on socializing the overall program objectives while also identifying clearly each involved department's membership and participation responsibilities.

The marketing materials group will usually include the high-level program objectives, the enterprise envisioned goals and measurements for success, and messaging from senior leadership as to the importance and priority of the initiatives. Many times these materials are located in a centralized web or shared materials type site making them universally accessible to the enterprise. Following are some examples for a decommissioning initiative:

1. Slide presentations and recordings from senior management on initiative importance to the enterprise.
2. High-level objectives for return on investment and impacts to enterprise bottom line: for both hard costs (storage and equipment) and process improvement impacts.
3. High-level objectives for compliance-related concerns such as information retention and disposition.
4. Clear guidance on program, working group, and stewardship responsibilities including guidance on regular meetings, discrepancy, and remediation processes and chain of command.

The meeting and process guidance can take the shape of procedures (supported by policies) or workflow type materials that make it clear how organization resources are to engage and support the ILM initiative activities. A critical message of this guidance must be the precedence or priority of any identified activity including any ramifications of not following the prescribed guidance. Within our decommissioning capability initiative these could include:

1. A list of specific meetings and respective functional area responsibilities/owners for each key meeting or interface articulated in the program workflows and processes (such as): the initial application assessment meeting, technical assessment meetings, decommissioning project kick-off meeting, and each of the Optim IMUM documented project-phase collaboration and sign-off interfaces.
2. Guidance for each meeting should include not only personnel required to attend, but also details as to the responsibilities of each role such as information or materials each is expected to contribute.

Another set of materials that can assist with the ILM program socialization process are high-level solution architectural diagrams or blueprints. This material can bridge the gap between existing physical infrastructure, assets and stakeholders, with any new assets, processes, and stakeholders being introduced into the enterprise with the new IG capability initiative.

IBM InfoSphere BluePrint Director is a graphical design tool that can be used to map these concepts and can additionally be linked to internal systems and documentation resources to help ease the transition. Figure 8-5 on page 151 and Figure 8-6 on page 152 are examples of the ILM Blueprint template that is included out of the box when you install IBM Blueprint Director, and can be modified and adapted to your specific ILM/Decommissioning program implementation.

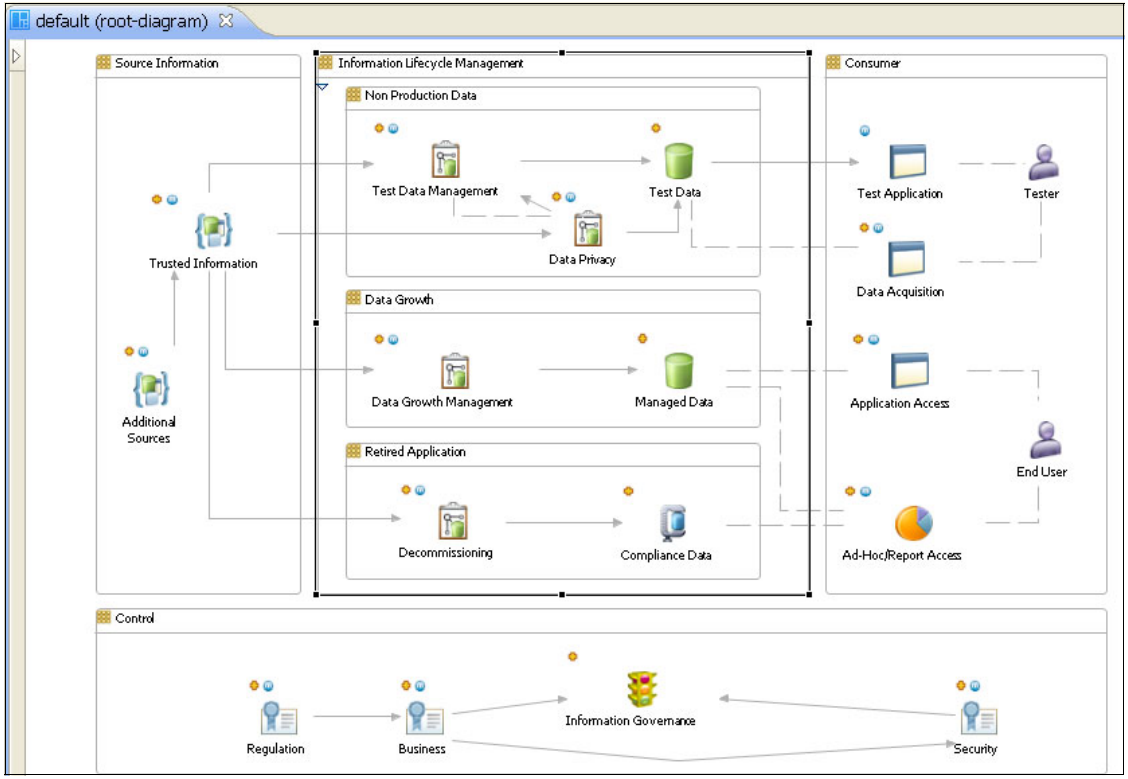


Figure 8-5 IBM InfoSphere Blueprint Director: ILM overview

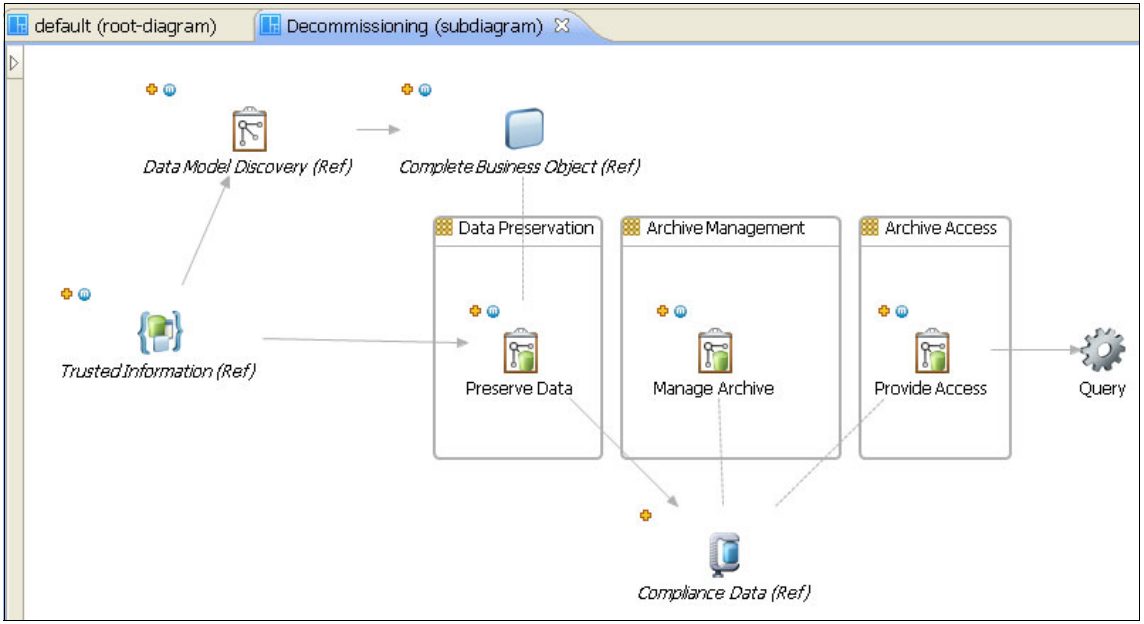


Figure 8-6 IBM InfoSphere Blueprint Director: ILM Decommissioning Blueprint

Each of these categories of materials should be documented and assigned owners in a formal ILM program Communications Plan that should be one of the standing topics of review in the regularly scheduled ILM program steering committee meeting.

ILM program success, measures, and scorecards

The value of setting goals with measures is not a new topic, numerous “balanced scorecard” and similar enterprise management advice is pervasive. The simple truth however is that very few organizations do an adequate job of both designing measures of success and then regularly reporting (and acting) on them. Measuring and monitoring is part of foundation IG principles. For our enterprise decommissioning example scenario, Table 8-4 on page 153 lists some examples of objectives and measures that could support regular ILM program dashboard type monitoring and reporting.

Table 8-4 Example ILM decommissioning program objectives and measures

Objective	Sub-Objective	Measure	Sub-Measure
Reduce IT costs	Reduce overall storage expenditures	Total storage size and cost	► Numbers of applications retired ► Total (net) storage size and cost: before and after decommissioning program ► Total storage size and cost with decommissioning: retired versus non-retired ► Average storage size and cost: retired versus non-retried ► Total storage size and costs: reclaimed/re-purposed versus removed
	Reduce overall (production) processing infrastructure	Total numbers of servers/ processing units	► Numbers of application servers retired and associated costs ► Numbers of application servers repurposed - and cost impact ► Total Servers/PUs (and costs) for all applications: retired versus non-retried
	Reduce overall existing processing infrastructure support costs	Total costs for infrastructure support and maintenance	► Total infrastructure costs (non-maintenance) ► Total infrastructure costs: retired versus non-retried ► Total infrastructure costs: top 10 non-retired ► Total infrastructure maintenance costs (non-hardware/lease) ► Total infrastructure maintenance costs: retired versus non-retired (non-hardware/lease)
Higher value IT focus	Increased investment in new IT assets and technologies	Total percent *all IT costs*: old versus new technologies	► Total Infrastructure Costs (include maintenance): all existing applications and technologies ► Total Infrastructure Costs (include maintenance): new applications and technologies (big data, streaming, social media)

Objective	Sub-Objective	Measure	Sub-Measure
Higher value and lower risk-records management and legal processes	Improve records and legal management processes	Number of applications migrated to standardized, centralized, and automated processes	► Number of applications - retirement candidates - mapped to record classes ► Number of applications - non-retirement candidates - mapped to record classes ► Number of applications - retirement candidates - using new standard processes/tools/automation ► Number of applications - non-retirement candidates - using new standard processes/tools/automation ► Percent of retirement candidate applications in non-compliance (retention management) ► Percent of non-retirement candidate applications in non-compliance (retention management)
	Reduced costs of legal hold and discovery	Cost to process legal hold or legal discovery	► Total and average costs of all HOLD processing - decommissioned applications ► Total and average costs of all discovery processing - decommissioned applications ► Total and average costs of all HOLD processing - non-decommissioned applications ► Total and average costs of all legal discovery processing - non-decommissioned applications

These example objectives and measures would need to be supplemented with some specific, initial baseline and target objective *values* for the proposed measures for each reporting milestone time frame. It is unfortunately common to see the establishing of high-level enterprise objectives without detailed measures and associated target values. This makes it extremely difficult to establish initial benchmark points of reference and to subsequently gauge any sort of incremental progress.

As an example, for the objective - measure “Reduce overall storage expenditures” -“number of applications retired”, you would want to set a current and realistic target for the number of applications to decommission. This would need to be based to some degree on the initial “understand data” analysis, specifically the numbers of viable decommissioning candidates, an understanding of how complex each decommissioning execution would be, and the proposed archiving execution capacity for the respective time frame.

Centralized solutions inventory (data collection forms)

In the 8.2.2, “Understand the information: Assess the application landscape” on page 126 section, we discussed the critical need to gather key, governance-oriented data characteristics about the landscape of applications across the enterprise. When IG foundation disciplines are still maturing, the *where* and *how* and by *who* these collected details of information are managed might not be clear. In fact, the establishing of a centralized repository of not only application characteristics and data definitions, but also related policies and responsible owners is a key evidence of higher levels of IG maturity. So in early phases of IG maturity, this might not be well-defined.

IBM Business Glossary (also known as *Business Information Exchange*) is part of the InfoSphere Information Server family of Information Governance solutions and provides one approach to managing this centralized content. In the 8.2.6, “Implement technologies to support the new IG capability” on page 160 section, we highlight Business Glossary and other IBM InfoSphere technologies that are used to support this IG initiative.

Whatever the determined tool to manage this foundation content, it might be necessary to build data collection tools, forms and questionnaires, especially in the early stages of IG maturity to facilitate the process of collecting and organizing this detail (up until the time that this exercise becomes integrated into the normal operations and corporate technology). Figure 8-7 on page 156, Figure 8-8 on page 157, and Figure 8-9 on page 158 provide some examples of the types of solution detail that these questionnaires might ask of the owners and stewards of the applications that are targeted for decommissioning. The assumption is that this initiative is focused on structured data.

Question	Description
Solution Name	Description - any special internal coding
Solution Business Purpose	Provide brief description of functionality and what business processes it supports.
Line of Business (LOB)	What line of business supports and funds this solution
Line of Business Owner	Overall LOB owner
Solution Owner	Specific solution owner - capable of making ILM-oriented design decisions
CIO Technology Owner	Division CIO or IT director solution falls under
Solution Data Specialists	LOB or IT assigned data specialist or DBA
Solution Data Stewards	LOB or IT assigned data steward (Information Governance role)
Records Management (RM) Coordinator	Who is the RM division assigned person to this solution
Solution Confidentiality or Risk Rating	Identify if the solution contains sensitive or confidential content requiring privatization
Any special Compliance or Audit Rating	Identify if the solution is subject to any corporate or industry compliance or audit rules/reporting
Solution Start Date	When did the solution go into production use - and/or start collecting data (alternatively what is oldest data in system)?
Solution Retirement Date	Identify if the solution is currently at end of life, in process of a migration, approaching retirement. Describe in detail.
Solution Database Platform	Identify if the solution is running on a distributed type DBMS or Mainframe type system.
Business Subject Matter Expert (SME)	Identify assigned business SME and vital contact information
Technical Subject Matter Expert (SME)	Identify assigned technical SME and vital contact information

Figure 8-7 General solution questions

Question	Description
Record Classes assigned to Solution	Identify the specific record management Record Classes assigned to content in the solution.
Record Classes and Information Retention and Disposition	For each Record Class identified, provide its associated information retention requirement, and any special disposition treatment (for example - must be reviewed before disposal, automatic disposal, legal review only)
Have Record Classes been mapped to solution business objects of information? Provide examples.	Identify if there exists any mapping or association of Record Classes to business objects of information (complete sets of related entities). If yes, please list the details/ attach additional details.
Record Retention Event	For each Record Class associated to the solution, identify if there are any business or legal events that trigger the record class associated retention requirement.
Are multiple geographic or country based Record Classes / Retention Schedules applicable to solution content?	Identify the matrix of Record Class by Geo, and specify any varying retention of disposition rules.
Non-Record Retention Period	For each Non-Record data object identified please provide its applicable Retention Period.
Is the solution or any of its content under any current Legal hold?	If yes , explain
Identify any solution content that might be candidate for any Legal Hold.	Identify if specific sets or classes of information can potentially fall under a legal hold or legal discovery status.
If solution content has potential for legal hold, identify all solution preservation levels that might apply.	As an example: 1. Legal hold requires preserving the data only 2. Legal hold requires preserving the data and any solution application/code 3. Legal hold requires preserving the data, solution application and supporting infrastructure

Figure 8-8 Solution records and legal management questions

Question	Description	Score
Technical		
Size	Total size of the database to be archived (GBs)	5 = under 50 1 = over 50
Number of tables	Total number of tables to be archived	5 = under 50 3 = 50 - 200 1 = over 200
Complex data types	Does the solution store any complex data types (LONG, CLOB, BLOB, XML, TYPES, BFILES columns or attachments)?	5 = No 3 = Yes, single 1 = Yes, multiple
Referential Integrity	Is referential integrity designed into the database or in the application? Are Primary/foreign Key constraints defined and enforced?	5 = Yes 3 = No 1 = Unknown
Data Sources	Is the data source a non relational database such as CUBE, VSAM, Sequential?	5 = No 3 = Yes, single 1 = Yes, multiple Non-relational sources
ERD	Is an up to date Entity Relationship Diagram available?	5 = Yes 1 = No
Connectivity	What type of database is to be archived: Mainframe, AS400, Oracle, SQL Server, DB2, Other	5 = DB2, Oracle, SQL Server 1 = Mainframe, Other
Codepages	The database code page determines what characters you can store in the database. For example, if the database code page is 819, then only English and western European characters can be stored in the database.	5 = Yes 3 = No 1 = unknown
Reporting and Access		
Queries	Are there any pre-defined reports or SQL statements available to support decommissioned data access requirements.	5 = Provided 3 = Not Provided- but can be defined 1 = All Ad-Hoc access
Documentation	Does documentation for any business, legal and/or compliance reporting requirements exist for decommissioned data	5 = Yes 3 = partial 1 = No
Service Level Agreements (SLA)	What is the Service Level Agreement (SLA) response times for the delivery of reports of the decommissioned data?	5 =None/negotiable 3 = Average (1 hour to 1 day) 1 = Immediate/(Mins)

Figure 8-9 Solution technical detail questions

The questions in Figure 8-9 include an additional column, “score”, a commonly included attribute that attempts to assess levels of complexity or difficulty associated with the data point being collected. These can be leveraged to develop a total complexity score per application, useful in planning and scheduling individual decommissioning project executions. This information also provides input into the broader landscape understanding regarding complexity. This “data understanding” helps to guide the types and volumes of overall enterprise decommissioning technology and capacity that may be required, short term and over the projected course for full landscape executions. This detail can also be analyzed and filtered to look for common requirements where prescriptive (lower cost/higher ROI) solution techniques can be leveraged for the overall capability deployment.

IBM Information Management Unified Methods for implementing InfoSphere Optim Solutions

The High Level ILM Decommissioning Project Work Flow web link provided in 8.2.4, “New decommissioning policies and processes” on page 145 identifies a category of materials that supports the execution of each (Optim) decommissioning project. These are primarily project management deliverable and project execution guidance materials. The Optim IMUM version of these are what IBM software solutions services teams have developed over the course of numerous IBM InfoSphere Optim implementations and what these teams would bring to the initiative when IBM GBS or SWG services are included in the ILM initiative team. These materials should be familiar to any project management organization and include items such as:

1. Project preparation materials, such as Business and IT team readiness Checklists, that confirm all parties are prepared to engage, have set up required systems and systems access, and similar pre-requisite tasks.
2. Implementation process support materials, such as templates for conducting kick-off meetings, solution overviews, white boarding sessions.
3. Standard project deliverables templates - based on solid project management principles:
    a. Scope, Objectives, and Approach verification and sign off
    b. Solution Requirements Mapping template
    c. Solution Design template
    d. Solution Unit and Systems Testing templates/guidance
    e. Project WBS and management templates
    f. Inter-team Reporting and Sign Off templates

Some additional project implementation-type materials noted in this list include cross department interaction and sign-off forms. These effectively become “contracts” of participation. These will often include materials that outline participating line-of-business charge back details. It is not uncommon for costs of the enterprise decommissioning initiative to be shared across the organization. The ILM Steering Committee and or Program Office should determine and publish the approach via a documented policy, broadly communicated via the appropriate techniques that are identified in the ILM program communication plan.

Additional information about solution project methods and approaches for data archiving can be found in the IBM Redbooks publication “Implementing an InfoSphere Optim Data Growth Solution,” SG24-7936-00 at the following URL:

http://www.redbooks.ibm.com/abstracts/sg247936.html?Open

8.2.6 Implement technologies to support the new IG capability

Our discussion has focused on structured data applications or business solutions with data stored in relational databases or similarly structured data management systems. Though we provide some guidance on IBM solutions that are focused on unstructured data in this section, we continue primarily with a focus on structured data-based applications as part of an enterprise ILM decommissioning capability.

Though technology can include any infrastructure, networking, and software that is required to implement your desired capability, we focus on the solution software. Architecture and related systems discussions are beyond the scope of this chapter.

Generally, we can break the technology selections into two categories or roles:

- Software that supports the core or main function of the capability, which in our case is the execution and management of decommissioning and resulting archived data.
- Software that supports all of the surrounding information governance aspects of the initiative.

IBM InfoSphere solutions family members can support these two roles. Some key advantages of selecting IBM solutions are the built-in integration between solutions. IBM understands and embraces how information management without the guidance and control of good governance can often, at best, be an exercise in “creating the same wheel” many different ways.

Given our ILM decommissioning focus in this chapter, Table 8-5 and Table 8-6 on page 163 list a number of the key, desired characteristics of technology solutions that would support enterprise decommissioning. Specific IBM related solutions are noted along with any comments on cross solution integration between technologies.

Table 8-5 Core decommissioning software solution characteristics

Core capability functionality	IBM solutions supporting ILM Decommissioning	Desired solution characteristics
Core archiving and archive data use	InfoSphere Optim Solutions - including Optim Manager, Optim Connection Manager (OCM) and Optim Service Interface, InfoSphere Optim Connect, InfoSphere Optim Search, IBM Federation Server	Support all major operating systems and platforms

Core capability functionality	IBM solutions supporting ILM Decommissioning	Desired solution characteristics
		Native support for all major databases/data stores, all major data codepages
		Scalable and flexible, heterogeneous platform service delivery (execution) support
		Multi-tier storage support for archived data
		Flexible archive process execution - manual, scheduled, batch, parameterized, service oriented
		Flexible and secure - master archive directory control (what archived, when, archive data locations)
		Flexible and configurable, complete business object focused archiving
		Automatic discovery of archive business objects (relationships and constraints) instantiated in DBMS
		Archive support for complex data types, associated file system files
		Unified (or federated) archive support for cross-data store business object content
		Preservation of relational/source data definition constructs within archive
		Controls/constraints on altering archived content (unalterable content)
		Centralized and secured repository for archive design metadata and archive processing history
		Access and security controls for archive data and archiving development and execution
		Restoration of archive data capability- to origin or redirected
		Familiar and flexible archive data access approaches
		Cross archive data inquiry capability

Core capability functionality	IBM solutions supporting ILM Decommissioning	Desired solution characteristics
Lifecycle management	IBM Atlas eDiscovery solutions family with InfoSphere Optim (includes retention schedule management integration)	Flexible retention assignment to archives, based on variety of content characteristics and driven by assigned record classes for content
		Automatically assignable retention schedules
		Automatic (and auditable) archive data retention and disposition
		Flexible integration capabilities with external records management and legal discovery processes
		Flexible legal hold and related retention schedule adjustment capabilities (hold, resume, split)
Monitoring, auditing, and reporting	InfoSphere Optim Solutions - including Optim Manager, Optim Connection Manager (OCM) and Optim Service Interface	Archive process execution details - monitoring, tracking, and reporting
		Comprehensive audit reporting capability - primary archive processing and management including archive data access and retention/disposition processing
		Archive data/file validation - integrity checking
		Archive directory metadata - reporting and inquiry
		Auditable archive data retention and disposition

Table 8-6 IG related to decommissioning software solution characteristics

IG-related capability functionality	IBM solutions supporting ILM decommissioning	Desired solution characteristics
IG Definitions: Application LOB	IBM InfoSphere Business Glossary and InfoSphere Metadata Workbench (and Metadata Asset Manager). Also, InfoSphere Discovery (integrated with Business Glossary for sensitive/data classification)	Data types, domains, sizes - all technical characteristics about content in data store
		Data lineage, source of content, federation, or shared ownership
		Managed subject areas or business objects-mapping of included content and composition
		Data element sensitivity or “classified” status
		Report and viewing designs for archived content access
IG Definitions: Records Management	IBM Atlas eDiscovery family of solutions and InfoSphere Business Glossary	Record Class-to-Application Content/Business Object mappings
		Retention schedules - per record class, by geography/department
IG Definitions: Legal	IBM Atlas eDiscovery family of solutions and InfoSphere Business Glossary	Guidance or templates for legal hold and discovery and evidence data provisioning (by application/business object or record class)
IG Definitions: IT Systems and Storage	IBM InfoSphere Information Server family of solutions, including Metadata Workbench, Data Architect, Business Glossary	All application storage types, allocations, and utilizations
		Detailed application-to-storage device mappings
		Storage management plans - road maps for consolidation, lease schedules, storage

IG-related capability functionality	IBM solutions supporting ILM decommissioning	Desired solution characteristics
IG Policies and Procedures: Application LOB	IBM Atlas eDiscovery family of solutions and InfoSphere Business Glossary	Rules and procedures guiding what data definitions to collect and where to record
		Rules and procedures guiding how to identify subject areas/business objects that applications manage and how and where to record
		Rules and procedures guiding how to formulate archive viewing requirements with application-related stakeholders and how and where to record
		Rules and procedures guiding how to work with application owners to determine appropriate SLAs for archive data access and how and where to record
IG Policies and Procedures: Records Management	IBM Atlas eDiscovery family of solutions and InfoSphere Business Glossary	Rules and procedures guiding how to work with application data owners to map record classes to application managed business objects and how and where to record
		Rules and procedures guiding how to evaluate record class-to-business object assignments with respect to retention and disposition requirements, make cost benefit decisions on over-retention, or express-type archive retention, and how and where to record
IG Policies and Procedures: Legal	IBM Atlas eDiscovery family of solutions and InfoSphere Business Glossary	Rules and procedures outlining how to provision archived data content for legal hold and evidence management and how and where to record
IG Policies and Procedures: IT Systems and Storage	IBM Atlas eDiscovery family of solutions and InfoSphere Business Glossary, InfoSphere Server Data Architect/Metadata Workbench	Rule and procedures outlining the types of application storage specifications to collect and how and where to record
		Rule and procedures outlining the steps to reclaim or repurpose storage that is made available as part of archiving/decommissioning and how and where to record

Following are brief overviews of some of the IBM solutions noted in Table 8-5 on page 160 and Table 8-6 on page 163.

IBM InfoSphere Optim

The IBM InfoSphere Optim (Optim) suite of solutions supports all of the fundamental or core functionality that would be required within the ILM capability disciplines: Archiving/decommissioning, test system data subsetting, and test system data masking. It has the ability to connect to and extract/archive content from all of the major database platforms and data stores using source defined database catalog metadata intelligence. It can additionally be driven by user designed (Optim level) definitions to accommodate sources with no formally implemented referential integrity or constraint-based rules. It can be scaled to support small and large implementations cost effectively. It can be implemented stand-alone or within a complex scheduling environment.

Optim also supports IG foundation principles with support for the capture and recording of all detailed processing results. It integrates with other InfoSphere family tools (such as Discovery, Atlas, and Business Glossary) to leverage pre-identified and in process defined rules and policies to guide the correct designs in extracting, masking, and long-term management of Optim produced artifacts.

Some of the most recent improvements expand on Optim's IG-related capabilities, adding deeper levels of integration with InfoSphere IG foundation solutions and providing more and improved centralized dashboards for visibility to processing metadata and results.

For more information about IBM Optim, refer to the IBM product information at the following URL: `http://www-01.ibm.com/software/data/optim`

IBM InfoSphere Discovery

IBM InfoSphere Discovery is a core IBM InfoSphere solution that automates the discovery and analysis of data, including relationships and data classifications (such as sensitive or private types of information) both within and across heterogeneous systems. InfoSphere Discovery is a key component in any information governance program - no matter what the core governance capability objective (Data Integration or MDM or ILM) because it addresses one of the ubiquitously required first steps in any of these initiatives: Understanding your data. By automating the complex process of data discovery, IBM InfoSphere Discovery replaces error-prone, manual data inspection methods. This enables you to gain insight and data level understanding of:

- Business objects
- Sensitive and private data
- Overlaps, transformation patterns, and rules

For all of the information included in your landscape, speeding time to value for critical initiatives such as data integration and system consolidation, protection and archiving, or decommissioning, governance initiatives.

The key advantage that it provides is to accelerate the investigation and analysis time frames by potentially multiple factors, with greater accuracy, and higher levels of visibility into potential data problems.

For more information, refer to the IBM InfoSphere Discovery product pages at the following URL:

http://www.ibm.com/software/products/en/infospherediscovery

IBM Atlas eDiscovery

The IBM eDiscovery Solutions family delivers rigorous and efficient electronic discovery processes to meet evolving and complex legal obligations. These products can assess costs and risk more accurately to help organizations make better decisions regarding case strategy. The IBM eDiscovery Solutions family improves alignment among eDiscovery stakeholders and links legal holds to information repositories to streamline legal processes and help reduce litigation costs. These core capabilities of aligning stakeholders and enabling better and more accurate/informed decisions is at the core of what a foundation governance solution should be intending to support.

The family of Atlas eDiscovery solutions includes the following:

- Atlas IT eDiscovery Process Management: Communicates discovery information automatically between IT and legal staff.
- Atlas eDiscovery Process Management: Automates legal holds, collection, and audit processes.
- Atlas eDiscovery Cost Forecasting and Management: Provides more control over discovery costs for improved legal outcomes.
- eDiscovery Analyzer: Provides case review, search, and analysis capabilities to dramatically reduce electronic discovery costs.
- eDiscovery Manager: Enables authorized IT and legal staff to search, cull, hold, and export case-relevant content for more cost-effective discovery processes.

For more information, refer to the IBM website at the following URL:

http://www-03.ibm.com/software/products/en/atlasediscoveryprocmgmt

IBM Business Glossary

IBM InfoSphere Business Glossary is a module in the family of the InfoSphere Information Server family of governance solutions that supports the creation and management of a centralized, enterprise vocabulary and classification system for all information collected and utilized by the organization.

By design, this promotes cross department understanding and alignment, and results in trusted information and consistent business use throughout the organization.

Business Glossary is a perfect starting point for IG-oriented programs. It naturally supports the needs of information stewards by providing the tools to establish and manage business term development and instantiation, as well a lineage and historical source verification.

It supports easy, multi-access mechanisms, including integration with existent organization portals and related tools, thereby empowering all information governance responsible and accountable stewards and users in the execution of their daily governance roles.

From within InfoSphere Business Glossary, users can define terms and categories (information definitions) and information governance policies and rules.

More details of specific BG use are identified in Chapter 2, “Information Governance organizational structures” on page 19, Chapter 3, “Business definitions and policies in IBM InfoSphere Information Server” on page 41, Chapter 4, “Workflows and business specifications” on page 57, and Chapter 5, “Metrics and measurements” on page 75.

Additional information can also be found at the following URL:

http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/topic/com.ibm.swg.im.iis.productization.iisinfsv.overview.doc/topics/cisoproductsinthesuite.html

8.2.7 Recruit and enable resources

In section 8.2.5, “Refine program and communicate” on page 149, we discussed the need to build and present initial socialization and program support collateral. This is an initial form of recruiting and enablement. The goal is to foster long-term support for the enterprise initiative and put cross functional department personnel “on alert” that their participation and active involvement is required to make the initiative successful.

Once underway, and leveraging the roles and responsibilities outlines discussed in 8.2.3, “Organization structures, roles, and responsibilities” on page 134, you are able to start zeroing in on specific training and enablement for the cross functional team roles.

Generally, there are two major skill sets or implementation knowledge areas that you focus on:

- General decommissioning program sessions focusing on program objectives, goals, and enterprise responsibilities to make them a reality
- Specific Optim decommissioning implementation project process and skills training

The first enablement can be a series of regular meetings, typically with a “launch” type meeting once the steering committee and program working groups are established. These are often web or virtual meetings where the focus is to explain the corporate executive vision, describe the organizational level structures and functions that have been defined to support the initiative, and a statement of objectives with targeted measures. These objectives and measures can then be the “dashboard” for regular company-wide reporting.

The second enablement needs to include more core technical or product use type training for whatever specific technical software and solutions are being deployed. These can be supplemented with workshop type efforts that can aide in shaping best practices and methods to follow in these project executions.

For IBM software solutions, there are a number of face-to-face and web-based or self-paced training alternatives.

IBM product enablement delivery is now available via a set of worldwide accessible global training providers. This new initiative called *Global Skills Initiative* (GSI) includes four world class IBM training partners who are the primary client interface for contracting and securing delivery of IBM information management designed and built enablement materials. For more information, refer to the following URL:

http://www-304.ibm.com/jct03001c/services/learning/ites.wss/zz/en?pageType=page&c=a0011023

The project implementation workshop type training is best delivered via IBM product and services specialists. Typical workshops focus less on the products and more on explaining and demonstrating the cross functional project team collaboration required to effectively deliver each individual archiving execution (for example, via the walking through of a mock project exercise).

Workshop sessions also enable program leads to begin identifying the most likely proponents or early adopters of the new program, and can help to further zero in on the best initial candidate to commence on the initial rounds of project executions.

8.2.8 Decommission applications

The ultimate goal of building an enterprise ILM-decommissioning capability is to commence with the decommissioning. Assuming that your efforts at establishing a framework of foundation IG capabilities (organizational structures and roles, information definitions and supporting policies, specifications, and workflows) have been successful (or are at least in the process of being shaped and matured), and you have selected technologies to support the capability and enabled the respective resources. Following are the next steps towards actual execution.

Review candidates and select initial wave of applications.

The first step is to zero in on the first wave of applications that can be submitted into the newly established decommissioning program workflows. If you recall from the 8.2.4, “New decommissioning policies and processes” on page 145 section, these new processes include an Initial Application Assessment and a more detailed Technical Application Assessment workflow. These processes seek to first qualify and verify the application as a proper fit for decommissioning and evaluating at a high-level best approach for execution, and then assessing the key technical details and validations to prepare for ingestion into the archiving project workflow.

Ideally, some representative number of applications within the targeted landscape of all retirement candidates has been processed through the first Initial Assessment workflow. This not only enables the program team to better identify the best set for the first wave, but also rules out any “clearly not fit for decommissioning” applications. It is not uncommon to see some organizations implement an abbreviated, “acid test” type of assessment that looks to quickly separate the best viable candidates (for example, that can be quickly express decommissioned) from the “still need more assessment processing”. This shorter set of questions might look something like those in Figure 8-10 on page 170, and may be more informally conducted (than the Initial Assessment workflow process as outlined in the examples).

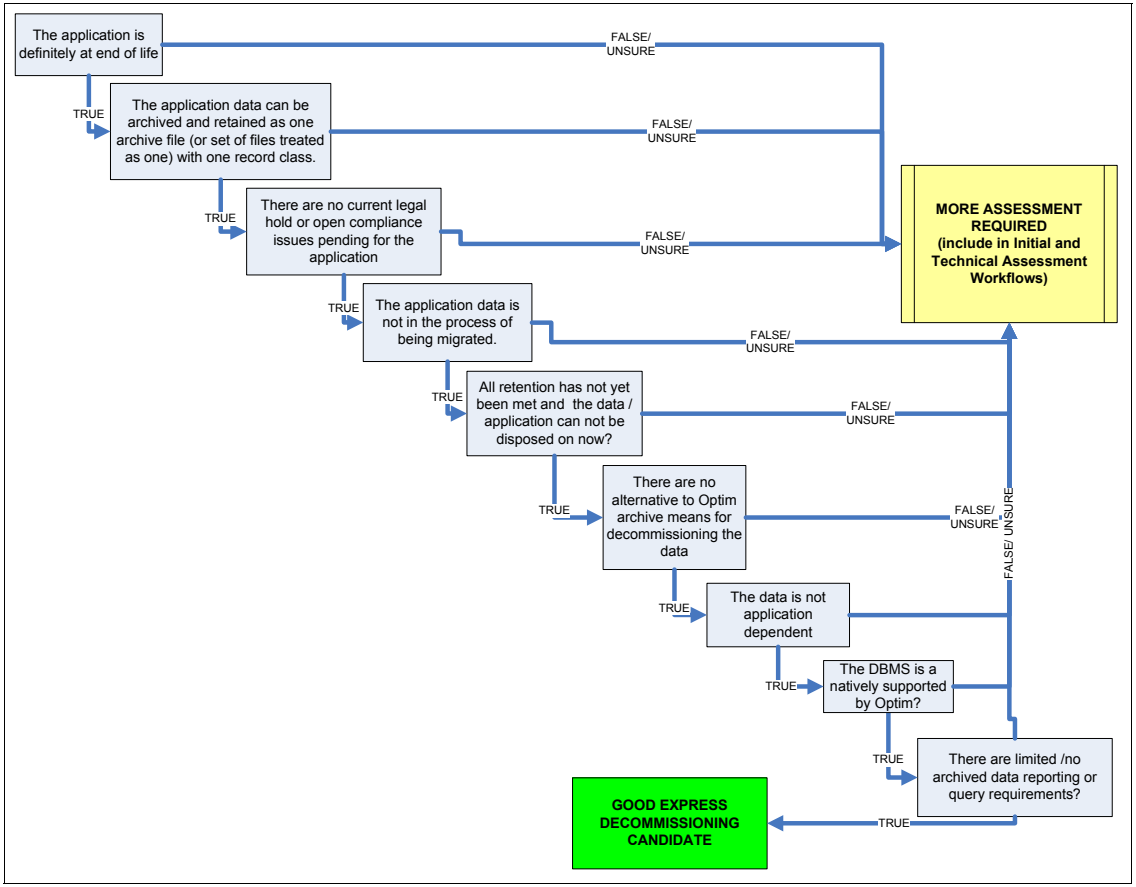


Figure 8-10 Quick "acid test" decommissioning evaluation

This initial application landscape analysis should look for candidates that can easily fit into as many prescriptive use case approaches for decommissioning that can accelerate execution project throughput rates for a first wave initiative.

Establish and ready the technical execution framework

That framework consists of infrastructure, software, systems, and systems support.

At this stage, you should have established, at a minimum, a high-level architecture design for the Optim and related InfoSphere solutions deployment. It is also common to see some initial development environments that are established because these are often utilized in the assessment phases, especially for InfoSphere Discovery based analysis.

Making final decisions and selections of supporting technologies and their implementation design is based on:

- A (hopefully) better understanding at this stage of the ILM program objectives, including any prescriptive behaviors to be promoted in the execution and ongoing management of decommissioned application content.
- A more detailed and documented assessment the larger landscape of applications within scope of this initiative and the initial targeted (first wave) of applications to be put through the new processes.

Earlier established, high-level architecture designs will map out overall flow of processing and archived data movement. These can now be expanded to include specific environment and machine level addresses and detailed specifications. In the case of InfoSphere Optim and Discovery solutions, for example, the better understanding you have regarding the sizes and volumes of the content and the amount of potential initial data analysis and archive execution throughputs required, the more confidently you can finalize your initial and projected architecture decisions.

More information about infrastructure sizing and decision factors can be found in the Implementing an “Implementing an InfoSphere Optim Data Growth Solution” Redbooks publication, SG24-7936 at the following URL:

http://www.redbooks.ibm.com/abstracts/sg247936.html?Open

Estimate first wave work and validate schedule and capacity

Once the first wave of applications is selected, the core (Optim and other InfoSphere-related technology) execution team work efforts can be estimated. For each individual project (and collectively), standard project management type collateral, such as project plans and work breakdown structures, are used to plot out the work and schedules. The IBM Optim implementation method materials identified in 8.2.5, “Refine program and communicate” on page 149, include templates that can assist here.

As identified in Figure 8-2 on page 127, the comprehensive “execution team” will include members from the cross stakeholder areas pictured. At the center of this is the core ILM decommissioning team: The primary Optim implementation solution team members. This will often be referred to as the Optim or Decommissioning Center of Excellence or Optim COE (though in some organizations the entire program may be referred to as the COE). This core team can be any combination of internally enabled and skilled organization resources or IBM or partner experienced players. It can also be implemented as a shared (viably outsourced) service. The correct or best approach will include cost to implement considerations, but must also consider the overall effectiveness and efficiency that the determined make-up and approach will impose on the need for cross stakeholder collaborative efforts.

The degree of involvement and work effort required from the non-core team members will be directly related to the amount of time these stakeholders (business, legal, and records management) have invested in building and documenting the foundation information governance-related details for each targeted application. This applies to all ILM discipline type projects.

For example, the level of effort estimation worksheet excerpt in Figure 8-11 on page 173, for core project team man-days, shows the details of the Analyze phase of an Optim archiving project (effectively same for decommissioning), and in blue highlights the areas where the non-core team members (generally labeled “business” in this worksheet) would typically be involved. Tasks such as Analyzing and Documenting Business Object Discovery, Confirm Responses from Questionnaires and Defining Functional Requirements (a series of workshops) emphasize that cross stakeholder involvement is required.

What is not as transparent is that the level of effort estimates (in man days) for the core team members is based on some assumptions that all of the processes and workflows of IG (previously discussed in this chapter) have been executed for this targeted application. If this assumption is not valid, or if the documenting and alignment of IG-related definitions and policies that apply to this type of ILM initiative are not complete, not only will the core team involvement need to be increased, but the supporting stakeholder involvement will also increase, requiring analysis and investigation and decision making around the domain of the IG principles we have previously discussed. This effort would need to be completed before accurate and appropriate implementation decisions can be discussed in these analyze tasks and used to reach final design conclusions.

Task	Business Typically Involved	Complexity Benchmarks Archiving		
		Low	Medium	High
Project Task Mandays		**60.75**	**77.25**	**103.25**
Project Management		9.11	15.45	25.81
Percent Total		*15%*	*20%*	*25%*
ILM / Optim Program Management (CoE) Involvement		3.04	3.86	10.33
Percent Total		*5%*	*5%*	*10%*
Total Project Mandays		**72.90**	**96.56**	**139.39**
Project Initiation		**0.50**	**0.50**	**0.50**
Planning Phase		**5.75**	**6.75**	**7.75**
Analyze Phase		**22.25**	**31.75**	**44.75**
Perform Initial Analysis				
Request/Review Data Model and Statistics		1.25	2.25	3.25
Perform Data and Business Object Discovery		2.50	3.00	5.00
Analyze and Document Data and Business Object Discovery	**Yes**	2.50	3.00	5.00
Test Connectivity/Credentials		0.50	0.50	0.50
Create/Update Optim Database Aliases		0.50	0.50	0.50
Collect Volume and Statistics Data		1.00	1.00	1.00
Confirm/Review Responses from Questionnaire	**Yes**	0.50	0.50	1.00
Obtain/Review Records Management Requirements		0.50	0.50	1.00
Define Business Requirements				
Conduct Business Overview Workshop				
Application Overview	**Yes**	1.50	1.50	1.50
Technical Sessions	**Yes**	1.00	1.00	1.00
Document Business Requirements (SOA)	**Yes**	1.50	2.00	3.00
Review and Approve Business Requirements	**Yes**	0.50	0.50	1.00
Define Functional Requirements				
Conduct Functional Requirements Workshop	**Yes**			
Application Architecture		1.50	3.50	3.50
Data Model Review	**Yes**	1.00	2.00	4.00
Technical Attributes		1.50	3.50	3.50
Conduct Post-Archive Data Access Workshop	**Yes**	1.00	2.00	3.00

Figure 8-11 Example Optim project level tasks: Effort estimating worksheet

A simple example: When these key workshop meetings are held and business and records management team members are unable to identify correct record classes for the application, or are unable to define specific mappings of these classes to business objects of content managed in the application data, the only choices are to delay the decision on design until this IG type detail is understood or settle for a ‘less than informed” or generalized (viably more risky) approach.

Hopefully this point emphasizes two key elements of level of effort and schedule planning:

- Complete level of effort planning must take into account the comprehensive (cross stakeholder) team member involvement
- Planning estimates will only be as good as the IG preparedness that underlies and supports them.

Ideally, if the initial and technical application assessment processes are a front loaded set of activity to the actual level of effort and schedule planning, applications that have gaps in their IG understanding would be held back from the execution queue.

With regards to core team resource planning, with straight application decommissioning, it is not uncommon to see one full-time resource on the core project team assigned to one or possibly two applications at a time, usually supported by an overall COE implementation manager and classic project manager. Both provide key technical and project oversight, inter-team negotiations, and guidance. This is important to understanding the amount of total resources that are required in relation to any parallel project execution throughput you are considering.

Finally, regarding system capacity planning, once the initial wave of applications has been identified, one simple way to look at capacity is to focus on the most common processing capacity bottleneck or constraint areas (personnel resources aside). For a decommissioning use case, these will typically be:

- The numbers of processors of the Optim “archive create” servers (the servers used to access the application database data and create the archive files)
- The amount of temporary or final storage locations for retaining the archived content
- The number of processors, memory, and storage necessary to support any archived data access - given approved requirements for frequency and concurrency

The assumptions here are that the implementation of this infrastructure is meeting or exceeding any minimum requirements specified for the InfoSphere Optim solutions. For more information, refer to “Detailed System Requirements for v9.1 of the IBM InfoSphere Optim solutions” at the following URL:

http://pic.dhe.ibm.com/infocenter/optsol/v9r1/index.jsp?topic=%2Fcom.ibm.nex.optsol.doc%2Ftopics%2Fwelcome_dg.html (also see any supporting solutions such as Discovery or Data Explorer).

It is also assumed here that frequency and concurrency of archived data access are minimal or controllable, so the focus of the discussion here is on the archive create process and the storage.

The capacity planning mock-up worksheet that is shown in Figure 8-12 highlights some approaches that you might use for both initial planning as well an on-going capacity evaluation and monitoring. This particular example demonstrates both a live application archiving (DG) and decommissioning (RET) set of execution inputs. The plan seeks to identify required processing windows and potential overlaps that might cause you to exceed current system processing capacity, and the accumulated, one time (RET) and ongoing (DG) file system/storage impacts. The net goal is to create a pre and post dashboard to alert program administrators where and when bottlenecks might occur, providing time to scale environments proactively.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
CAPACITY PLANNING		PROD CREATE STORAGE (MBs)							LONG TERM ACCESS STORAGE (MBs)					PROD CREATE CPU (cores)						
		Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7		AF (mb)				Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Key	Allocated	2,500.00	2,500.00	2,500.00	2,500.00	2,500.00	2,500.00	2,500.00	Allocated	500,000.00			Allocated	16	16	16	16	16	16	16
B1=12am-4am	Required - B1	30.00							Required	133,065.00			Required - B1	5						
B2=4am-8am	Required - B2	30,000.00											Required - B2	8						
B3=8am to12pm	Required - B3												Required - B3							
B4=12pm-4pm	Required - B4												Required - B4							
B5=4pm-8pm	Required - B5												Required - B5							
B6=8pm-12am	Required - B6												Required - B6							
DETAILS																				
APPLICATION - BUSINESS OBJECT	Year Started Archiving	DG or RETire	Processing Slot	Avg AF +AFX size per Freq. (mb)	No. of Processes per Request	No. of Jobs per Slot	Application Data Start Year	Standard AF File Retention (yrs)	Archive Older than (Yrs)	No of Years of Archive Files Kept	Avg AF Total Size Per Year (mb)	Avg AF Index Size Per Year (mb)	Estimated Total Archive Data Size in Storage							
APP 1 - Bus Object 1	2012	DG	1-B1	10.00	1.00	1.00	2007	8	2	5	1000	1	5005							
APP 1 - Bus Object 2	2012	DG	1-B1	20.00	4.00	1.00	2001	7	3	4	2000	5	8020							
APP 2 - ALL Bus Object	2012	RET	1-B2	30,000.00	8.00		2007	6	2	4	30000	10	120040							
TOTALS:													133065							

Figure 8-12 Mock-up of a capacity planning worksheet

With decommissioning versus live application archiving, there is usually more flexibility in the processing schedules - both from timing (avoiding processing execution overlaps) and duration (windows of available processing time) perspectives. Even so, given that many application decommissioning initiatives eventually expand into live application archiving, it is prudent for the core program leads to build and maintain production processing schedules, similar to a manufacturing or factory environment, where raw resources (permanent and temporary processing storage) and processing capacity (server cpu availability) can be tracked and managed as new applications are fed into the execution phase. The exercise is an ongoing COE program management task, requiring

updating and adjusting initial estimates post processing, as well as post disposition of archived content.

A number of systems utilization and storage management utilities can be leveraged to assist with “feeding” the ongoing capacity planning process, such and nmon and perfmon. Additionally, new capabilities in the version 9.1 (FP4 and above) release of Optim will begin supporting the aggregation and dashboard presentation of a large set of execution details (request names and executions, processing timings and statistics, archive file and index sizes, archive access metrics) that will be essential input into this capacity management process.

For more information about infrastructure sizing and decision factors, see the “Implementing an InfoSphere Optim Data Growth Solution” IBM Redbooks publication SG24-7936 at the following URL:

http://www.redbooks.ibm.com/abstracts/sg247936.html?Open

Run first wave, manage projects, collect, and report results

The final step is the processing of each targeted candidate through the newly designed processes, leveraging the implemented policies and enabled resources. Depending on the approach and the horizons and milestones established for the initiative, application candidates can be worked through the Initial and Technical assessment workflows and then fed into the Optim decommissioning project.

In some organizations, the up-front workflows and processes will be executed across a series of applications in a build-up phase, almost an initiative of its own (assess phase). This is ideal as it creates a solid backlog of pre-analyzed, execution ready candidates. In other situations, the process will be a just-in-time approach, perhaps one or only a few applications at a time being run through the entire set of sequential workflows.

In either case, any IG foundation type work focused on these applications will contribute to the efficient movement through the decommissioning execution.

As applications are decommissioned, other new workflows and processes (similar to those highlighted in Table 8-3 on page 146) emerge and mature. Records and legal management processes, storage management, and reclamation procedures, will all be highlighted as needed components for effective decommissioning.

As applications are processed, details of the execution are collected and reported, as per the scorecards and dashboards identified in section 8.2.5, “Refine program and communicate” on page 149, and this feedback is reported up through the program management and steering committee levels.

8.2.9 Build into an operational capability

Operationalizing a decommissioning capability is the process of transitioning your capability from a series of consecutively executed projects to an internalized and formally established (and supported) set of policies and procedures. Some other evidence would include the shifting of the costs to operate the capability to more of an overhead or departmental budget line, and the focus on measurement of net achievements across the entire enterprise, versus individually scrutinized costs/benefits for each project/application.

But the most profound evidence is the synthesizing of the rules and practices of good information governance into the day-to-day operations of the organization. This would include the very up front, on ongoing activities of application assessment and information classification through the execution and eventually disposition management of each application, such that each involved department and key stakeholder is now treating the program as part of the corporate fabric of business as usual.

A project orientation to executing a governance type activity can have some early successes because (good) project management enforces meeting project proposed and documented requirements. Long-term IG program success needs to blend the short term, individual business unit/information owner requirements ("I need to decommission *this* application"), with the program level governance behaviors and outcomes that can ensure ongoing and persisted IG results ("we need to understand why, when, and how to decommission *this and all of the applications* for which I have responsibility").

For example, in the initial rounds of the enterprise decommissioning program execution, the processes and the collateral that support the program will be looking for requirement details that many times the owners will not know. Information such as record classes and mappings to business objects in the application will often not be clear. Even intelligence such a "what makes up a business object of managed information" in the application may not be known. This information should be identified at those junctures as foundation governance details that generally must be identified and recorded as good IG for all information sources in some central, metadata, or glossary resource. Specific policies that guide the owners on what, why, and when they should record this detail, and processes for how it should be initially and ongoing managed should also be an outcome. Ideally many of these sources and mechanisms will already be in place. Operationalizing them will be incorporating them seamlessly into the wider IG-related initiatives.

In our decommissioning example, a similar exercise should occur around the systems and storage management policies and processes. Guidance and rules for existing infrastructure and storage management teams to follow for

eliminating or repurposing excess equipment capacity, or reclaiming and reusing storage, should become part of operational procedures, in alignment with, the goals and objectives of the related IG capability. In our case, this would translate to the systems and storage teams having visibility to lifecycle characteristics of each application as early as the initial provisioning phase. Using this intelligence, when they are included in the requirements and design workshops that seek alignment of decommissioning treatment and impacts, their system design will be capable of supporting the retention and disposition goals of the program. And in the most mature cases, automated feedback from the IT systems resource management area might even make this type of detail available for capacity planning and predictive outcome analysis.

A favorite diagram, Figure 8-13 (and quote), that sums up this operationalizing concept is documented in the “Information Governance Lifecycle Reference Guide “(material published by Compliance Governance and Oversight Council), in the section on “Operationalizing the Strategy”:

“Translating strategy into tactics and turning goals into results requires clear connection between the business objectives, the processes and actions required to achieve them, the capacity to execute those actions, and measurement for accountability.”

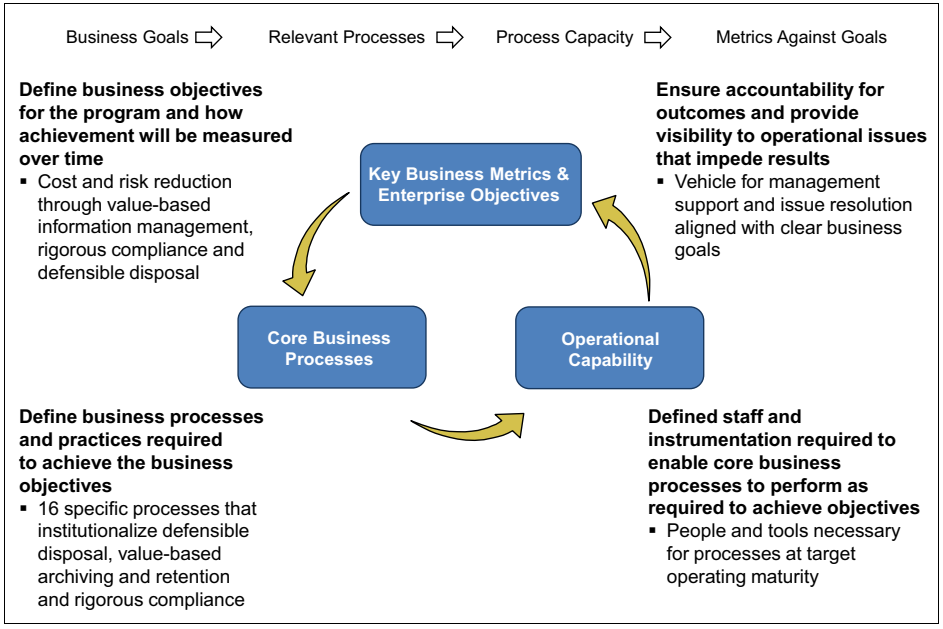


Figure 8-13 Information Lifecycle Governance

This diagram demonstrates that operationalizing an ILM capability, which brings with it more consistent and cost effective executions, is progressively matured. It requires holistic organization support evidenced via dedicated structures and responsible resources, policy supported processes and analysis (audit) and measurement of IG applied principles and overall program accomplishment.

Program level leadership must be clear on the impacts to existing ways of doing business (changes required to the carrying out the current day-to-day operations) and exert the control and mandates to enforce the required changes in behavior.

With a decommissioning capability, though it might take a series of months or even years to catch up on backlogs of retired applications, at some point this ingestion backlog will shrink and stabilize. A common maturing characteristic at this point is to expand the program into live application archiving. The levels of ILM-decommissioning process maturation and operationalization achieved can then be easily extended into ILM-archiving (live applications).

8.2.10 Monitor, measure, and report

In Chapter 6, “Business drivers for information governance” on page 91 we delve deeper into the more general need for sound monitoring and feedback mechanisms wrapped around all information governance initiatives. These should include both direction for establishing measures and goals, how and by who they will be reported, and how any feedback on adjustments might be actioned.

For ILM-Decommissioning, some of the typical techniques and tools for feeding and managing these governance monitoring mechanisms include:

- Include regular project level (ILM and program) meeting and reporting details - standard project management measures of on-time task execution, overruns, general efficiency measures.
- Leverage standard Optim processing outputs (statistical reports), and Optim Manager details such as average throughput processing rates, sizes of data, timings, numbers of tables to support regular achievement readouts.
- Provide candid assessments on any new or existing IG supporting and enabling roles and workflow execution effectiveness, indicating gaps or inefficiencies, and recommend any adjustments required.
- Audit and provide feedback on the use of foundation governance definitions and policies to support the capability execution process.
- Leverage new functionality in Optim solutions (Optim Service Interface and Optim Manager) to automate or standardize the process of archiving content and the feeding of monitoring and measurement details into reporting systems.

Building automated outputs and centralized dashboards to be reviewed regularly by the ILM specific and overall leadership is an advanced level maturity goal in any IG program.

Read more on monitoring, measuring, and reporting in Chapter 6, “Business drivers for information governance” on page 91.

9

Test data management

Provisioning and supporting test data environments can be a challenge for many organizations without guidance or direction. Over the years, discipline and best practices have been developed, tested, measured, endorsed, and published. These practices have confirmed the benefits and inefficiencies associated with certain test data management strategies. In 2008, IBM published a white paper titled “Enterprise Strategies to Improve Application Testing”. This paper highlighted the impracticalities, and prohibitive costs associated with cloning techniques used to provision application data for a testing. Aside from the obvious reasons associated with database volumes, and time, risk has become a major factor. Application databases have grown to the point where the application database includes hundreds or thousands of tables with data volumes in the terabyte or petabyte range. While you may be able to support increases in production volumes, the ability to support those increased volumes in tests is getting more difficult. It is very rare to find organizations with these volumes in production to have test environments with the same capacity. It is just too costly for every test environment to be an exact replica of the production machine, especially when organizations have 10 - 20 test beds. In organizations that can support these volumes, their testers may suffer from data quality issues. Just think about the time it takes to perform root cause analysis. The more data the longer it will take. If cost and support complexity are not compelling enough reasons to rethink your provisioning strategy, you should consider the risk of exposing sensitive data in these test environments. Cloning creates potential vulnerabilities and also violates domestic and international regulatory requirements for privacy.

Developing effective data provisioning strategies and privatization are integral aspects of test data management. These strategies should be complemented with administrative and operational capabilities especially for organizations who use communal test environments. These types of environments are effective and prone to some common issues. The probability of being an unwitting victim of another tester's activities is very high and requires the capability to reset the environment to a known state. Some other capabilities to consider when developing a test data management provisioning strategy are the following:

- The ability to limit the scope and volume of application test data while preserving the integrity and contextual meaning of the test data is paramount, which means the data subsets need to focus on a specific business scenario that is enforced at the application and database layers.
- The ability to identify and operate on data as a complete business object. This includes identification of relationships that use compatible but not identical data types, as well as data elements that extend across system boundaries.
- The ability to de-identify sensitive data in development and testing environments to lower risk and avoid consequences that are associated with disclosure
- The ability to modify or manufacture data required to force specific error conditions or to validate specific processing functionality.
- The ability to compare test data before and after privatization tests is essential to establish confidence, and arbitrate issues that are related to the de-identification process.

The capability to support provisioning and managing different test paradigms is common to many organizations who require support for different testing paradigms such as the following:

- Unit testing
- Functional testing
- Application/system testing
- Integration testing
- Regression testing
- Load and performance testing
- User Acceptance Test (UAT)

Each of the testing paradigms supports different testing communities that include:

- Application developers
- Line of business (LOB)
- Technical infrastructure
- Quality assurance (QA)
- Training

Each of these communities has different requirements for content, volume, and currency of data. Some examples of these differences include the following:

- The LOB and training organizations who typically need specific data for an application feature to function properly.
- Application developers who require baseline data necessary to operate the system. They also require specific sample sets of data to use for functional and system testing.
- Quality assurance also requires baseline data, specifics sample sets for functional and system tests. They also will need larger volumes of current and historical data to use for load and regression testing.

Managing the process to support these testing requirements is often referred to as *test data management* (TDM). Over the years TDM has been practiced and evolved to support legislative and market requirements for protecting sensitive data. Test data management/data privacy (TDM/DP) provides organizations the ability to reduce time and effort used to provision and maintain test environments. It also has proven to reduce testing storage costs, reduce application testing errors from bad data, and reduce the overall risk of a data breach.

Protecting sensitive data includes privatization of personal, financial, and health information, as well as the organization's intellectual property. These data types are often shared and are related to other data elements within and across applications. Simply overwriting values with series of values like **** or xxxx was once considered a viable technique until someone discovered **** or xxxx was useless for application testing. In fact, many applications would fail if they encountered an account number of xxxx. A better alternative would be a technique that did not compromise the context or semantic meaning of that data. Some examples of these special data elements would include:

- National identifiers for domestic and international countries
- Credit cards
- Drivers licensee
- Automotive vehicle identifiers
- Bank routing numbers
- Names
- Addresses that maintain the logical association of a city, state, postal code
- Email addresses

It would be easy to inadvertently compromise the context and integrity of the data and application. One illustration of this could be how you handle de-identification of a credit card number. Card numbers have semantic meaning that is used to validate a card number and identify the card type and issuer. Altering certain values in a credit card could not only render the card invalid, it could also transform a VISA card to be an American Express or MasterCard. One might think that if a credit card number has been masked and passes a Mod-10 check,

it is valid and acceptable. This may work in some scenarios but not in others, especially if the application was used at Olympic events where only Visa cards were accepted. Provisioning and privatization of data must ensure that sensitive data is protected without compromise to the context of the data.

9.1 Provisioning

Several years ago, techniques of restoring data from a disaster recovery copy, or cloning production data were considered viable techniques used to provision application data for testing. However, regulatory and compliance requirements have jeopardized the viability of those techniques. Newer object-oriented provisioning techniques leverage smaller volumes of data that are contextually complete and meaningful to the business and infrastructure teams. These new objects or perspectives are Complete Business Objects (CBO) and have become an effective technique used to manage the provisioning and privatization of sensitive data for application testing.

Assembling an effective TDM/DP framework with CBO requires the ability to support common provisioning scenarios that include:

- *Initial State*: This type of provisioning is used for planned activities and is used to establish a new environment with sample data that can be used by business and infrastructure teams to begin their testing efforts. Communal test environments will require the ability to return to a known base state periodically. A good goal is to offer the ability to regularly refresh these environments and create the capability to refresh test environments on demand.
- *Beginning of a Known State*: Occasionally in a communal test environment we forget about sharing and do something that has a negative impact on others. Rather than spend time doing root cause analysis, we might require the ability to provision to a known state. One example might be to provision to the beginning of a specific billing cycle.
- *Specific Point within a Cycle or State*: From time to time, organizations require the capability to address unplanned or “out of cycle” tests, especially if they need to address a production issue. Waiting for an entire environment to be cloned or provisioned is extremely inefficient and can have a negative impact on service level agreements and support contracts. The ability to rapidly provision from an initial or known state is good and requires the ability to provision with specific data to isolate and test certain scenarios. An example of this might be the ability to test the effects of a particular change in a client’s features or functions at a specific time such as the last day of a specific billing cycle.

Effective provisioning of test data requires the ability to provision a new environment, refresh an existing environment to a logical starting point, and provision with specific data.

9.2 Privatization

Today's economic landscape and legislation are changing the way organizations view reducing time, cost, and risk associated with provisioning application test data. Organizations begun to outsource development and testing with organizations located outside of their native country and need to comply with domestic and international laws related to handling of sensitive data. Some commonly known laws include the following:

- The Health Information Portability and Accountability Act (HIPAA), which mandates that an individual's personal and health information needs to be protected.
- The Payment Card Initiative (PCI) which mandates how an individual's personal information needs to protect clients and merchants who use or process credit card transactions. This law is designed to protect both the client and merchants.
- The Patriot Act in the United States prohibits personal identifiable information from crossing country borders.
- The European Parliament Directive 95/46/EC is designed to address the details of processing and movement of personal data.

Organizations have realized that privatization is more than a just a technique used to reduce exposure and risk of a data breach. If done properly, not only will costs and risk go down, but testers would not be able to distinguish live production data from obfuscated data. No one would know that the data had been de-identified until someone tried to use it inappropriately.

Privatization techniques need to accommodate different types of data stores and formats, within and across applications, systems, and platforms. Some scenarios require privatization to be done on data at rest while others may require the data to be processed dynamically while the data is in flight. This includes the ability to support structured data stored in a database as well as data stored in related documents, data marts, warehouses, and Hadoop-based structures.

Implementing privatization rules is a great beginning but it does not guarantee perpetual protection. A good paradigm for privatization of sensitive data is similar to how organizations handle user IDs and passwords. While user IDs are assigned and passwords are provided, the password is something that needs to

be changed periodically. Similarly, you will want the ability to alter the privatization details periodically.

Almost everyone has heard or experienced a situation where privatization compromised the data or the contextual meaning of the data. This is problematic and creates frustration when someone is doing all the right things and getting the wrong results. Be sure to leverage tools, techniques, and best practices to identify and document the requirements and objective success criteria. It may take a few iterations to get the process for provisioning and protecting sensitive information transparent, consistent, and meaningful to business users and technical resources.

9.3 Approach

The Data Governance Unified Process, discussed in section 1.3.2, "Understanding why implementing IG can be hard" on page 7, has 14 major steps or milestones, 10 of which are considered required and cross multiple areas of Information Governance. These 10 steps are necessary to implement the Master Data Management, Governed Analytics, Security and Privacy, and Information Lifecycle Management. Some consider TDM/DP as an aspect of Security and Privacy, others might view this practice as part of the Information Lifecycle Management. Regardless of the perspective, a common approach that is used to establish the TDM/DP core capability includes:

- Assessment of applications and testing environments.
- Analysis and design/redesign of existing and desired testing processing environments including the policies, rules, and processes for provisioning and managing test data.
- Alignment of the existing organizational structures (roles, responsibilities, and communications) to support the new governance-oriented approach.
- Design, building, and enforcement of TDM and DP rules, policies, and processes.
- Alignment of the technological capabilities to support the new governance-oriented testing approach with objective measurable criteria.
- Operationalizing the new approach with governance foundations integrated into project and software development lifecycles that withstand compliance and audit scrutiny while gaining efficiencies and reducing risk.
- Monitoring and measurements to compare, control, and perfect the process and outcomes against the initial and evolving governance objectives.

When you examine governance framework taxonomies with the intention of building an information governance program, it is not always clear how and where to get started. The process of establishing these foundational capabilities can appear to be like a sequential or waterfall approach but should not be approached or implemented this way. Some capabilities can and should be approached and executed in parallel. The following sections of this chapter describe an approach used to define, automate, and enforce a foundational governance discipline for provisioning application data with de-identified test data used internally or externally to perform many testing paradigms and scenarios.

9.4 Assessment

Generally speaking, a maturity assessment with a focus on the current state of affairs, is the place to get started. It is rare to begin a project or effort that has never been done before and have all knowledge and requirements in place. The best place to begin is to begin with a particular application selected for the initial TDM implementation. The focus is less on the specifics of the applications and more on collecting and evaluating the existing and future objectives for the test environments. This includes setting expectations for provisioning test data, and includes the data privacy policies, practices, and requirements. To ensure an enterprise perspective, it is important to interview the testers, subject matter experts, and trainers of the targeted application. These perspectives help pave the way for gaining acceptance of new processes. Once this has been completed, it needs to be repeated at least twice with other teams and applications.

Developing this initial understanding of the information requires knowledge and definition of a complete business object and criteria. It is similar to the steps outlined in section 8.2, “ILM: Decommissioning” on page 124. Many of the application characteristics, rules, and business objects can be reused and augmented with privacy and measurements specific to the application testing rules, goals, and policies. If not, some form of forensics will be needed to define and classify complete business objects within and across systems.

The by-product of an assessment should be uncovering gaps in understanding and definitions within the existing foundation governance areas related to provisioning and privatization of application test data. Information collected includes capturing required resources, personnel, processes, and policies. Considerations should be made that keep your scope above an individual application or department. Be sure to include multiple applications and departments to ascertain an enterprise perspective. It is not uncommon for different departments and applications to use a different vocabulary, or have a different definition of terms such as customer, or have policies and rules that are

not consistent. It is not uncommon to uncover gaps between this current state and the desired state. These gaps help identify the need for new policies, practices, roles, and responsibilities.

In order to adequately understand the test environment types and expected outcomes, identification of various application-specific use cases is necessary and needs to be aligned with the business goals. This assessment process needs to get the cross-organization perspectives ranging from business, security, legal, and compliance, as well as the information technologists supporting the infrastructure. A great by-product of the assessment is a common vocabulary, and understanding of terms, definitions, processes, and rules. It is also something that is mandatory for organizations seeking to increase their organizations levels of information governance maturity.

The more assessments that are done, the easier they become and can often be developed into surveys that can be given during an interview process. Some questions that can be used to capture general application and database specifics might include:

- Describe the data source (database) to be processed. Can you provide details such as the platform, vendor, version, code page, schema/data model. It is desirable to capture the number of tables, size of each table (rows, and bytes) and categorize them as data, application, reference, system, or log).
- Describe the size of the production database size, and identify the number of test copies, and owners of these copies.
- Describe the origin of the application. Is it a packaged application or something that was custom built? If it is a packaged application, can you provide the vendor, module, and version of this application?
- Describe any integration points that this application has with other databases, applications, and systems.

This information is helpful for identifying system-related specifics, however, some questions should be raised that help drive provisioning requirements. Some sample questions might be:

- Describe your process for provisioning and refreshing test environments. How long does this take? Does this include de-identification? How do you know if an application contains sensitive data?
- Describe how you determine if any of the sensitive data elements are shared or integrated across applications?
- Describe how do you handle communal test environment refreshes today and how frequently are they done?

The last part of a TDM assessment should focus on organizational policies and processes. Some sample questions might be as follows:

- Describe your existing test data management provisioning and privatization policies, rules, and processes. Elaborate on how they are enforced and measured.
- Describe how you determine a state of readiness for testing. How does that ensure that the integrity of the test data is technically, logically, and contextually correct?
- In an audit situation, how would you demonstrate that provisioning and privatization policies have been performed and continue to be enforced?

Keep in mind that this list is not complete; it is only designed to stimulate conversations and thoughts that define a current state of affairs, and help expose gaps that could negatively impact the implementation of core TDM/DP capabilities.

9.5 Gap Analysis: Aligning capability and maturity

After a thorough assessment it is always a good practice to review the findings with governance leadership to understand the level of capability that the organization is targeting to achieve. Documenting the current levels of maturity combined with the fundamental requirements for a test data management capability need to be logically grouped and reviewed by many stakeholders that would be impacted. Once these are classified and grouped, they need to be prioritized into a practical deployment approach known otherwise known as a *crawl-walk-run approach*. The prioritization should be simple. Categories of *must have*, *should have*, and *nice to have*, are common mechanisms used especially when you socialize them with a wider audience from development, quality assurance, the line of business, and the regulatory and legal communities.

A likely by-product of the Gap Analysis is that some of the desired TDM capabilities will be dependent upon a supporting and enabling governance discipline. This is expected and should not be overlooked as they will be critical to overall success. There might be some requirements that can be satisfied with current roles, policies, and procedures (and technology); invariably the need for increased maturity will be concluded. This will drive investigation of process, roles, and supporting technology required to establish realistic program and project milestones. The assessment process will help to qualify expectations and quantify the level of effort and time frame required to define, implement, and support enterprise governance and compliance.

During this phase, many things come to the surface that will affect the design phase ranging from existing tools to the requirements and techniques used to perform those activities. Some privatization of sensitive data may be done when data is at rest and others may be dynamically invoked at the stored procedure or SQL level. Other techniques may be done outside of the target environment as in the case of processing proprietary loader files, or for processing data in Hadoop-based environments.

9.6 Organizational alignment

Previously in section 8.2.3, “Organization structures, roles, and responsibilities” on page 134, we discussed the need for additions or adjustments to organizational roles and responsibilities to align with the new or improved organizational policies, processes, and technologies that support the desired governance initiative. This is also true with efforts to enable a TDM/DP governance initiative. In both cases, companies need to consider the stewards of the infrastructure and the data, as well as the producers and consumers of this information from multiple perspectives.

These TDM and DP roles of governance include the application line-of-business users and owners, IT personnel involved with development that need to perform various tests ranging from unit, functional, and user acceptance testing personnel, and all roles responsible for institutionalizing risk and compliance. Each area brings a different perspective to the governance capability requirements and have varying expectations of the results.

These new or modified roles need to be supported by policies and rules that drive procedures to support their work. Executive and cross-functional working group level governance structures will be as essential here as with any governance initiative to make the decisions on what rules and procedures will be defined and enforced, and to establish the measures and reporting mechanisms that will be required for visibility into aggregated results. If done properly, users will be able to correlate what is being done, how it is being done, and the frequency of activity.

For example, an organization might initially require multiple policies to address provisioning data as the capability maturity grows. However, each set of policies and procedures will still require some basic alignment on the highest level. Common questions to ask include, “What is your process for provisioning test data?” and “What mechanisms are in place to ensure sensitive data protection?”.

In addition to the executive and governance working group structure and roles (see section 8.2.3, “Organization structures, roles, and responsibilities” on page 134), some typical TDM /DP roles are listed below:

- *Architect*: Works with the line-of-business and security organizations to identify high-level policies, rules, classifications, and enforcement to address and protect multi-use data fields. The architect will also work with regulatory or compliance teams to ensure that policies, standards, and rules, such as PCI, personally identifiable information (PII), HIPAA, and the Patriot Act are understood, and defensible. The architect is focused, ensuring the business and technology needs are met. This includes helping to design the infrastructure, process, and to evangelize the solution. They are the primary catalyst to assist others cross organization boundaries that include research and development, quality assurance, and the LOB.
- *Provisioning and privacy specialist*: Works with the architect and LOB and is focused on implementing rules to enforce defined policies, and standards. This includes the capability to perform analysis used to identify, define, and implement complete business objects that can be subset and privatized. They are responsible for artifact creation and maintenance. This individual is responsible for getting the process to run as scheduled and impromptu activities, and will help troubleshoot issues that arise. They become the TDMP/DP Infrastructure subject matter expert (SME).
- *Masking and privacy developer*: An optional role used to address additional automation or functionality that is not available as an out-of-the-box solution. This may be a simple as developing batch scripts necessary to integrate processes, interpretive scripts, or external programs used to leverage existing assets such as encryption or decryption of data, or special account number generators.
- *Administrator*: An infrastructure person that is focused on installation and maintenance of the TDM/DP environment. They work closely with the architect and developers to ensure that the environment is working properly and stays in a supportable configuration. They maintain the environment and enforce that internal and external security and access requirements are being met. The are the primary contact to any of the TDM/DP infrastructure product vendors.
- *Application SME*: Ensures that the application and supporting TDM/DP processes are defined and functioning at the business and application layer as expected. They will do this before and after implementation and after any changes to the application or process are made. They work closely with the architect to identify practical usage of the application, and identify any specific functional requirements at the application and data layers and are responsible for testing and analyzing the results after provisioning or de-identification. They provide the architect and the provisioning and privacy specialist the necessary knowledge that is required to perform basic application tests.

- *Project manager:* Responsible for accomplishing the TDM/DP project objectives. Their responsibilities include creating clear and attainable objectives, and manage the project's cost, time, scope, and quality. They are responsible for herding all parties that are involved with a TDM/DP governance strategy.

In many cases a single person can function in one or more of these profiles. In the beginning, some part-time assistance might be needed from a network, security, and database administrator to ensure connectivity, and permissions are appropriate for each of the TDM/DP roles to complete their assigned tasks.

9.7 TDM definitions, policies, and processes

The TDM/DP capability distills into designing, building capabilities to provision right-sized logical data sets that are referential intact subsets of privatized data based on specific requirements and priorities documented in the assessment process. At a minimum, some basic level of rules for provisioning test data and for privatizing sensitive data for test environments, must be recorded centrally and incorporated into the work efforts of the operation to succeed with TDM. One effective technique is to have this information stored in centralized repository to create a glossary of terms, policies, and rules. Creating a system of record for these policies and rules is much easier today via the leveraging of business glossary solutions software (refer to Chapter 4, "Workflows and business specifications" on page 57). The IBM Business Information Exchange is an InfoSphere family product suite that offers the capabilities to document, enforce, and measure the foundation governance definitions, policies, and rules.

The process for prioritizing and building the definitions and policies that will support the TDM/DP capability can begin even without any formal tools or repositories. Getting started can be as straight forward as reaching out to the current sets of TDM stakeholders and the SME, application users, testing teams, and even training organizations to gather the details of what they do frequently in order to better understand and develop a baseline awareness of basic application and testing usage. Capturing, aggregating, and aligning these into a common set of use cases will help develop the desired outcomes for what the TDM definitions and policies must direct and guide. Definitions include terminology that is used to describe business object definition, execution, and administrative management. Securing end-user involvement as early as possible is essential and an effective way to gain support from this community. Another good place to begin to determine definitions, policies, and process are with the practitioners who are key stakeholders from the business and technology teams. This audience will be the first one exposed to the new governance order and is also the first to issue their opinions and influences the ratings and acceptance of

the new and enhanced TDM/DP processes. Another group to consider working with early on is the legal and compliance stakeholders, who may add a completely different perspective on the definition and policy. Gathering input from these communities will help build a holistic view that can be used to prioritize the essential needs. Keep in mind that policies and process will need to be governed.

In terms of TDM definitions, begin with the concept of a complete business object such as a customer, order, or subscription. These definitions include:

- The physical table names
- Logical and physical relationships
- Data traversal techniques
- Naming conventions for the business objects
- Administration of the business objects
- Data privacy policies and rules
- Security

In some cases, the business objects are known and available in machine-readable forms such as data models, or data definition language (DDL). In other scenarios, forensic tools will be needed to identify and classify the data. IBM InfoSphere Discovery is one of these tools that can be used to put the topology and interrelationships of the data into a Complete Business Object (CBO).

Each type of testing and use case may require different types and volumes of data. There are no formulas that define “the right size test database” although there are some standard approaches used to control the scope and volume of data for these objects. The scope of a business object falls into one of following three categories where:

- Scope is controlled based on some set of character values. This may be something as simple as a status, or demographics like country or state.
- Scope is controlled based on numerical criteria. Some common uses of numerical criteria would be for specific account balances, or data from a specific time period.
- Scope may also be driven by a combination of character and numerical values. For example, data may be needed for all accounts whose status is active during the first quarter of last year.

These scoping techniques imply a good working knowledge of the business object. If you do not have this type of information, a mathematical approach can be used. Statistical sampling is one of these techniques and can be used to get a high degree of test coverage with a minimal amount of data and maintain the same data distribution patterns as the production data itself.

TDM definitions have three aspects, which include the business object, its scope, and the privatization rules for elements within the business object. Privatization

rules are techniques that are used to alter specific elements within the business object. Definitions and policies, for de-identification of sensitive data often are driven by industry standards, and laws and range from essential guidelines and mandates that include PCI, HIPAA, Safe Harbor, and the Patriot Act to internal initiatives used to protect intellectual property. The Security and Privacy discipline requires protection of sensitive information without compromising the integrity of the application or affect the meaning and usage of the data.

Privacy policy definitions range from simple string manipulations, to techniques that include lookup tables, hashing functions, user-defined functions, scripts, and intelligent transformations. Intelligent transformations are used to accommodate special data elements whose values have semantic meaning such as credit card numbers, national identifiers, and email address, just to name a few.

You need to be thorough and practical with these definitions. A best practice will initially limit the number of fields to the common ones required to safely de-identify personally identifiable information so that no one or more fields can be used to determine their original identity. At some point the privatizing of description or comment fields will have to be evaluated to be sure that those fields do not contain sensitive data. Rather than spending time and effort to find and replace these values, some practical methods can be used. Simply replacing these comment and description fields with blanks, literal, or values from a lookup table can be an effective way to reduce exposure and vulnerability with a minimal amount of cost and effort.

After defining a business object and its corresponding scope controls, you need to raise the questions of what constitutes being complete and how could we measure and validate it from both the business and technological perspectives? One approach to consider is locating a couple of credible users from the line-of-business and training departments and have them provide a few of the most commonly used transactions or use cases to use for testing.

The users and requirements for test data management in your organization require that TDM capabilities include the ability to configure, execute, and compare results. These also need to support different platforms ranging from relational databases and files, to “big data” used in Hadoop environments. When defining your foundation governance policies and processes, these variations should be factored into the mix, to include making recommendations on the best techniques for different classes of test systems or responding to different levels of service level agreements (SLAs) for provisioning systems with privatized subsets of data. The assessment includes collecting provisioning and privatization SLAs for each test system and each testing environment. Assuming the technical approaches are all validated and feasible (such as, from cost benefit and technology perspectives), some of the TDM governance policies and procedures might provide a “best approach with these SLA” type breakdowns or

guidance. One example might be to use a static approach to extract and de-identify prior to load, versus dynamic masking - when volumes of data to be masked are large and SLAs for creation are rigorous and time sensitive.

Table 9-1 provides some examples of some typical processes and the related policies that would be reasonable candidates to add to the foundation governance framework for an ILM-test data management capability.

Table 9-1 Processes and policies for a foundation governance framework

TDM/DP process	Related policy	Policy details
Evaluate current application candidates for TDM/DP.	Application ownership, responsibility, policies, and rules.	Rules and guidelines to determine who has the authority for making decisions related to solutions and the definition of the comprehensive scope of these decisions.
Evaluate current application.	Identify and document application-related obligations.	Guidance to Business, IT, Legal departments on why and how test data management fits into organization-wide IG objectives. Should also include specific department participation requirements and objectives as target achievement goals. Examples would include when and how departments should participate in solution inventory exercises and the test data management evaluation meetings.
Evaluate each application's technical complexity and incorporate planning and scheduling steps for each TDM/DP project.	Identify and document technical information governance obligations related to TDM /DP.	Various department team members are required to contribute to technical meetings and analysis efforts. Policies in this area identify required participation or IG-related metadata management obligations that each stakeholder is responsible for maintaining on an ongoing basis (as part of regular IG data management) or for specific ILM efforts.
Evaluate current application.	Identify and document ILM program level guidance and prescriptive approaches for TDM/DP.	ILM leadership establishes guidelines for the approved set of test data management approaches that include prescriptions for alignment across the organization that increase value and lower the total cost of ownership.
Execute TDM/DP solution and manage test data.	Identify and document project implementation specifics.	Monitor and measure standard and repeatable test data management implementation steps and standards, including supporting artifacts and collateral.

TDM/DP process	Related policy	Policy details
Execute one or more processes to satisfy and ensure that TDM/DP policies are applied, and getting desired results.	Identify and document enterprise and project level TDM/DP policies.	Monitor and measure organizational rules and standards for applying test data management to existing test data management processes, policies, and rules.
Capture, time, and effort to implement and execute TDM/DP process.	Establish a Business model used for Inter-departmental funding and cost accounting (charge back).	Apply rules and models for assessing and assigning costs for any test data management efforts to the appropriate stakeholder departments.
Collect and evaluate TDM/DP outcomes and success criteria.	Ensure auditing/reporting objectives are being met and measured against predefined success criteria measures.	Standard ILM program direction and guidance on reporting outputs and accomplishments of individual test data management efforts and for overall test data management objectives and success measures.
Environment Provisioning.	TDM/DP policies.	Policy specifics for provisioning and identifying specific policies and rules that require enforcement TDM/DP automation.
Manage development, application testing, and training requirements for all solutions and aligned with IT, Privacy, and Audit processes/policies.	Policies supporting procedures for everything from legal inquiry and discovery, to acceptable methods for managing ownership, stewardship, provisioning, and auditing of all information systems.	The IBM Information Lifecycle Governance (ILG) solutions team in alignment with the non-profit Compliance and Governance and Oversight Council have established a set of (currently) 16 processes that organizations should both assess and mature to achieve higher levels of ILM-related information governance and control.

Documenting and centralizing access to sensitive data definitions and classifications would fall under the larger “Classification and Metadata Discipline” of foundation governance. You can find more about this topic in Chapter 4, “Workflows and business specifications” on page 57.

The IBM InfoSphere Discovery will locate and classify sensitive data types, which can be used as a good starting point for your foundation data definition and masking policy repositories. An example of specification, such as PII, personal health information (PHI), or HIPAA would include the classification and privatization rules that are shown in Table 9-2 on page 197.

Table 9-2 Classification and privatization rules

Data classification	Data privacy rule	Privatization technique
Names	Protect the true name of an individual or business.	Obfuscate using techniques that are gender and country neutral or specific with hash lookups.
Addresses	Protect the true address of an individual or business.	Produce real state and country addresses with corresponding postal codes with intelligent lookups that leverage hashing algorithms.
Phones	Protect the exposure of an individual or business telephone number.	Privatize existing telephone numbers using simple or intelligent string manipulations, scripts, or substitutions using lookup tables. Technique needs to support privatization of data at rest stored in databases, or files, or performed dynamically using user-defined functions or ETL tools.
Credit Cards	Protect the exposure of an individual or business card number.	Privatize existing telephone numbers using semantics transformation to obfuscate card numbers that are past the mod 10 check that can be adjusted to be card issuer and type dependent or independent. Technique needs to support privatization of data at rest stored in databases, or files, or performed dynamically by using user-defined functions or ETL tools.
Emails	Protect the exposure of an individual or business email address.	Privatize email address by altering address at both the name, and domain level using a semantics transformation. Technique needs to support privatization of data at rest stored in databases, or files, or performed dynamically by using user-defined functions or ETL tools.
National Identifier	Protect the exposure of an individual's National Identifier.	Privatize domestic and international identifiers that include USA -SSN, United Kingdom NINO, Canada 's SIN, France's - INSEE. Italy's Fiscal Code (CF), and Spain's NIF/NIE. Technique needs to support privatization of data at rest stored in databases, or files, or performed dynamically using user-defined functions or ETL tools.
Comment Fields	Protect the exposure of sensitive data placed in comment fields.	Privatize comment fields using simple or complex mechanisms to eliminate, or substitute using literals or replace based on values that are derived from a lookup table.

Data classification	Data privacy rule	Privatization technique
Description Field	Protect the exposure of sensitive data placed in description fields.	Privatize description fields using simple or complex mechanisms to eliminate, or substitute value using literals or replace based on values derived from a lookup table.
Attachments	Protect the exposure of sensitive data contained in attachments.	Privatize attachments using an elimination, substitution, using simple or complex mechanisms to eliminate, or substitute with generic attachments stored in a lookup table. Another alternative is to redact or substitute sensitive data within an attachment.
Intellectual Property	Protect the exposure of intellectual property.	Privatize intellectual data using a combination of out-of-the-box policies, and rules with or without augmentation from scripting tools, external service, program, scripts.

Benefits begin with the reduction of time, effort, and test data storage costs that also reduce risk for individual application/business departments that helps organizations achieve different levels of governance maturity.

9.8 Alignment of technology capabilities

Aligning the practice of performing TDM/DP policies and processes with technology is the next step, which includes the physical subsetting and masking execution capabilities with the ability to enforce and measure TDM/DP activities. Technological alignment is a key element to be considered during an assessment to address gaps and inefficiencies. Proper governance requires incorporating tools and techniques that:

- Manage requirements for the system, source code, content, rules, and defects.
- Manage automation for building systems, provisioning data, leveraging automated testing for load and regression tests.
- Provide the ability to objectively measure, monitor, and render aggregated and correlated results.

Different testing aspects require different uses of data. For example, a developer might need to see the ability to add, modify, or delete parts and quantities. Another developer might be testing aspects of order control so that an item ordered does not exceed the quantity on hand. This can be problematic if the first

tester has deleted or renamed a known part name or number. These scenarios are one of the most significant factors that drive the need and capability to reset an entire test environment to a known state.

Mature organizations have evolved and extended the capability to request or initiate a test data refresh from browser based or work flow systems by leveraging API's or command-line interfaces. This effort can help to establish self service or regular refresh process. Some considerations need to be in place for this to be effective. Without coordination or planning, it is possible for a refresh to compromise someone's testing. We have seen a trend that a refresh is not a single job but several jobs that include the ability to create a baseline from initial execution and be complemented with specifics required to testing specific scenarios.

Once organizations develop their processes for provisioning test data, they will want to confirm process execution and results. These results will be used to establish baselines used to gauge duration and frequency allowing the organization to develop or refine processes that meet or exceed SLA's to a larger audience ranging from the business executives, subject matter experts, process and technology experts, to the individuals responsible for manual and automated testing.

TDM/DP testing strategies need to support the volume and expected frequency demand of different types of testing ranging from functional, application, and user acceptance testing. They will also want to recycle these test scenarios for practical load and regression testing. Understanding the number of variations of these test cases by application, and across applications will assist organizations and help them understand operational and throughput testing requirements.

Table 9-3 on page 200 lists desired characteristics and technology solutions that support enterprise test data management capabilities. Specific IBM related solutions are noted along with any comments on cross-solution integration between technologies.

Table 9-3 Desired characteristics and technology solutions

Core capability functionality	IBM solutions supporting test data management	Desired solution characteristics
Optim test data management	InfoSphere Optim Solutions for test data management	► Ability to subset referentially intact business objects across platforms and technology ► Ability to control volume of CBO for unit, functional, system, and user acceptance testing ► Ability to support bulk and selective refresh of a test environment ► Integration with external glossary, automated testing tools, quality centers ► Auditable results for security and operations ► The ability to have “use case” driven subsets of production data, smaller and more manageable. The ability to define a repeatable process while maintaining user flexibility to dynamically alter criteria as needed.
Finding data based on usage patterns that cross system boundaries	Guardium	► The ability to understand what and where your enterprise data resides ► The ability to monitor and report on database access for internal measurements and audit scenarios ► The ability to perform sensitive data discovery, and data audits ► The ability to redact sensitive data retrieved using SQL ► The ability to redact sensitive data from structured documents
Monitoring, auditing, and reporting	Governance Dashboard	► The ability to capture and render definitions of terms, policies, and enforcement by rule, policy, and owner (steward)
Process documentation	InfoSphere Blueprint Director	► The ability to create and share abstract information flows or implementation plans, including reference of the data governance, data integration, and data quality requirements and initiatives.

Core capability functionality	IBM solutions supporting test data management	Desired solution characteristics
Core Optim TDM Data Privacy	De-identify data and files	► The ability to de-identify sensitive data stored in test databases and files using intelligent functions to produce contextual correct data. For example, keeping gender names correct, altering names, address to be real but factious values. Even the ability to perform functions on credit cards, National Identifiers required to comply with PII, HIPAA, and other regulations concerning one's right to privacy. This capability can be done statically while the data or files are at rest or dynamically when the data is accessed using user-defined functions or ETL tools. Privatization can be done on extract, insert, or as a stand-alone process.
	De-Identification Development utilities	► The ability for one to run a test cycle and compare changes between the data helps identify and refine testing process and arbitrate disputes related to data changes. ► The ability to relationally browse and edit data required to perform specific test cases or scenarios, very helpful when testing boundary conditions.
	Ability to alter the data to be masked statically or dynamically	► Out-of-the-box policies to address privatization of personal and business data that can be executed statically and dynamically on data stored at reset and dynamically "on the fly".
	Data Privacy Policies	► The ability to apply user and role-based object and functional based security.

9.9 Operationalizing your TDM/DP approach

Implementing the TDM/DP capabilities into functional processes is driven by the business and technology policies and definitions that are controlled by the Governance Council and enforced by a TDM Center of Excellence.

User and organization objectives and outcomes should be reviewed from a requestor and organization perspectives before and after implementation. The requestor is very focused on when and what is required to begin their testing. They require the test data to be available and correct, while the organizational

perspective is focused on understanding the process. The organization needs to know if the policies and rules are being followed, by whom, and how frequently are the requests for TDM/DP occurring. Governance requires these perspectives to get a holistic view, which is necessary to support compliance, and continuous process improvement efforts.

Implementing protection and privatization today does not ensure it is still protected universally. Statistics show that the greatest threats come from within the organization so one cannot assume that their data is safe. So objectively measuring the process and techniques used to move and privatize the data, even for existing TDM/DP executions built, should be part of the quality control operations. Organizations need to consider altering or skewing privatization algorithms to protect the art of “protecting”. That is, they should implement security around the test data subsetting and masking design and development processes, including any collateral that is built to support this work, such as privacy tables, development tools, and secured test data landing zones.

9.10 Value-based measurements and monitoring

Organizations getting started with information governance or those that have already begun to implement governance programs will want to determine if they are achieving quantifiable benefits to the business. Unfortunately metrics by themselves can be misleading or meaningless, especially if there is not consensus around the definition, classification, measurement, or results. Defining metrics without objectivity or value to the business is a futile exercise. Governance and compliance metrics need to satisfy audiences from internal and external audiences with multiple views. These views typically include a producer and consumer perspective and include the context of administration, monitoring, and enforcement. They will also include internal governing and external governing bodies especially for organizations that perform internal pre-audit activities to ensure their ability to pass an external audit for regulatory or compliance purposes.

Previously, we discussed the key components of a policy-based governance strategy and their inclusion in the definition of policies, specifications, and rules necessary to enable measurements, correlation, and viewing with a scorecard or dashboard style approach. A common design that is used in factory and plant automation is the traffic light icon. Red, yellow, and green all carry a common definition and meaning that could be assimilated into the dashboard. For example, an executive might want to see a “stop light display” to indicate that an application is ready for off-shore testing while the risk and compliance officer will want the ability to drill down into the policies in order to garner more details regarding rules to ensure that they are consistent; for example, to determine if

the types of rules and enforcement are being done with the capability for decomposition into the specifics of standards, rules, and metrics. Dashboards like these need to be designed such that various users can understand the status of their governance programs. See Figure 9-1.

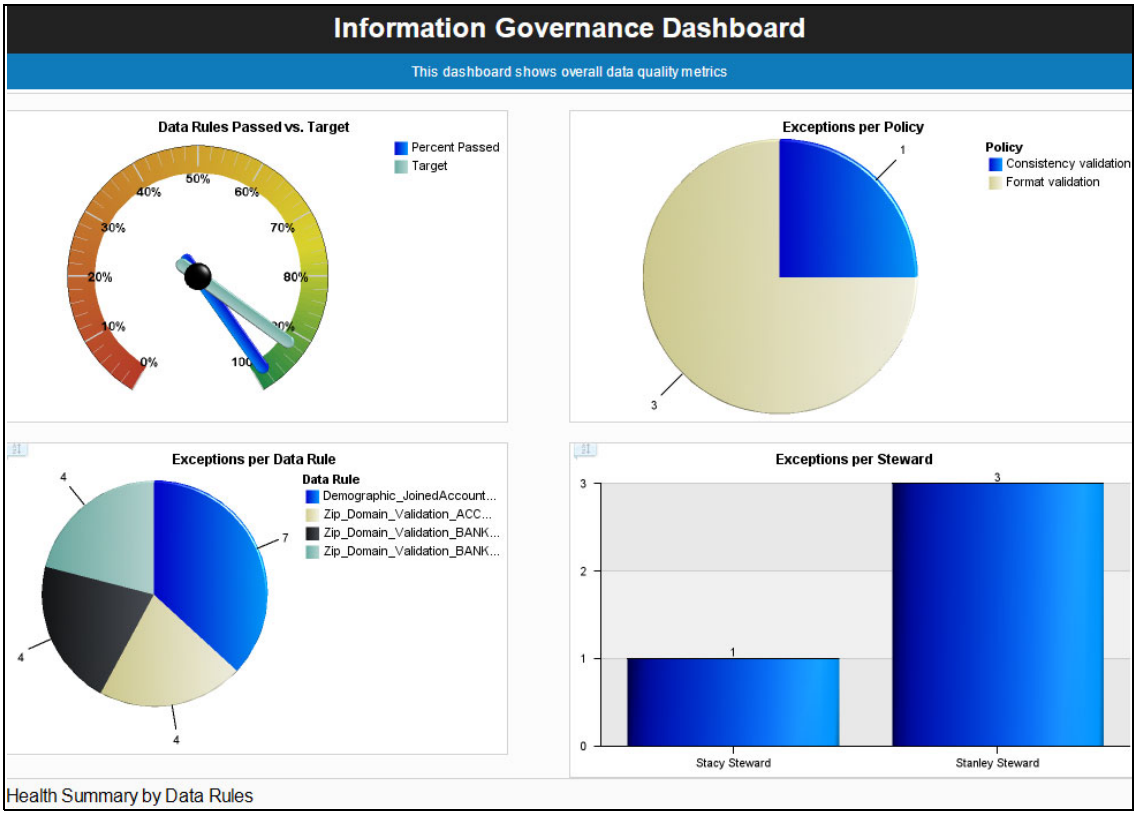


Figure 9-1 Sample Information Governance Dashboard

Information Governance maturity for test data management is an evolving process that is supported by monitoring policies, standards, rule enforcement, reuse that leads to refinement and optimization techniques. For example, after automating processes, the concept and capability for self service may be provided. When this occurs, one will need to track identification of the requestor, the frequency of the request, and the rationale behind the request. Patterns or metrics will emerge. If one can see that the frequency of a particular request occurs every four weeks, one might consider instituting an automatic refresh to occur automatically every four weeks.

Measurement perspectives vary based on the audience. For example, owners of the data might be focused on providing data based on defensible policy, while

stewards focus on policy, specification, process, technology, content, and context. Custodians, alternately, are focused on continuing to get predictable results in a predictable time frame. In general, metrics should be continually monitored for accuracy, and checked to ensure that they are trustworthy from a few perspectives that include the business users, technologists, and risk and compliance users and executive perspectives.

A critical goal for a TDM governance program is to control the policies, standards, and rules that are associated with provisioning and protecting sensitive data with the premise of transparency to the end users, reducing and eliminating any impact to the business. These methods need to support the business in ways that reflect that key policies, standards, and rule objectives are met. This implies measuring results through the entire lifecycle of the test data from provisioning and privatization for testing to demonstrate compliance and enforcement.

10

Master Data Management

In this chapter, we introduce Master Data Management (MDM) which is one of the core applications that forms part of the IBM InfoSphere Platform. The purpose and structure of MDM is explained and then we follow through an MDM project, using MDM in a real life business change scenario. The core components are depicted in Figure 10-1.

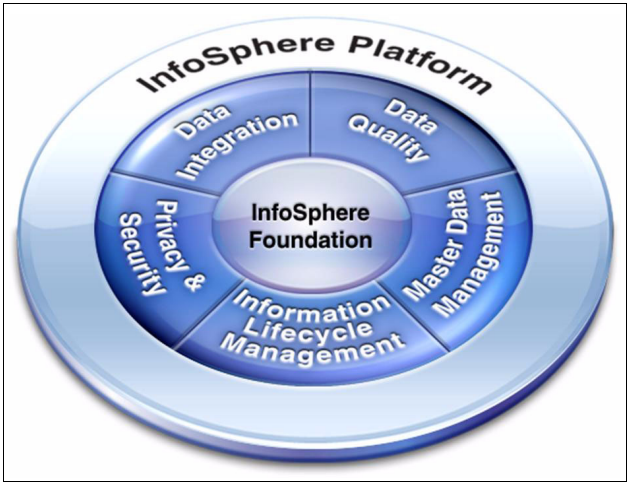


Figure 10-1 InfoSphere Platform core components

10.1 What is Master Data Management

IBM InfoSphere MDM is a comprehensive suite of products and capabilities that are used to manage master data in an organization. Master data covers many different types of data and because of this, it is divided into data domains to help break down the complexity. The principal domains are: Party, Account, and Product. MDM enables organizations to get the maximum value from master data by centralizing these data domains and providing a large set of prebuilt business services that supports a full range of MDM functionality.

The party domain allows management access and distribution of data that is related to parties such as customers, vendors, and suppliers, and it maintains a single, consistent version of this data.

The product domain manages the definition and authorship of products. The collection of products makes up any number of catalogs that are accessible to other systems across the enterprise.

The account domain is an operational-styled hub that manages all account administrative master data, relationships between parties and accounts and products and accounts. It does not extend to account transactions.

What is the actual data mastered

Here we concentrate on the party data. Party data is always sensitive data regarding either persons or organizations. Party data contains information including addresses, contact methods, contracts and accounts held, internal structures of parties, and relationships between parties of all kinds. Combined with this is a large amount of reference data in the form of code tables or look ups, which are used to qualify the master data and provide a constraint on what values are allowed to be used when persisting master data. A simple example is a code table called “Product”, which would only allow predefined products from a controlled list to be included in the detail of any accounts or contracts.

Functionality that can be applied to the data in these domains includes holding virtual views of a record, maintaining a single unified record, and applying workflow and lifecycle management abilities in administering the data.

MDM provides data stewardship tools that help ensure quality and security and most importantly allow for the detection and de-duplication of records. The data stewardship function is essential for an MDM implementation and is closely bound with the requirements of an Information Governance program. This is depicted in Figure 10-2 on page 207.

Every MDM implementation is tailored for the specific needs of the client organization and a highly developed set of common services, and a unified workbench for customizations are at the core of the architecture.

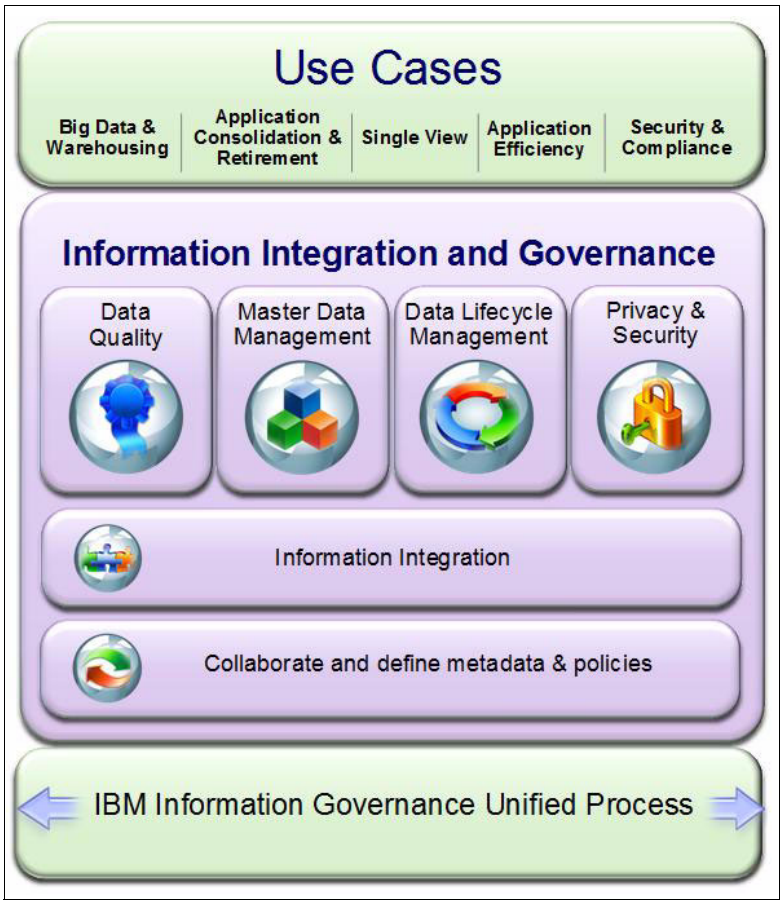


Figure 10-2 Information Governance with Master Data Management

A sound MDM practice in any organization has a major synergy with the practices and benefits of an Information Governance program.

Between MDM and Information Governance initiatives there are shared areas of analysis and design in the whole of any company's IT landscape. Both MDM implementations and the ongoing requirements of information governance need to achieve a detailed and common understanding of the business processes and business rules surrounding the use and care of sensitive personal and organizational data.

The aim in MDM is to persist high quality cleansed data and make it available to every other relevant application and business process in the organization. MDM is implemented to configure data rules and workflows supporting the business processes.

The aim of information governance is to create demonstrably robust and repeatable procedures that ensure the quality, availability, and security of data. Information governance can then refer to the MDM documented system for repeatable practices of data processing and a clear authoritative source of what is minimum and complete data, depending on the purpose any data set.

Staff who work on the MDM applications are also likely to be involved in information governance initiatives as well. The relationship between MDM and information governance is that of a sophisticated application (MDM) whose practical operation answers both a computing need and a business requirement to implement information governance. There is a marked trend of merging the aims of MDM with information governance because of their complementary nature and this will increase with the widening adoption of business process definition and workflow functionality into MDM.

Information governance is becoming increasingly important for organizations where compliance with external regulations and the ability to demonstrate a high degree of care and stewardship of data is a concrete business advantage. The data in MDM is, by its nature, always personal and very often sensitive. Handling large volumes of data of this type in a well developed information governance framework will give a company measurable advantages through enhanced capabilities such as:

- Being regarded as a trusted organization by the public
- Having the ability to recognize opportunities to cross-sell and up-sell
- Having the ability to detect risks and liabilities
- Being compliant with regulatory and industry standards
- Being able to extract valuable insights from data rather than only manage it

There is a strong bond between information governance and MDM where the first principle of MDM is to make high-quality data consistent and available to all relevant enterprise business processes and applications, and the demands of information governance that require data to be regulated in quality, use, and security.

10.1.1 MDM functionality that can support information governance

This section takes some of the major features of MDM and compares their effective support for an information governance program.

The aim of information governance is to establish practices to steward data consistently and to a high standard and this is often hindered by the multiplicity of applications found in many companies. It is difficult to identify where the accurate and complete data exists among systems where there is no close synchronization and where these systems exist for discrete business purposes.

With a well developed MDM much of the data subject to information governance initiatives is well defined in one application. MDM gives a single instance of master data and at the same time a repository of reference data (code sets) and metadata, the definitions used throughout the company to define the meaning and structure of their data.

Data held in MDM

This is a high-level description of the data that you would expect to find in a mature MDM application.

Parties in MDM

Parties in MDM are natural persons and organizations. The obvious “telephone book” information such as name, address, and contact numbers are kept for each party. In addition, there is data about identifying the parties, links to and from external computer systems, party roles and party preferences. In all cases, multiple entries can be held, marked by start and end dates, so that for example, a party can have an unlimited number of telephone numbers and an unlimited number of names.

The party itself can be subtyped, so that departments within an organization can be identified uniquely, for example all cost centers. Unlimited numbers of relationships can be stored between any types of party, so for example a cost center may have an “owned by” relationship with an organization; an organization may have an “owned by” relationship with a natural person; a natural person may have a “married to” relationship with another natural person.

Accounts in MDM

Accounts and contracts are stored in MDM. All the details about contractual agreements and their dependent account numbers are structured in MDM so that they can exist as separate objects and be associated with as many parties as necessary. For example, a bank account may be associated to one party who “owns” the Account, and also be associated to a different party who has a credit card paid off from that account.

Accounts and contracts can have relationships such as “this contract supersedes that contract” or “this base agreement covers that and those accounts”.

Any number of links can be held to contracts from external computer systems and especially to separate transactional systems. MDM does not contain the

transactions records of activity in an account. Transaction data is fast moving and not considered master data.

Products in MDM

MDM can be used to author and manage the definition of the products of an enterprise. The products can be of many types, for example, a goods product or financial product. The collection of products can be entered into an unlimited number of categories, for example “savings account” may appear in “core banking category” and also in “mortgage facilities category”.

Products may have terms and conditions, individual specifications, and equivalent products. Unlimited numbers of links to identifiers in external computer systems can be held.

The product data in MDM can be managed separately for the authoring and management of products and categories, but also in a mature MDM implementation a link can be made from any product to either or both of parties and accounts. From this, it is possible to go from a party and see what products are used, and also to go from an account or contract to see what products are involved.

Functions in MDM

This is a description of the major functional areas in MDM and a discussion on how these functions can be built into business processes and reporting schemes to facilitate aspects of an information governance initiative.

Groups and hierarchies

All the data types in MDM can be put into groups or hierarchies. Groups and hierarchies must only contain one type of data, that is, “all members are natural persons” or “all members are contracts” or “all members are addresses”.

Common examples of groups could be “gold customer group” or “high credit risk group”. Hierarchies are often used to represent organizational structures or sales regions or product bundles.

For information governance, there is the enforcement of business rules implicit in setting up hierarchies and groups. These business rules, or referential integrity conditions are designed into the services layer so that parties can only be included in groups and hierarchies where certain conditions are met, for example the business rule “only persons with an employee type relationship can be included in discount sales group”. This can be established as a consistency check applied to all parties going through a business process of being considered for discounts.

De-duplication

One of the major problems in databases with large numbers of records is the frequency of duplicate records being persisted. Duplications can exist because of slight variations of spelling on data entry or from merging data sets from different systems. However, when the duplicates are created they pose a problem for information governance and for an efficient MDM implementation. There are well developed de-duplication possibilities in MDM based on either deterministic matching or probabilistic matching with configurable capabilities for merging duplicate records automatically or manually.

Deterministic matching typically searches for a pool of candidate duplicates and then compares values found in specified attributes between all pairs of possible duplicates. Allowances are made for missing data. The results are given a score and the scores used to decide if the records should be considered as the same or different. There is a gray area where the scores indicate uncertainty and these duplicates are usually referred to a data steward for investigation and decision.

Probabilistic matching looks at specified attributes and checks the frequency that these attributes occur in the dataset before assigning scores. The scores are influenced by the frequencies of existing values found.

Whichever method of de-duplication is chosen, one great strength of MDM is to detect and resolve duplicates.

Many business problems arise from the presence of duplicates, for example, in a set of persons where there are duplicates:

- An individual may be sent two sets of marketing material, which is wasteful and shows that the company has bad records or processes.
- Shipment and statements may go to the wrong address or SMS number.
- Credit and risk ratings may be inaccurate because they are not based on a complete picture of the individual.
- A company might give incomplete answers to queries from customers, which is always embarrassing and can lead to legal penalties.
- Sales and service opportunities might be missed because the full records of persons activities are not available.

One of the important reasons for having an information governance program is to have the certainty that you, as an organization, can communicate with the correct business partner or customer. There are boundless numbers of stories about the bad effects of having duplicate records, from embarrassment at sending multiple mailings or e-mails to the same person, to continuing to use an old address because only one record has been updated. Sales and marketing departments have been known to block out large post code sectors to avoid multiple mailings

because of public criticism, but the downside was to reduce their marketing coverage by 120 dwellings per sector ignored. The most serious stories from a bad publicity point of view are when a person dies, and the database is updated by one record being closed. Because a duplicate exists, mailings and sometimes demands are still generated and sent to the deceased person.

Knowing your customer and compliance

Know your customer (KYC) and compliance are two things that are combined in functionality in MDM. This allows the administration of setting compliance conditions to satisfy business or legal requirements, and then ascribing actual data to every customer, so that the company can demonstrate adherence to compliance on an individual customer basis. Compliance is general to any data held in MDM and is modeled with a common data model specifying the compliance targets and the documentation necessary to fulfill these compliance targets. Briefly the MDM compliance expects that some data recorded in MDM against a party needs to exist and can be tested to be correct.

For example, a user of an MDM system wishes to verify the date of birth of a particular customer. The compliance element is stated as the business rule: “A lower tax pension may only be paid to a customer over 70 years of age”. The data quality rules for customers already have a mandatory element <DoB>, so a date of birth exists. The compliance part for pensions payments further requires that the date of birth must be proven by sight of a birth certificate.

Sight of the birth certificate becomes the compliance target. This condition can be prescribed and stored in the database as a “compliance document” and linked to one or many business rules. When the customer produces the document, it can be recorded that this customer has fulfilled the compliance requirement and the details, or an image of the document can be stored with the customer records. In this way, MDM KYC links the party to the compliance target and can provide a full picture of all conditions where this customer is compliant. The same rules for compliance are applied to every person or party that shares a given product offering or business relationship.

MDM can also be used for applying detailed privacy requirements to parties and their information, controlling conditions of business, for example, reaching agreement on the set of products or services that they are entitled to purchase. Party privacy preferences can also be recorded against the party or to specific addresses, phone numbers, email addresses, or accounts.

Information governance is put in place to demonstrate the good stewarding of data and there are numerous and increasing numbers of legal regulations, from many countries, which must be observed. Using MDM to record compliance gives a system supported base for reporting and checking compliance and also

for reporting of outstanding actions that must be taken, now or in the future, for example, validity periods of documents, time elapsed since last verified.

10.2 MDM scenario - outsourcing and acquisition

In this scenario, we discuss bringing outsourced systems back in-house and acquiring another company. This is depicted in Figure 10-3 on page 214.

Scope

Build an in-house MDM solution that can fulfill the requirements of mastering customer data that is sourced from other systems and merged into one unified MDM system.

Objectives

Greatly improve the accuracy and quality and completeness of customer data by having all records administered in one application. Make comprehensive customer data available to all the business processes where and when it is needed. Remove dependency on external IT Supplier Companies. Support the initiatives of an information governance program.

Goals

Reduce direct costs of storing and providing customer data. Gain direct control over development and integration of solution. Improve data security.

Description

The scenario involves building a new MDM solution for mastering party and customer data and to make this data available throughout the newly enlarged enterprise. The new solution will host the migration of data from an old enterprise customer information system (CUST1) into the new MDM and concurrently, will assimilate the customer database of a newly acquired company (CUST GREEN) into the same new MDM. In this scenario, the existing CUST1 is outsourced and was built on an old-technology platform and file structures. It is therefore incapable of effective economic development for supporting more demanding business requirements. The CUST GREEN application support is provided by agreement from the parent company from which it was purchased, for a limited time period, after which the support will cease, with no possible option of a time extension.

The enterprise has limited control and knowledge of the data from the outsourced system's current state of development and capability. This is depicted in Figure 10-3.

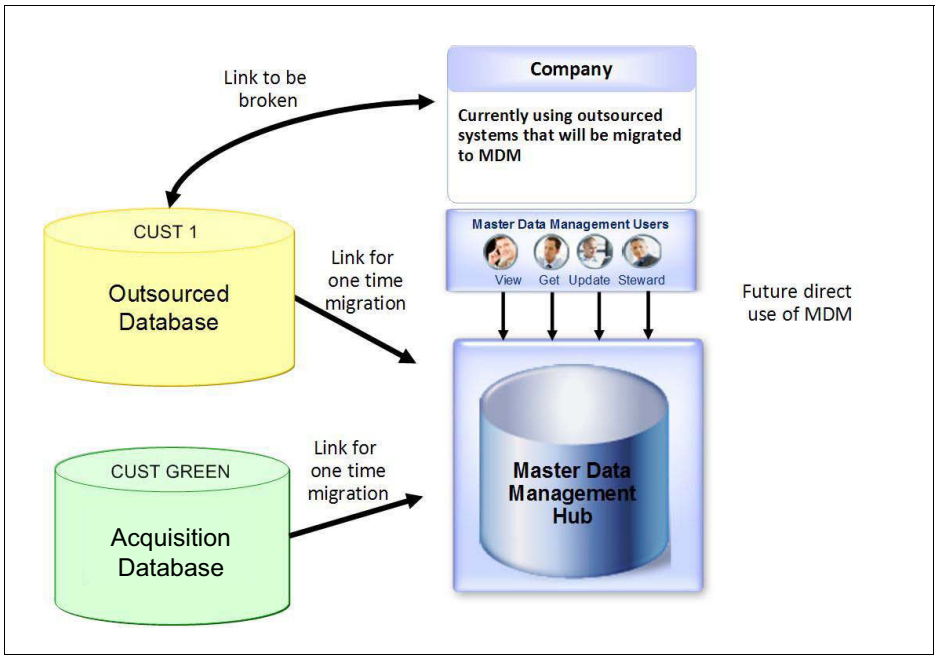


Figure 10-3 Simple architecture diagram for outsourcing and acquisition

This combination of scenarios has been repeatedly seen in companies worldwide, especially in the banking and insurance sectors.

The reasons for bringing systems back in-house from a previously out-sourced host include: Lack of ability to develop systems that are governed by long-term contracts. Typically, customer systems that have been outsourced are using older technology and based on designs that simply record data and provide reports. The process of development is made more difficult by additional contractual obligations and the indirect way that requirements are analyzed and fed through to development teams. This results in outsourced applications that are never changed or improved much but are used to feed many home-built local systems, working at department or process level.

Acquisition of another company, or a department of another company will always require decisions about the IT landscape and support changes that must take place. Often the acquired enterprise is taken over with their existing systems continuing to be supplied for a limited period, by the seller, or the seller's service provider. This gives a fixed time within which to migrate the acquired enterprise

systems into the new owner's IT applications, or build larger capacity applications (naturally with enhanced functionality) to cater for the enlarged business.

10.2.1 Analysis

The data found in both sources (CUST1 and CUST GREEN) is the typical data that is persisted and administered in an MDM application. It is sensitive information about persons and organizations and their interactions with the company, their internal structures, and relationships with other businesses. The MDM implementation must give system support to safeguard and steward the information by building repeatable processes and in-built rules to allow consistent decision support throughout the customer or party lifecycle.

Requirements

From a high level, analysis works down from the business requirements that are discovered by the development team. Many of the requirements can be stated in terms of information governance. These are a sample set of representative functional business requirements considered for this scenario:

- Only one record will be persisted for each party.
- All business processes and systems in the company will use the core party data found in MDM.
- Persons and organizations may have many roles when interacting with us (the enterprise), for example customer, broker, prospect, supplier.
- Solution must support simple customer lifecycle.
- Data indicating the worth and potential worth of a customer will be stored securely.
- Persons and organizations who are customers may be assigned to marketing sectors.
- Privacy preferences expressed by the party must be observed in all interactions with that party.
- Access to party records must be possible through multiple keys from external systems.

Data

At a low level, analysis considers the data elements found in a specific file system or systems and details the semantic meaning and the basic metadata such as record layout, data format, cardinality, frequency of values. This low level is necessary for complete understanding when specifying and building extract, transform, and load (ETL) and messaging jobs, which are a core part of all developments of this nature.

As well as understanding the metadata, the other aspect of data is examining the quality of the actual data that resides in the source systems. Data can be difficult to isolate, for example the statement “we record all our Prospects in the system” may be true but on examination it is discovered that the person and organization records have the word “Prospect” entered in the “second name field” or worse, appended to an existing name in a name field or other field.

The problem this gives for MDM and for data governance is that the rules and processes that should be applied to the data cannot be automated or supported until the data is isolated. There will always be a set of rules around identifying a party as a sales prospect. Correctly identifying the party is a pre-requisite then there follows the reasons to consider this party a prospect. Consider if they made an inquiry about products and services, do they have a product that is near its end date, or if they have a single product that is often bundled with or associated to other products. These can be good indicators of a prospect but that party may have stated to the organization that they do not wish to be contacted for the purpose of marketing.

Once the data of prospect has been successfully extracted from the source data there follows the question of the validity of that information. It should be trusted as valid because we rely that the individuals who made the business decision to make them a prospect, have checked all the information available, and observed the business rules and processes. Our aim in the design part of the application, however, will be to have the system capable of easily querying and retrieving all the available data that could be referenced for supporting such a decision.

In this example, the data may include age of the person, addresses and contact methods, income, number of dependents, and any privacy preferences they have declared. This is a long way from automating the decision-making process. In most organizations there is a minimum set of data from which to work, but the final decision is made by a person, and that person may choose to make a party a prospect, without having the whole set of data available in the system, or indeed overruling the indication that is given by a certain set of data.

When a party has many roles it affects the decision on whether or not the party can also be a prospect. There may be, for example, exclusion rules that forbid combinations of roles, for example a broker is not allowed to be a supplier too. The next requirement, for a customer lifecycle, is affected too because the lifecycle status may be derived from a set of roles through time, or may be a different structure altogether but with conditions, for example, an inactive customer by reason of debt default, cannot be a prospect.

10.2.2 Design

Now that the functional requirements are known, this section on design will only consider the data and services and business processes designed to fulfill the business requirements. The aspects concerned with functional requirements, as examples, response times, numbers concurrent users, and volume metrics, do not have a direct bearing on information governance beyond providing a working application. This section builds the data model, a map of data sources and conversions for ETL, service transactions, and workflow steps. The objective is to develop both a sound MDM design and to support information governance objectives.

Logical model

From the requirements above, the model design will have identified the persons and organizations tables (party tables), and those dependent child tables from party, that accommodate unlimited roles and reference keys to external systems. Tables that contain data that may have mutually exclusive entries such as the roles, will have start and end dates that enforce separation in time. For each table, every data element will be listed along with their data types, lengths, and whether they are mandatory or optional.

Mapping source data

In this case, there are two sources (one database brought in-house and one database acquired through purchase of another bank). There are many detail differences in the presentation and meaning of the data, which have to be reconciled to a common meaning and stored in the most appropriate place in the MDM database. Some examples of resolving the real meanings of data are:

- An element described as “Archive Flag” really meant that the person was no longer regarded as a customer. Similar data from the second source system was an end lifecycle status on person. In this case in MDM the presence of an end date in the person record preserved the meaning from both systems.
- A code table for gender was found to have three values, which had roughly the same frequency in a large data sample. The business meaning was found to be that two of the values indicated male or female and the third value indicated that the record was of type “organization.” This required a specific translation in the ETL job to load data into MDM.
- An element described as “Department” from one source system was actually a free-form description of where a person in an external company worked. In the other source system, the same element name “Department” referred to where an employee worked internally and was a fixed code list of values.

Design of transactions

Transactions or services are the XML-based messages that allow data to be added and updated in the MDM database. Simple services can quickly be modified to do two things: To be linked and executed in a certain order that reflects the business process, and to be modified so that some behavior is introduced before of after the service executes.

Examples of this from the preceding requirements are:

Requirement: *Persons and organizations who are customers may be assigned to marketing sectors*.

The service that would add a person to a marketing sector (in this case a grouping) checks first to ensure that the person has an active role of “customer” before allowing the association with a marketing group.

Integration with consuming systems

This organization is running multiple databases and systems that are specific to a business process and none of these applications contain all the enterprise data about a customer. Each application may have a partial view, and in very many cases different values and formats are used for what should be the same data. This situation has come about through a history of mixed application development, company mergers, and acquisitions and evolving business processes satisfied by local point-solutions that are specific to one or few business processes. Examples for this scenario include:

- Prospect database
- Account billing database
- Customer/product utilization file
- Customer database retail
- Customer database wholesale
- Order tracking customer system

10.2.3 Summary of information governance support

Earlier in chapter two, you saw the diagram in Figure 10-4 on page 219, which shows how the information governance foundations can be related to core governance capabilities in the applications of the InfoSphere Platform.

These umbrella foundations, (definitions, policies and rules, processes and workflows, roles, and stewards), are all found and expressed in the applications below the umbrella (shown in Figure 10-4 on page 219).

The applications (MDM, ILM, Data Quality, Security and Privacy) should all share as far as possible the same definitions and policies, which form the foundation,

and be designed in such a way as to and support existing business processes and data stewardship functions.

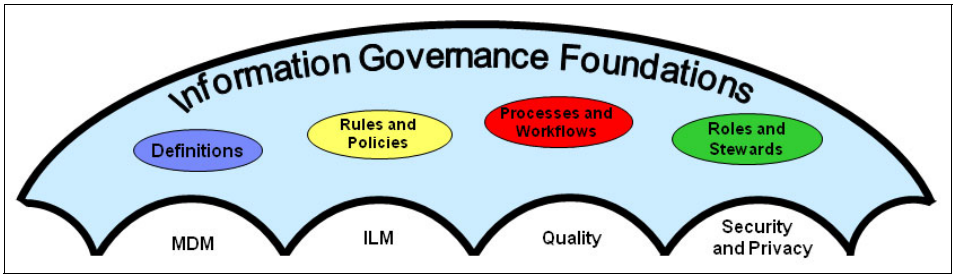


Figure 10-4 Foundation and functional governance components

Examples follow of MDM providing direct support for the information governance foundations.

Definitions

When setting up MDM, there is a physical mapping between the source systems attributes and those held in the MDM database. In most implementations, there are also a set of mappings or a canonical model in the enterprise service bus (ESB) that allows translation of messages into different formats.

In the case of MDM, there is a direct link to Business Glossary, where the semantic meanings of attribute names from the glossary can be linked to the actual table and field names used in the database.

Policies and rules

In the MDM database, there are minimum constraints to enable high performance. When designing the services layer, however, the services are used to constrain data to follow business rules at a low level. For example, a service for adding an account might be modified (called a *pre-behavior extension*) to check the date of birth of an applicant before adding an account of type “pension”. The service would fail, for example, if the applicant was under 65 years of age.

For more complete sets of business rules, which are held outside of MDM, we are able to call out to a rules engine at any step when executing an “add” or “update” service to ensure that the transaction complies with the policies and rules of the organization.

Processes and workflow

At a user interface level, the services in MDM are chained, or made into composites, to support the flow of information that is expected by the operator. These services can be tailored to reflect the most common use cases of process.

If there is a requirement for more specialized workflows, there is a business process management (BPM) module within MDM that can be designed to implement workflows and direct tasks to different users and groups.

Roles and stewards

MDM data stewardship presents data stewards with the ability to retrieve any data and to create structures such as hierarchies and groups. From an information governance point of view, the most important data stewardship support is managing the de-duplication and searching functions in MDM.

The MDM matching and de-duplicating functions can be modified and tuned by data stewards and set to report duplicates. Typically perfect matches can be collapsed into one record, and duplicates with a degree of uncertainty reported to stewards for resolution. Those suspected duplicates that do not match well enough are ignored. The levels of match, or thresholds, can be tuned to suit the implementation.

MDM and information governance

MDM exists to provide trusted customer data to the enterprise. This aim, and the process of implementing and running MDM, complements information governance where the drive is to understand and improve process integrity concerning the stewardship of data. Conversely, information governance is vital to the success of MDM projects. Unifying MDM and information governance is essential to overcome both organizational (data stewardship) and technical (systems support) issues and helps address the need for evaluating the entire data ecosystem, including processes, resources, and deliverables.

11

Data protection and security scenario

News headlines about the increasing frequency of information and identity theft have focused awareness on data security and privacy breaches, and their consequences. In response to this issue, regulations have been enacted around the world. Although the specifics of the regulations may differ, failure to ensure compliance can result in significant financial penalties, criminal prosecution, and loss of customer loyalty.

In addition, the information explosion, the proliferation of endpoint devices, growing user volumes, and new computing models such as cloud, social business, and big data have created new vulnerabilities. To secure sensitive data and address compliance requirements, organizations need to adopt a more proactive and systematic approach.

Since data is a critical component of daily business operations, it is essential to ensure privacy and protect data no matter where it resides. Different types of information have different protection requirements; therefore, organizations must take a holistic approach to safeguarding information. In this chapter, we discuss and describe this scenario to implement data security for audit reporting and compliance purposes.

11.1 Scope, objectives, and goal

The scope of this publication is to describe sensitive data, both structured and unstructured in enterprise database systems, data warehouses, and file shares.

You should protect data via real-time monitoring of all data traffic activities, including actions by privileged users. Also, protect sensitive data repositories against new threats or other malicious activity and continually monitor for weakness.

The goal is to safeguard sensitive data by detecting unauthorized or suspicious activity and then alerting key personnel, and to meet and comply standards and regulations for data protection, such as Sarbanes-Oxley (SOX), Payment Card Industry (PCI), Health Insurance Portability and Accountability Act (HIPAA), and to reduce operational costs via automation, centralized cross-DBMS policies, and audit repositories.

11.2 Approach

In this section, we describe the goals and objectives through the approach taken for database security and compliance, as depicted in Figure 11-1.

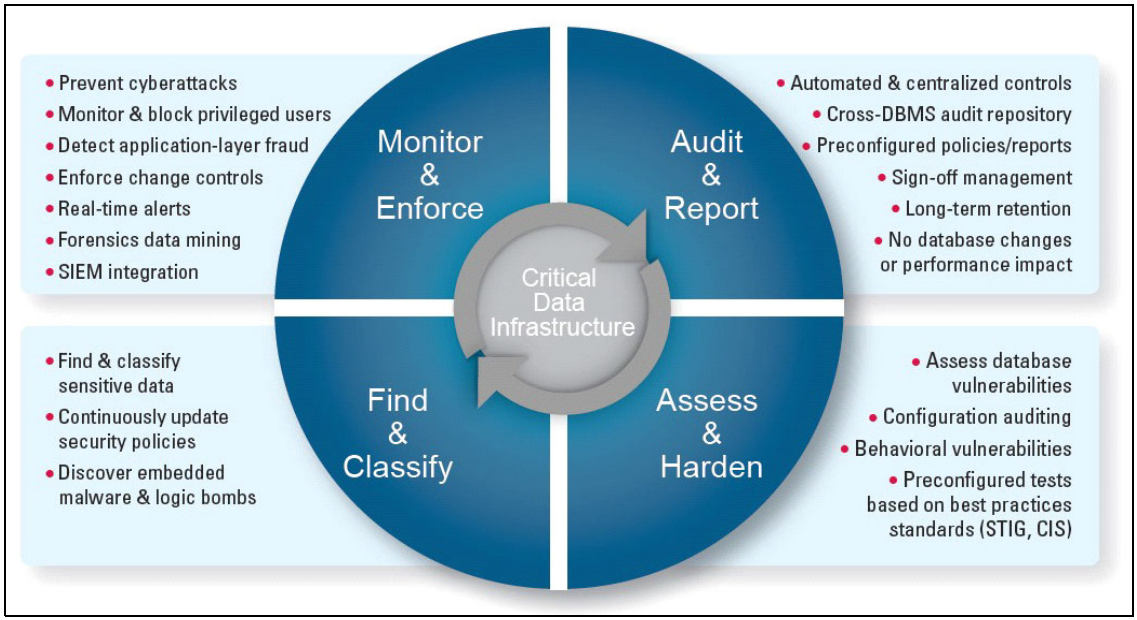


Figure 11-1 Database security and compliance

11.2.1 Discover your data

Sometimes a new database is introduced into a production environment outside of the normal control mechanisms. For example, the new database might be part of an application package from a software vendor. In older installations, some databases may have been left unmonitored and “forgotten” because the data or activities performed on it were not seen as a risk when the database was implemented.

Or in another case, a rogue DBA might create a new instance of the database and use it as desired, without being monitored. IBM InfoSphere Guardium's Auto-discovery application can be configured to probe the network, searching for and reporting on all databases discovered.

After an auto-discovery process has been defined, it can be run on demand or scheduled to be run on a periodic basis. There are two types of jobs that can be scheduled for each process:

- A scan job scans each specified host (or hosts in a specified subnet), and compiles a list of open ports from the list of ports specified for that host. A scan job must be run before running the second type of job.
- A probe job uses the list of open ports compiled during the latest completed scan only. The probe job determines if there are database services running on those ports. You can view the results of this job on the Databases Discovered predefined report (described later).

The two jobs can be scheduled individually, or the auto-discovery process can be defined to run the probe job as soon as the scan job completes.

Because the processes of scanning and probing ports can take time, the progress of an auto-discovery process can be displayed at any time (by clicking the Progress/Summary button).

After the jobs are completed, the results can be viewed using predefined reports.

To summarize, the following steps outline the procedure for using the Database Auto-discovery application:

1. Configure one or more Auto-discovery processes to search specific IP addresses or subnets for one or more ports.
2. Run the Auto-discovery process on demand or on a scheduled basis.

You can then view Auto-discovery reports, or create custom reports.

11.2.2 Identify and classify sensitive data

As the size and organization of the corporate database grows, sensitive information like credit card numbers and transactions, or personal financial data, may be present in multiple locations, without the knowledge of the current owners of that data. This frequently happens in corporations that have experienced mergers and acquisitions and in older corporations where legacy systems have outlasted their original owners. Even in the best of cases, integration and enhancement projects between disparate systems can easily leave sensitive data unknown and unprotected.

IBM InfoSphere Guardium provides the Classification feature to discover and classify sensitive data so that you can make and enforce effective access policy decisions.

A Classification Policy consists of a set of rules that are designed to discover and tag sensitive data elements. Sensitive data is looked for based on the type of discovery and an action is performed when sensitive data is found.

A Classification Process is a job consisting of a classification policy and one or more data sources. The process can be scheduled to run on a periodic basis as a task in a compliance workflow automation process.

The Classification feature secures information and manages risk when the sensitivity of the information is not known. By applying security policies to groups of objects with similar properties, you can group sensitive objects from multiple data sources. It assists in ensuring compliance when it is not clear which information is subject to the terms of particular regulations.

11.2.3 Secure and protect

When the sensitive data has been discovered and classified, the next step should be to determine the technologies that are needed to secure and protect it. As shown in the Figure 11-2 on page 225, you can divide it into the following types:

- Structured data
- Unstructured data
- Offline data
- Online data

The strategy and tools to secure each of these data types are different and we discuss them further in this section.

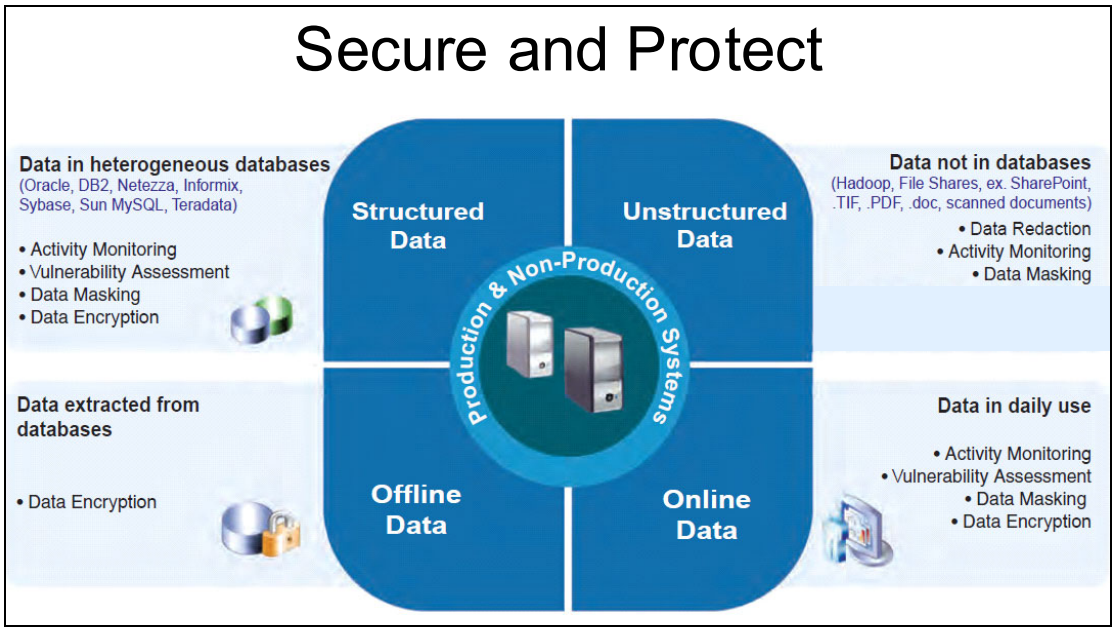


Figure 11-2 Secure and protect

Vulnerability Assessment

The InfoSphere Guardium Vulnerability Assessment application enables organizations to identify and address database vulnerabilities in a consistent and automated fashion. The Guardium assessment process evaluates the health of the database environment and recommends improvement by:

- Assessing system configuration against best practices and finding vulnerabilities or potential threats to database resources, including configuration and behavioral risks.

  As examples, identifying all default accounts that have not been disabled and checking public privileges and authentication methods chosen.

- Finding any inherent vulnerabilities present in the IT environment, such as missing security patches.

- Suggesting and prioritizing an action plan based on discovered areas of most critical risks and vulnerabilities. The generation of reports and suggestions provide guidelines on how to meet compliance changes and elevate security of the evaluated database environment.

  Guardium provides over 200 predefined tests to check database configuration parameters, privileges, and so forth. This applies to structured data and online data.

Data redaction: Structured data

In production database environments, sensitive information such as credit card numbers and SSN numbers should not be visible to outsourced database administrators (DBAs), application developers, or personnel who do not have a need to know. The InfoSphere Guardium policy actions feature allows a customer to mask portions of database query output (for example, credit card numbers) in reports for certain users. The selection Replacement Character in the Data Pattern/SQL Pattern section of the extrusion rule menu choices defines the masking character. Should the output produced by the extrusion rule match the regular expression of the data pattern, the portions that match subexpressions will be replaced by the masking character. Predefined regular expressions can also be used.

Data redaction: Unstructured data

The IBM InfoSphere Guardium Data Redaction solution provides automated redaction that works in two ways, depending on whether the document is free-text or a structured form. For free-text documents, the redaction engine automatically identifies and extracts relevant units of information, as shown in Figure 11-3.

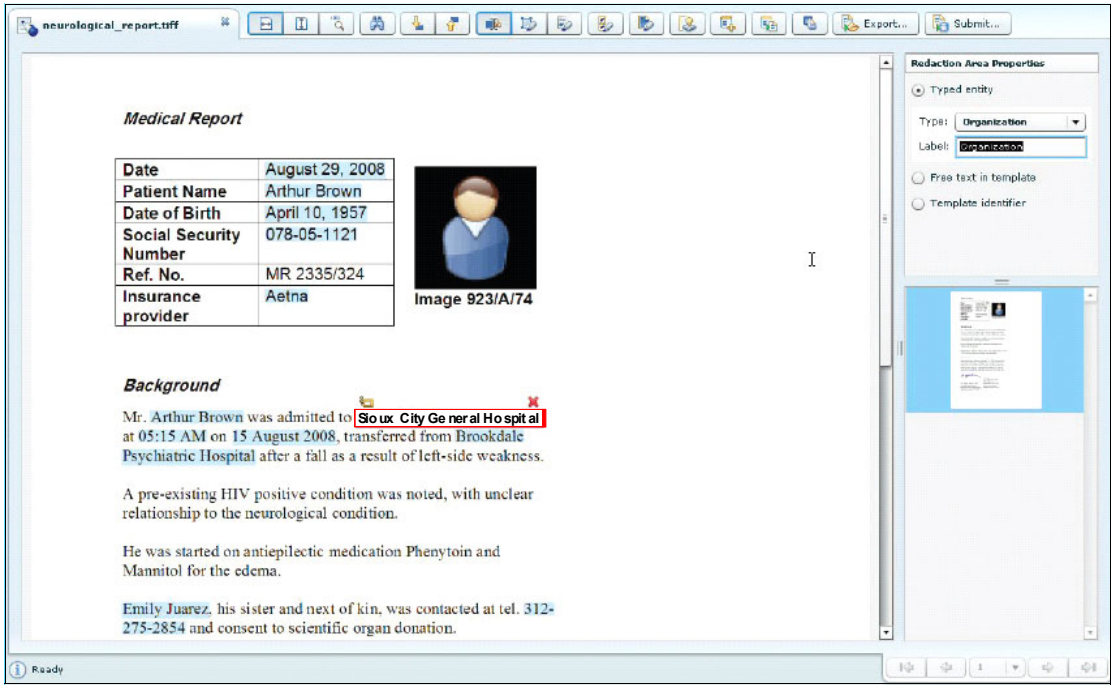


Figure 11-3 Free-text document with highlighted text redacted

Simply using text patterns is not enough—there is no formal pattern, for example, that captures personal names. Dictionaries are not sufficient either, since homonyms can disguise meanings (for example, is “bush” a plant or a former U.S. president?). Instead, it is necessary to combine regular expressions and dictionaries with a syntactic analysis of the text surrounding the relevant information. Structured forms, however, require a different technique, in which the known form layout is leveraged for accurate redaction. This allows even low-quality scans with hand written text to be processed; if they are accidentally skewed or resized, they can be straightened and aligned with a template. To accomplish this, a reviewer begins by redacting a sample form (for example, a blank) and marking the sensitive fields to be redacted, together with elements that identify instances of the form, such as the form title or identification number. This creates a template for subsequent forms. The software redaction solution matches templates to forms, eliminating the costly presorting of different form types. Next, it applies a template to each form, precisely deleting the marked fields that are based on their position, as depicted in Figure 11-4.

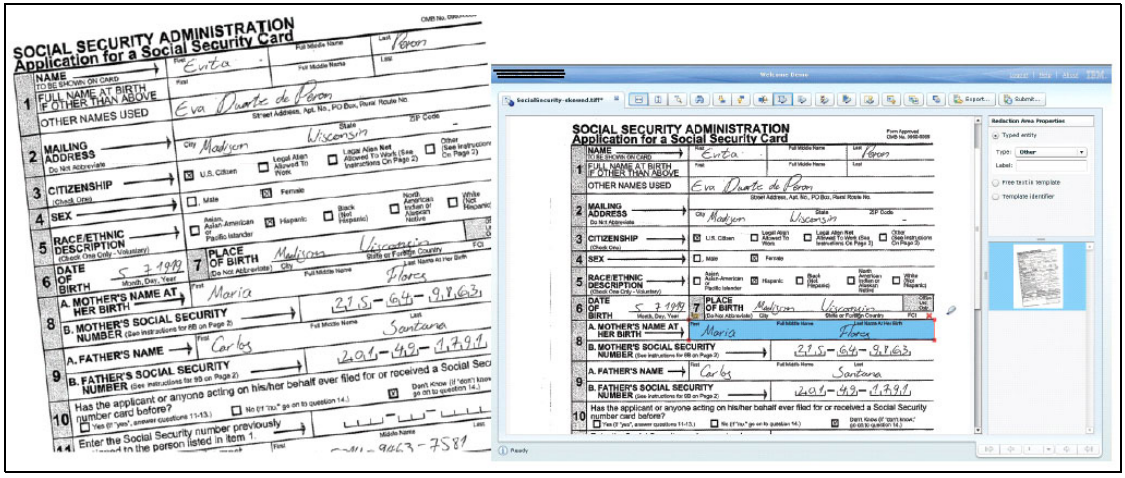


Figure 11-4 Skewed scan of a form

InfoSphere Guardium Data Redaction processes documents in many formats, such as PDF, TIFF, Microsoft Word, plain text, and XML files.

Some input documents, such as Microsoft Word and many PDFs, carry text in them, but others such as TIFF and some PDFs are pure images. For image files, the solution applies high-quality optical character recognition, and then processes the text. If there are any photographs or other graphics in a document, the solution preserves them as such.

Though most sensitive information arrives as text, images too can contain sensitive information. For example, an X-ray image may identify a patient’s name,

a portrait photograph may betray an identity or a satellite image may expose the location of a military unit. In the InfoSphere Guardium Data Redaction system, sensitive images can be located in a form using templates, or marked by a reviewer in the web-based Redaction Manager.

Each of the document formats, PDF, TIFF, Microsoft Word, and plain text and XML files, can serve not only as input but also output, and the choice of output type is configurable. In some cases, regulations or business needs require the redacted document to be in “native” format, the original format of the input. In other cases, it is necessary to output all documents, regardless of the input, into standard graphical formats preserving the precise layout, such as TIFF or PDF. Alternatively, if further machine processing is needed, plain-text output can be specified for all input formats.

Finally, the InfoSphere Guardium Data Redaction system automatically removes the wide variety of hidden information that is often stored in PDFs and Microsoft Word documents, even without the user’s knowledge. This includes hidden layers, comments and scripts, white text and tiny fonts, metadata such as the names of document editors and the creation date, and historical contents of a document preserved with editing features such as Undo and Track Changes. All these are safely deleted as part of the redaction workflow.

Efficient workflow: With thousands of documents to redact, workflow management is essential. Simply printing out the pages and deleting the sensitive text will not do it—it would be impossible to keep track of the stacks of paper—and the same is true with ad hoc redaction of masses of electronic documents. The only solution is for redaction to fit into Enterprise Content Management (ECM) processes. The InfoSphere Guardium Data Redaction system supports various such workflows out of the box, and can read and write documents in ECM systems such as IBM FileNet® P8 and IBM Content Manager 8.

Batch redaction: This workflow automatically redacts large numbers of documents in a repository. Depending on regulatory requirements, a reviewer can then examine 0 - 100% of the redacted documents and approve, reject or refine the redaction as needed. In this way, the redaction solution combines the strengths of machine processing and human domain knowledge.

On-demand redaction: A workflow used when individual documents must be processed as needed. For example, a business user may need to cleanse private information from a document before emailing it to a business partner. The sender can open the document in Redaction Manager, which instantly suggests text to be redacted. The sender can then refine this redaction before releasing the document.

Secure document viewing: For this workflow, InfoSphere Guardium Data Redaction provides a document viewer. All documents, regardless of original format, are displayed to the user in a uniform way in the browser, with no need to download a document that could subsequently be leaked. Sensitive information in the document is securely deleted according to the recipient's job role. In accordance with regulations, this data is typically deleted in a way that does not allow it to be viewed; for some types of information, users may have permission to securely retrieve the redacted units after specifying their need to know.

In Figure 11-5, a user is providing a business justification to access redacted SSN information; the user's permission level determines whether the requested information is revealed.

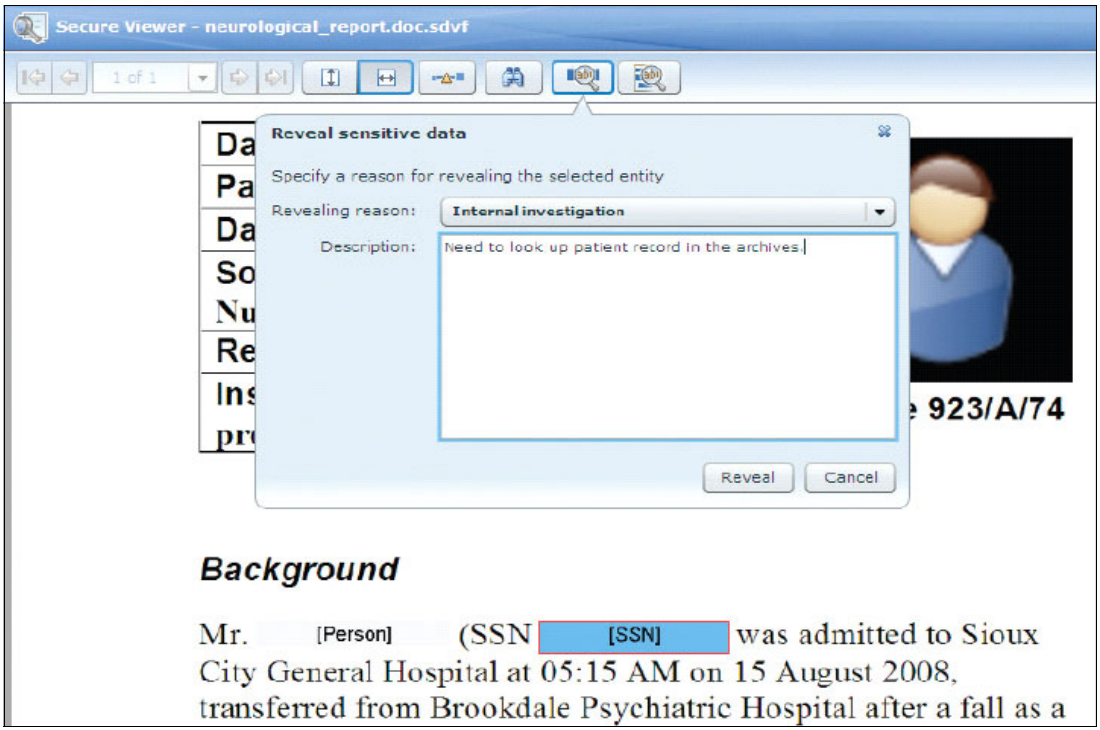


Figure 11-5 Secure document viewing

Policy-based redaction: To gain maximum business value in the redaction process while also minimizing deployment costs, the redaction solution must have the ability to quickly and easily implement the policies defined by regulatory frameworks, typically by cross-referencing the recipient's role against the type of information to be redacted.

The relevant roles are already defined in many enterprises. The redaction solution leverages these roles and links them to fine-grained permissions drawn from regulations, creating a privacy compliance system that directly meets requirements at minimum cost.

Thus, a physician might be allowed to see a patient's medical information, but not sensitive financial information, while the reverse is true for the hospital's billing clerk.

Likewise, in eDiscovery (electronic discovery) as part of legal cases, the United States Federal Rules of Civil Procedure specify that a litigant's attorney can see all client documents in full, including privileged information, while the opposing counsel can see the documents minus the attorney-client privileged information.

Data encryption

Main drivers for using encryption and access controls to protect sensitive data include compliance mandates such as PCI Data Security Standards (DSS), HIPAA, the Health Information Technology for Economic and Clinical Health (HITECH) Act, the Gramm-Leach-Bliley Act, data breach disclosure requirements, and the rapidly growing number of domestic and international data protection legislative acts.

While passing audits and minimizing risk are the primary goals for implementing encryption, selecting the right encryption solution for enterprise file shares, databases and storage systems requires specific focus on performance, extensibility, key management, and low administrative overhead.

InfoSphere Guardium Encryption Expert provides a comprehensive solution for key management and encryption of data at rest. It offers strong data security controls through policy-based access controls, separation of duties and auditing capabilities, all of which can be maintained from a centralized management console.

InfoSphere Guardium Encryption Expert consists of two major components:

- InfoSphere Guardium Encryption Expert Manager
- InfoSphere Guardium Encryption Expert Agents

The flexibility and scalability of the solution's design stems from its separation of the manager from the agents. InfoSphere Guardium Encryption Expert Manager provides centralized administration of encryption keys and data security policies, while InfoSphere Guardium Encryption Expert Agents help protect structured and unstructured data stores such as database and file server files, folders, documents, image scans, voice recordings, and logs.

InfoSphere Guardium Encryption Expert Manager

The InfoSphere Guardium Encryption Expert Manager integrates key management, data security policy management, and event log collection in to a centrally managed cluster that provides high availability and scalability to thousands of InfoSphere Guardium Encryption Expert Agents. This enables data security administrators to easily manage standards-based encryption across Linux, UNIX, and Windows operating systems in both centralized and geographically distributed environments. As shown in Figure 11-6, the InfoSphere Guardium Encryption Expert Manager stores the data security policies, encryption keys, and audit logs in a hardened appliance that is physically separated from the InfoSphere Guardium Encryption Expert Agents. Security teams can enforce strong separation of duties by requiring the assignment of key and policy management to more than one data security administrator so that no one person has complete control over the security of data. The InfoSphere Guardium Encryption Expert Manager is accessed from a secure web-management console and supports multiple Encryption Expert Agents. The InfoSphere Guardium Encryption Expert Manager functions as the central point for creating, distributing, and managing data encryption keys, policies, and host data security configurations.

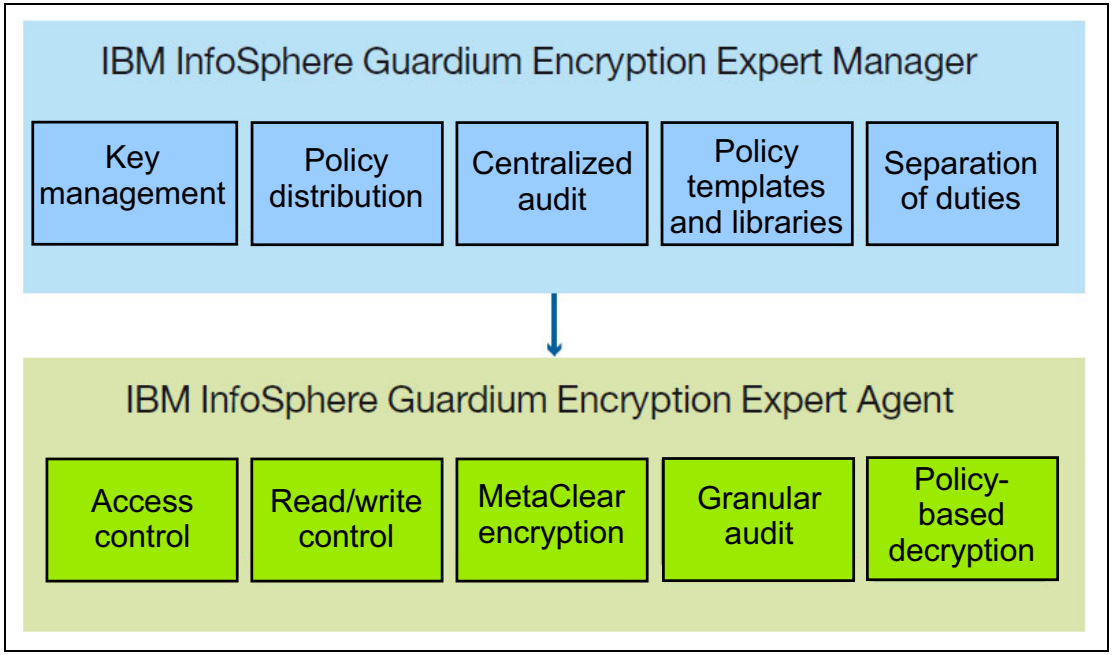


Figure 11-6 Components of InfoSphere Guardium data encryption

InfoSphere Guardium Encryption Expert Agents

InfoSphere Guardium Encryption Expert Agents are software agents that sit above the file system logical volume layers. The agents evaluate any attempt to access the protected data and apply predetermined policies to either grant or deny such attempts. They maintain a strong separation of duties on the server by encrypting files and leaving their metadata in the clear so IT administrators can perform their jobs without directly accessing the information. The agents perform the encryption and decryption, and access control works locally on the system that is accessing the data at rest. This enables encryption to be distributed within the data center and out to remote sites, while being centrally managed via the InfoSphere Guardium Encryption Expert Manager cluster as shown in Figure 11-7 on page 233. InfoSphere Guardium Encryption Expert Agents are installed on each server where data requires protection. The agents are specific to the OS platform and transparent to applications, databases (including Oracle, IBM, Microsoft, Sybase, and MySQL), file systems, networks, and storage architecture. Current OS support includes Windows, Linux, Sun Solaris, IBM AIX®, and HP-UX.

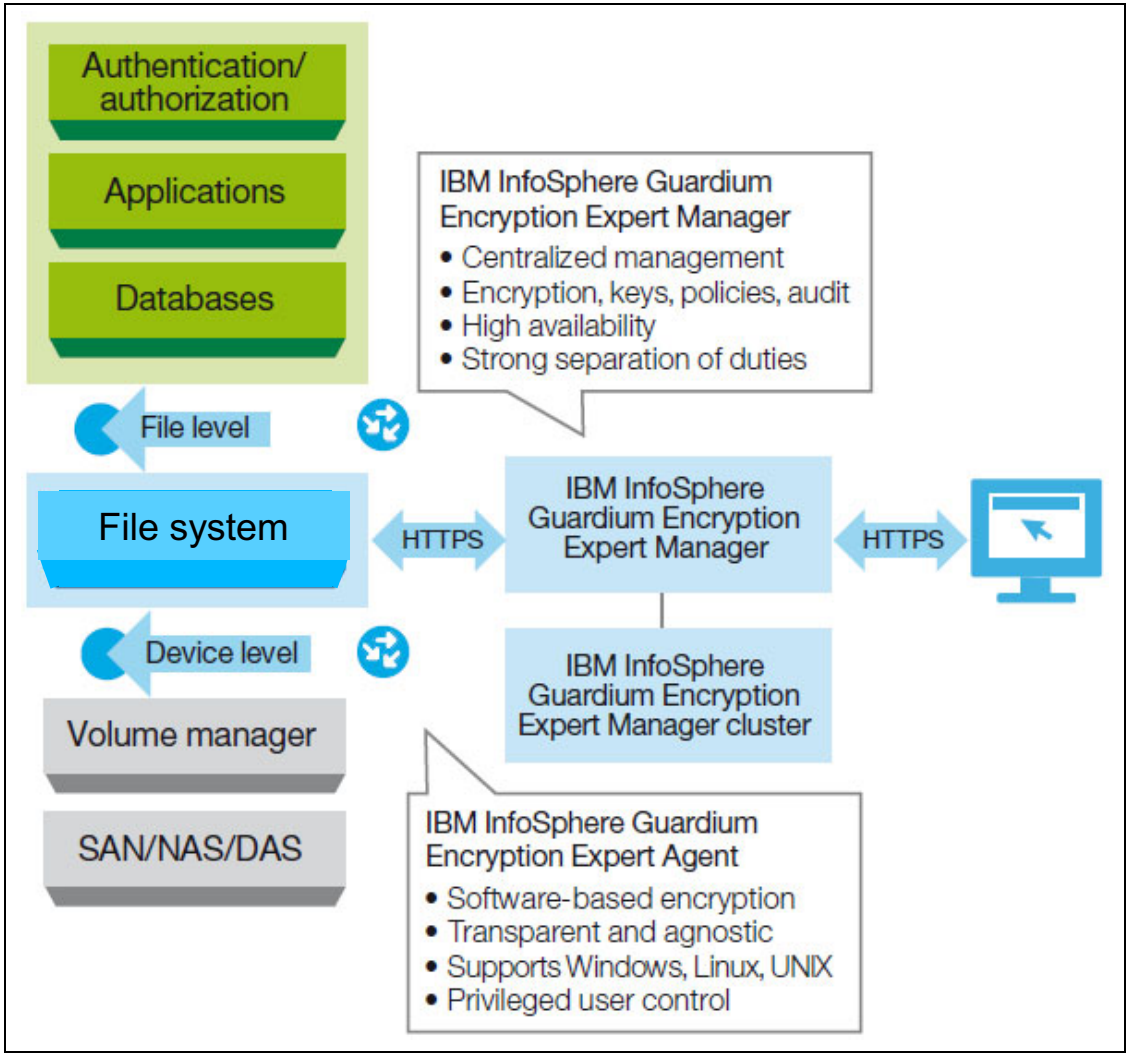


Figure 11-7 Encryption distribution

The advantages of this approach are as follows:

- Transparent implementation
- High performance
- Centralized key and policy management
- Strong separation of duties
- Role-based administration and domains
- Encryption and access control
- Scalability

- Distributed IT environments
- High availability
- Fine-grained auditing

11.2.4 Monitor and audit

After the sensitive data has been found and classified across the organization and preventive measures have been taken to encrypt and redact data, it is important to monitor database, data warehouses, file shares, and big data environments while automating compliance.

Activity monitoring

The InfoSphere Guardium Database Activity Monitoring (DAM) solution prevents unauthorized activities by privileged insiders or hackers while monitoring users to identify fraud, without any changes to the data infrastructure and applications or impacting performance. It adds the ability to monitor and enforce policies with alerting and blocking for sensitive data access, privileged user actions, change control, application user activities, and security exceptions such as failed logins. It also provides centralized aggregation and normalization of audit data from across the data infrastructure for enterprise-wide compliance auditing and reporting, correlation, investigations, and forensics with a secure, tamper-proof audit trail that supports the separation of duties (SOD) required by auditors. Integrated Compliance Workflow Automation enables to automate the entire compliance auditing process, including report distribution to oversight teams, sign-offs, and escalations with preconfigured reports relating to SOX, PCI DSS, and data privacy.

Gathering requirements

The first step is to gather and define requirements. Requirements are often determined by an organization's auditors, especially in SOX or PCI implementation, but can also be determined by internal security rules.

If clear definition of requirements does not exist, try to answer the following questions:

- Logging and real-time alerting:
  - Who needs to be monitored? Privileged users, DBAs, everyone?
  - What types of actions must be monitored? DDL, DML, selects on specific tables?
  - What type of information can safely be ignored?
  - What type of activity should prompt alerts?

  - What type of activity should prompt more verbose logging (logging the full SQL string, including values)?
  - What are the sensitive objects?
- Reporting:
  - What reports do I need?
  - What fields do I need in my reports?
  - What should prompt an action to appear on reports (query conditions)?
- Audit Review:
  - Who needs to receive monitoring reports?
  - How frequently should reports be delivered?
  - Should users be required to sign reports or is reviewing reports sufficient (sign off of reports can be configured on a per user basis)
  - Should the delivery of reports stop at any receivers until they have reviewed or signed off on them or should they be delivered to all users at once?
- Threshold (correlation) alerting:

  What type of activities over a period of time should prompt threshold (correlation) alert and what mechanism should be used to send these alerts and who should be the recipient of these alerts?
- Data level access control (blocking)
  - Who needs to be blocked?
  - What type of activities should be blocked?

Building groups

Groups simplify policy and query creation by allowing users to organize Guardium data elements that are based on their reporting requirements. It is much easier to create reports and policy after the groups have been defined. For example, assume that the company has 25 separate data objects containing sensitive employee information, and a report is required on all access to these items. A very long query testing for each of the 25 items can be formulated. Alternatively, a single group called *sensitive objects* can be defined, containing those 25 objects. That way, in queries or policy rule definitions, only the group needs to be included in the where clause, instead of each separate object.

There are six ways to populate groups. Generally, manual entry is sufficient but if many members need to be loaded, or a group needs to be updated on a scheduled basis, one of the other methods might be more appropriate.

The following are methods to populate groups:

- Manual Entry
- LDAP: Imports group members from your Lightweight Directory Access Protocol (LDAP) or Active Directory Servers
- Populate from Query: Adds group members based on data in the Guardium database (this data can be imported from an external data source)
- Auto Generated Calling Prox: Analyzes stored procedures and updates an object group based on what the procedure does or the data it accesses
- Classification: Analyzes databases and updates groups based on patterns in the data. For example, it can search for text patterns in tables that could contain credit card numbers and then add the tables to a sensitive object group
- Guard: Allows to import many group members from a flat file via Secure Shell (SSH)

Defining policy

A policy is a set of rules and actions that are applied against SQL traffic as it is captured by the Guardium appliance in real time. Each rule within the policy contains a set of criteria and one action. For example, send an alert via email any time DML is executed against a sensitive object. Each rule is applied in sequence as the data is being collected in real time.

This is where you ensure that activity is logged based on the monitoring requirements as defined by the logging and real-time alerting questions. Policies define what traffic should be ignored, what activities require more detail, and which actions should prompt real-time alerts.

The order and logic of the policy is very important. Also, there are options that can completely change the methods used to log data. These methods include Selective Audit Trail, Flat Logging, and Baselines. Refer to the user manual for details about these methods.

Creating reports

Now that the groups and policies have been defined, queries and reports need to be created. The following are required elements in creating queries/reports:

1. Main Entity: The main entity defines the data type that will be the main focus of the report. Generally, one of the following four main entities will be chosen:
   a. *Session*: Used when reporting on successful logins to the database server. This main entity provides one line per login, with no detail on the activity performed by the user.

b. *Command*: Used when user actions are the main focus of the report. Each individual command that a user issues will have its own line on the report.

c. *Object*: Used if the actual object name accessed is required. Each object accessed will appear on separate line. Generally, this will result in multiple lines per SQL requests (one for each object referenced in the SQL requests).

d. *SQL (or Full SQL if logging full details)*: Used to provide one line per unique SQL statement. This is appropriate if the SQL statement is required in the report. *Note:* a complete SQL statement can be hundreds of lines long and can make reports very difficult to read.

2. Query Attributes: Query Attributes are the fields that appear in the report. The most commonly used attributes include:

   a. Time Stamp

      i. From Access Period if using Command or Object main entity and you are not logging full details

      ii. From Full SQL if you are logging full details

   b. Session Start

   c. Client/Server: Server IP

   d. Client/Server: Client IP

   e. Client/Server: DB User Name

   f. Client/Server: Source Program

   g. Session: Database Name

      i. (Client/Server: Service Name if Oracle)

   h. Command: SQL Verb (If using a Main Entity of Command or lower)

   i. Object: Object Name (If using a Main Entity of Object or lower)

   j. SQL: SQL (or Full SQL if logging full details)

3. Query conditions: The query conditions filter the data that will appear on reports (the where clause of query). Because groups have been defined, creating the where clause is very simple.

   Use Groups or Run-time Parameters, instead of hard coding values, whenever possible. This allows for much more flexibility later, if reports need to be changed. Runtime parameters also allow to produce multiple result sets from a single query.

Developing workflow

Guardium workflow functionality allows reports to be scheduled and the result to be delivered to end users for review and signoff. Workflow results can be

delivered to individual users, group of users or roles. The suggestion is to deliver workflow results to roles, which allows more than one user to review and sign off on a result and it is easier to manage employee's absence and turnover.

In order to benefit from this functionality, users and roles should be defined in Guardium:

- Define roles in the organization by answering the following questions:
  - Who should receive reports and what is the job function of each receiver (infosec, DBA manager or internal audit)
  - What users have the same job function and can provide an equivalent review and signoff
- Created defined roles through UI (access management) or utilize roles that are predefined in Guardium
- Create users through UI and assign appropriate role through access management

The Audit Process Builder is used to define:

- Who receives the reports
- Which reports are delivered
- The frequency of delivery
- The workflow, which includes:
  - The order of delivery
  - Whether sign-off is required
  - Whether the delivery should stop at any user or role until they have reviewed or signed off on the audit process

A

Additional material

This book refers to additional material that can be downloaded from the Internet as described in the following sections.

Locating the Web material

The Web material associated with this book is available in softcopy on the Internet from the IBM Redbooks Web server. Point your Web browser at:

`ftp://www.redbooks.ibm.com/redbooks/SG248164`

Alternatively, you can go to the IBM Redbooks website at:

`ibm.com/redbooks`

Select the **Additional materials** tab and open the directory that corresponds with the IBM Redbooks form number, SG248164.

Using the Web material

The additional Web material that accompanies this book includes the following files:

File name	*Description*
WorkFlow.zip	Contains High Level Work Flow Diagrams for Application Assessment, Decommissioning Project Process, and Technical Application Assessment.
Spreadsheets.zip	Contains two spreadsheets from Chapter 8, “Information Lifecycle Management” on page 121 regarding “Roles and Responsibilities with Foundation Principle Area” and “Understand Data with Foundation Principle Area.”

Downloading and extracting the Web material

Create a subdirectory (folder) on your workstation, and extract the contents of the Web material .zip file into this folder.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

You can search for, view, download, or order documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

This publication is also relevant as a further information source:

- *Three guiding principles to improve data security and compliance, IBM White Paper IMW14568-USEN-05*

Online resources

These websites are also relevant as further information sources:

- IBM InfoSphere Guardium Data Redaction: Reconciling openness with privacy

  http://public.dhe.ibm.com/common/ssi/ecm/en/imw14307usen/IMW14307USEN.PDF

- IBM InfoSphere Guardium Tech Talk: Guardium 101

  https://www.ibm.com/developerworks/community/files/form/anonymous/api/library/a7d92be4-beb0-4b24-b1c3-eaa8f29e91c9/document/e2ecd990-937d-40ff-bc07-abbfcfddb0f8/media/guardium101_slidesonly.pdf

- DIBM InfoSphere Guardium Encryption Expert: An overview

  http://public.dhe.ibm.com/common/ssi/ecm/en/imw14581usen/IMW14581USEN.PDF

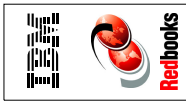
Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



IBM Information Governance Solutions

(0.5” spine)
0.475”<->0.875”
250 <-> 459 pages

IBM Information Governance Solutions

New approaches for developing long lasting and trusted information assets

A strong foundation of data policies and practices

Scenarios to help guide your development

Managing information within the enterprise has always been a vital and important task to support the day-to-day business operations and to enable analysis of that data for decision making to better manage and grow the business for improved profitability. To do all that, clearly the data must be accurate and organized so it is accessible and understandable to all who need it.

That task has grown in importance as the volume of enterprise data has been growing significantly (analyst estimates of 40 - 50% growth per year are not uncommon) over the years.

All this has brought with it a renewed and critical need to manage and organize that data with clarity of meaning, understandability, and interoperability. What is described here has been and is being done to varying extents. It is called “information governance.”

This IBM Redbooks publication focuses on the building blocks of a solid governance program. It examines some familiar governance initiative scenarios, identifying how they underpin key governance initiatives, such as Master Data Management, Quality Management, Security and Privacy, and Information Lifecycle Management.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-8164-00 ISBN 0738439517