

IBM Information Server Integration and Governance for Emerging Data Warehouse Demands

Information Quality and Governance

Self-service Data Integration

Big Data Integration



Chuck Ballard
Manish Bhide
Holger Kache
Bob Kitzberger
Beate Porst
Yeh-Heng Sheng
Harald C. Smith



International Technical Support Organization

**IBM Information Server: Integration and
Governance for Emerging Data Warehouse
Demands**

July 2013

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

First Edition (July 2013)

This edition applies to Version 9.1 of “IBM InfoSphere Information Server for Data Warehousing” (product number 5725-C80), “IBM InfoSphere Information Server Packages” (product number 5725-G05), and “IBM InfoSphere Information Server” (product number 5724-Q36).

© Copyright International Business Machines Corporation 2013. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices	vii
Trademarks	viii
Preface	ix
Authors	xi
Now you can become a published author, too!	xiv
Comments welcome	xv
Stay connected to IBM Redbooks	xv
Part 1. Overview and concepts	1
Chapter 1. Overview of IBM InfoSphere Information Server	3
1.1 Packaged Editions	4
1.2 Information Server Components	6
1.2.1 InfoSphere Blueprint Director	8
1.2.2 InfoSphere Discovery	9
1.2.3 InfoSphere Metadata Workbench	11
1.2.4 InfoSphere Data Architect and IBM Industry Data Models	12
1.2.5 InfoSphere Business Glossary	15
1.2.6 InfoSphere QualityStage	16
1.2.7 InfoSphere Information Analyzer	18
1.2.8 InfoSphere Data Quality Console	18
1.2.9 InfoSphere Information Services Director	19
1.2.10 InfoSphere FastTrack	19
1.2.11 InfoSphere DataStage	20
1.2.12 InfoSphere DataStage Balanced Optimization	21
1.2.13 InfoSphere Change Data Delivery	22
1.2.14 InfoSphere Data Click	22
Chapter 2. Using Information Server to design and implement a Data Warehouse	25
2.1 How the capabilities fit together	26
2.2 Method and proven practices: Business-driven BI development	28
2.3 Phases	30
2.4 Information Server components by Phase	30
2.4.1 Plan	31
2.4.2 Discover	32
2.4.3 Analyze	32
2.4.4 Define	33

2.4.5 Develop	33
2.4.6 Deploy	34
Part 2. Meeting the increasing demands of workloads, users, and the business	35
Chapter 3. Data Click: Self-Service Data Integration	37
3.1 Motivation and overview	38
3.1.1 Benefits of Data Click over traditional approaches	40
3.1.2 Data Click details	42
3.2 The two-click experience for a self-service user	44
3.2.1 Running and feedback	46
3.2.2 Advanced user configuration	47
3.3 Summary and more resources	51
Chapter 4. Incorporating new sources: Hadoop and big data	53
4.1 Big Data File Stage	56
4.2 Balanced Optimization	59
4.3 Balanced Optimization for Hadoop	61
4.3.1 Complete pushdown optimization	62
4.3.2 Hybrid pushdown optimization	64
4.4 IBM InfoSphere Streams Integration	67
4.5 Oozie Workflow Activity stage	69
4.6 Unlocking big data	72
Chapter 5. SPSS: Incorporating Analytical Models into your warehouse environment	73
5.1 Analytics background	75
5.2 Motivating examples	75
5.2.1 A banking example	75
5.2.2 A telecom example	76
5.2.3 A customer care example	77
5.3 End-to-end flow	78
5.3.1 Model building by using IBM SPSS Modeler	78
5.3.2 Model scoring within IBM InfoSphere DataStage	79
5.4 Integrating IBM SPSS Models with external applications	80
5.4.1 Publishing a Stream	81
5.4.2 Running a Stream	82
5.5 Building SPSS Stage in IBM InfoSphere DataStage	84
5.5.1 Extending IBM InfoSphere DataStage	84
5.5.2 Other features of SPSS stage	86
5.6 Summary	87
Chapter 6. Governance of data warehouse information	89
6.1 Information and expectations	91

6.1.1	Business drivers	91
6.1.2	Using the information	92
6.2	Information Governance: The Maturity Model	93
6.2.1	Elements of an Information Governance Maturity Model	95
6.2.2	Business terms: The language of the business	95
6.3	Business terms: Enablers of awareness and communication	97
6.3.1	Sources of business terms	98
6.3.2	Standard practices in glossary development and deployment	100
6.3.3	Examples of glossary categories and terms	101
6.4	Information Governance policies and rules	103
6.4.1	Definition and management of information policies	104
6.4.2	Definition and management of Information Governance rules	106
6.4.3	Standard Practices in Information Governance policy and rule development	106
6.5	Information stewardship	107
6.6	Information Governance for the data warehouse	109
6.7	Conclusion	114
Chapter 7. Establishing trust by ensuring quality		115
7.1	Moving to trusted information	117
7.1.1	Challenges to trusted information	118
7.1.2	Impact of information issues	118
7.2	Mission of information quality	119
7.2.1	Key information quality steps	120
7.3	Understanding information quality	121
7.3.1	Data Quality Assessment	121
7.3.2	Expanding on the initial assessment	124
7.4	Validating data with rules for information quality	125
7.4.1	Incorporating business value and objectives	125
7.4.2	Defining the primary requirements	126
7.4.3	Designing the data rules	127
7.4.4	Example of data rule analysis	129
7.4.5	Setting priorities and refining conditions	129
7.4.6	Types of Data Rules	129
7.4.7	Examples of rules	130
7.4.8	Considerations in Data Rule design	131
7.4.9	Breaking requirements into building blocks for Data Rules	132
7.4.10	Evaluating Data Rule results	134
7.5	Measuring and monitoring information quality	136
7.5.1	Establishing priorities	136
7.5.2	Setting objectives and benchmarks	137
7.5.3	Developing the monitoring process	138
7.5.4	Implementing the monitoring process	139

7.6 Information Quality Management	141
7.6.1 Lifecycle and deploying Data Rules	141
7.6.2 Publishing data rules for reuse	143
7.6.3 Deploying data rules to production	144
7.6.4 Running Data Rules in production	145
7.6.5 Delivering and managing data quality results	145
7.6.6 Developing Information Quality reports	147
7.6.7 Retaining and archiving old results	149
7.6.8 Improving ongoing processes	149
7.7 Conclusion	151
Chapter 8. Data standardization and matching	153
8.1 Use cases	154
8.1.1 Conditioning and standardization	156
8.1.2 Address verification	165
8.1.3 Matching and de-duplication	167
8.1.4 Consolidation and enrichment	169
8.1.5 Summary	170
Related publications	171
IBM Redbooks	171
Other publications	172
Online resources	172
Help from IBM	172

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:

BigInsights™

Cognos®

DataStage®

DB2®

developerWorks®

IBM PureData™

IBM®

InfoSphere®

Optim™


PureData™

QualityStage®

Rational Team Concert™

Rational®

Redbooks®

Redbooks (logo) ®

SPSS®

The following terms are trademarks of other companies:

Netezza, and N logo are trademarks or registered trademarks of IBM International Group B.V., 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.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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

The data warehouse can be a complex and expensive IT investment. Despite the expense, organizations continue to invest heavily in data warehouse efforts because warehouses are the essential underpinnings of analytical capabilities that are required to help organizations deeply understand their business, their customers, and their markets. Gaining deeper customer insight from the information requires incorporating an ever-increasing volume and variety of information, with a higher quality. Regulatory requirements require greater transparency and insight into how the information flows throughout an organization. Competitive pressures require a wider variety of employees across the organization to nimbly access and analyze the information in the form and time they require.

As these needs increased, so have the demands that are placed on development teams and on the tools they use to help them design, construct, and populate a data warehouse. Ad hoc processes might suffice to deliver on earlier warehouses, but fall short under today's demands. Individual tools for data movement and data quality provide tremendous value, yet often lack the capabilities that are required to maximize your return on investment. And information that is accessible and understandable only by a select few specialists prevents timely responses to business challenges. To meet the needs of today's and tomorrow's data warehousing demands, organizations must take a business-driven approach, adopt proven development practices and processes, provide self-service capabilities to their internal customers, and fully use the capabilities of a comprehensive data integration platform.

This IBM® Redbooks® publication assumes that the reader has a basic working knowledge of data warehousing design and construction, and that the organization already has a data warehouse in place. It provides an overview of IBM Information Server's capabilities, sufficient to understand how it supports data warehouse design, construction, and population. Most importantly, it describes how organizations can maximize the value of their data warehouse investment by using some of the following key capabilities of IBM InfoSphere® Information Server:

- ▶ Enable business users to create their own data marts
- ▶ Incorporate big data with the warehouse environment
- ▶ Incorporate statistical and analytical models into the warehouse
- ▶ Establish governance of data warehouse information for downstream use
- ▶ Establish trust in the quality of information
- ▶ Improve location-based information in the warehouse, such as addresses

In this book we describe the usage of many of the following product modules and components of InfoSphere Information Server:

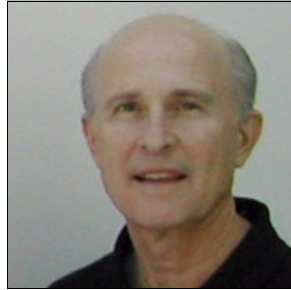
- ▶ IBM InfoSphere Blueprint Director
- ▶ IBM InfoSphere Business Glossary
- ▶ IBM InfoSphere Change Data Delivery
- ▶ IBM InfoSphere Data Architect
- ▶ IBM InfoSphere Data Click
- ▶ IBM InfoSphere Data Quality Console
- ▶ IBM InfoSphere DataStage®
- ▶ IBM InfoSphere Discovery
- ▶ IBM InfoSphere FastTrack
- ▶ IBM InfoSphere Information Analyzer
- ▶ IBM InfoSphere Metadata Workbench
- ▶ IBM InfoSphere QualityStage®

Who this book is for

This book is intended for business leaders and IT architects who are responsible for building and extending their data warehouse and Business Intelligence infrastructure. It provides an overview of powerful new capabilities of Information Server in the areas of big data, statistical models, data governance and data quality. The book also provides key technical details that IT professionals can use in solution planning, design, and implementation.

Authors

This book was produced by the following team of product and solution specialists from around the world working with the 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 a Bachelors degree and a Masters degree in Industrial Engineering from Purdue University.



Manish Bhide is a senior architect responsible for driving the big data integration initiatives for IBM Information Server. He also is responsible for the key initiative DataClick, which aims to simplify the use of technology for data scientists and self-service users. Manish has been with IBM for more than 12 years and has diverse experience in different areas of Information Management spanning IBM Research and IBM Software Group. He has a passion for innovation and is credited with creating several technologies that made an impact on the marketplace. Manish has filed more than 32 patents and has more than 25 publications in IEEE/ACM conferences. Manish has a PhD in Computer Science from IIT Bombay.



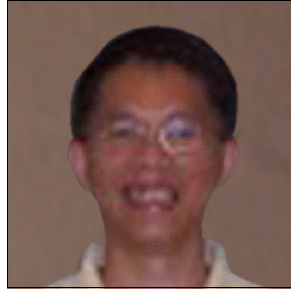
Holger Kache is a Technical Solutions Architect in IBM Information Management in the US. He has 12 years of product engineering experience with a focus on Data Governance and Metadata. Holger holds a masters degree in Computer Science from the Brandenburg Technical University in Cottbus, Germany. His previous publications include research papers and books on Federated Database Technology, Database Query Optimization, and Data Governance.



Bob Kitzberger is a senior product portfolio manager in the IBM InfoSphere business, focused on the strategy for the IBM InfoSphere Information Server family of data integration and data quality software. In previous positions, he was responsible for product management of the IBM Rational® modeling and requirements management products, and led development teams for complex software development tools and systems software. Bob is a graduate of the University of Oregon's MBA program and the University of California San Diego's computer science program.



Beate Porst is the Product Manager for IBM InfoSphere QualityStage and the Data Quality Console located at the IBM Silicon Valley Lab in San Jose, CA. As the Product Manager, she is responsible for the entire product lifecycle from strategic planning to tactical execution. Before becoming a Product Manager, Beate was an architect in the IBM Information Management Engineering and Solution group that focused on architecture and development of reusable assets to support a richer metadata integration among IBM Information Management products. Beate has more than 10 years of experience in Information Management, holding engineering and product management roles in IBM DB2®, InfoSphere Federation Server, and InfoSphere Information Server. Beate holds a Masters Degree in Computer Science from the University of Rostock, Germany.



Yeh-Heng Sheng is a Software Architect for InfoSphere Servers, Information Management in San Jose, California. He has over 25 years of Database and Data Warehouse development experience. He holds a PhD degree in Computer Science from State University of New York, Stony Brook. His expertise is in the areas of database query optimization, client/server connectivity and logical programming. He has written extensively on these subjects, and presented at conferences worldwide.



Harald C. Smith is a Software Architect in IBM Software Group, in Littleton, MA, specializing in information quality, integration, and governance products. He has over 30 years experience working in product and project management, software and application development, technical services, and business processes, and is IBM certified in delivering IBM Information Management solutions. Harald is the co-author of *Patterns of Information Management* by IBM Press and is an IBM developerWorks® Contributing Author for which he has written extensively. He has a Bachelors degree in History and a Masters degree in Social Science from the University of Chicago.

Other Contributors

In this section, we thank the following people who contributed to this IBM Redbooks publication, in the form of written content, subject expertise, and support:

- ▶ IBM locations worldwide:
 - Tony Curcio
Product Manager, IBM Software Group, Information Management,
Charlotte, NC, US
 - Robert J. Dickson
Worldwide Information Server Technical Sales, IBM Software Group,
Worldwide Sales, Schaumburg, IL, US

- Sriram Padmanabhan
Distinguished Engineer, Chief Architect for InfoSphere Servers,
Information Management, Silicon Valley Lab, San Jose, CA, US
 - Barry Rosen
Senior Certified Global Executive Architect, Information Management,
Westford, MA, US
 - Guenter Sauter
Product Manager, IBM Software Group, Information Management,
Somers, NY, US
- From the International Technical Support Organization:
- Mary Comianos, Publications Management
 - Ann Lund, Residency Administration
 - Linda Robinson, Graphics Support

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 this website:

<http://www.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 by using one of the following methods:

- ▶ Use the online **Contact us** review Redbooks form found at:
<http://www.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>



Part 1

Overview and concepts

In this part we, introduce the following concepts:

- ▶ IBM InfoSphere Information Server packaged editions and capabilities
- ▶ An iterative process for designing and expanding your data warehouse, and how Information Server's capabilities support that process

The following background information provides context for better understanding Information Server's capabilities:

- ▶ Chapter 1, “Overview of IBM InfoSphere Information Server” on page 3
- ▶ Chapter 2, “Using Information Server to design and implement a Data Warehouse” on page 25



Overview of IBM InfoSphere Information Server

IBM InfoSphere Information Server is a market-leading data integration platform that helps you understand, cleanse, transform, and deliver trusted information to your critical business initiatives, such as Business Intelligence, big data, Master Data Management, and point-of-impact analytics.

Information Server provides a broad set of capabilities that are built on a common platform to ensure high levels of team productivity and effectiveness. These end-to-end capabilities help you understand and govern data, create and maintain data quality, and transform and deliver data. At the core of these capabilities is a common metadata repository that stores imported metadata, project configurations, reports, and results for all components of InfoSphere Information Server.

This chapter includes the following topics:

- ▶ Packaged Editions
- ▶ Information Server Components

1.1 Packaged Editions

Information Server's capabilities are available in four different editions, which provide a comprehensive set of capabilities for common data integration, quality, and governance tasks, as shown in Figure 1-1.

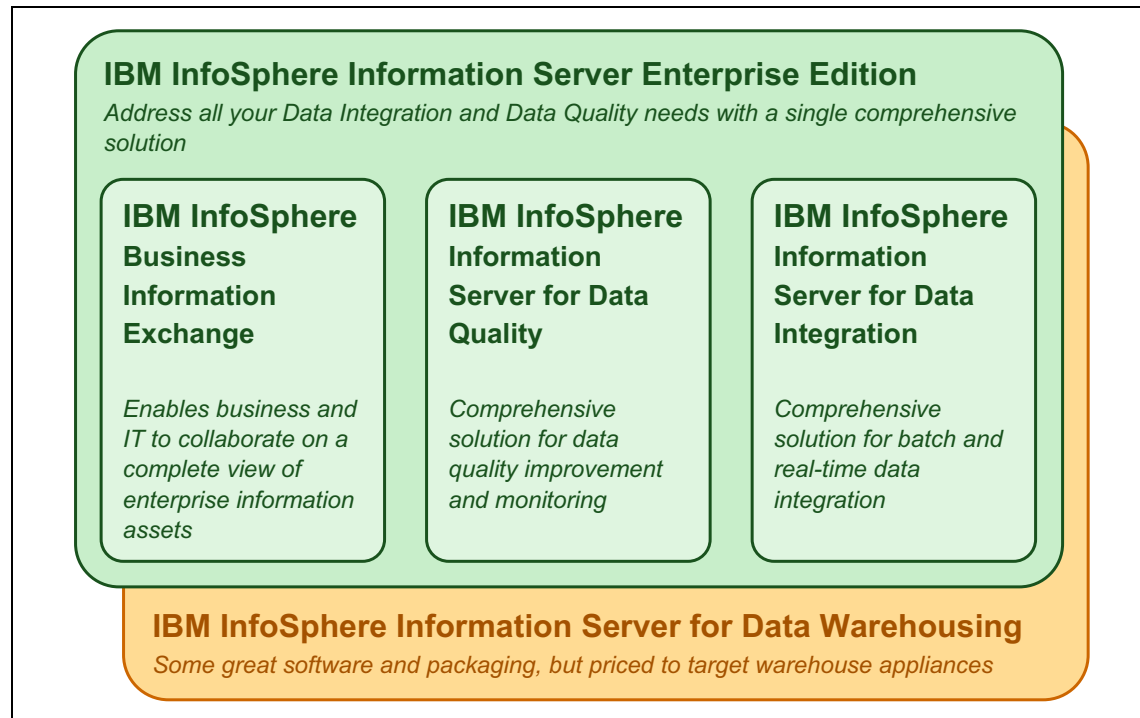


Figure 1-1 Information Server packaged editions

IBM InfoSphere Business Information Exchange encourages a standardized approach to discovering your IT assets and defining a common business language, and helps you to complete the following tasks:

- ▶ Create a well-documented, end-to-end information blueprint to ensure you aligned your business requirements with your enterprise and reference architectures before you start your strategic project.
- ▶ Establish a common business language and manage business perspectives about information and align those views with the IT perspective.
- ▶ Automate the discovery of relationships within and across data sources to accelerate project deployment and manage and explore data lineage to create trusted information that supports data governance and compliance efforts.

- ▶ Provide a solid foundation for different types of information and governance projects, including information integration, lifecycle management, and security and privacy initiatives.

IBM InfoSphere Information Server for Data Quality provides rich capabilities that enable you to cleanse data and monitor data quality, which turns data into trusted information. By analyzing, cleansing, monitoring, and managing data, you can make better decisions and improve business process execution. Use this product for the following tasks:

- ▶ Automate source data investigation, information standardization, and record matching, which are all based on business rules that you define.
- ▶ Use comprehensive and customizable data cleansing capabilities in batch and real time.
- ▶ Enrich data and make sure the best data across sources survives.
- ▶ Match records to eliminate duplicates, householding, and many other operations.
- ▶ Monitor and maintain your data quality. Establish data quality metrics that are aligned with business objectives that enable you to quickly uncover data quality issues and establish a remediation plan.
- ▶ Properly manage and support a data governance program.

By using IBM InfoSphere Information Server for Data Integration, you can transform data in any style and deliver it to any system, which ensures faster time to value. Built-in transformation functions and a common metadata framework help you save time and expense. InfoSphere Information Server also provides the following options for delivering data, whether via bulk (extract, transform, load), virtual (federated) or incremental (data replication) delivery:

- ▶ Use unique project blueprinting capabilities to map out your data integration project to improve visibility and reduce risk.
- ▶ Use unique discovery capabilities to understand the relationships within and across data sources before moving forward.
- ▶ Deliver faster time to value by deploying an easy-to-use graphical interface to help you transform information from across your enterprise.
- ▶ Integrate data on demand across multiple sources and targets, while satisfying the most complex requirements with the most scalable run time available.

- ▶ Save time and expense by using hundreds of built-in transformation functions, and promote collaboration through a common metadata framework. Select from multiple data delivery options whether through bulk data delivery via extract, transform, and load (ETL) or incremental data delivery (Change Data Delivery).
- ▶ Benefit from balanced optimization capabilities and choose the deployment option that works best for you, such as ETL and extract, load, and transform (ELT).
- ▶ Take advantage of connectivity to database management system (DBMS), big data sources, messaging queues, Enterprise Resource Planning (ERP) and other packaged applications, industry formats, and mainframe systems, all with unique native API connectivity and parallelism.

IBM InfoSphere Information Server Enterprise Edition provides all of the capabilities of the other three packaged editions, which provide your team with the full set of capabilities that are required to support trusted and governed information for your warehouse.

1.2 Information Server Components

Information Server consists of numerous product modules or components. These components are combined into the packaged editions that are described in 1.1, “Packaged Editions” on page 4, as shown in the table in Figure 1-2 on page 7.

Product Module	Capability	Business Information Exchange	Information Server for Data Quality	Information Server for Data Integration	Information Server Enterprise Edition
Blueprint Director	Blueprinting and Best Practices	✓	✓	✓	✓
Discovery	Data Discovery	✓	✓	✓	✓
Metadata Workbench	Metadata Management and Lineage	✓	✓	✓	✓
Data Architect	Logical and Physical Data Modeling	✓		✓	✓
Business Glossary	Business Glossary and Workflow	✓			✓
Quality Stage	Data Cleansing and Enrichment		✓		✓
Information Analyzer	Data Quality Validation and Monitoring		✓		✓
Data Quality Console	Data Quality Exception Management		✓		✓
Information Services Director	SOA Deployment		✓	✓	✓
FastTrack	Data Specification Mapping			✓	✓
DataStage	Extraction, transformation, load (ETL)			✓	✓
DataStage Balanced Optimization	Extract, load, transform (ELT) and other workload balancing modalities			✓	✓
Change Data Delivery	Change Data Delivery			✓	✓
Data Click	Simple data integration tasks such as offloading data to a data mart			✓	✓

Figure 1-2 Information Server product modules and capabilities

This section describes each of the product modules with their general functionality, which gives you a broad perspective of the range of Information Server capabilities. This Redbooks publication does not attempt to provide a full explanation of how these capabilities are used to design, construct, and populate a data warehouse, but rather how to maximize the value of your warehouse after it is constructed. In “Related publications” on page 171, we provide a list of other resources to help you with warehouse design, construction, and population by using the broad set of Information Server capabilities.

1.2.1 InfoSphere Blueprint Director

IBM InfoSphere Blueprint Director is aimed at the Information Architect designing solution architectures for information-intensive projects. Blueprint Director goes beyond traditional diagramming tools and white boards that are typically used to map an information solution. It enables a unique paradigm for information integration projects where teams can define, document, and manage end-to-end information flows. Blueprint Director enables the predictability and success of information projects by linking blueprints to reference architectures, reusable best practices and methodologies, and business and technical artifacts. With several useful features, such as built-in templates and methodology, the milestone feature, and deep integration into the IBM Rational Team Concert™ platform, the evolution of architecture blueprints over their initial project lifecycle and beyond is well-supported from vision to execution and completion.

InfoSphere Blueprint Director helps you to jump-start your project by using blueprints. It is a graphical design tool that is used primarily for creating high-level plans for an InfoSphere Information Server based initiative. Such initiatives can be in information governance, information integration, business intelligence (BI), or any other information-based project. To make the task simpler, InfoSphere Blueprint Director comes bundled with several, ready-to-use, and project-type-based content templates that can be easily customized to fit the project. Alternatively, a blueprint can be created as needed but is discouraged. A blueprint for building a new data warehouse is included, as is shown in Figure 1-3.

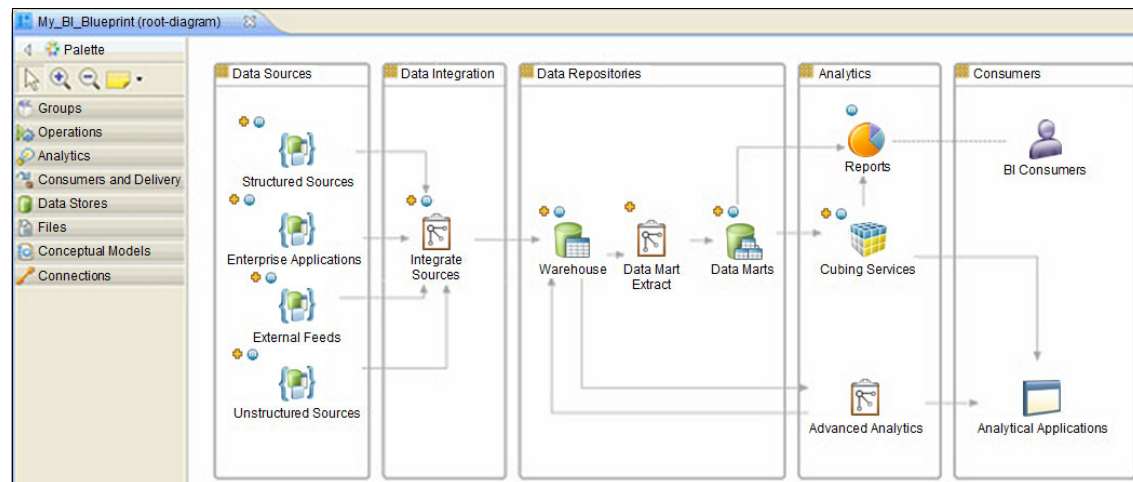


Figure 1-3 Blueprint Director's Business-Driven BI template

InfoSphere Blueprint Director has a design canvas onto which standard graphical objects that represent processes, tasks, or anything else, are dragged. Objects can be connected one to the other, which implies a sequential order to the events or dependencies between them. Each graphical object has a label that indicates its purpose. However, the object can optionally be linked to content that was produced and published in IBM Rational Method Composer. When a single object represents several tasks or processes, the object can drill down to create or link to a more detailed layer of the blueprint diagram. This way, the final blueprint is likely to contain several hierarchical levels of processes (and subprocesses). The hierarchical blueprint diagram, combined with the methods (text descriptions), forms the basis of the project plan as built from top-to-bottom (high-to-low level).

InfoSphere Blueprint Director is a unique component among the InfoSphere Information Server product modules and components. It is a stand-alone, Eclipse-based, client-only application that does not have any dependencies on the InfoSphere Information Server infrastructure for persistence, authentication, or other shared services. This component provides useful flexibility for planning the project at an early stage before all of the infrastructure is ready and available.

1.2.2 InfoSphere Discovery

InfoSphere Discovery is an automated data relationship discovery solution. It helps organizations gain an understanding of data content, data relationships, and data transformations; to discover business objects; and to identify sensitive data within and across multiple heterogeneous data stores. The automated results derived from InfoSphere Discovery are actionable, accurate, and easy to generate, especially when compared to manual (non-automated) data analysis approaches that many organizations still use today.

InfoSphere Discovery works by automatically analyzing data sources and generating hypotheses about the data. In the context of an information integration project, this process provides an understanding of data and their relationships. It can be used for governance or to help in source-to-target mapping as a planning aid to data integration specification.

There are two work flows you can take through the InfoSphere Discovery product. Both workflows start with profiling, primary-foreign key discovery, and understanding individual systems, and can be used for data archiving, test data management, and sensitive data discovery. However, both paths then proceed to use the resulting information for the following kinds of data-intensive projects:

- ▶ Unified Schema Builder

A complete workbench for the analysis of multiple data sources and for prototyping the combination of those sources into a consolidated, unified target, such as an MDM hub, a new application, or an enterprise data warehouse. Unified Schema Builder helps build unified data table schemas by accounting for known critical data elements and proposing statistic-based matching and conflict resolution rules before you write ETL code or configure an MDM hub.

- ▶ Transformation Analyzer

This workflow is used when two existing systems are mapped together to facilitate data migration, consolidation, or integration and delivers the most advanced cross-source transformation discovery capabilities that are available in the industry. Transformation Analyzer automates the discovery of complex cross-source transformations and business rules (substrings, concatenations, cross-references, aggregations, case statements, arithmetic equations, and so on) between two structured data sets. It also identifies the specific data anomalies that violate the discovered rules for ongoing audit and remediation.

The Transformation Analyzer component discovers the de facto business rules that relate two data sources in your existing distributed data landscape and then outputs actionable transformation logic that can be used by IBM InfoSphere DataStage to move data from a source to a target, as shown in Figure 1-4 on page 11.

Transformation	
CASE WHEN AGE <=25 THEN Youthful_Driver = 'Y' ELSE 'N' END	
Hit Rate = 90%	
Application A	Application B
AGE	Youthful_Driver
17	Y
24	Y
55	N
28	N
40	N
33	N
Exception 83	Y
29	N
36	N
42	N

Figure 1-4 InfoSphere Discovery's Transformation Analyzer

1.2.3 InfoSphere Metadata Workbench

InfoSphere Metadata Workbench plays an integral part in Information Server by providing data lineage and impact analysis reports on various assets that are stored in its metadata repository, including business intelligence reports, databases, and established data assets. One of the core features in Metadata Workbench is its built-in support to automatically deduce linkages among metadata that Information Server components create or import.

By using these functions, business users can get answers to questions such as “What are the sources for this report that I'm looking at?” and “I do not trust the data in this report. Which transformations were applied to the data on which this report is based?” The answers to such questions are critical to the compliance laws currently in place in many industries, including finance, banking, and manufacturing. Similarly, IT analysts can get answers to questions regarding impact of change, including “Which assets and processes would be affected if I change the constraints on this column?”

Upon delivery, Metadata Workbench provides various lineage reports about the following assets:

- ▶ Imported databases (tables, columns) and table-structured files and their fields
- ▶ DataStage and QualityStage ETL jobs and contained stages
- ▶ Imported BI reports and report fields

In Figure 1-5, we show an example data transformation and movement process in which data from two sources (RUT_UT and RUT_EMEA) are moved by two separate DataStage ETL jobs into a staging database (CRT_STAGING). From there, another DataStage ETL job moves the data into a data warehouse called ENT_DWH. Finally, a fourth DataStage ETL job (TO_MART) builds out a data mart (MART) that three IBM Cognos® BI reports are then using.

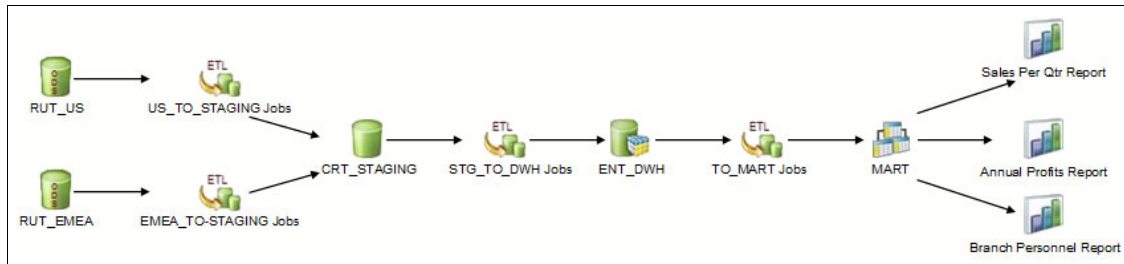


Figure 1-5 Sample data flow topology in enterprise

In this particular data flow, Metadata Workbench supports all participating components (jobs, database tables, and BI reports). Therefore, a data lineage report that was started on one of the BI reports, such as the SalesPerQtr report, successfully reports the entire flow all the way back to the two data sources.

For the example, all metadata that is describing how the data moves from one place to the next is stored in the metadata repository, which enables Metadata Workbench to build out the complete lineage report by deducing the chain of source-to-target relationships.

1.2.4 InfoSphere Data Architect and IBM Industry Data Models

IBM InfoSphere Data Architect is an enterprise data modeling and integration design tool. You can use it to discover, model, visualize, relate, and standardize diverse and distributed data assets, including dimensional models.

From a top-down approach, you can use InfoSphere Data Architect to design a logical model and automatically generate a physical data model from the logical source. Data definition language (DDL) scripts can be generated from the data model to create a database schema based on the design of the data model.

Alternatively, InfoSphere Data Architect can connect to the RDBMS and instantiate the database schema directly from the InfoSphere Data Architect physical data model. This generation facility works both ways in that you also can reverse engineer an existing database into an InfoSphere Data Architect data model for modification, reuse, versioning, and so on.

Rather than designing models from the beginning, you can purchase one of the IBM Industry Models in a format that is usable by InfoSphere Data Architect. In this manner, you can jump start the database design phase of the project and benefit from data modeling expertise in the specific industry. Standard practice is to scope the industry standard logical model to fit your requirements and build an appropriate data model that combines industry standards with customer specifics. An added advantage of the IBM Industry Models package for InfoSphere Data Architect is that it includes an Industry standard glossary model. This model populates the InfoSphere Business Glossary, complete with relationships (assigned assets) to the InfoSphere Data Architect logical model and generated physical data model, as shown in Figure 1-6 on page 14.

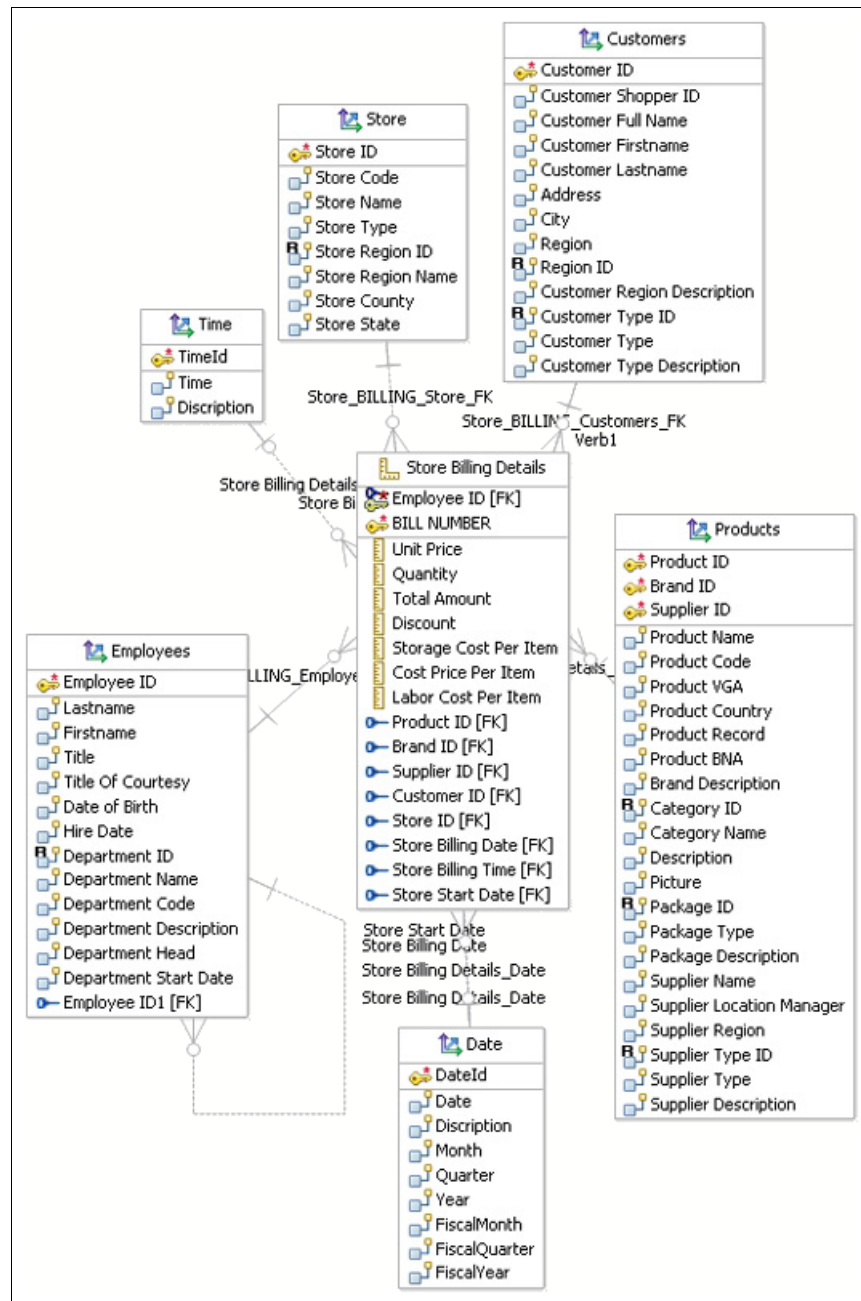


Figure 1-6 A star schema dimensional model in InfoSphere Data Architect

1.2.5 InfoSphere Business Glossary

IBM InfoSphere Business Glossary is an interactive, web-based tool that enables users to create, manage, and share controlled vocabulary and information governance controls in a repository called a business glossary. The vocabulary and governance controls define business semantics and enable business leaders and IT professionals to manage enterprise-wide information according to defined regulatory or operational business requirements. IBM InfoSphere Business Glossary Anywhere, its companion module, augments InfoSphere Business Glossary with more ease-of-use and extensibility features.

Business Glossary, Business Glossary browser, and Business Glossary Anywhere support complex enterprise development environments with a unique set of the following capabilities:

- Manage business terms and categories

Business Glossary provides a dedicated, web-based user interface for creating, managing, and sharing a controlled vocabulary, including batch editing capabilities. Terms represent the major information concepts in your enterprise and categories are used to organize into hierarchies.

- Manage stewardship

Stewards are people or organizations with the responsibility for a given information asset. By using Business Glossary, administrators can import steward profiles from external sources, generate and edit profiles in the web interface, and create relationships of responsibility between stewards and business terms or any of the artifacts that are managed by Information Server.

- Customize and extend

The needs around business metadata tend to differ from one enterprise to the next. For this reason, there is no “one-size-fits-all” meta-model. In addition to the ability to customize the entry page to the application, administrators can extend the application with custom attributes on business categories and business terms.

- Collaborate

It is not enough to simply document business metadata. This information is active in the enterprise with open access to all members of business and development teams. IBM InfoSphere Business Glossary provides a collaborative environment in which users can evolve this important information asset as the business changes and adapts to market conditions, shifting customer needs and competitive threats.

- Contextual search and visibility business term definitions

Business Glossary Anywhere is an application independent search window that can be called from any application (such as Microsoft Excel, data modeling tools, reporting applications, and Microsoft Word) that provides instant access to Business Glossary terms, taxonomies, and stewards.

- Simply Browse

Business Glossary browser is an intuitive, read-only web-based interface that requires no training to use. Business users can search and explore the common controlled vocabulary and relationships, identify stewards that are responsible for assets and provide direct feedback.

1.2.6 InfoSphere QualityStage

IBM InfoSphere QualityStage provides data cleansing capabilities to help ensure quality and consistency by standardizing, validating, matching, and merging information to create comprehensive and authoritative information for multiple uses, including data warehousing.

InfoSphere QualityStage uses predefined, customizable rules to prepare complex information about your business entities for transactional, operational, and analytic applications in batch, real time, or as a web service. Information is extracted from the source system, measured, cleansed, enriched, consolidated, and loaded into the target system.

Your organization can use InfoSphere QualityStage to complete the following data quality tasks:

- Data investigation

You use InfoSphere QualityStage to understand the nature and extent of data anomalies and enable more effective data cleansing and matching.

Investigation capabilities give your organization complete visibility into the condition of data at any moment. Data problems in established sources can be identified and corrected before they corrupt new systems.

Investigation uncovers potential anomalies, metadata discrepancies, and undocumented business practices. Invalid values and default values are identified so that they can be corrected or added to fields that are proposed as matching criteria.

► Data standardization

Creating a standardized view of your data enables your organization to maintain accurate views of key entities such as customer, partner, or product. Data from multiple systems is reformatted to ensure that data has the correct, specified content and format. Standardization rules are used to create a consistent representation of the data.

With data standardization, IBM InfoSphere QualityStage Standardization Rules Designer provides capabilities to enhance standardization rule sets. You can add and modify classifications, lookup tables, and rules. You also can enhance information by completing global address cleansing, validation and certification, and geolocation, which is used for spatial information management. Longitude and latitude are added to location data to improve location-based services.

► Data matching

The matching process ensures that the information that runs your enterprise is based on your business results, reflect the facts in the real world, and provide an accurate view of data across your enterprise.

Powerful matching capabilities detect duplicates and relationships, even in the absence of unique identifiers or other data values. A statistical matching engine assesses the probability that two or more sets of data values refer to the same business entity. After a match is confirmed, InfoSphere QualityStage constructs linking keys so that users can complete a transaction or load a target system with quality, accurate data.

► Data survivorship

Survivorship ensures that you are building the best available view of related information. Business and mapping rules are implemented to create the necessary output structures for the target application. Fields that do not conform to load standards are identified and filtered so that only the best representation of the match data is loaded into the master data record.

Missing values in one record are supplied with values from other records of the same entity. Missing values also can be populated with values from corresponding records that were identified as a group in the matching stage.

1.2.7 InfoSphere Information Analyzer

InfoSphere Information Analyzer provides capabilities to profile and analyze data to deliver trusted information to your organization.

Data quality specialists use InfoSphere Information Analyzer to scan samples and full volumes of data to determine their quality and structure. This analysis helps to discover the inputs to your data integration project, ranging from individual fields to high-level data entities. Information analysis enables your organization to correct problems with structure or validity before they affect your data integration project.

After data is analyzed, data quality specialists create data quality rules to assess and monitor heterogeneous data sources for trends, patterns, and exception conditions. These rules help to uncover data quality issues and help your organization to align data quality metrics throughout the project lifecycle. Business analysts can use these metrics to create quality reports that track and monitor the quality of data over time. Business analysts can then use IBM InfoSphere Data Quality Console to track and browse exceptions that are generated by InfoSphere Information Analyzer.

Understanding where data originates, which data stores it lands in, and how the data changes over time is important to develop data lineage, which is a foundation of data governance. InfoSphere Information Analyzer shares lineage information with the rest of Information Server by storing it in the metadata repository. Other Information Server components can access lineage information directly to simplify the collection and management of metadata across your organization.

1.2.8 InfoSphere Data Quality Console

IBM InfoSphere Data Quality Console is a browser-based interface that you can use to track and browse exceptions that are generated by InfoSphere Information Server products and components.

The data quality console provides a unified view of data quality across products and components. For example, you can use the data quality console to assess how the data quality of a particular table is affected by multiple Information Server components. If you identify problems with data quality, you can collaborate with other users to resolve the problems.

In the data quality console, you can browse exceptions that are generated by the following products and components:

- ▶ InfoSphere Discovery
- ▶ InfoSphere Information Analyzer

1.2.9 InfoSphere Information Services Director

IBM InfoSphere Information Services Director provides a unified and consistent way to publish and manage shared information services in a service-oriented architecture (SOA). By using InfoSphere Information Services Director, information specialists can design and deploy reusable information integration tasks including data cleansing, data transformation, and data federation services.

For example, consider the need to match customer data entry against a “golden” customer record that is stored in the warehouse. Information Server's QualityStage module can be used to construct a lookup/match rule against the warehouse. Not only can this QualityStage job be used in batch mode, but it can be shown as a service. Information Services Director is used to show this QualityStage job as a service, which allows the front-end data entry application to match the customer record at point of entry and helps ensure downstream data quality.

1.2.10 InfoSphere FastTrack

IBM InfoSphere FastTrack streamlines collaboration between business analysts, data modelers, and developers by capturing and defining business requirements in a common, familiar format and then transforming that business logic directly into DataStage ETL jobs.

The completed mapping specification can be output in several formats, including an annotated InfoSphere DataStage job that is generated directly by InfoSphere FastTrack. This format is useful for the InfoSphere DataStage developer because the specification is delivered in a manner in which the developer is familiar. In addition, this delivery format provides a job template that can be used as the basis for creating a job, including design artifacts that can be copied to the new job as is.

As shown in Figure 1-7 on page 20, the use of InfoSphere FastTrack for mapping specification documentation includes the following advantages:

- ▶ Centrally stored and managed specifications
- ▶ Simple drag-and-drop functionality for specifying source and target columns

- Accuracy of source and target column names (that exist in the repository) with assured correct spelling
- Discovery of mappings, joins, and lookups assistance, based on published data profiling results, name recognition, and business-term assignment








































	Source		Target		Transformation	
	Field	Busi	Field	Rule	Function	
1	 CHECKING.STATE		 CUSTOMER.STATE			
2	 CHECKING.SS_NUM		 CUSTOMER.TAX_ID			
3			 CUSTOMER.GENDER		setNull()	
4	 CHECKING.ADDR1		 CUSTOMER.ADDR1			
5	 CHECKING.CITY		 CUSTOMER.CITY			
6			 CUSTOMER.YEARS_CLIENT		setNull()	
7	 CHECKING.CUSTOMER_ID		 CUSTOMER.CUSTOMER_ID			
8	 CHECKING.ADDR2		 CUSTOMER.ADDR2			
9	 CHECKING.NAME		 CUSTOMER.NAME			
10			 CUSTOMER.LEVEL		setNull()	
11	 CHECKING.ACCOUNT_BALANC		 CUSTOMER.ACCOUNT_BALAN			
12			 CUSTOMER.ONLINE_ACCESS		setNull()	
13	 CHECKING.ZIP		 CUSTOMER.ZIP			

Figure 1-7 A mapping specification in InfoSphere FastTrack

1.2.11 InfoSphere DataStage

InfoSphere DataStage is a data integration tool that enables users to move and transform data between operational, transactional, and analytical target systems.

Data transformation and movement is the process by which source data is selected, converted, and mapped to the format required by target systems. The process manipulates data to bring it into compliance with business, domain, and integrity rules, and with other data in the target environment.

InfoSphere DataStage provides direct connectivity to enterprise applications as sources or targets, ensuring that the most relevant, complete, and accurate data is integrated into your data integration project.

By using the parallel processing capabilities of multiprocessor hardware platforms, InfoSphere DataStage enables your organization to solve large-scale business problems. Large volumes of data can be processed in batch, in real time, or as a web service, depending on the needs of your project.

Data integration specialists can use the hundreds of prebuilt transformation functions to accelerate development time and simplify the process of data transformation. Transformation functions can be modified and reused, which decreases the overall cost of development and increases the effectiveness in building, deploying, and managing your data integration infrastructure.

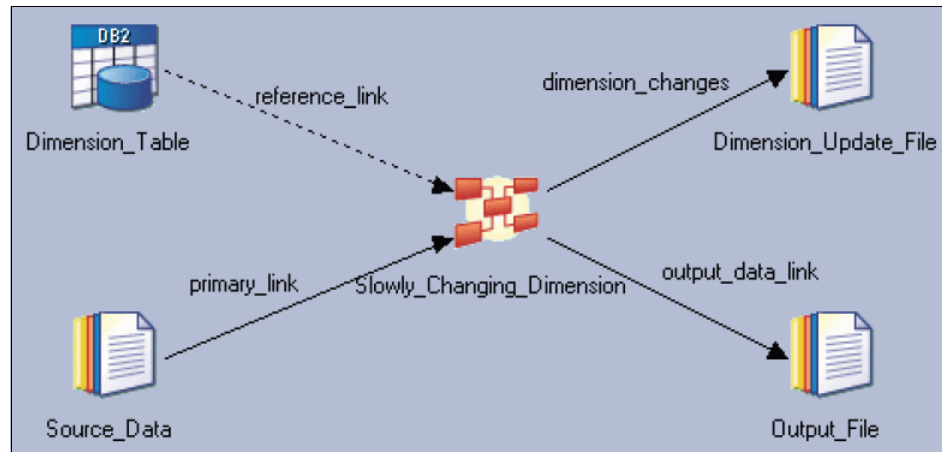


Figure 1-8 DataStage job as shown in InfoSphere DataStage and QualityStage Designer

1.2.12 InfoSphere DataStage Balanced Optimization

An IBM InfoSphere DataStage job consists of individual stages that are linked together and describe the flow of data from a data source to a data target. Balanced Optimization allows you to maximize job performance and optimize resources usage, which enables you to balance the workload across source and target systems. This allows Information Server to support not only the Extract-Transform-Load paradigm, but alternatives such as Extract-Load-Transform, where transformation tasks are performed on the target system, such as an IBM PureData™ for Analytics data warehousing appliance.

Balanced Optimization helps to improve the performance of your InfoSphere DataStage job designs that use connectors to read or write source data. You design your job and then use Balanced Optimization to redesign the job automatically to your stated preferences.

For example, you can maximize performance by minimizing the amount of input and output (I/O) that are used, and by balancing the processing against source, intermediate, and target environments. You can then examine the new optimized job design and save it as a new job. Your root job design remains unchanged.

You can use the Balanced Optimization features of InfoSphere DataStage to push sets of data integration processing and related data I/O into database management systems (such as an IBM PureData System for Analytics warehousing appliance) or into a Hadoop cluster.

1.2.13 InfoSphere Change Data Delivery

InfoSphere Change Data Delivery captures and delivers data to your warehouse in real time. This helps enable businesses to gain immediate awareness of market landscape and operational statistics to streamline business processes, improve customer service, and capture time-sensitive opportunities.

InfoSphere Change Data Delivery offers rapid and timely delivery of data changes for InfoSphere DataStage extract, transform, and load (ETL) processes; InfoSphere QualityStage data quality processes; and PureData for Analytics warehouse appliances, to help ensure that these systems have updated and timely data to make informed business decisions.

1.2.14 InfoSphere Data Click

InfoSphere Data Click is an exciting new capability that helps novices and business users retrieve data and provision systems easily in only a few clicks. InfoSphere Data Click helps improve the timeliness of InfoSphere PureData for Analytics (Netezza®) data warehouse environments by delivering data in real time.

IBM InfoSphere Data Click simplifies data movement and eases data placement. You can use InfoSphere Data Click to offload warehouse databases or offload select schemas and tables within warehouse databases.

You can use InfoSphere Data Click to retrieve data and work on that data in a test environment. You can move data from an operational database to a private sandbox. You can isolate the data to experiment with data transformations or you can create reports from subsets of the data. By isolating and analyzing the data in a test environment, you do not jeopardize the integrity of the business information in the production environment.

InfoSphere Data Click relies on IBM InfoSphere DataStage and also can use IBM InfoSphere Change Data Capture to provide efficient extraction, high throughput, and with minimum risk to your production system, as shown in Figure 1-9 on page 23.

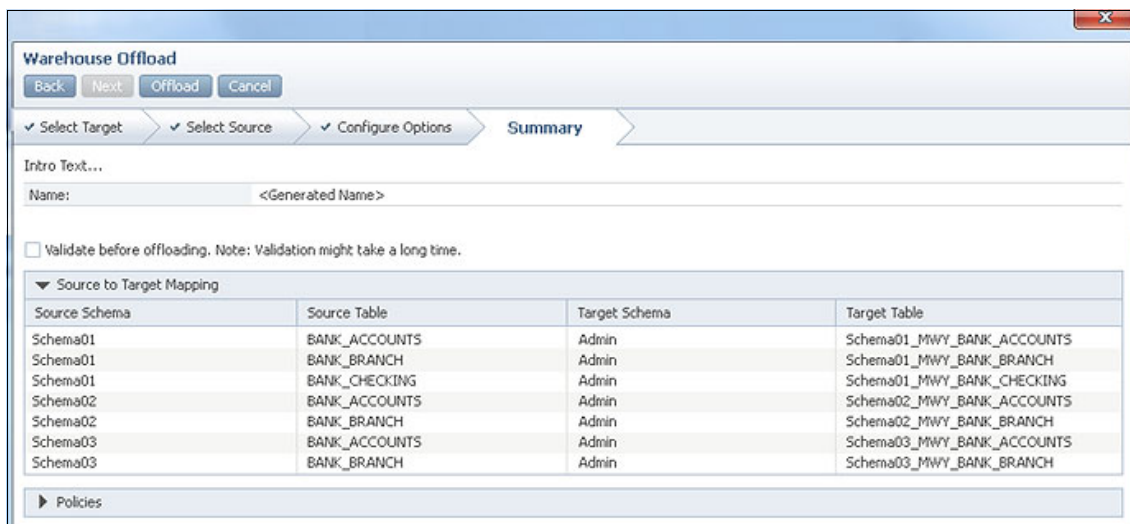


Figure 1-9 Warehouse offload in only a few clicks with InfoSphere Data Click



Using Information Server to design and implement a Data Warehouse

In this chapter, we provide a brief overview of how Information Server's capabilities can be used together to plan, design, and implement a data warehouse. It follows a basic plan-analyze-design-develop-deploy flow, intended to touch upon key aspects of a warehouse development effort. It is not intended to be a complete process, but sufficient to introduce how your warehousing development team might use Information Server's capabilities in combination.

The chapter includes the following topics:

- ▶ How the capabilities fit together
- ▶ Method and proven practices: Business-driven BI development
- ▶ Phases
- ▶ Information Server components by Phase

2.1 How the capabilities fit together

Figure 2-1 shows a simplified view of how the major components of Information Server work together to create a unified data integration solution. A common metadata foundation enables different types of users to create and manage metadata by using tools that are optimized for their roles. This focus on individualized tooling makes it easier to collaborate across roles, which makes your team more efficient and reduces the risk of your project.

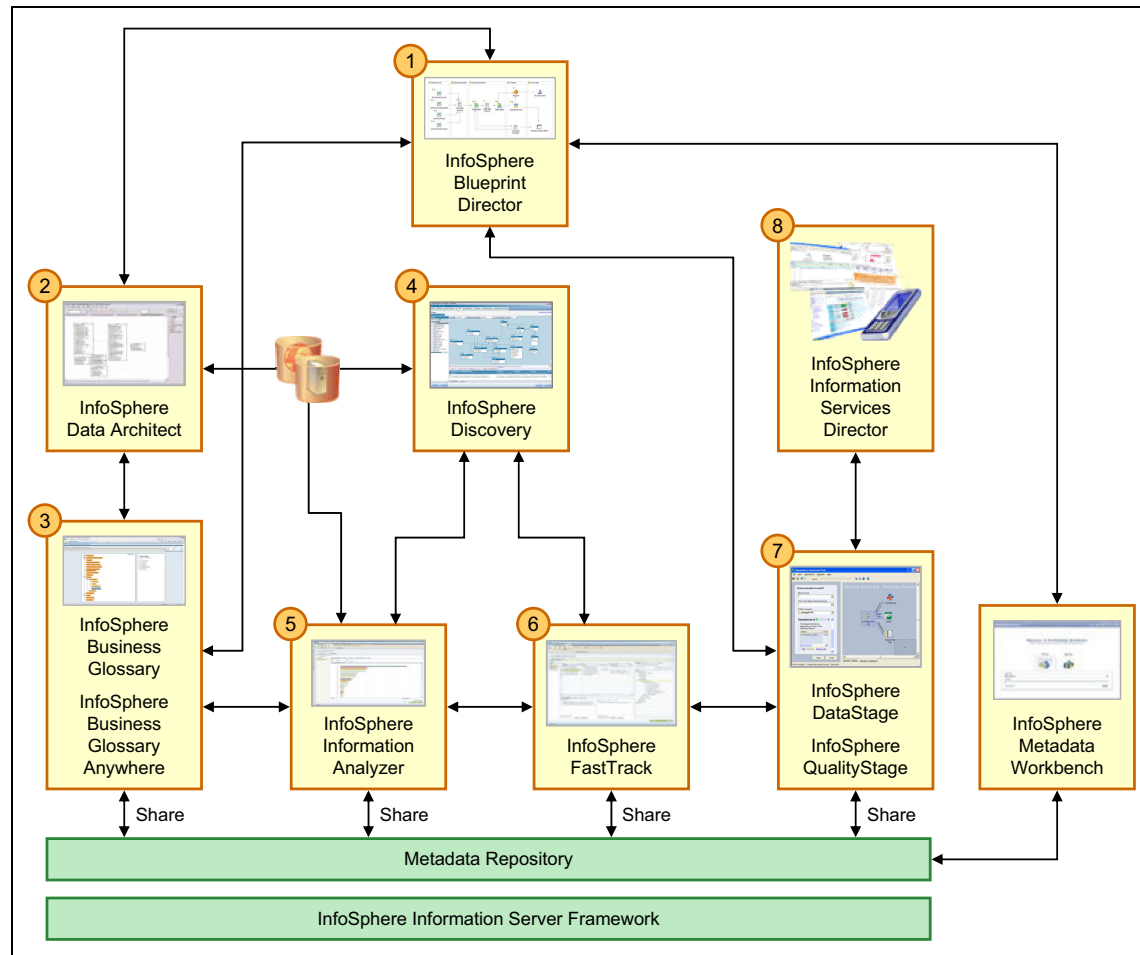


Figure 2-1 How Information Server's Components Work Together

The following components are shown in Figure 2-1 on page 26:

1. Enterprise Architects use InfoSphere Blueprint Director to plan and manage your project vision.
2. After a blueprint of your information project exists, Data Architects can use InfoSphere Data Architect to discover the structure of your organization's databases, reverse engineer and visualize physical data models from existing source systems, derive or define new logical data models for key information assets, and create physical data models for target systems, including dimensional structures in the warehouse.
3. This (meta)data can be input to InfoSphere Business Glossary, where Business Analysts and Data Analysts define and establish a common understanding of business concepts.
4. Data Analysts also can use InfoSphere Discovery to automate the identification and definition of data relationships, feeding that information to InfoSphere Information Analyzer and InfoSphere FastTrack.
5. Data Quality Specialists use InfoSphere Information Analyzer to design, develop, and manage data quality rules for your organization's data to ensure data quality. As your organization's data evolves, these rules can be modified in real time so that trusted information is delivered to InfoSphere Business Glossary, InfoSphere FastTrack, InfoSphere DataStage and QualityStage, and other InfoSphere Server components.
6. Data Analysts can use InfoSphere FastTrack to create mapping specifications that translate business requirements into business applications. Data Integration Specialists can use these specifications to generate jobs that become the starting point for complex data transformation in InfoSphere DataStage and QualityStage.
7. By using the InfoSphere DataStage and QualityStage Designer, Data Integration Specialists develop jobs that extract, transform, load, and check the quality of data. These jobs are deployed on InfoSphere DataStage, while data quality jobs are deployed by using InfoSphere QualityStage. DataStage Balanced Optimization also can be used to optimize performance of integration jobs, which allows you to balance and optimize the usage of your source, target, and Information Server engine resources (implementing, for example, alternative modalities such as ELT). This capability is particularly useful for getting the highest possible performance in IBM PureData for Analytics warehousing appliances (which is powered by Netezza).
8. SOA Architects use InfoSphere Information Services Director to deploy integration tasks (such as matching and lookup) from the suite components as consistent, reusable information services.

9. InfoSphere Metadata Workbench provides end-to-end data flow reporting and impact analysis of your organization's data assets. Business Analysts, Data Analysts, Data Integration Specialists, and other users interact with this component to explore and manage the assets that are produced and used by InfoSphere Information Server. InfoSphere Metadata Workbench enables users to understand and manage the flow of data through your enterprise, and discover and analyze relationships between information assets in the InfoSphere Information Server metadata repository. You use InfoSphere Metadata Asset Manager to import technical information into the metadata repository, such as BI reports, logical models, physical schemas, and InfoSphere DataStage and QualityStage jobs.

2.2 Method and proven practices: Business-driven BI development

The design, creation, and evolution of a data warehouse is a complex software development undertaking. Like any software development project, you can reduce risk and improve the end quality of your warehouse if you define an appropriate set of processes for your team to follow, a method for designing and developing your warehouse, and a set of best practices for usage of your particular tools.

The InfoSphere Blueprint Director includes a template for Business-Driven BI Development, which provides diagrams, methods, and best practices for creation of a data warehouse. It not only provides a rich visual environment for understanding the landscape of your data warehousing solution, but integrated guidance based on best practices. For example, you can see which activities are required to deploy a business intelligence solution, which roles perform these activities, which tools they might use, and what the recommended solution landscapes are. This template (and others) can be used to define new blueprints based on best practices to reduce risk and the time to completion of your solution. Figure 2-2 on page 29 shows Blueprint Director with the Business-Driven BI Development template loaded.

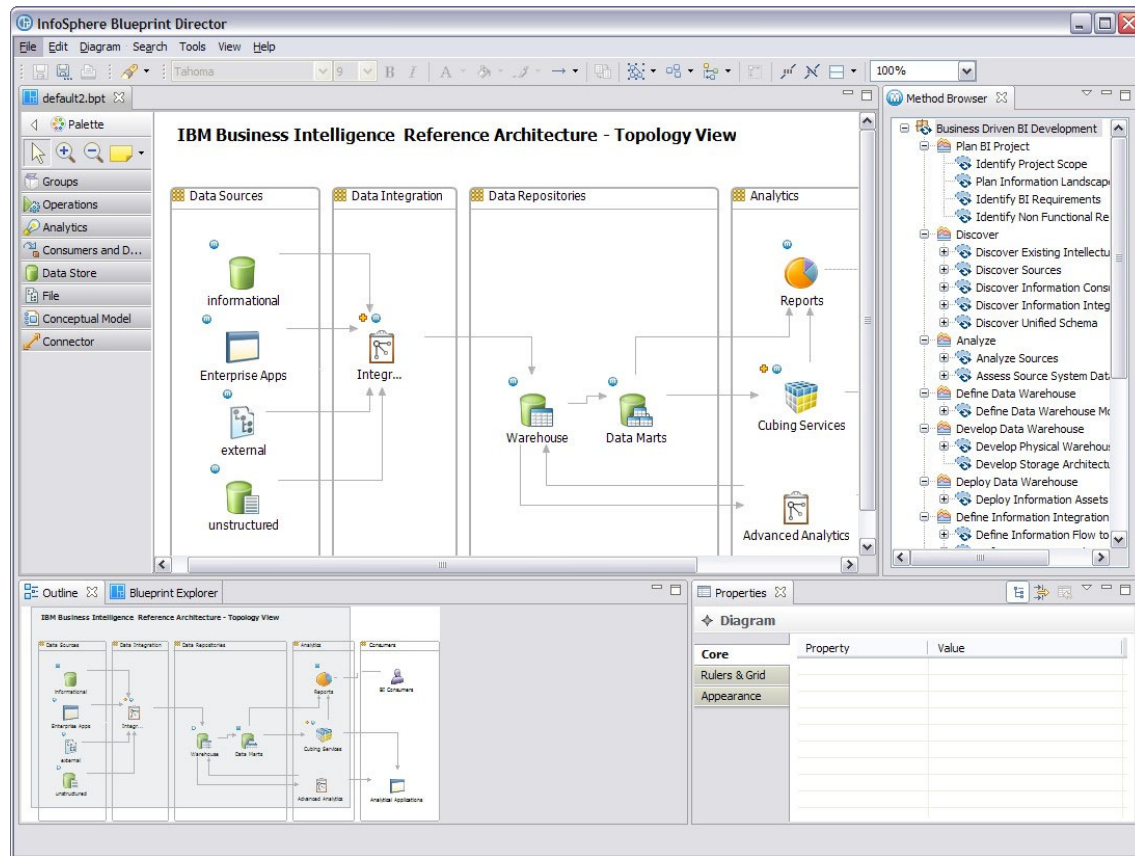


Figure 2-2 Blueprint Director's Business-Driven BI Development template

The remainder of this section provides a high-level overview of the major tasks of data warehouse design and development. Therefore, it does not provide an in-depth description of the much richer Business-Driven BI Development method and practices which come with Blueprint Director. We encourage you to explore Blueprint Director and the Business-Driven BI Development template to improve the productivity of your team.

2.3 Phases

Information-intensive projects such as data warehousing design and construction are often developed as a set of phases, with work often going on in parallel across various members of the team, and work on one phase overlapping with work on another. Figure 2-3 shows a high-level workflow from phase to phase of a project, starting with Discovery and ending with a deployed system that is managed, optimized, and governed.

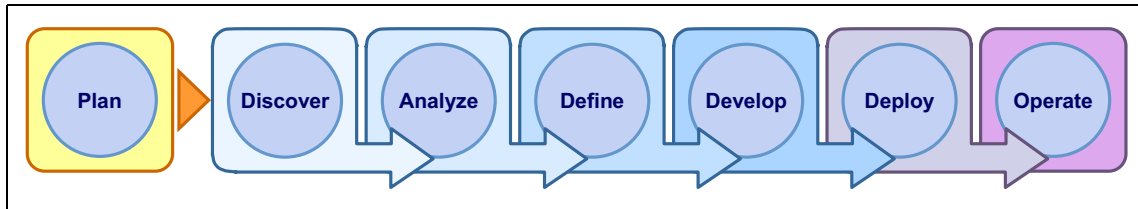


Figure 2-3 Example phases of an information-intensive project

Teams can develop a warehouse in iterations, where a set of use cases (or stakeholder requirements) is taken through the development phases to a functioning system. The system might not yet be deployed for production use until a sufficient number of use cases have been implemented and the quality and performance of the system meet expectations.

Further, any long-lived system goes through further iterations where new functionality is added. Part 2 of this Redbooks publication describes several such system enhancements, such as incorporation of big data and analytic models into the warehouse environment. By using the metadata-rich Information Server environment, these and other enhancements can be performed with lower risk and higher efficiency because much of the work from previous iterations can be directly reused in the new effort.

2.4 Information Server components by Phase

A common metadata foundation enables different types of users to create and manage metadata by using tools that are optimized for their roles. This focus on individualized tools makes it easier to collaborate across roles. For example, data analysts can use analysis and reporting functions to generate integration specifications and business rules that they can monitor over time. Subject matter experts can use web-based tools to define, annotate, and report on fields of business data.

The common metadata foundation allows for smoother transition of work from one user to the next, from one phase to the next, and from one iteration to the next.

2.4.1 Plan

In the planning phase, you perform the following tasks:

- ▶ Capture and document information requirements (InfoSphere Data Architect, InfoSphere Business Glossary)
- ▶ Use predefined industry data models (IBM Industry Data Models)
- ▶ Capture and document functional and non-functional requirements
- ▶ Sketch a solution outline (InfoSphere Blueprint Director)
- ▶ Understand and document the business concepts of interest to the project (InfoSphere Business Glossary)

InfoSphere Information Server includes capabilities that you can use to manage the structure of your information project from initial sketches to delivery. By collaborating on blueprints, your team can connect the business vision for your project with corresponding business and technical artifacts. To enhance your blueprint, you can create a business glossary to develop and share a common vocabulary between your business and IT users. The terms that you create in your business glossary establish a common understanding of business concepts, which further improves communication and efficiency.

Requirements can be documented in various forms. Data requirements, for example, can be captured at the conceptual (or business) level by using InfoSphere Business Glossary, at the logical data model level by using InfoSphere Data Architect, or as use cases or non-functional requirements (by using, for example, tools such as IBM Rational Requirements Composer, not a part of InfoSphere Information Server).

To jump start projects and to help ensure industry best practices, an organization can import information from IBM Industry Data Models (via InfoSphere Data Architect), which includes a glossary, logical, and physical data model. The glossary models contains thousands of industry-standard terms that can be used to pre-populate IBM InfoSphere Business Glossary. Organizations can modify and extend the IBM Industry Data Models to match their particular business

2.4.2 Discover

In the Discover phase, you perform the following tasks:

- ▶ Identify information sources (InfoSphere Data Architect, InfoSphere Discovery)
- ▶ Discover key details of data sources (InfoSphere Discovery)

Information sources are a key aspect of your warehouse. Enterprise logical models and physical data models of existing systems can provide insight into which system (or systems) provides the needed information, or highlight gaps; that is, information that is needed to satisfy system requirements yet whose source is not yet identified.

InfoSphere Information Server can help you automatically discover the structure of your data, and then analyze the meaning, relationships, and lineage of that information. By using a unified, common metadata repository that is shared across the entire suite, InfoSphere Information Server provides insight into the source, usage, and evolution of a specific piece of data.

After the data models are defined, business context is applied, and information sources are identified, the analyst runs a data discovery process against the source systems that are used to populate the new target data model. During the discovery process, the analyst can identify key relationships, transformation rules, and business objects that can enhance the data model, if these business objects were not previously defined by the IBM Industry Data Models.

2.4.3 Analyze

In the Analyze phase, you perform the following tasks:

- ▶ Analyze information sources (InfoSphere Information Analyzer)
- ▶ Determine data quality, redundancy gaps, usage, and trust (InfoSphere Information Analyzer)

From the discovered information, the analyst can expand the work to focus on data quality assessment and ensure that anomalies are documented, reference tables are created, and data quality rules are defined. The analyst can link data content to established glossary terms to ensure appropriate context and data lineage, deliver analytical results and inferred models to developers, and test and deploy the data quality rules. When the data quality rules are applied to data from source systems, exceptions to the rules can be tracked in IBM InfoSphere Data Quality Console.

2.4.4 Define

In the Define phase, you perform the following tasks:

- ▶ Define and refine enterprise logical model for data warehouses and data marts (InfoSphere Data Architect)
- ▶ Define and refine business glossary (InfoSphere Business Glossary)
- ▶ Derive physical data models from logical data models (InfoSphere Data Architect)
- ▶ Define source-to-target mapping specifications (InfoSphere FastTrack)

In the Plan phase, initial information requirements were captured in the form of logical data models and the business glossary. In the Define phase, those requirements are elaborated and transformed into specifications for the target data warehouse and marts. Use of IBM Industry Data Models helps ensure adherence to industry standards and helps speed this challenging aspect of warehouse system development.

The logical models for the warehouse and marts are transformed into physical models by using InfoSphere Data Architect's forward engineering capability. These physical models form the target system definition for ETL jobs.

The analyst is now ready to create the mapping specifications, which are input into the ETL jobs for the new application. By using the business context, discovered information, and data quality assessment results, the analyst defines the specific transformation rules necessary to convert the data sources into the correct format for the data warehouse's physical data model target. During this process, the analyst not only defines the specific business transformation rules, but can define the direct relationship between the business terms and their representation in physical structures. These relationships can then be published to IBM InfoSphere Business Glossary for use and to enable better understanding of the asset relationships.

2.4.5 Develop

In the Develop phase, you perform the following tasks:

- ▶ Develop data movement and transformation (InfoSphere DataStage and QualityStage Designer)
- ▶ Define data standardization, matching, and survivorship rules (IBM InfoSphere QualityStage Standardization Rules Designer and InfoSphere QualityStage Match Designer)

The business specification in InfoSphere FastTrack now serves as historical documentation and direct input into the generation of the IBM InfoSphere DataStage ETL jobs. The defined business rules are directly included in the ETL job as code or annotated To Do tasks for the developer to complete. InfoSphere DataStage and QualityStage Designer provides a rich canvas for designing data transformation and movement jobs, including high-performance parallel jobs which can fully use multi-core, multi-processor environments.

In this phase, QualityStage Standardization Rules Designer is used to define rules to ensure data conforms to your organization's standardization requirements through cleansing, validation, certification and geolocation. For example, geolocation rules can augment location data with longitude and latitude information to enable location-based services and analysis.

QualityStage Match Designer is used to detect duplicates, and to find when two or more sets of data values refer to the same entity.

After it is deployed and operational, exceptions to matching and standardization rules can be tracked with IBM InfoSphere Data Quality Console.

2.4.6 Deploy

In the Deploy phase, you perform the following tasks:

- ▶ Generate Data Definition Language (DDL) to construct target warehouse structures (InfoSphere Data Architect)
- ▶ Deploy DataStage or QualityStage jobs as SOA services (InfoSphere Information Services Director)
- ▶ Deploy and run ETL jobs to populate warehouse and marts

The target physical data models can be transformed into DDL. The DDL can then be used by the DBA to create the target tables in the warehouse and marts.

When the InfoSphere DataStage job is ready, the developer also can decide to deploy the same batch process as an SOA component by using IBM InfoSphere Information Services Director. For example, if the warehouse is the “golden” customer master, you can define a lookup and matching rule by using QualityStage and deploy that rule as a service by using Information Services Director. The service can be used by front-end systems to validate new customer information against the customer master.

When ETL jobs are ready, they can be deployed and run to populate the warehouse and marts.



Part 2

Meeting the increasing demands of workloads, users, and the business

In Part 2, we provide more information about how to use Information Server's capabilities to solve problems around self service, big data, statistical models, information governance and data quality. Numerous other IBM Redbooks publications provide dmore detailed information about the broader set of Information Server capabilities for data integration, data quality, and metadata management. For more information about those Redbooks publications and online information sources, see Appendix , "Related publications" on page 171.

Part 2 includes the following chapters:

- ▶ Chapter 3, “Data Click: Self-Service Data Integration” on page 37
- ▶ Chapter 4, “Incorporating new sources: Hadoop and big data” on page 53
- ▶ Chapter 5, “SPSS: Incorporating Analytical Models into your warehouse environment” on page 73
- ▶ Chapter 6, “Governance of data warehouse information” on page 89
- ▶ Chapter 7, “Establishing trust by ensuring quality” on page 115
- ▶ Chapter 8, “Data standardization and matching” on page 153



Data Click: Self-Service Data Integration

Information Server introduced a new feature called InfoSphere Data Click (Data Click) in its latest version 9.1. Data Click provides a new data movement paradigm for non-technical and business users and is described in detail in this chapter. More information about Data Click that reaches beyond the scope of this chapter can be found in the user documentation that ships with the Information Server 9.1 product.

This chapter includes the following topics:

- ▶ Motivation and overview
- ▶ The two-click experience for a self-service user
- ▶ Summary and more resources

3.1 Motivation and overview

To understand the basic principles of Data Click, consider the diagram that is shown in Figure 3-1 on page 39. A business user with no experience in designing database operations is interested in building a set of data analytics reports.

The diagram depicts IBM Cognos Business Intelligence (Cognos BI) as the analytics environment but any other relational reporting environment can be used here. The key is that an IBM PureData System for Analytics powered by IBM Netezza is used in the analytics environment to host the analytical data mart. While the data marts have fairly complex design and runtime requirements, the business user is only interested in the data and wants to be self-sufficient, without the constant help from an IT partner.

Data Click provides that capability and allows the business users to create their own set of data marts with a simple two-click action and without any technical expertise. The data for the IBM Netezza data marts can be used from Oracle Database or IBM DB2 (DB2) for Linux, UNIX, and Windows data stores in the current version of Data Click. Figure 3-1 on page 39 shows how Data Warehouses in DB2 or Oracle Database can be used as sources for the analytical IBM Netezza data mart.

Data Click is independent of a model and any kind of database can be used as a source, including Master Data Management systems and transactional or operation databases.

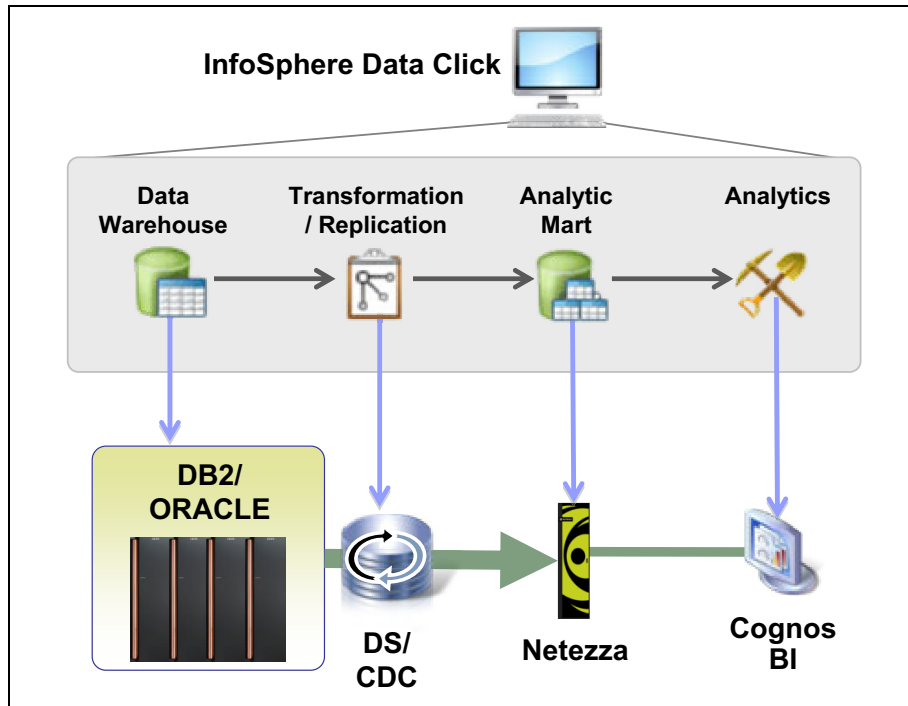


Figure 3-1 InfoSphere Data Click

While the business user requires an agile and simple mechanism to construct data marts in IBM Netezza, the data owner or data steward has different requirements. The goal is to keep track of the quality of the data and make sure that it is not copied, moved, or otherwise manipulated without their consent.

Data Click has this sort of governance aspect built-in and strikes the right balance between business agility and governance. Visibility on the source data and the ability to copy data to the IBM Netezza data marts are all constrained in Data Click and provide a managed and controlled environment, keeping the data stewards in charge of their data. Other governance aspects including the traceability and lineage of the data copies are all built into Data Click.

For the actual data movement, Data Click uses InfoSphere DataStage (DataStage) and InfoSphere Data Replication based on policies selected and source system capabilities.

Data Click is intended for the following applications:

- ▶ Be a tool to support non-technical (self-service) users to generate sandbox databases in IBM Netezza systems, which are populated from Oracle Database or DB2 sources.
- ▶ Provide a simpler mechanism to get the user started on basic data movement requests.
- ▶ Hide some of the technical complexities and challenges that are associated with the native tooling.
- ▶ Provide a mechanism to identify database assets that are available to the self-service users and define the policies and constraints about how the data can be offloaded.

Data Click is not intended to perform the following tasks:

- ▶ Replace native tooling interfaces or the value they provide, especially for complex off load scenarios.
- ▶ Provide a new engine technology to create data marts or provision data.

3.1.1 Benefits of Data Click over traditional approaches

Data marts and analytical sandboxes are commonly created by using a set of SQL scripts for data extraction, a set of DDL scripts to create the data mart schema, and another set of load scripts. This is certainly an agile way to construct and deploy a sandbox but lacks any data governance and is a suitable approach only for database administrators or engineers.

Another common approach is to build Extract Transform Load (ETL) jobs, for example, with InfoSphere DataStage. Unlike SQL scripts, ETL jobs can be designed and run with the consideration of data governance. Access control, data lineage, operational metadata, and operational reporting all can be made available for ETL jobs. The business user, however, is challenged to design ETL jobs or even put ETL jobs into action. Without the ability for self-service users to construct and run their own set of ETL jobs, the burden again falls onto the database administrator or ETL developer to deliver the jobs.

Data Click features the following benefits over traditional approaches:

- ▶ Lower total cost of ownership

The SQL or ETL approach requires a direct communication between the business user with the analytical requirements and the technical user with the expertise. However, this model does not scale in a corporation with hundreds of business users and a hand full of IT experts.

One of the benefits of Data Click is to provide the same business agility that is achieved with SQL or ETL-based execution but at lower total cost of ownership (TCO). A single database administrator can enable a number of business users to construct their initial data marts and progressively update or refine them. The IT administrator does not have to be involved with the users beyond the initial setup of Data Click. For the SQL or ETL developer, it means no involvement in a process that was traditionally time-consuming and cumbersome.

From an operations cost perspective, this translates into reduced time-to-value for newly provisioned Pure Data System for Analytics and significant reduction in ETL or SQL engineering resources.

- Using the best technology for the workload

With the traditional use of ETL or SQL, a developer must decide what technology to use at what time. This is determined by the functional requirements or the performance requirements for the workload. SQL scripts might be the better approach for a data warehouse with only a few million facts and a small set of dimensions. ETL is certainly the choice for the throughput requirements of a large data warehouse. However, SQL can be used for log-based replication, a capability not readily available with ETL. In the end, a technology decision must be made based on functional requirements, performance requirements, and the properties of the workload.

Data Click has these decisions built in with a deterministic approach that always chooses the right technology for the right task. Data Click loads data via batch with the highest scalability and top performance or propagates changes in real time at the highest speed. At the same time, Data Click does not depend on the availability of a particular type of engine. If one type of engine was not configured or should not be used, Data Click can compensate for the lack of functionality and reduces the UI to show only the available functions.

- Pro-active and built-in governance

Some of the following commonly asked questions are often asked concerning the use of ETL jobs or SQL scripts:

- Who ran this job or script?
- When was it last run?
- Where does the data in a data mart come from?
- Is the data current or stale?
- Are there errors or warnings with the load process?

Any SQL or ETL developer finds it difficult to answer these questions individually for the business users but it is impossible to answer them when they are trying to keep a perfect record of all of the runs. Consider the number of business users demanding new sandboxes and the turnaround time for the developer. What is needed is an automated way of capturing the metadata surrounding the design and running of each job or script.

Data Click automates the metadata generation and allows any user (IT or business) to track and audit the offload requests. This can be done starting from the target asset (the sandbox or data mart) or from the operational process (the offload request). Metadata is created for every target, each offload request, and linked with the engine information, the user record, and the operational errors and logs. Only the combination of all these records allows a complete answer for to a comprehensive list of questions. Data Click has governance built in.

Data Click has the following advantages over traditional SQL- or ETL-based sandbox generation:

- ▶ Business Agility at lower TCO
- ▶ The use of the best technology for the workload
- ▶ Pro-active and built-in governance

3.1.2 Data Click details

Following the motivation and use cases previously described, Data Click is designed and implemented as a solution to a common and recurring set of problems. As such, Data Click is not a new product nor is it a feature within just a single product. Data Click is a solution using a number of components of the InfoSphere Information Server Data Integration Edition. While the user works with only a simple browser interface (as you see in the next section) the core of Data Click relies on the following enterprise level capabilities that support the levels of automation and governance inherent in this solution:

- ▶ InfoSphere Metadata Asset Manager
- ▶ InfoSphere Blueprint Director
- ▶ The metadata repository of InfoSphere Information Server
- ▶ InfoSphere Metadata Workbench (Metadata Workbench)
- ▶ InfoSphere DataStage
- ▶ The operations console of InfoSphere DataStage and InfoSphere QualityStage (QualityStage)
- ▶ Optionally InfoSphere Change Data Capture (CDC)

Some of the products are used in the following functions throughout the lifecycle of a Data Click solution:

- ▶ InfoSphere DataStage Administrator to configure the InfoSphere DataStage project used by Data Click
- ▶ Asset Manager to import source and target metadata (including connections) to Metadata Server
- ▶ Blueprint Director to design blueprints, associate metadata, configure the scope of Data Click, and publish blueprints to Metadata Server
- ▶ Metadata Workbench to find published blueprints, launch blueprints, and start the Data Click UI
- ▶ Operations Console or InfoSphere DataStage Designer to monitor the execution of Data Click jobs
- ▶ Metadata Workbench to inspect and review the data lineage created by Data Click

If CDC is configured with Data Click, CDC Management Console can be optionally added to configure data store connections and user privileges that are used for Data Click and to monitor or alter subscriptions.

For more information about individual products, see the following resources:

- ▶ IBM InfoSphere Information Server Version 9.1 Information Center at this website:
http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp?topic=%2Fcom.ibm.swg.im.iis.productization.iisinfsv.home.doc%2Ftopics%2Fic_homepage_IS.html
- ▶ IBM InfoSphere CDC version 6.5.1 Information Center at this website:
<http://pic.dhe.ibm.com/infocenter/cdc/v6r5m1/index.jsp?topic=%2Fcom.ibm.cdc.doc.homepage.doc%2Fic-homepage-cdc.html>

3.2 The two-click experience for a self-service user

Before taking a more detailed look at an end-to-end Data Click solution, we start with the experience a self-service user gets from a fully configured Data Click environment. In the best case, the business user must use only a couple of mouse clicks to create a sandbox data mart in an IBM Netezza database.

The scenario starts out with a fully configured blueprint that is published by an administrator. The business user obtained the link to the blueprint and upon authentication, can work with the blueprint similar to the one shown in Figure 3-2. A single link on the right side of the window becomes active when the Data Click Activity element in the blueprint is selected. By using this link, you can see the Data Click configuration UI.

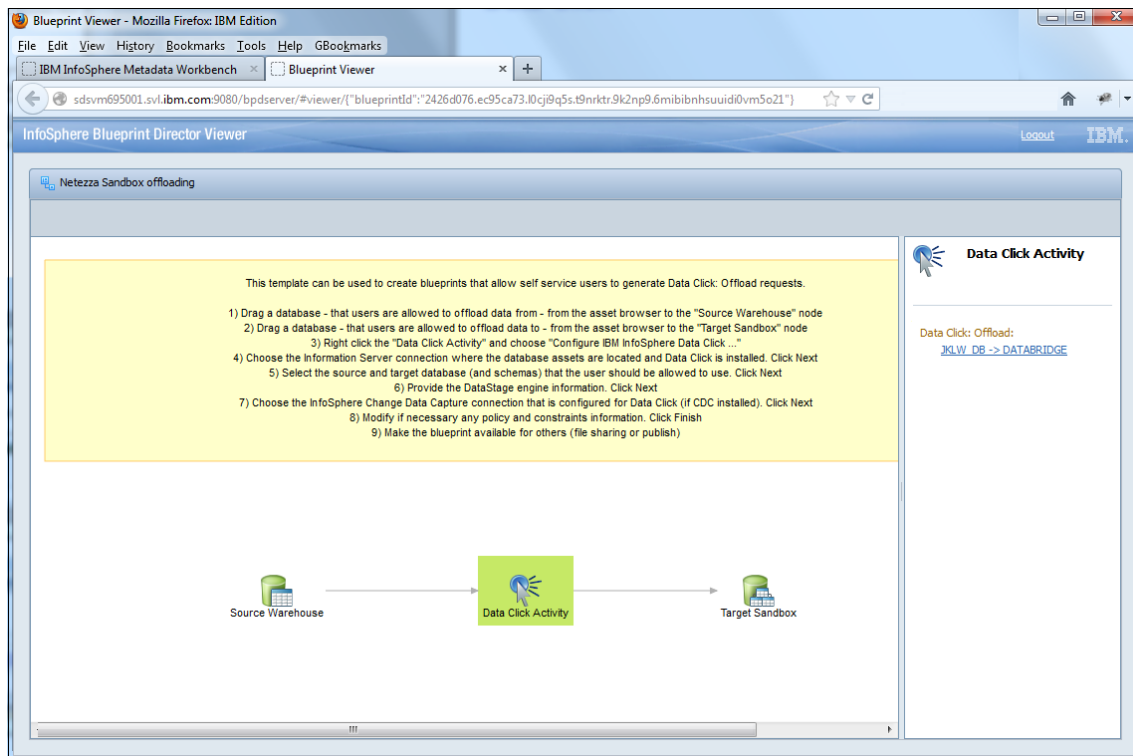


Figure 3-2 Web blueprint

In Figure 3-3, the Data Click configuration wizard shows the UI as started from the blueprint. The UI is implemented as a simple wizard with four pages to perform the following tasks:

- ▶ Select the target
- ▶ Select the source
- ▶ Configure options
- ▶ Review a summary and run

In a fully configured scenario, the first three pages are pre-populated and valid as indicated by the green check mark next to the tab name. The self-service user does not have to inspect these pages and can immediately finish the configuration with a first mouse click.

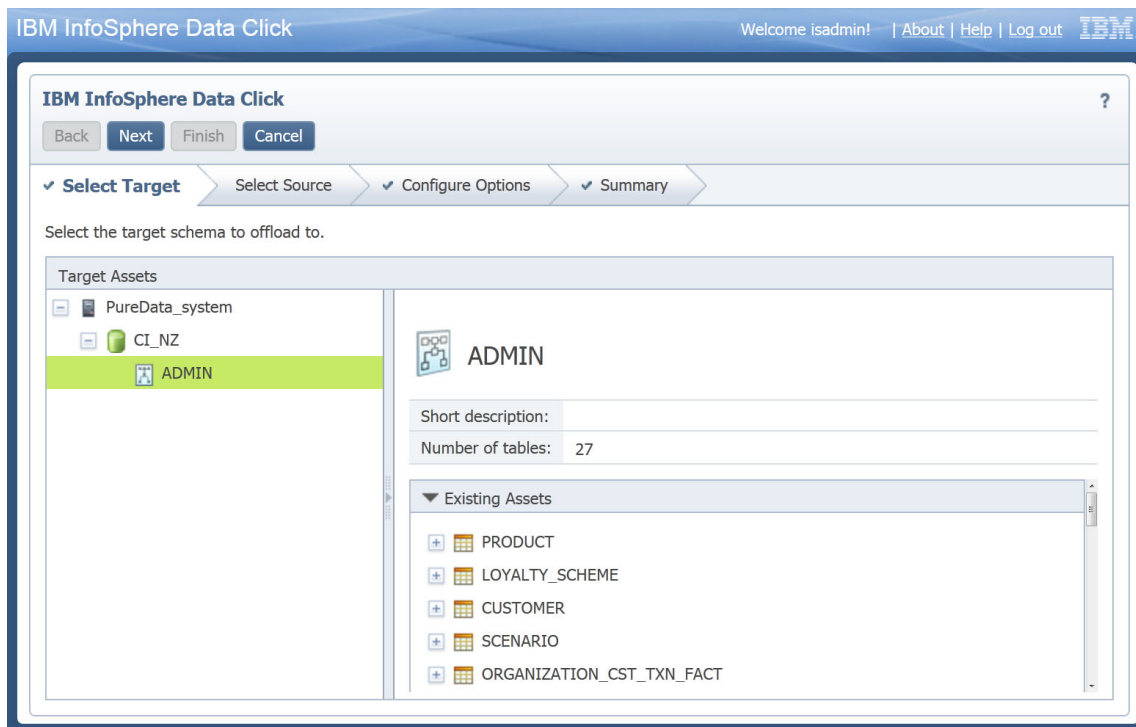


Figure 3-3 Data Click configuration wizard

In Figure 3-4 on page 46, the Data Click summary page shows the final wizard page as presented to the self-service user. The list of source tables and mappings to target tables is shown with a number of policies and constraints. The business user can assume that all of these settings are correct and trigger the run by clicking **Finish**.

IBM InfoSphere Data Click

Welcome isadmin! | [About](#) | [Help](#) | [Log out](#) **IBM**

IBM InfoSphere Data Click

Back Next Offload Cancel

✓ Select Target ✓ Select Source ✓ Configure Options ✓ **Summary**

Review the settings for the offload request.

Name: default_isadmin_30May13_114329

☐ Validate the offload request before offloading the data

▼ Source to Target Mapping

Source Schema	Source Table	Target Schema	Target Table
CSTINSIGHT	INDIVIDUAL_CST_LOYALTY_FACT	ADMIN	INDIVIDUAL_CST_LOYALTY_FACT
CSTINSIGHT	INDIVIDUAL_CST_TXN_FACT	ADMIN	INDIVIDUAL_CST_TXN_FACT
CSTINSIGHT	INDIVIDUAL_CST_CHURN_FACT	ADMIN	INDIVIDUAL_CST_CHURN_FACT

▼ Policies and Constraints

Engine:	DATASTAGE
Maximum number of records to be extracted for each table:	1000000
Maximum Number of Parallel Jobs:	3
Maximum Number of Tables to Extract:	5
Frequency of Updates at the Source:	LOW

Figure 3-4 Data Click summary page

This concludes the Data Click activity for the business user, who can then close the Data Click UI, the blueprint, or even the web browser. Data Click runs entirely asynchronously and in the background.

3.2.1 Running and feedback

When Data Click runs in the background, it performs the following tasks, depending on configuration and product capabilities:

- ▶ Create the target tables (if not yet created).
- ▶ Write schema metadata (for the created target tables).
- ▶ Populate the target tables (insert or update or replace based on configuration).
- ▶ Log operational metadata for construction and population of target tables.
- ▶ (Optional) Continuously replicate source changes to the target (depending on configuration).

- ▶ Write lineage metadata for source to target mappings (for continuous replication only).
- ▶ Log replication events (for continuous replication only).

There is no immediate feedback from Data Click to the self-service user. This is especially difficult when the web UI was closed before the background tasks were completed. To verify that Data Click finished running, the user has the following options:

- ▶ Validate that the target tables exist and are fully populated.
- ▶ Inspect the operational metadata or the event logs for continuous replication.
- ▶ Inspect the schema or lineage metadata, which is written after the job run completes.

3.2.2 Advanced user configuration

While the two-click experience can be the right experience for a new user, the Data Click UI offers a few more advanced options for repeated use.

Figure 3-5 shows the first page of the Data Click configuration wizard where the user can select a single IBM Netezza target database schema from a list of available targets. The page lists the tables that are already created in the selected schema to guide the user with the selections. Data Click automatically adds to this table browser upon running a new request.

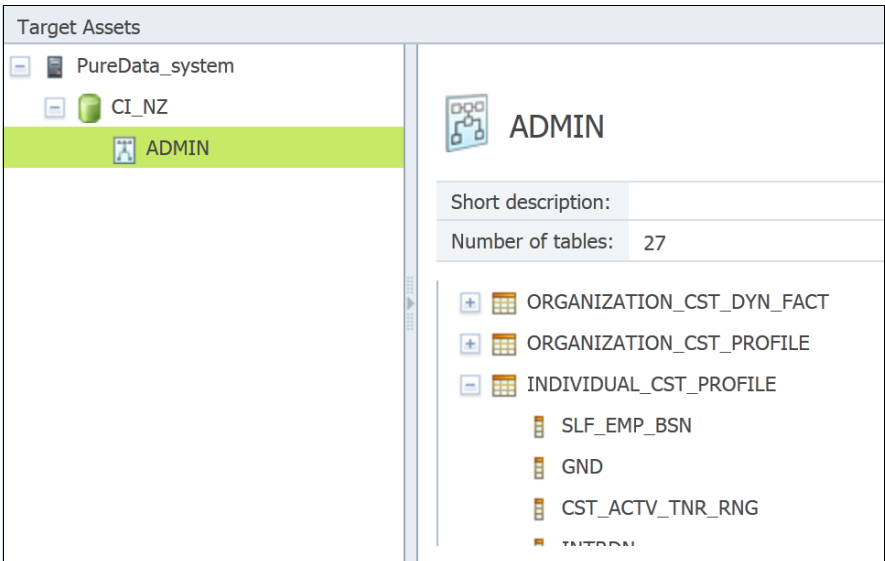


Figure 3-5 Data Click UI: Target selection

Figure 3-6 shows the source selection page in the wizard. On this page, the user is presented with a list of databases, schemas, tables, and columns that can be accessed. The list can be a subset of the entire physical database schema and is filtered based on the following criteria:

- The metadata that is available in Metadata Server.
- The constraints put around the source Data Click activity by the administrator.

The business user can further reduce the number of tables and columns to be copied to the IBM Netezza sandbox through multi-selection in the tree browser.

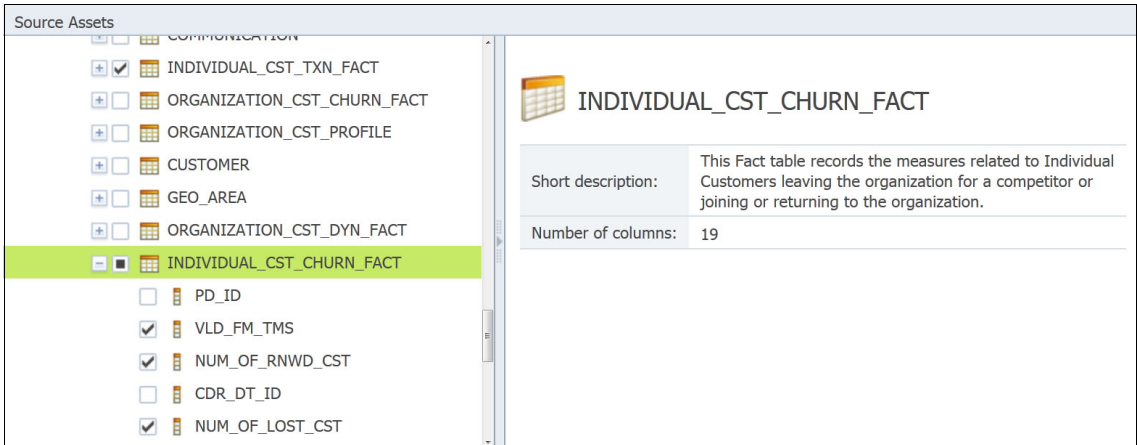


Figure 3-6 Data Click UI: Source selection

Figure 3-7 on page 49 shows the third wizard page where the user has options concerning the running of the Data Click request. The most important selection here is whether to create a one-time snapshot of the selected source tables or use a refresh schedule.

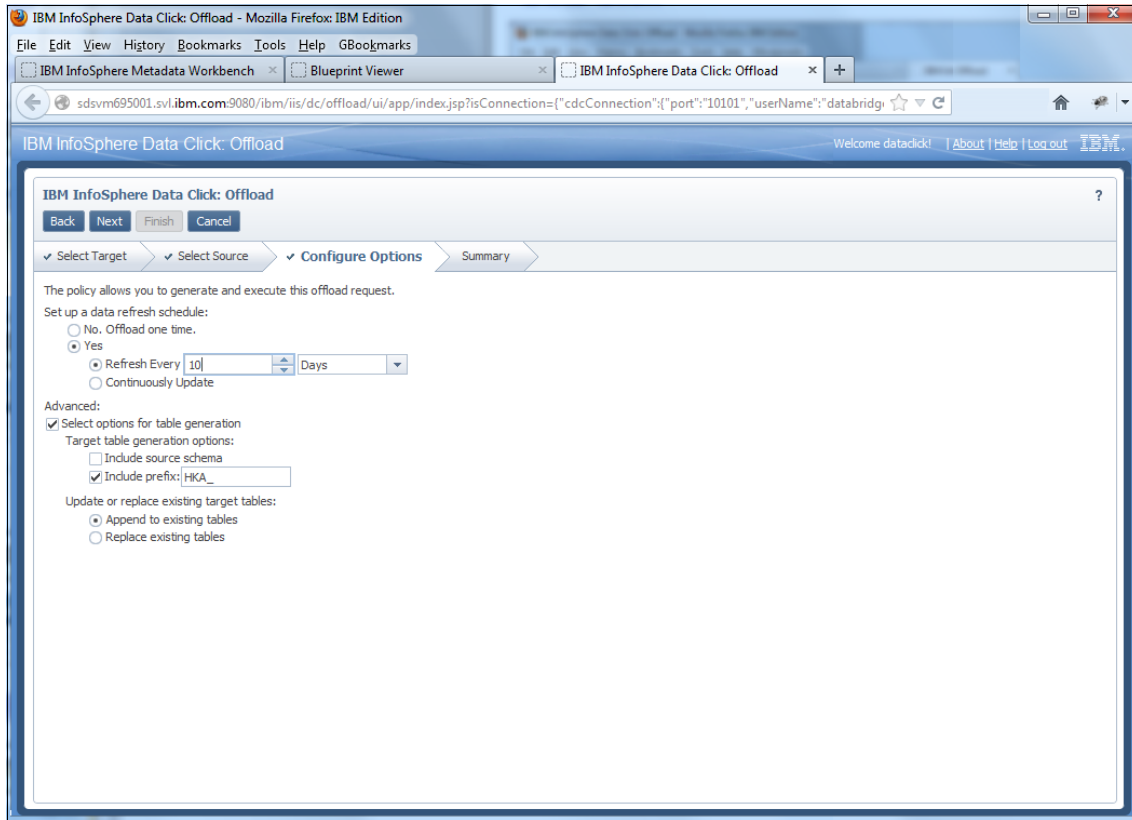


Figure 3-7 Data Click UI - Configuration Options

With the refresh schedule, the user has an option to perform a batch update once a week on a given day, or replicate source changes in real time. Some of the options are available only if Data Click was configured by the administrator with InfoSphere DataStage and CDC support. For more information, see 3.1.1, “Benefits of Data Click over traditional approaches” on page 40.

Depending on the selection that is made for the refresh schedule, the user has the following advanced options available to configure the run:

- Use of table name prefixes

This is relevant for an IBM Netezza data mart as IBM Netezza databases support only a global namespace. Table names must be unique within the database and schemas do not serve as namespaces. For convenience, Data Click offers to prefix the created table name with the source schema name or with a user-defined prefix.

- Define the update policy on the target table
Should a target table already exist, Data Click can append or replace the records in that table. This option is relevant only for InfoSphere DataStage based operations because Replication always updates records.

The user is presented with the summary page, as shown in Figure 3-8. The page lists the selected source tables with a mapping to the respective target table. The target table names are prefixed with an optional prefix. In a collapsed section of the page, the Policies and Constraints become visible to the business user but cannot be edited. The following settings are made by the administrator:

- The runtime engine to be used for the run.
- A record limit for the number of records to be loaded into the target table.
- A limit on the number of InfoSphere DataStage jobs that are used for the run (if InfoSphere DataStage is the runtime engine).
- A limit on the number of tables that can be created and copied.
- Other options concerning the expected frequency and volume of the runs.

Name: default_isadmin_30May13_122257

☐ Validate the offload request before offloading the data

▼ Source to Target Mapping

Source Schema	Source Table	Target Schema	Target Table
CSTINSIGHT	INDIVIDUAL_CST_LOYALTY_FACT	ADMIN	CSTINSIGHT_HKA_INDIVIDUAL_CST_LOYALTY_FACT
CSTINSIGHT	INDIVIDUAL_CST_TXN_FACT	ADMIN	CSTINSIGHT_HKA_INDIVIDUAL_CST_TXN_FACT
CSTINSIGHT	INDIVIDUAL_CST_CHURN_FACT	ADMIN	CSTINSIGHT_HKA_INDIVIDUAL_CST_CHURN_FACT

▼ Policies and Constraints

Engine:	DATASTAGE
Maximum number of records to be extracted for each table:	1000000
Maximum Number of Parallel Jobs:	3
Maximum Number of Tables to Extract:	5
Frequency of Updates at the Source:	LOW
Enabled Request:	GENERATE_AND_EXECUTE
Estimated Size of Data at the Source:	SMALL

Figure 3-8 Data Click UI: Summary

The two most relevant bits of information are found at the top of the page. The name of the request is generated and encodes the following information:

- The name of the project that is used.
- The user ID with which to generate the request.
- The time stamp for the generated request.

This name is unique for a given system and can be used to monitor the running of a Data Click request, validate the operational logs and events, and associate the lineage metadata with the generating request.

The user can chose to deploy the run request only for validation without an actual run. In this case, the engine creates the InfoSphere DataStage job or the CDC subscription without running on it. This option can be enforced by the administrator when the Data Click activity is configured.

Upon review of all the information in the summary page, the self-service user activates the Data Click request as in the two-click scenario by clicking **Finish**. The session then can be closed and the user can exit out of Data Click.

3.3 Summary and more resources

In this chapter, we described how Data Click defines a new paradigm for data movement and sandbox creation without the need for technical skills. With DB2 or Oracle Database databases as sources, Data Click can create and load a PureData System for Analytics powered by IBM Netezza. Support for other source and target systems is planned.

While there is a significant overhead for deployment and configuration of Data Click, the savings are in the scale of use. Any initial overhead is offset by a large number of self-service users or the need for repeated data movement.

Governance is built into Data Click through the use of Information Server and especially the use of metadata for a controlled and governed run. Privileged users can work with their data sources in a controlled and constrained manner without the need for database credentials. User activities are managed by IT with constraints built directly into the blueprint. Operational and lineage metadata is created for every run.

For administration and running monitoring tasks, the IT staff can continue to work with the native tooling that is available for DataStage or CDC. Data Click requires no specific engine or data source capabilities.

IBM continues to invest in Data Click development and new documentation is expected to become available with more functionality. For the most current user documentation, see the IBM InfoSphere Information Server Version 9.1 Information Center at this website:

<http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/topic/com.ibm.swg.im.iis.dataclick.nav.doc/containers/dclkintrocont.html>

Product offerings are available through the IBM InfoSphere Information Server software portal at this website:

http://www-947.ibm.com/support/entry/portal/Overview/Software/Information_Management/InfoSphere_Information_Server

The portal contains links to more material for Information Server, including papers and presentations for Data Click.

There continues to be new developerWorks articles and Tech Notes for Information Server and individual Information Server features. Search the IBM developerWorks portal for the current set of technical documents on Data Click at this website:

<http://www.ibm.com/developerworks/>



Incorporating new sources: Hadoop and big data

Hadoop provides a fault-tolerant distributed processing environment for managing and processing massive semi-structured and structured data, such as social data, web logs, sensor data, and images. Collectively, these forms of data and the volume and speed at which they are generated in today's world has led to the coinage of the term *big data*. With the explosion in big data, Hadoop has become a critical, scalable platform for processing it. Furthermore, it is imperative for data warehouse environments to integrate with Hadoop and to incorporate these new big data sources. IBM InfoSphere Information Server provides a comprehensive integration with Hadoop and other big data sources.

In the strictest sense, Information Server extends its integration and governance capabilities to include Hadoop and big data sources. In addition, Information Server can use the new big data processing capabilities in these new platforms and extend their value to data warehousing and business intelligence. Information Server acts as a glue to bring Hadoop and big data into the warehousing and enterprise ecosystem.

In the integrated DataStage/Hadoop environment, big data can be extracted into Data Warehouses for complex analytic query processing. Some processed data or metadata can be stored in the Data Warehouse or written to big data sources. In this chapter, we describe the DataStage/Hadoop integration methodologies, which include the following major functionalities:

- ▶ Big Data File Stage: Used for loading and extracting files from Hadoop systems
- ▶ Balanced Optimization for Hadoop: Makes use of the Hadoop platform for processing
- ▶ IBM InfoSphere Streams Connector: Used for integration with real-time, low-latency analytics processing
- ▶ Oozie Workflow Activity Stage: Used for end-to-end big data processing coordination

InfoSphere DataStage integrates Hadoop files with traditional databases. Big data can be extracted, processed, and loaded into traditional data warehouses. Conversely, data that is in traditional data warehouses can be extracted, processed, and loaded into big data sources. Figure 4-1 on page 55 shows that the integration methodologies enable InfoSphere DataStage as an ETL hub for big data and traditional data warehouses. In this chapter, all the references to Hadoop also apply to IBM BigInsights™ because the later is a big data platform that is powered by Hadoop.

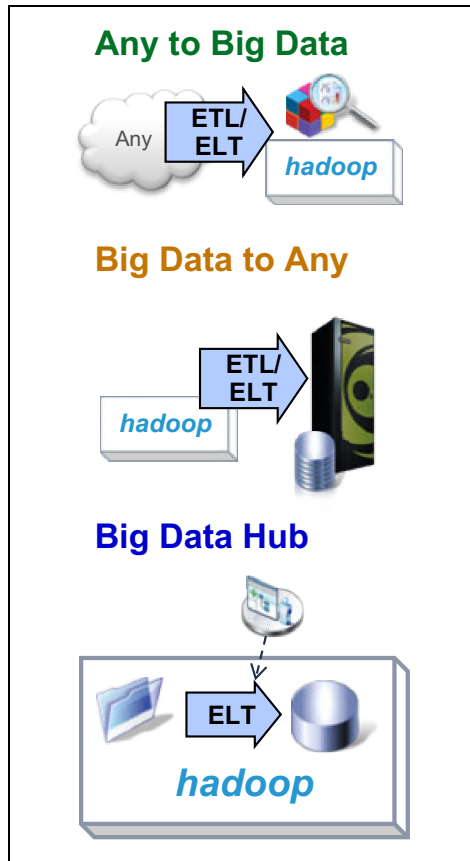


Figure 4-1 Integration methodologies

This chapter includes the following topics:

- ▶ Big Data File Stage
- ▶ Balanced Optimization
- ▶ Balanced Optimization for Hadoop
- ▶ IBM InfoSphere Streams Integration
- ▶ Oozie Workflow Activity stage
- ▶ Unlocking big data

4.1 Big Data File Stage

Big Data File Stage (BDFS) specifically enables load and export from file systems. It was first released in Information Server V8.7. The BDFS stage provides parallelism and scalability capability to address the big volume requirements and provides metadata and data lineage capabilities.

In an integrated Data Warehouse/Hadoop environment, data can be extracted from Hadoop and eventually loaded into various database systems or file systems. In the other direction, data can be extracted from databases or files and loaded into Hadoop file systems. In both cases, IBM InfoSphere DataStage is at the center of the data movement. With the BDFS stage support, DataStage enables a full spectrum of big data movement across all supported data sources including big data, as shown in Figure 4-1 on page 55. The BDFS stage also uses the parallel run support that is inherently built in the Parallel Engine to seamlessly scale up the data extraction and loading.

Consider a use case as shown in Figure 4-2 on page 57. On the left side there are enterprise sources and enterprise data that might be staged in large file systems. This data must be brought in on a regular basis into IBM InfoSphere BigInsights for the analytics to be applied on the new data.

On the right side, there are a number of downstream systems that might be Data Warehouses, Data Marts, or other analytic systems. These systems need summaries or subsets of the types of data that are now in BigInsights so they can apply more processing and deliver business reports and other analytics that the business needs.

For both of these capabilities, there is a need to move the data into and out of these systems as efficiently as possible and must be able to scale. This is where IBM InfoSphere DataStage becomes most useful with the scalable BDFS stage implementation.

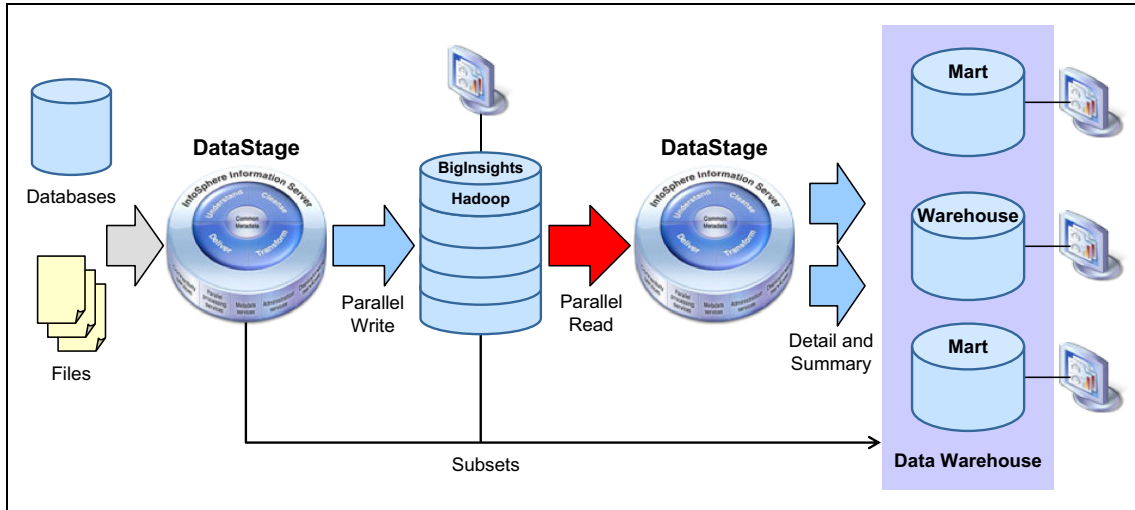


Figure 4-2 Use case

When BDFS stage is used, Information Server automatically generates the metadata of the Hadoop files and the job metadata. This metadata is stored in the metadata repository and can be used to show the lineage and effect of data by using Metadata Workbench. The lineage diagram in Figure 4-3 shows how a particular field is generated from the source fields.

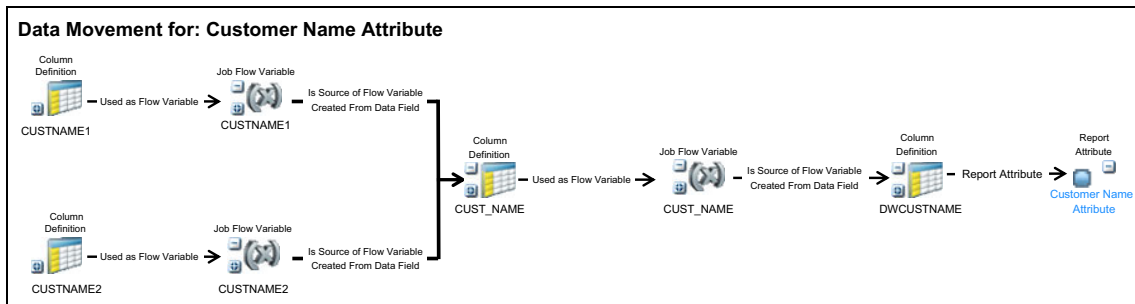


Figure 4-3 Lineage diagram

In addition to the support for parallel execution, the BDFS stage supports multiple input files (when used as a source stage) and multiple output files (when used as a target stage). Figure 4-4 on page 58 and Figure 4-5 on page 59 show the properties of the BDFS stage. In Figure 4-4 on page 58, a File Pattern is used that specifies multiple input HDFS files that are read by the BDFS stage (used as a source stage).

The properties BDFS Cluster Host and BDFS Cluster Port Number specify the Hadoop cluster host and port number. In Figure 4-4, the property Root File String is used to generate target file names. This can be a fully qualified path or only the root string of the file name, in which case the files are created in the current working directory. For example, by default, with the root string /tmp/outputFile, the generated file names are /tmp/outputFile.part00000, /tmp/outputFile.part00001, and so on, where part00000 and part00001 are file partition names.

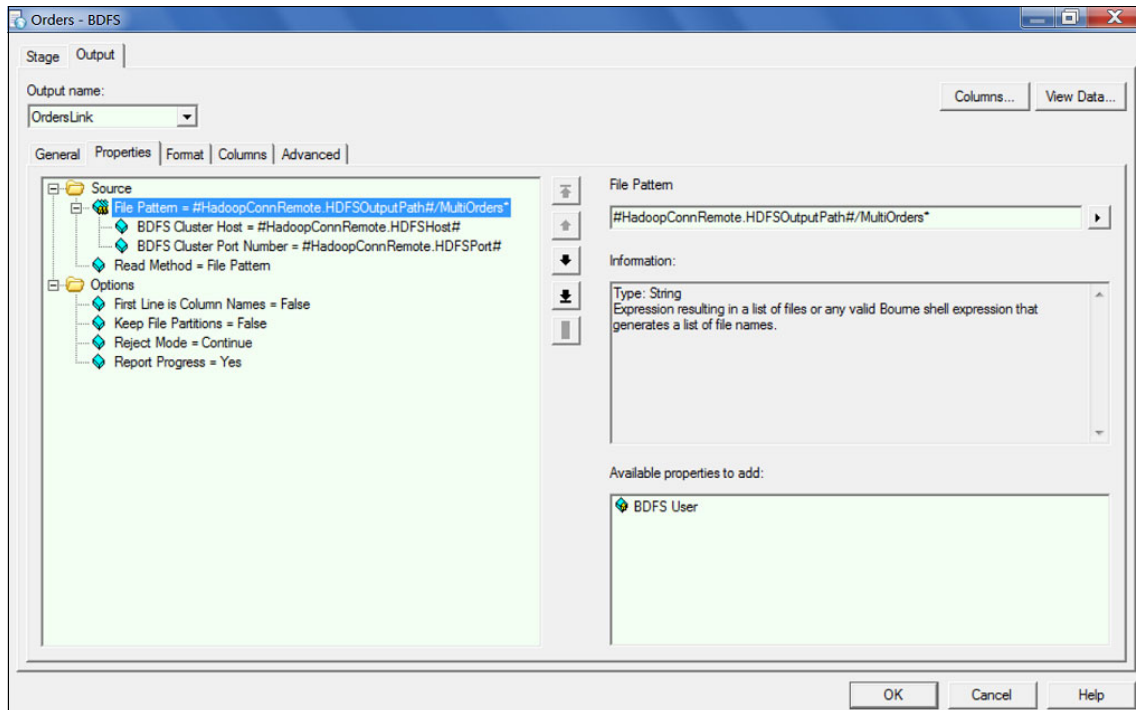


Figure 4-4 File Pattern

In addition to the support for parallel running, the BDFS stage supports multiple input files (when used as a source stage) and multiple output files (when used as a target stage). Figure 4-4 and Figure 4-5 on page 59 show the properties of the BDFS stage. In Figure 4-4, a File Pattern is used, which specifies multiple input HDFS files that are read by the BDFS stage (used as a source stage). The properties BDFS Cluster Host and BDFS Cluster Port Number specify the Hadoop cluster host and port number.

In Figure 4-5, the property Root File String is used to generate target file names. This can be a fully qualified path or just the root string of the file name, in which case the files are created in the current working directory. For example, by default, with the root string /tmp/outputFile, the generate file names are /tmp/outputFile.part00000, /tmp/outputFile.part00001, and so on, where part00000 and part00001 are file partition names.

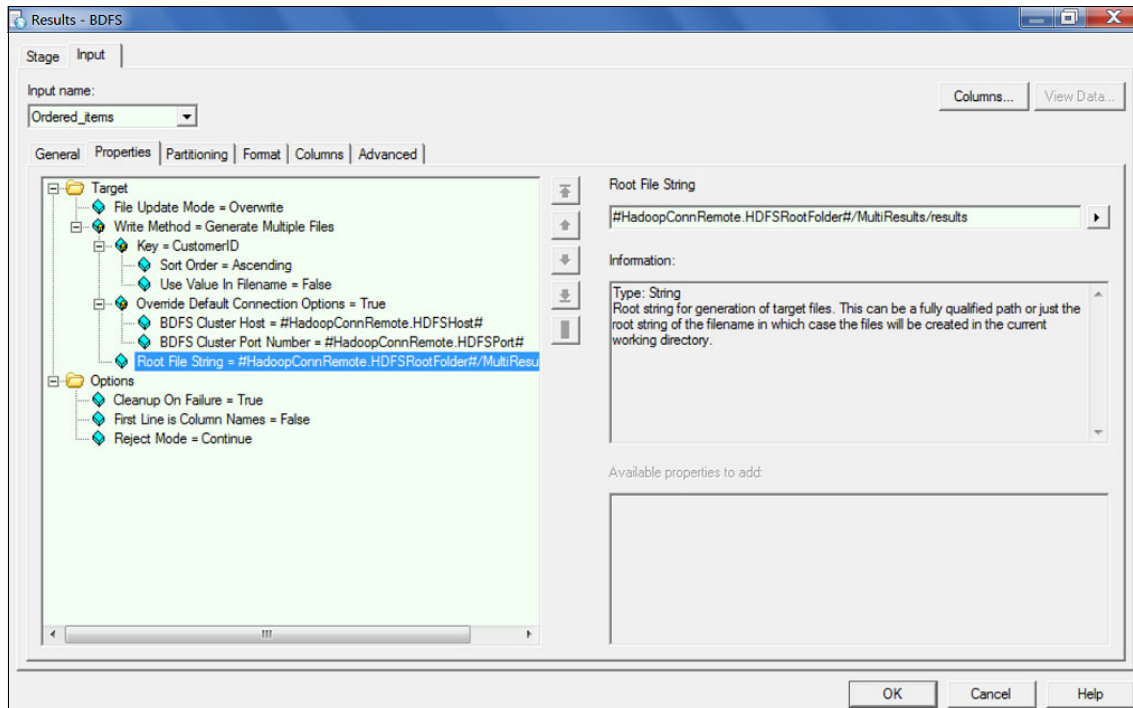


Figure 4-5 Root File String

4.2 Balanced Optimization

InfoSphere DataStage jobs provide connectivity, data manipulation functionality, and highly scalable performance. The InfoSphere DataStage visual flow-design paradigm is easy to use when you are designing simple-to-complex data integration jobs. Better performance might be achieved, however, if the processing load can be shared or redistributed among InfoSphere DataStage and the source or target databases. With InfoSphere DataStage Balanced Optimization (BalOp), DataStage users can control where the intensive work is done: in source databases, InfoSphere DataStage, or target databases.

For job designs that use connectors to read or write data from data sources, BalOp gives you greater control over the job. You first design a job in DataStage then use BalOp to redesign the job automatically to your stated preferences. This redesign process can maximize performance by minimizing the amount of input and output that is performed, and by balancing the processing against source, intermediate, and target environments. You can then examine the new optimized job design and save it as a new job. The root job design remains unchanged. BalOp enables you to take advantage of the power of the databases without becoming an expert in native SQL.

The following principles can lead to the better performance of parallel jobs:

- ▶ **Minimize I/O and data movement**
Reduce the amount of source data read by the job by performing computations within the source database. Where possible, move data processing to the database and avoid extracting data just to process it and write it back to the same database.
- ▶ **Maximize optimization within source or target databases**
Make use of the highly developed optimizations that databases achieve by using local indexes, statistics, and other specialized features.
- ▶ **Maximize parallelism**
Take advantage of default InfoSphere DataStage behavior when databases are read and written to. Use parallel interfaces and pipe the data through the job so that data flows from source to target without being written to intermediate destinations.

BalOp uses these principles to improve the potential performance of a job. You influence the job redesign by setting options within the tool to specify which of the principles are followed.

Optimization pushes processing functionality and related data I/O into database sources or targets, depending on the optimization options that you choose.

When a job is optimized, BalOp searches the job for patterns of stages, links, and property settings. The patterns typically include one or more of the supported database connector stages (DB2, Netezza, Oracle, or Teradata). When a candidate pattern is found, BalOp combines the processing into the corresponding source or target database SQL and removes or replaces any stages and links that are no longer needed. It then adjusts the remaining stages and links.

After a pattern is found and the job design modified, the process is repeated. Optimization stops when none of the patterns match anything further in the optimized job, which indicates that there is no more work to be done. The process of combining data processing logic into specific database SQL statements to be run by the target database server is called *pushdown*, which essentially enables extract, load, and transform (ELT).

4.3 Balanced Optimization for Hadoop

BalOp pushdown pushes data transformation into data servers. In this way, data processing is performed within the data servers. This fits well with one of Hadoop's main assumptions that the cost of moving applications is far less than the cost of moving massive data. In this section, we describe how BalOp pushdown is extended to big data sources.

MapReduce (MR) is a programming model for processing big data in Hadoop. MR code is usually processed locally in Hadoop cluster nodes because the cost of moving applications is far less than the cost of moving massive data. Developed by IBM, Jaql is one of the high-level MR query languages. It was selected as the first MR language for BalOp/Hadoop integration. There are other high-level MR languages, such as PIG and HIVE, which are potential candidates for more Map-Reduce language support in BalOp in the future. By extending BalOp with MR query generation in a way similar to SQL pushdown, it becomes possible to move ETL-like data extraction and processing logic to Hadoop systems. To this end, a new stage type called MapReduce Stage is created to handle remote query running for MR queries that are generated by BalOp. A MapReduce stage is automatically generated by BalOp when data processing logic is converted to MR queries. Detailed BalOp pushdown examples are given in the subsequent sections of this chapter.

Figure 4-6 on page 62 shows big data platforms are integrated with traditional Data Warehouses by using the same DataStage ETL and BalOp pushdown framework. Hadoop becomes another platform that can be used for data transformation. The use of DataStage artifacts permits common language and migration to and from Hadoop platform. Furthermore, DataStage Designer offers an easy-to-use data transformation authoring tool for Hadoop. Instead of writing scripts and MR queries, users can now create DataStage jobs by using built-in stages to perform equivalent data extraction, transformation, and load.

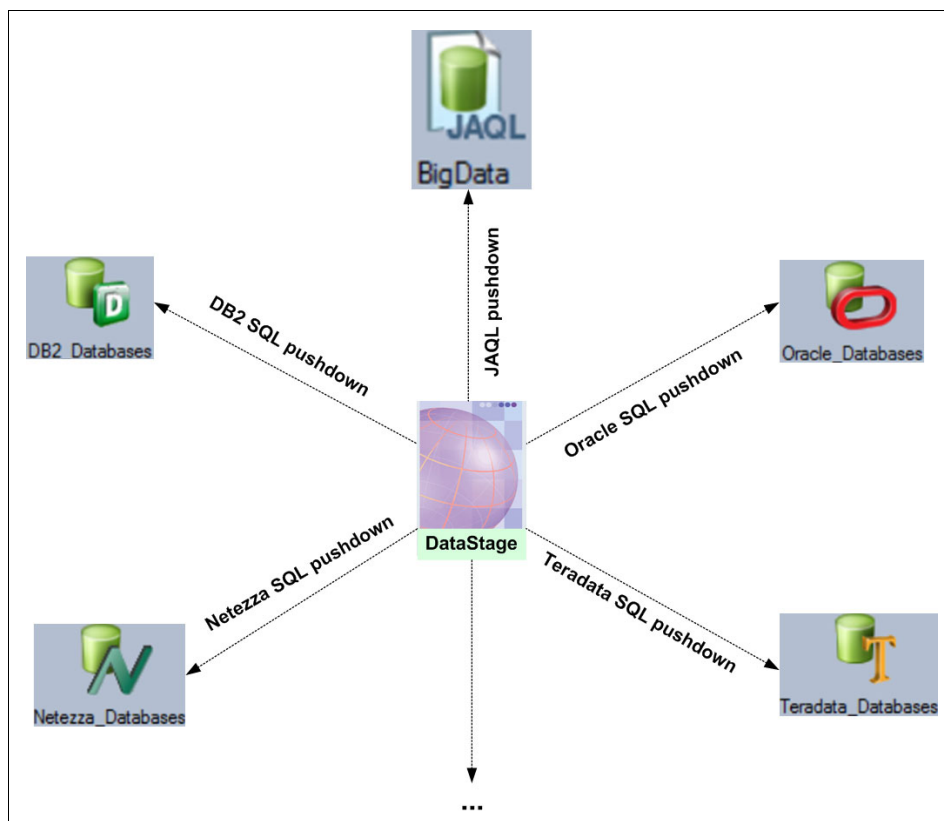


Figure 4-6 Big data platforms

4.3.1 Complete pushdown optimization

For DataStage jobs that access only Hadoop data, BalOp combines all the processing logic in the original job into one or multiple target MapReduce stages in the optimized job. In this process, a source BDFS stage is turned into a high-level read statement, while a target BDFS stage is turned into a write statement. All other supported stage types¹ in the job are translated into their equivalent query statements in the generated MR queries, as shown in Figure 4-7 on page 63.

¹ Not all DataStage stage types are supported for MR generation in the current release. See InfoSphere DataStage product documentation for all supported stage types.

In this job, a main HDFS file is accessed by a BDFS stage on the left side of the job. A lookup stage is used to check the incoming rows from the main HDFS file against data in eight different reference HDFS files. The intermediate result then goes through a Transformer stage that filters out some unwanted rows. The final result is written to a target HDFS file. By using BalOp, the entire job can be completely pushed down to Hadoop. The optimized job that is shown on the right in Figure 4-7 is the result of the optimization.

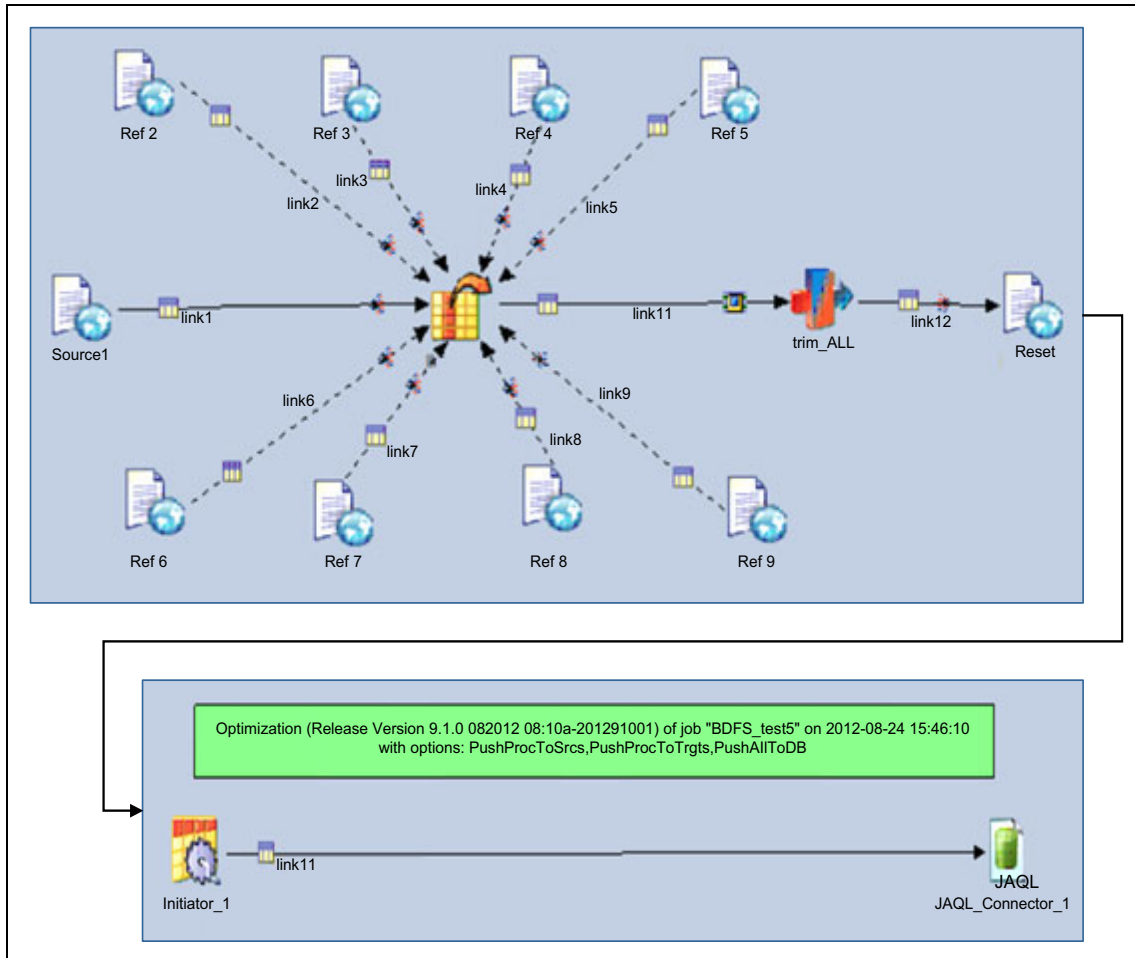


Figure 4-7 Example MR Query

Figure 4-8 shows the properties of the target MapReduce stage in the optimized job. The MapReduce Query property in the MapReduce stage contains the generated MR queries.

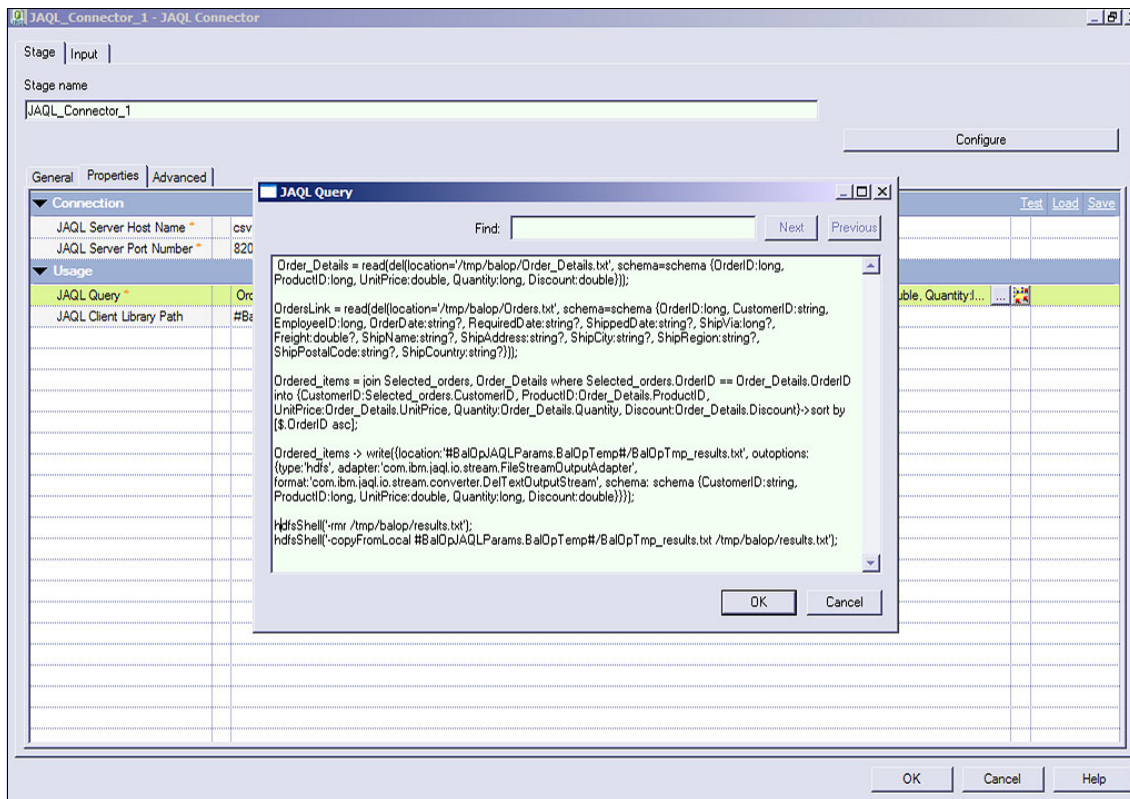


Figure 4-8 Properties of the target MapReduce stage

4.3.2 Hybrid pushdown optimization

A DataStage job can access Hadoop data and traditional databases. In this case, BalOp can generate a hybrid pushdown. Some processing stages can be pushed into a MapReduce stage, while others can be pushed into some database stages. BalOp provides a number of optimization options to help control which processing stages get pushed into a source stage instead of a target stage. The examples that are shown in Figure 4-9 on page 65 and Figure 4-10 on page 66, are jobs that involve databases and HDFS files.

In Figure 4-9, two DB2 tables are joined after some data transformation. The intermediate result then goes through two lookup operations on DB2 tables. After further data transformation and aggregation, another lookup operation on an HDFS file generates pre-sorted final result, which is then written to an HDFS file. By using only the default rules, all the stages to the left of the last lookup are pushed down to DB2, while the remaining stages get pushed down to Hadoop. BalOp provides a fine granular control, called *Stage Affinity*, that can be used to shift the source-target dividing line to the left or to the right.

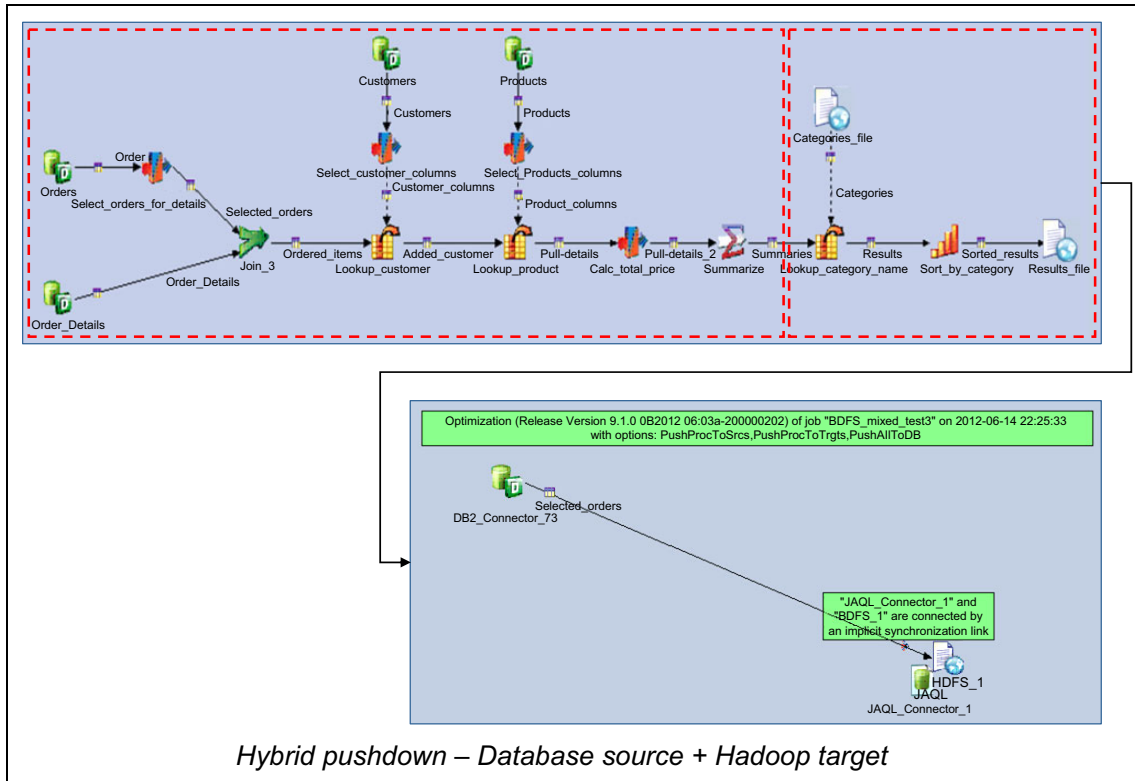


Figure 4-9 Hybrid pushdown optimization: Example 1

The job that is shown in Figure 4-10 is the same as the job that is shown in Figure 4-9 on page 65 except that HDFS files and DB2 tables are switched. We refer to these kinds of optimizations as *hybrid pushdown optimizations*. BalOp can perform such complex hybrid pushdown scenarios and support a comprehensive integration of data warehouse and big data.

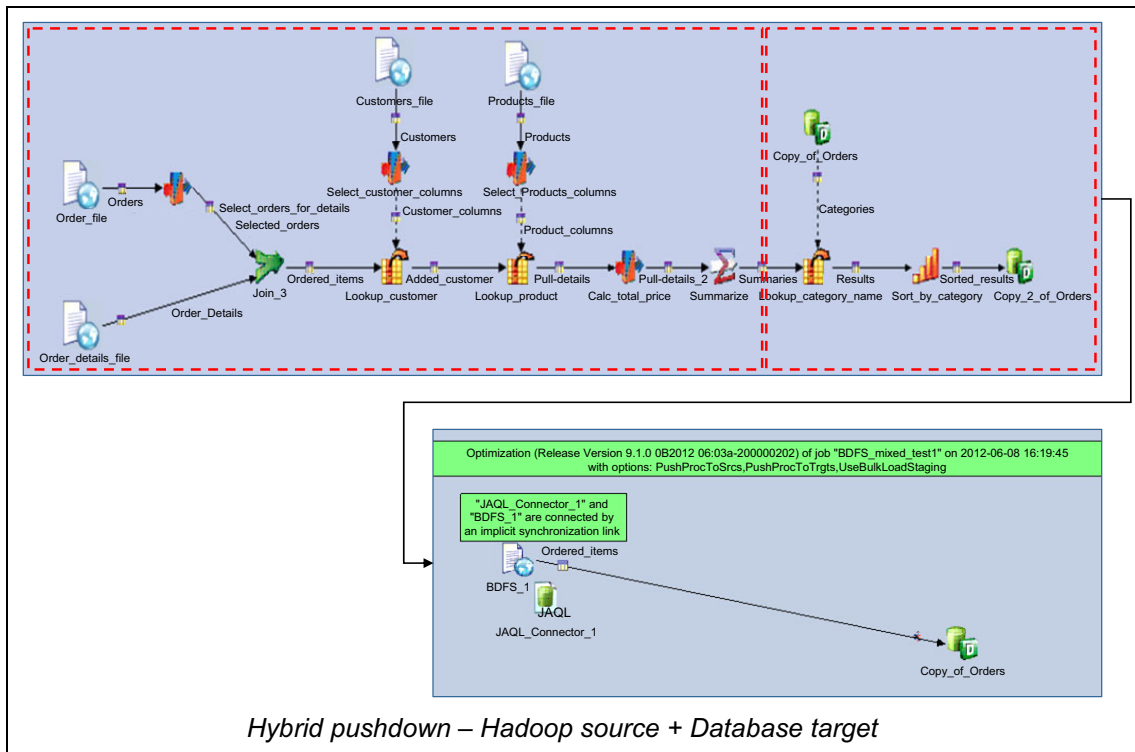


Figure 4-10 Hybrid pushdown optimization: Example 2

Data transformation can be pushed into database servers, Hadoop clusters, or stay in DataStage, depending on the overall needs. In this way, data can flow bidirectional between databases and HDFS file systems and processed in a wanted environment where the particular processing can be most efficiently done.

4.4 IBM InfoSphere Streams Integration

IBM InfoSphere Streams is a software platform that enables the development and running of applications that process information in data streams. Streams enables continuous and fast analysis of massive volumes of moving data to help improve the speed of business insight and decision making. The IBM InfoSphere DataStage Integration Toolkit provides operators and commands that facilitate integration between IBM InfoSphere Streams and IBM InfoSphere DataStage. To accomplish the integration, it is also required to configure InfoSphere Streams connectors.

Integration of InfoSphere DataStage and InfoSphere Streams applications involves flowing data streams between the applications and configuring them to use the data. The integration occurs through an InfoSphere Streams connector, a Streams DSSource operator, and a DSSink operator, as shown in Figure 4-11 on page 68. The DataStage job in Figure 4-11 on page 68 receives input data from a DB2 source table. A Transformer stage splits the data into two sets, the first set of rows are sent to a Streams application through a DataStage Streams connector, which communicates through TCP/IP with a DSSource operator in the Streams application. The Streams application processes the requests and sends results through a DSSink operator to the second Streams connector in the job. The Merge stage merges the data from the second output link of the Transformer stage with the result generated by the Streams application. The final result is then written to a Netezza table.

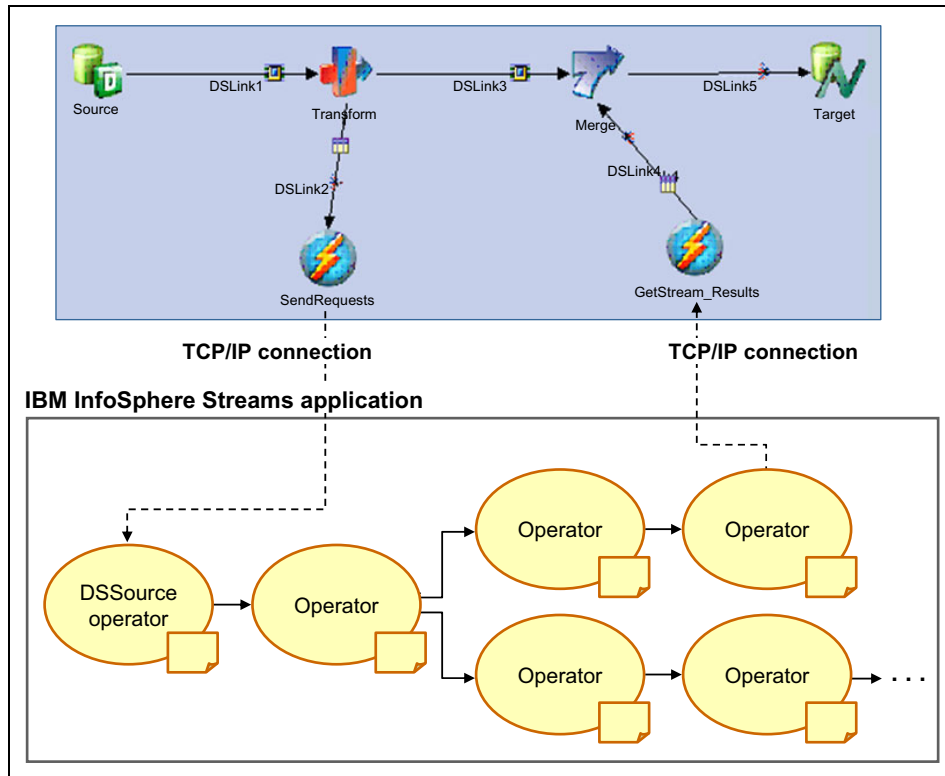


Figure 4-11 Integration of InfoSphere DataStage and InfoSphere Streams applications

InfoSphere Streams integration is a tightly coupled application integration that creates a combined application paradigm for InfoSphere DataStage and InfoSphere Streams.

4.5 Oozie Workflow Activity stage

Oozie is a workflow system that can be used to manage Hadoop jobs. Oozie workflows are a collection of actions that are arranged in a control dependency. These actions are computation tasks that are written in MapReduce, or other frameworks. They are used to write applications to process large amounts of data. The DataStage Oozie Workflow Activity stage enables integration between Oozie and InfoSphere DataStage. An InfoSphere DataStage Sequence job contains activities, which are special stages that indicate the actions that occur when the sequence job runs. The Oozie Workflow Activity stage is used to start Oozie workflows from a DataStage Sequence job.

Figure 4-12 shows a DataStage Sequence job with two activity stages. The first activity is a Job activity that starts a DataStage job that is named `Data_Gen_MultiOrders`, while the second activity is an Oozie Workflow activity that starts an Oozie workflow that is named `hdfs://myserver:9000/user/applications/DEMO/workflow`, which is specified in the Application Path field in the Oozie Workflow activity stage, as shown in Figure 4-13 on page 70. The application path points to where the workflow definition file is located. A workflow definition file is a programmatic description of a workflow in XML format.

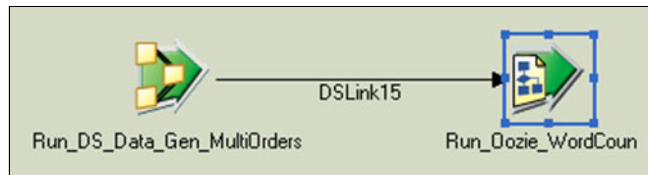


Figure 4-12 DataStage sequence job

In addition to the workflow definition file, the Oozie Workflow Activity stage contains fields for specifying Oozie server connection properties, workflow parameters and values, and some other control properties. For example, the `http://myserver:8280/oozie` URL specifies the target Oozie server, where `myserver` is the name of the Oozie server. The port number 8280 is used for InfoSphere BigInsights, but differs depending on the target Hadoop system.

Figure 4-13 shows the Oozie Workflow activity stage properties.

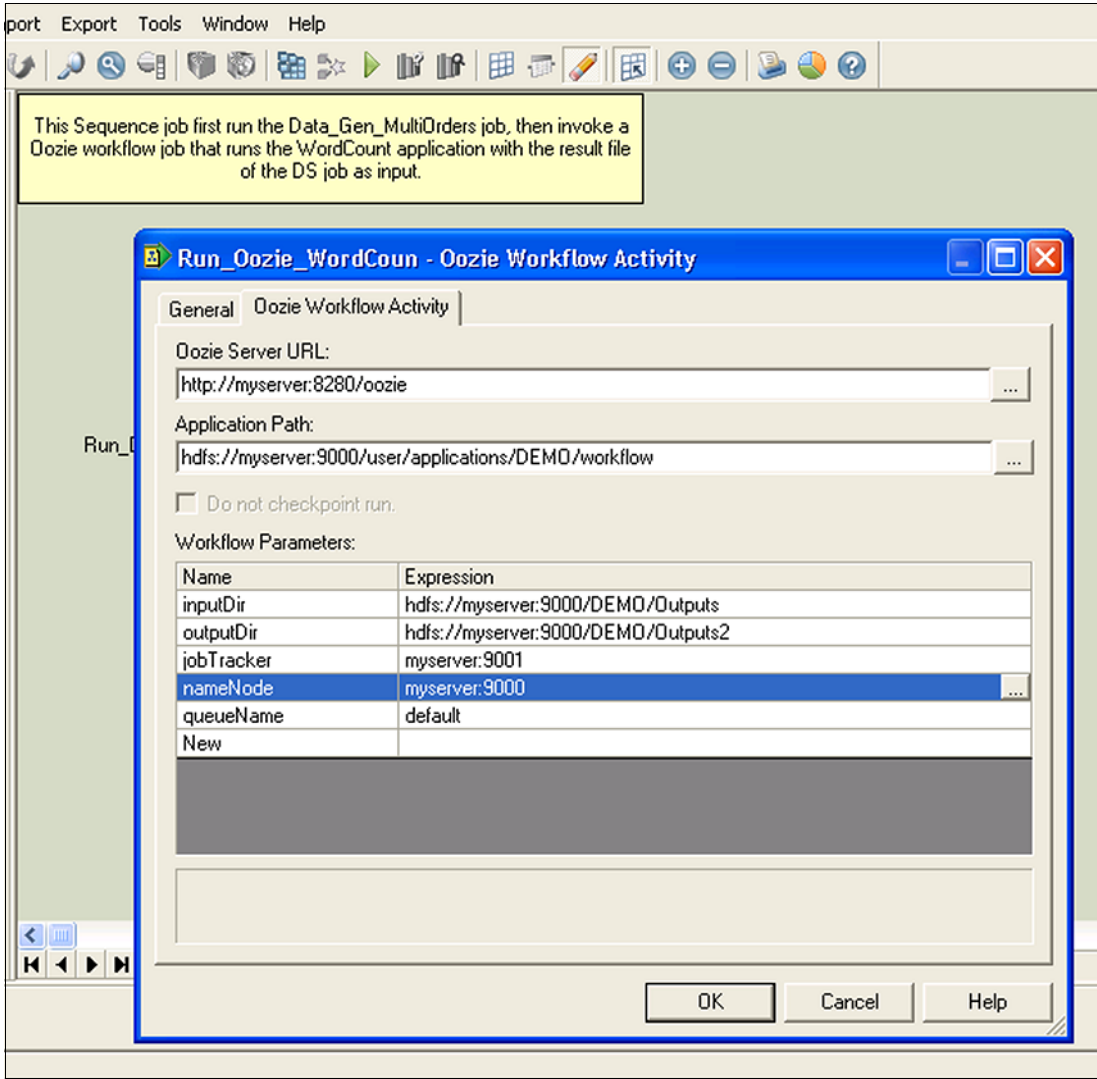


Figure 4-13 Oozie Workflow activity stage

Figure 4-14 shows the Job activity stage properties.

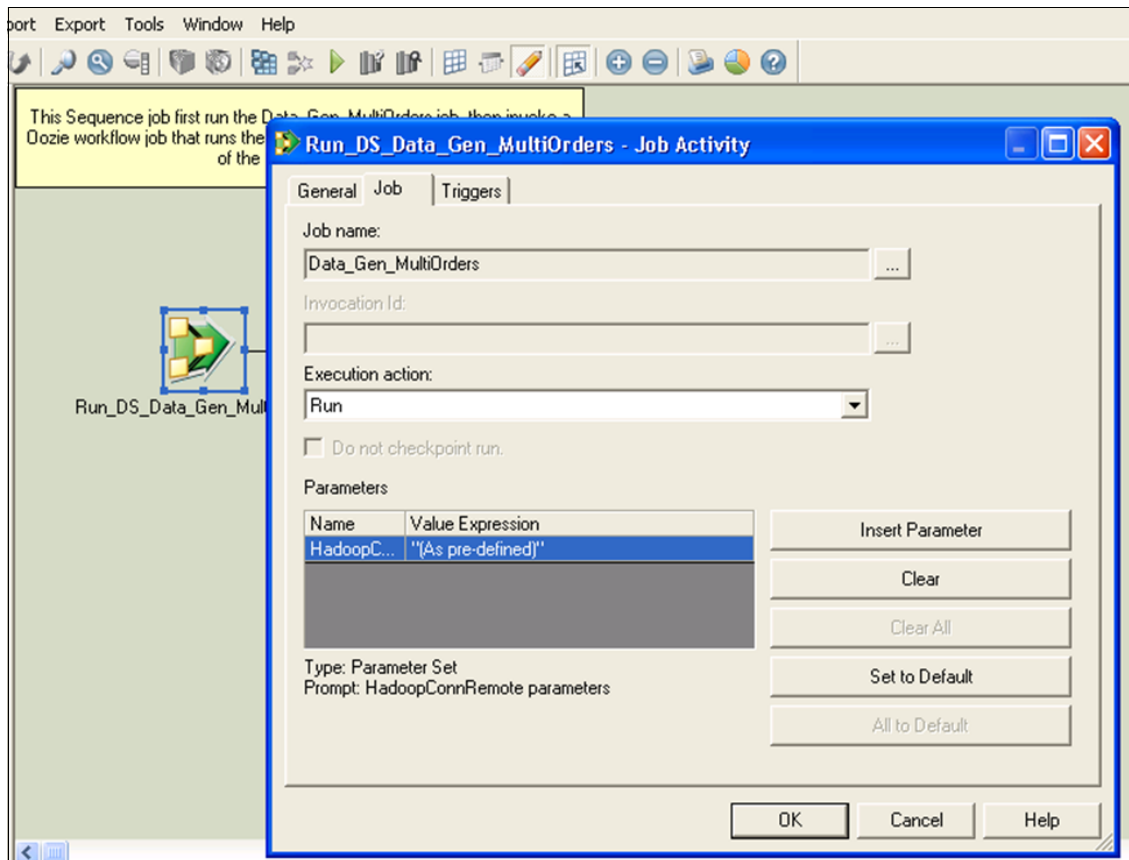


Figure 4-14 Job activity stage properties

The Oozie Workflow activity stage integrates Hadoop jobs with DataStage jobs. Data sharing can be accomplished through shared HDFS files or local files. For example, the DataStage job that is started by the Sequence job in the previous example might be a BDFS stage that writes data to an HDFS file. The Hadoop job that is run by the Oozie workflow then reads data from the HDFS file, processes the data row by row, and places the result in another HDFS file which can be accessed by other DataStage jobs. The results demonstrate the return on data, as shown in Figure 4-15 on page 72.



Figure 4-15 Return on data

4.6 Unlocking big data

IBM InfoSphere Information Server provides a comprehensive integration with big data sources. The BDFS stage allows DataStage jobs to read/write files from and to Hadoop systems. Oozie Workflow Activity stage further integrates Hadoop jobs with DataStage jobs in a seamless fashion. InfoSphere DataStage has full access to big data and Hadoop systems can make full use of InfoSphere Information Server. In addition to the bidirectional support, the InfoSphere Streams integration provides other support for real-time, low-latency analytics processing. Application integration that involved big data and traditional Data Warehouse is possible by the closed-loop integration framework. Moreover, as BalOp pushes data processing logic into Hadoop systems, DataStage applications get the full benefits of Hadoop's fault-tolerant massive distributed processing.



SPSS: Incorporating Analytical Models into your warehouse environment

Predictive Analytics is emerging as an important trend in Business Intelligence as enterprises seek to gain new business insights and value from their Data Warehouses. Increasingly, businesses are seeking to obtain valuable information from their different sources as quickly as possible to provide real-time responsiveness and promotions to their customers. To do this, they employ analytic capabilities such as the IBM SPSS® analytics suite, which is one of the foremost offerings for statistical and predictive analytics. The IBM SPSS products provide a comprehensive and easy-to-use set of statistical modeling, data mining, and decision support tools for business users, analysts, and statisticians. These products help to build statistical or data mining models to predict various business parameters such as the likelihood of a customer moving to the competition and customer segmentation.

In this chapter, we explain how IBM SPSS model prediction capabilities can be integrated with IBM InfoSphere Information Server's transformation capabilities to improve the real-time responsiveness of SPSS model predictions. IBM InfoSphere Information Server provides a set of robust extension capabilities by which the analytic models that are generated by IBM SPSS products can be started.

This chapter includes the following topics:

- ▶ Analytics background
- ▶ Motivating examples
- ▶ End-to-end flow
- ▶ Integrating IBM SPSS Models with external applications
- ▶ Building SPSS Stage in IBM InfoSphere DataStage
- ▶ Summary

5.1 Analytics background

A typical pattern for predictive analytics today is to extract data from data warehouses into a separate data mart and then apply the predictive models to obtain valuable insights. The results of the analytics are then fed to decision makers or back into operational systems. A key characteristic of running analytics software in such a manner is that it is a batch operation where the analytic model is built once and it is applied on large amounts of data in batch.

The main disadvantage of today's approach is that data is read and transformed multiple times before it is used by the end application once while it is loading the data into the data warehouse or data mart and later when the data is extracted from the data marts for processing in the analytic models.

A more efficient method of performing this type of end-to-end operation is to integrate the process of running the analytic models during the import (or export) of new data into (from) the warehouse. For this to be possible, there must be a mechanism to start the SPSS model from within the context of an InfoSphere DataStage job. By doing this, the analytical model can be applied on the data that is ingested into the warehouse or mart and the output can be stored directly into the resulting tables. Once the output of the statistical model is available in the data warehouse or data mart, business applications such as reporting tools and marketing campaigns can make use of this data readily without the need for a separate analytic step.

5.2 Motivating examples

To better appreciate the use and value of this approach to analytics, a few examples are described in this section.

5.2.1 A banking example

Consider a bank that wants to achieve more profitable results from marketing campaigns by matching the right offer to each customer. The bank wants to identify whether a customer is likely to respond favorably to a given marketing campaign so that it can generate a targeted mailing list of customers. The bank sends the specific marketing campaign material to this subset of customers that receive a better response to the campaign at reduced costs. Such a mechanism also improves the customer experience because they receive promotional campaign material that is of interest to them.

To ensure that the right offer is generated for the right customer, the bank wants to use key indicator characteristics of the customer such as the following:

- ▶ Response of the customer to previous campaigns
- ▶ Income level of the customer
- ▶ Number of transactions done by the customer per month

This information can be fed to an SPSS model which can then predict the propensity of a customer to respond favorably to a marketing campaign.

In a typical business intelligence environment, the data is first loaded into a data warehouse or data mart by using an ETL tool such as IBM InfoSphere DataStage and then it is extracted and processed by using IBM SPSS Analytics. The analytic model performs the prediction and the output of the prediction is stored into a target table. The information that is required by the IBM SPSS Analytics model is available to the ETL tool when the data is loaded into the warehouse. Ideally, if the SPSS model can be started from within the IBM InfoSphere DataStage job, the SPSS model can be started in the same job flow and populate the likelihood of the customer responding favorably to the marketing campaign directly into the data warehouse or data mart. Thus, the overall latency of the analytics can be reduced and the speed of targeted campaigns improved. You can readily make use of the data that is present in the warehouse to run the marketing campaign.

5.2.2 A telecom example

Consider another example from the telecom industry, where a company wants to predict the usage pattern of new customers so that it can generate the right offer for prospective customers. The company segments its customers by service usage patterns and categorizes them into the following groups:

- ▶ Basic Service
- ▶ E-Service
- ▶ Plus Service
- ▶ Total Service

The features and cost of each of the categories differs (Total Service indicates high cost and features while Basic Service incurs low cost and features). Proposing a Total Service to a price-sensitive customer likely leads to the customer moving to the competition, whereas proposing a Basic service to a high net worth individual leads to a dissatisfied customer and reduced profit for the company. Hence, it is important that the right service offer is made to the prospective customers.

Given a set of prospective customers, the company wants to make use of demographic data to categorize them into one of the selected categories so that individual offers can be made for each of them. Enterprises make use of IBM SPSS products such as Data Modeler to perform this type of categorization. They can build an IBM SPSS Analytics model, which takes as input the demographic information of the customer and predicts the category or segment to which the customer is likely to belong.

As in the previous example, the IBM SPSS Analytics model today is started on data that is in the data warehouse or data mart, which leads to reading and writing of the data multiple times. The data warehouse or data mart in this example consists of data about prospective customers. This information typically is stored in the marketing department's prospects data warehouse or data mart from where it is used to generate different offers. Data is loaded into the data warehouse or data mart by using IBM InfoSphere DataStage. If during this loading process the SPSS model is started on the data, it leads to a reduction in the overall process time. The category or segment then is stored in the data warehouse or data mart as a separate field. The business processes in the marketing department then make use of this field to generate the right offer for the prospective customers.

5.2.3 A customer care example

Consider an example from the customer care domain. Customers call in to the call center of an e-commerce company to check the status of their orders and complain about their services. To appease customers, the company provides different kinds of coupons to them. It is important to provide the right coupon to the right customer so that the customer is satisfied. For example, a customer who has a history of buying games should be offered a coupon of the latest gaming console. Similarly, a customer who has a history of buying cook books should be offered a coupon for grocery products. At the same time, not every customer must be offered a coupon. As an example, only those customers who are likely to migrate to competition should be provided one.

In this example, the following types of analytics can be performed:

- ▶ Identification of customers who should be offered coupons
- ▶ Identification of the right coupon for each customer

For the first case, models can be built based on the characteristics of customers who moved to the competition. For the second case, association rules mining can be performed by using IBM SPSS, which helps to identify the likelihood of a customer buying a product based on the other products that the customer bought in the past. Thus, the IBM SPSS Analytics model must be run on the transaction database to identify the two characteristics. This information is then stored in the data warehouse or data mart and is made available to the call center agent so that they can offer the coupon to the customer in case they are dissatisfied.

5.3 End-to-end flow

In this section, we describe the end-to-end flow for implementing the use cases that were described in 5.2, “Motivating examples” on page 75. The overall solution features the following phases:

- ▶ Model Building Phase: Consists of model building by using IBM SPSS Modeler.
- ▶ Model Scoring and Application Phase: Consists of model scoring by using IBM InfoSphere DataStage.

5.3.1 Model building by using IBM SPSS Modeler

The first step in the use cases is to build an analytic model by using IBM SPSS Modeler. In the first example, the model takes as input the characteristics of the customer and predicts the propensity of a customer to respond favorably to a marketing campaign.

In the second example, the model takes the demographic information of a prospective customer as input and predicts the category of the customer (from one of four categories).

In the final example, the model takes the transaction history of the customer and predicts the set of products that customer is likely to be interested in so that a coupon for the product can be offered to the customer.

These models can be built by using IBM SPSS Modeler. IBM SPSS Modeler is a data mining workbench that helps customers build predictive models quickly and intuitively, without programming. It helps to discover hidden relationships in structured data that is stored in databases, mainframe data systems, flat files, or within an IBM Cognos Business Intelligence environment and predict the outcome of future events and interactions.

Modeler's graphical interface puts the power of data mining in the hands of business users to discover new insight and increase productivity, which allows the organization to quickly realize a positive return on investment.

After the model is built, it must be started from within a DataStage job.

5.3.2 Model scoring within IBM InfoSphere DataStage

IBM InfoSphere DataStage is widely used to load data into data warehouses and data marts. It provides various custom operator mechanisms to start external applications from within IBM InfoSphere DataStage. To better understand this functionality, consider a simple example where there is an external Java application that takes two strings as input and concatenates them to produce the output. If this application is to be started from within a DataStage job, you can make use of the Java Integration Stage, which is part of the InfoSphere DataStage 9.1 release. For more information, see the online documentation that is available at the IBM InfoSphere Information Server Version 9.1 Information Center at this website:

http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp?topic=%2Fcom.ibm.swg.im.iis.productization.iisinfsv.home.doc%2Ftopics%2Fic_homepage_IS.html

By using the Java Integration Stage, you can integrate the external Java application and start it from a stage in IBM InfoSphere DataStage. Such a stage takes two string (varchar) columns as input and generates a single column as the output. The output column contains the string that is formed by concatenating the two strings that are received in the two input columns.

In section 5.2, “Motivating examples” on page 75, the external application is the runtime scoring component of IBM SPSS Modeler. You can build a custom stage (SPSS Stage) which takes the name of the IBM SPSS Analytics model and the necessary data that is required as input by the model. The data then is sent to IBM SPSS Modeler Runtime component with the model name, which is started on the data. IBM SPSS Modeler Runtime component, in turn, starts the model on the provided data and generates the prediction. This predicted data and the input data are returned for downstream processing.

Thus, in the first motivating example (5.2.1, “A banking example” on page 75), the SPSS stage takes as input the characteristics of the customer (along with the model name) and generates the likelihood of the customer responding positively to the marketing campaign as output. This output is in the form of a boolean column. The output of the stage is the original input data and this other boolean column.

In the second motivating example (5.2.2, “A telecom example” on page 76), the input to the SPSS stage is the demographic information (along with the model name) and the output is a category column that contains the predicted category for the customer.

In the final motivating example (5.2.3, “A customer care example” on page 77), the transaction history of the customer forms the input to the SPSS stage (along with the model name) and the output is a set of products that the customer is likely to buy.

In summary, to perform analytics from within an IBM InfoSphere DataStage job, you first must build a model by using the IBM SPSS Modeler and then start it by building a custom stage in IBM InfoSphere DataStage. In the remaining sections of this chapter, we provide more information about the interface between IBM SPSS Modeler Runtime and IBM InfoSphere DataStage.

5.4 Integrating IBM SPSS Models with external applications

In this section, we provide an overview of the features that are provided by IBM SPSS Modeler for interfacing with external applications. For more information, the same, see this website:

<http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp>

The integration of IBM SPSS Modeler with IBM InfoSphere DataStage can be accomplished by using IBM SPSS Modeler Solution Publisher. IBM SPSS Modeler Solution Publisher is a powerful tool for integrating data mining results into a business process to solve real-world problems. By using IBM SPSS Modeler Solution Publisher, you can create a packaged version of a stream that can be embedded in an external application or run by an external Runtime engine (such as IBM InfoSphere DataStage; the stream in SPSS is the equivalent of a job in IBM InfoSphere DataStage). This packaged version enables the deployment of data modeling streams into a production environment to support everyday business processes and to empower the organization’s decision makers with the knowledge that is gained from mining the data.

Deploying a solution by using IBM SPSS Modeler Solution Publisher involves the following phases:

1. Publishing a stream
2. Running a stream

These phases are described in the following sections.

5.4.1 Publishing a Stream

The first step is to build a model by using IBM SPSS Modeler that provides a good solution to the business problem (such as finding the likelihood of a customer responding positively to a marketing campaign). The Stream then must be published by using IBM SPSS Modeler Solution Publisher, which creates a detailed description of the Stream on the disk (as image, parameter, and metadata files, which are described next).

Publishing Streams is done directly from IBM SPSS Modeler by using one of the following export methods as target nodes:

- ▶ Database
- ▶ Flat File
- ▶ Statistics Export
- ▶ IBM SPSS Data Collection Export
- ▶ SAS Export
- ▶ Microsoft Excel
- ▶ Microsoft XML

The type of export node determines the format of the results to be written when the published Stream is run by using the IBM SPSS Modeler Solution Publisher Runtime or external application. For example, the use of a Database export node implies that the results are written to a database each time the published Stream is run. In our use case, we want the results to be sent downstream in the IBM InfoSphere DataStage job. (We explain how this is done later in this section.) When a Stream is published by using IBM SPSS Modeler, it creates the following files:

- ▶ Image File: The image file (*.pim) provides all of the information that is needed for the Runtime to run the published Stream exactly as it was at the time of export.
- ▶ Parameter File: The parameter file (*.par) contains configurable information about data sources, output files, and run options.
- ▶ Metadata File: The metadata file (*.xml) describes the inputs and outputs of the image and their data models. It is for use by applications that embed the runtime library and must know the structure of the input and output data.

After the model is built and Stream published, as shown in Figure 5-1, the next step is to use that model and start it from within an IBM InfoSphere DataStage job.



Figure 5-1 Publishing SPSS stream

5.4.2 Running a Stream

After the Stream is published, the process that is implemented in the Stream can be rerun by running the published Stream. The re-running of the Stream (as described in 5.2.2, "A telecom example" on page 76), consists of finding the category or segment of a customer given the demographic information. This is done by using the stand-alone IBM SPSS Modeler Runtime (modelerrun.exe) or by developing an application that uses the SPSS Modeler Runtime Library to run the Stream. It is more advantageous to use the SPSS Modeler Runtime Library for integration with IBM InfoSphere DataStage because it is specifically designed for integration of IBM SPSS Modeler with third-party applications. In this case, the application is the new SPSS stage in IBM InfoSphere DataStage.

To run the Streams outside of SPSS Modeler (by using the Runtime or a custom application), the IBM SPSS Modeler Solution Publisher Runtime must be installed on the machine where the application is running. This implies that the IBM SPSS Modeler Solution Publisher Runtime must be installed on all of the machines on which the model scoring must take place from within IBM InfoSphere DataStage. This might be a subset of the nodes where IBM InfoSphere DataStage is deployed.

IBM SPSS Modeler Solution Publisher provides a Runtime programming library (CLEMRTL) that other programs (such as IBM InfoSphere DataStage) can embed and use to control running the Stream. The CLEMRTL procedures can be called from client programs that are written in C and C++. Hence, the SPSS Stage in DataStage must be built in C/C++.

One key element that is missing in this description is a mechanism to ensure that the Stream can read data that is available on the input link of the SPSS stage and the results of running the Stream are sent to the output link of the stage. Recall that when the Stream is published, it is configured to read and write data from files and databases. You should override that setting to use the Stream in IBM InfoSphere DataStage. This can be accomplished by using the CLEMRTL library. It provides two APIs for configuration of inputs and outputs, viz., `setAlternativeInput` and `setAlternativeOutput`. The `setAlternativeInput` API replaces a file input source with an alternative input source whereas the `setAlternativeOutput` API replaces a file output target with an alternative output target. These APIs take an iterator as input. The iterator that is provided to the `setAlternativeInput` method produces the alternative input data. It is called once for each input record. Similarly, the iterator for the `setAlternativeOutput` method uses the image output. It is also called once for each result row that is produced by the image. By using these iterators, you can read data from the input link and write the results to the output link of the IBM InfoSphere DataStage job.

In this section, we described the features that are provided by IBM SPSS to interface with external application. In the next section, we describe the steps that are required for building the SPSS Stage in IBM InfoSphere DataStage.

5.5 Building SPSS Stage in IBM InfoSphere DataStage

There are two major components in building the SPSS Stage within IBM InfoSphere DataStage. The first is generic and involves the steps that are required for extending IBM InfoSphere DataStage to build a custom stage, as described in 5.5.1, “Extending IBM InfoSphere DataStage” on page 84. The features that are supported by the SPSS stage and techniques to implement them are described in 5.5.2, “Other features of SPSS stage” on page 86.

5.5.1 Extending IBM InfoSphere DataStage

As described 5.4.2, “Running a Stream” on page 82, the SPSS stage must be built in C or C++. You use of the Build Stage mechanism to create the SPSS stage. There are multiple mechanisms available for creating stages in IBM InfoSphere DataStage. For more information, see this website:

http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp?topic=%2Fcom.ibm.swg.im.iis.ds.design.doc%2Ftopics%2Fc_ddesref_Custom_Stages_for_Parallel_Jobs.html

When a Build stage is defined, you must provide the following information, as shown in Figure 5-2 on page 85:

- ▶ Description of the data that is input to the stage.
- ▶ Whether records are transferred from input to output. A transfer copies the input record to the output buffer. If you specify auto transfer, the operator transfers the input record to the output area immediately after running the per record code. The code still can access data in the output buffer until it is actually written.
- ▶ Any definitions and header file information that must be included.
- ▶ Code that is run at the beginning of the stage (before any records are processed).
- ▶ Code that is run at the end of the stage (after all records are processed).
- ▶ Code that is run every time the stage processes a record.
- ▶ Compilation and build details for actually building the stage.

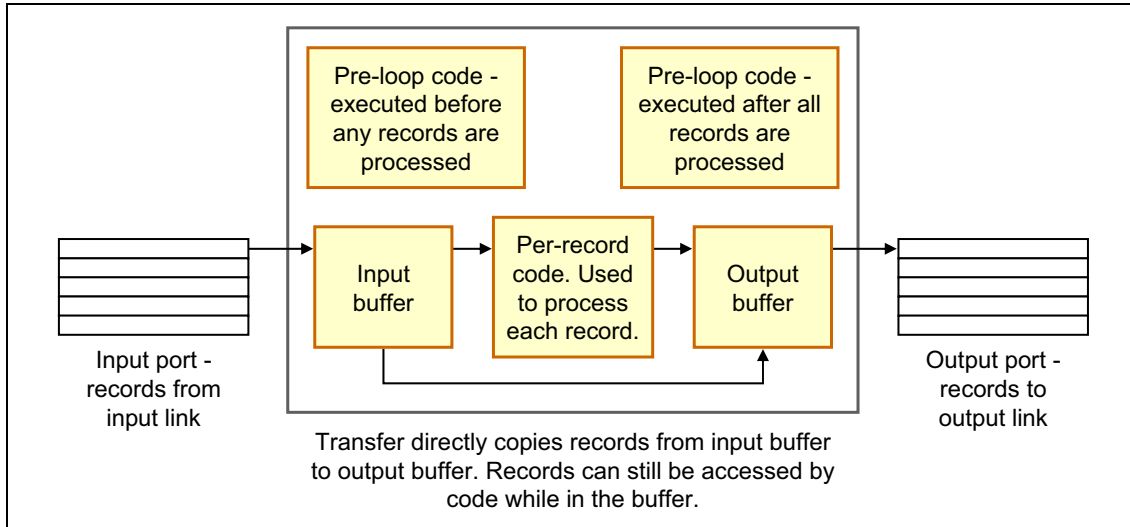


Figure 5-2 Components of a Build Stage

Out of the previous list, the most crucial item is the Per record code, which is used to process each record. In this code, the CLEMRTL library is used to run the SPSS Stream. The stage takes the image and parameter file name and its location as input and uses it to run the SPSS Stream. The stage makes use of the `setAlternativeInput` and `setAlternativeOutput` methods to specify input and output iterators. The iterator that is used by the `setAlternativeInput` reads data from the input buffer, whereas the iterator that is used by the `setAlternativeOutput` writes data to the output buffer. This ensures that data from the input link is used for running the Stream and that the output of the Stream is sent to the output link, as shown in Figure 5-3.

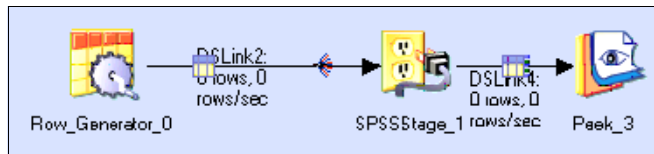


Figure 5-3 Sample IBM InfoSphere DataStage job with SPSS Stage

5.5.2 Other features of SPSS stage

In this section, we describe the following features that can be supported by the SPSS Stage:

- Mapping of Input Columns to Stream Columns

An IBM SPSS Analytics Stream processes data that conforms to a specific schema and generates data that conforms to the output schema. This schema information is present in the metadata file that is generated when the Stream is published. When the SPSS stage is used in a job, the user must specify the schema based on the metadata present in the metadata file. The per record code must use the metadata file and the metadata of the input columns to map the input columns to the columns that are expected by the Stream. This can be done based on the column names and the data types.

- Parallel Execution of Stage

Each record that is input to the IBM SPSS Analytics Stream is processed independently by the Stream from the other records. Hence, the SPSS stage can be easily parallelized. We can make use of any of the partitioning methods that are supported by IBM InfoSphere DataStage.

- Support for Reject Links

IBM InfoSphere DataStage supports the concept of a reject link whereby the data that fails processing in a stage is sent to the reject link. This helps to provide a mechanism to analyze the rejected data to correct the input data and fix issues. It is possible to support a reject link in the SPSS Stage to identify records that caused an error while running the IBM SPSS Model Stream. This reject functionality can be implemented by using the following features of the CLEMRTL library:

- The SPSS stage code can register an error handler that is called whenever an error occurs. The error handler can be registered by using the `clemrtl_setReportHandler()` function.
- Whenever an error occurs while a Stream is run, the SPSS stage can make use of the `clemrtl_getErrorDetail()` function to get the details of the error message.

Thus, by using the `clemrtl_setReportHandler()` function, we can register an error handler that starts the `clemrtl_getErrorDetail()` function to get further details of the error and send the error records to the reject link.

5.6 Summary

We described how the various features of IBM InfoSphere DataStage and IBM SPSS Modeler can be used to run an SPSS Stream from within an IBM InfoSphere DataStage job as part of a batch run. We also described how a custom SPSS stage can be developed for running a particular analytic model that was developed in IBM SPSS Modeler. The custom SPSS stage can be easily adapted to run any of the SPSS Modeler Runtime image files with customization of inputs and outputs as necessary. The benefit of the SPSS stage is that it can provide on demand scoring of the input data records as they are loaded into a warehouse or when they are extracted from the warehouse for downstream processing. This avoids the multiple read and writes that are required today before the output of analytics can be used by the end applications.



Governance of data warehouse information

With the rapid increase in the volume and variety of global information (and the increasing demand by organizations to use this information for effective competitive advantage), an organization's information becomes one of its key assets. However, and at the same time particularly because of the advent of big data (as examples, data from social media, logs, sensors, and so forth), this huge volume and variety of information becomes increasingly challenging for an organization to manage, control, and ultimately govern.

Big data is a driver for new data in, and in association with, the data warehouse, and the governance framework to support it. There is a need to quickly separate the interesting data from the un-interesting data in the many large or streaming data sources, and for that need governance policies and terms are an enabler.

With government regulations on the rise, from Sarbanes-Oxley (SOX) in the United States to the equivalent European Sarbanes-Oxley and the Japanese Financial Instruments and Exchange Law (commonly referred to as J-SOX), poor governance of information can have an enormous impact, including hefty fines, damaged reputations, and declining market share.

At the same time, innovative competition and global expansion created a critical need for trusted, relevant information. Now, more than ever, business decisions must be informed, and the quality of that information is a critical element in effective decision making.

This chapter includes the following topics:

- ▶ Information and expectations
- ▶ Information Governance: The Maturity Model
- ▶ Business terms: Enablers of awareness and communication
- ▶ Information Governance policies and rules
- ▶ Information stewardship
- ▶ Information Governance for the data warehouse
- ▶ Conclusion

6.1 Information and expectations

Expectations regarding the access, care, and use of a company's information assets are rising to unprecedented levels. For example, companies are expected to maintain complete, accurate information about their customers and the history of their relationship. They are expected to be able to share in the management of their data, and breaches of trust are not easily forgiven. As a result, reputations and revenues suffer the consequences¹.

6.1.1 Business drivers

Business drivers for investment depend on effective information that is governed and understood whether they are performing the following tasks:

- ▶ Empowering risk and compliance initiatives with the information they require.
- ▶ Optimizing revenue opportunities by ensuring effective and efficient interactions with customers, partners, and suppliers.
- ▶ Enabling collaborative business processes with consistent and trustworthy information.
- ▶ Reducing the total cost of ownership for maintaining consistent information across the enterprise.

As examples, consider the following business drivers:

- ▶ Chief Marketing Officers (CMOs) are looking to reduce customer churn, identify optimal sales channels, and improve the customer experience with the organization.
- ▶ Chief Financial Officers (CFOs) are looking to address the following challenges:
 - Risk-adjusted forecasting and risk-based resource allocation
 - Better financial risk management (as examples, market, credit, and liquidity risk)
 - Regulatory requirements around financial reporting (for example, SOX) or fulfilling compliance obligations (for example, Anti Money Laundering rules and regulations)

¹ IBM Data Quality Management solutions: Building opportunity, managing risk, pg.2, <http://ftp.software.ibm.com/software//data/sw-library/infosphere/whitepapers/LIW11882-US-EN-00.pdf>, IBM, ©2007.

- ▶ Chief Information Officers (CIOs) must address the following related challenges:
 - Ensuring regulatory compliance
 - Reporting on risk exposure against business objectives
 - Reducing risk exposure particularly to data access or privacy breaches

6.1.2 Using the information

User organizations often state that they need to do a better job using information. This problem manifests itself in many ways. As examples, it is sometimes referred to as information complexity, or a deluge of information. The primary issue is that a great deal of valuable information is locked away in various databases and systems throughout the business. The organization has no easy way to use this information to improve the business, compete more effectively, or to innovate.

For example, retail companies might be unable to use demand signals from their stores effectively to drive their supply chains. Across all industries, it is common to find that organizations are not using customer analysis to tailor their marketing and sales activities. In other cases, entire classes of information (such as free-form text fields) are being ignored because they seem to be too difficult and expensive to deal with.

Another information issue for many organizations is that they have multiple versions of the truth. That is, multiple versions of data accuracy regarding customers, products, and issues across their various systems. This prevents them from completely understanding their customers and tailoring their interactions accordingly.

It also leads to supply chain collaboration problems because suppliers and customers have differing concepts and definitions of products. It also causes difficulties when trying to comply with information-centric regulations, such as Sarbanes-Oxley or Basel II, which require definitive information with associated proof.

The Information Warehouse, as one of the central hubs and sources of organizational information, is heavily affected by these challenges. As examples, consider the following questions:

- ▶ Is customer demographic or geographic data available for new analysis to drive business models?
- ▶ Are financials appropriately rolled up for reporting?
- ▶ Is information about customers secured?
- ▶ Can data be obtained by unauthorized individuals?

As noted in the IBM Data Governance Maturity Model, “organizations must find a way to govern data in alignment with business requirements without obstructing the free flow of information and innovation.”²

6.2 Information Governance: The Maturity Model

Information Governance is a cross-cutting concern that complements other governance programs that are system or process focused. The Information Governance model, as shown in Figure 6-1 on page 94, shows the components that are needed to develop and maintain trusted information.³ It is centered on the notion of an information asset. An information asset is a piece of information that is important to the workings of the organization; for examples the details about a customer, or an investment, or a product description.

Information assets can be characterized according to the type of information they represent. Such characterizations are called *information asset types*. For example, the information asset type for customer details might define that customer details include the person’s full name, contact information, the products they bought, their loyalty status, and their credit status. The contact information might be broken down further into home address, phone number, and email address. The definition also includes a description of the valid values for these fields.

Information Governance was initially called *Data Governance* and these two terms are often used interchangeably. Be aware that there are other aspects to and domains of the governance of a business that are outside the term Information Governance, so we do qualify the term *governance* as specific to the Information Governance domain. Throughout this publication, we use the term Information Governance unless we are referring to a specific group (for example, the IBM Data Governance Council) or a specific work (for example, the IBM Data Governance Maturity Model).

² The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance, pg.3,
https://www-935.ibm.com/services/uk/cio/pdf/leverage_wp_data_gov_council_maturity_model.pdf, IBM, ©2007.

³ IBM Information Governance Model, version 10, IBM white paper by Mandy Chessell, pg.1, IBM, ©2010.

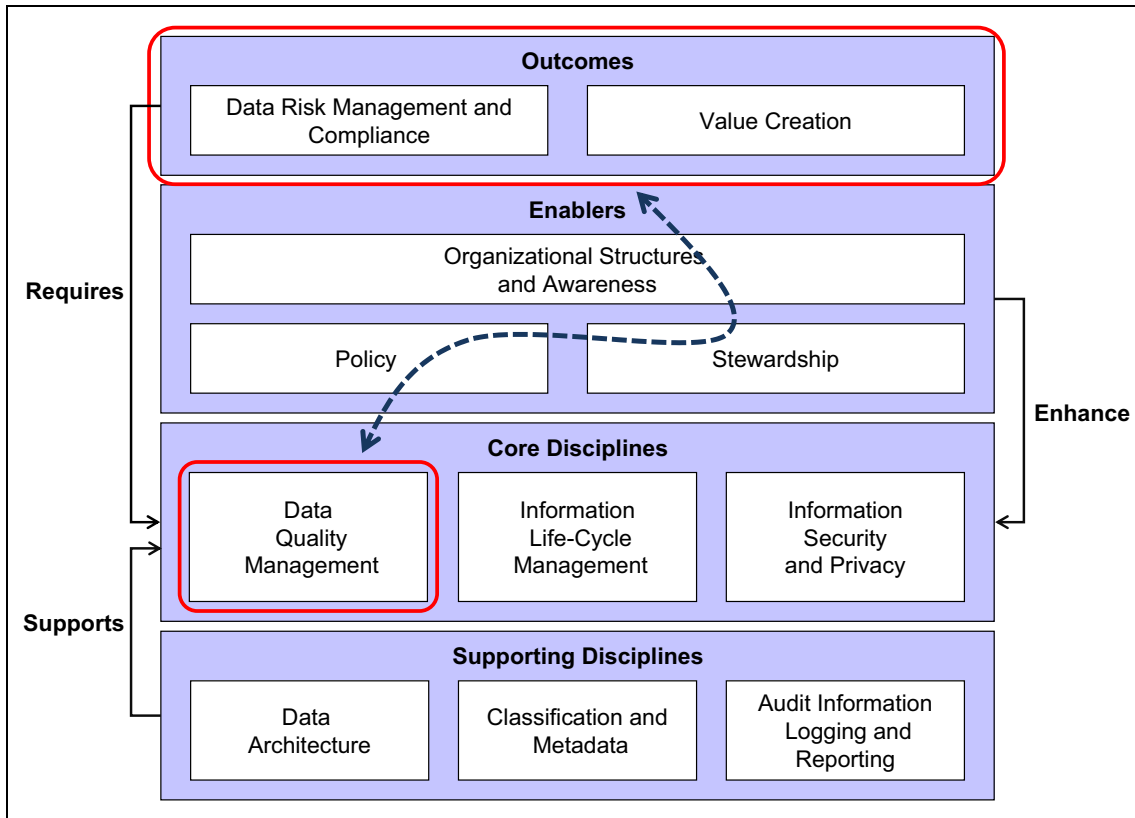


Figure 6-1 Elements of Effective Information Governance⁴

⁴ The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance, pg.8, https://www-935.ibm.com/services/uk/cio/pdf/leverage_wp_data_gov_council_maturity_model.pdf, IBM, ©2007.

6.2.1 Elements of an Information Governance Maturity Model

The IBM Data Governance Maturity Model highlights three core disciplines that are needed for effective Information Governance: Data quality management, Information Lifecycle management, and Information Security and Privacy. These disciplines not only require, but are driven by, business outcomes in the form of risk management and compliance or value creation. To achieve these business outcomes, there must be enablers in the form of organizational structures and awareness, policy, and data stewards.⁵

In relation to the Data Warehouse, in this chapter we focus specifically on how IBM InfoSphere Information Server addresses data quality management through the supporting disciplines of data architecture, classification, and information reporting, which are enabled through policy and data stewardship.

At a general level, you might say that the quality of all data within the Data Warehouse must be managed and governed. Given the volume, rate of change, and variety of information that enters the data warehouse, this level of quality control is likely to be impossible or of such cost that an organization could not feasibly support it. Instead, an organization must assess the risks, compliance requirements, and where it seeks to drive value (that is, it is expected outcomes), to define what information must be managed. They then establish or incorporate policies regarding which information, and to what level, the organization should maintain and the relevant targets or key performance indicators for quality, security, and retention.

6.2.2 Business terms: The language of the business

For individuals within an organization to understand the outcomes or the policies, such information must be communicated both broadly and on multiple levels. It also must be connected to the information with which it is associated. It is not sufficient to broadcast this information, there also must be context. A key driver to understanding context is a glossary of those terms that are critical to an organization. Such terms, which are available through a readily accessible glossary, allow for all members of the organization to obtain a common reference point.

⁵ *The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance*, pg.8,
https://www-935.ibm.com/services/uk/cio/pdf/leverage_wp_data_gov_council_maturity_model.pdf, IBM, ©2007.

These terms can be said to form the language of the business. They are used to perform the following tasks:

- ▶ Define authoritative meaning
- ▶ Increase understanding throughout the enterprise
- ▶ Establish responsibility, accountability, and traceability
- ▶ Represent business hierarchies
- ▶ Document business descriptions, examples, abbreviations, and synonyms
- ▶ Find relevant information assets
- ▶ Encourage use (and reuse) of correct terminology

Terms are continuously added, modified, and, in some cases, deprecated in use or meaning. That implies that they go through their own lifecycle and must be maintained and updated by the Data Stewards that are responsible for them across the different lines of business. This process allows the language of the business to stay current.

Thus, they are an ongoing enabler in the Data Governance Maturity Model. They facilitate communications between business and IT via a common business vocabulary over time and allow the linkage of the correct business terms to correct information assets, a critical aspect to information governance for the Information Warehouse and the information flowing into and out of the data warehouse.

For example, consider the following points regarding the term *tax expense*:

- ▶ Term category: COSTS
- ▶ Longer name: Tax to be paid on gross income
- ▶ Description: “The expense due to taxes calculated as the income prior to tax times the applicable tax rate...”
- ▶ Data steward: Jane Smith (who is responsible for updates)
- ▶ Status of the term: Current

This is what the business users who are reviewing reports, running operational processes, or analyzing trends understand and reference. They expect that the information that is delivered to them in these formats aligns to that definition.

From the standpoint of the database administrator (DBA) who is working with the Information Warehouse, though, information is not organized in this way. The DBAs understand that the Information Warehouse contains the following components:

- ▶ A DB2 database with a schema named NAACCT
- ▶ A Table named DLYTRANS
- ▶ A Column named TAXVL that has a data type of Decimal (14,2)
- ▶ The column has a Derivation SUM(TRNTXAMT)

This information (for example, TAXVL) makes no sense or has no context for the business user just as the term *Tax Expense* does not appear in the context of the implemented Information Warehouse. If there is no association between the term and the column where the actual information is, there cannot be effective governance of this information.

It is important to remember that understanding is abstracted as information systems are constructed, going from the original business semantics to concept to logical model to physical model and, finally, to implemented data structures, such as the data warehouse.

6.3 Business terms: Enablers of awareness and communication

The business glossary becomes a cornerstone for such knowledge. It starts with giving business concepts or objects names and definitions that are common and agreed upon by the community of users. By giving a concept or object a name, it can be located, tracked, assigned, and secured. Having these names created according to guidelines with a discipline of approved and chartered process creates consistency and instills confidence. Users who search for information rely on the authority of such a source to provide the correct and complete information.

A business glossary has many of the attributes that are required to support Information Governance for the Information Warehouse. In essence, all of these attributes depend on the ability to create, preserve, and disseminate knowledge about information across the organization. This knowledge includes awareness about what the information is, how it is used, and who uses it. It also includes awareness about where the information is coming from, what happens to it along the way, and where it ends up.⁶

A business glossary goes beyond a list of terms. Linking terms to IT assets (such as columns in the data warehouse or fields in a business intelligence report) establishes a connection between business and IT and enhances collaboration between parties and users. General business users have fast access to commonly used vocabulary terms and their meaning, often with more information, such as any constraints and flags that indicate special treatment.

⁶ Management with IBM Information Server, by Jackie Zhu et al, IBM Redbooks, <http://www.redbooks.ibm.com/abstracts/sg247939.html?open>, IBM, ©2011

A business analyst has a better understanding of the terms that are used in business requirements, which result in better and faster translation into technical requirements and specifications. By viewing the IT assets that are assigned to a term, data analysts and developers can be more precise in their job development or report design.

Business terms (and the categories that organize them) are a way to express the following business requirements examples:

- ▶ The meaning of a business concept.
- ▶ Technical instantiation of the business concept (via Assigned Assets). For example, the database column “acc_num” relates to the term “Account Number”.

While a Glossary can be built from scratch, given the range of terms that are used in an organization and the time that is needed to achieve consensus of understanding across multiple lines of business, this is unlikely to be practical. Instead, it makes sense to draw from existing resources to build the set of terms and then focus on customization, annotation, and stewardship for those items, perhaps centered on core business units.

6.3.1 Sources of business terms

Terms can be gathered and imported from the following sources:

- ▶ Existing lists of terms:
 - Requirements to IT from business personnel are good sources of terms.
 - Spreadsheets for different projects often contain common terminology. Comma-separated value (.csv) or Extended Markup language (.xml) files drawn from these can be imported, including terms, descriptions, custom attributes, stewards, and category assignments.
- ▶ Business Intelligence (BI) Reports:
 - BI Reports are Business Language and Data: These make excellent sources for term candidates who are already familiar to the business community.
 - BI reports that were imported to the Information Server can be converted into categories and terms. These might come from a number of different BI reporting solutions.
- ▶ Logical Data Models:
 - A well-documented Logical Model is a good first source of terms; a poorly-documented Logical Model is not.

- Logical Models that were imported to the Information Server can be converted into categories and terms.
 - It supports the transformation of dependencies from IBM InfoSphere Data Architect model assets on InfoSphere Data Architect Glossary Words to Assignments on Business Glossary InfoSphere Data Architect Eclipse Terms.
 - This enables the interchange of Asset Assignments between InfoSphere Data Architect model elements and Information Server physical schemas.
- Industry Models and Warehouse Packs
- If there is no existing data dictionary, IBM Industry Models or Warehouse Packs are a good start, as shown in Figure 6-2 on page 100:
- Contains authoritative, comprehensive terms and definitions.
 - Also includes data models for data warehousing.
 - Includes ready-for-use glossaries with relationships to the IT assets in the data warehouse models.

Do not publish a ready-for-use Industry Models glossary as-is to your enterprise because these models are large and likely contain more terms than are relevant to your organization.

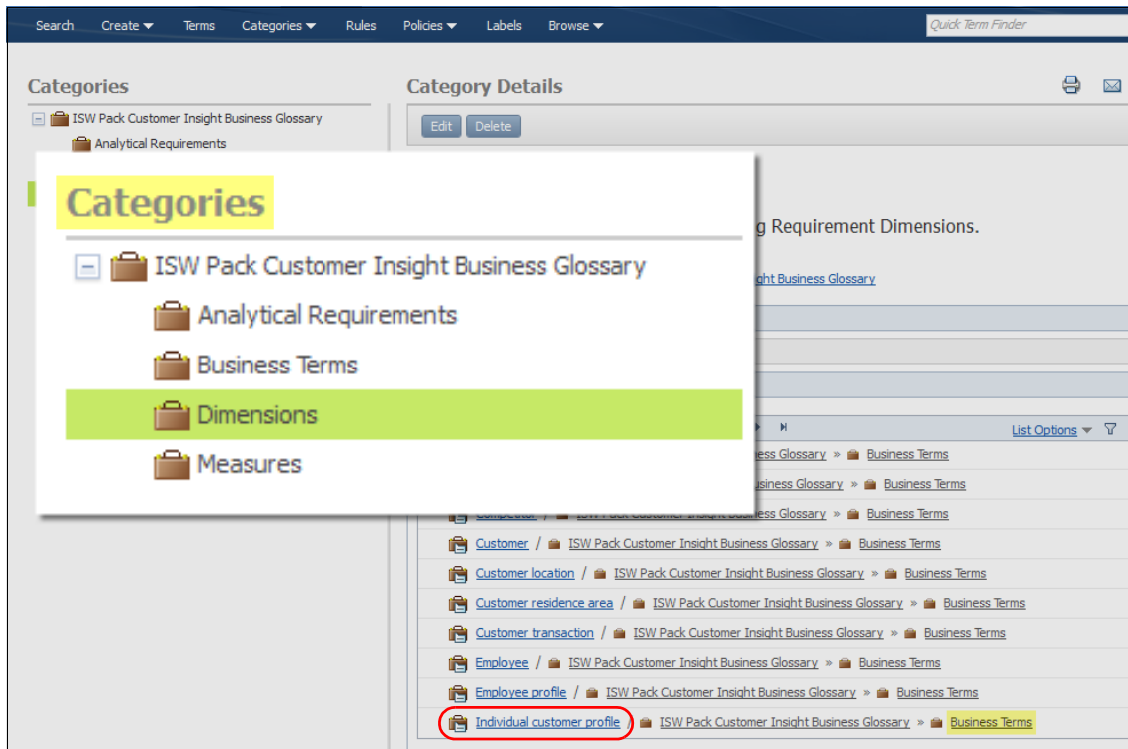


Figure 6-2 Categories and terms from the InfoSphere Warehouse Pack for Customer Insight

6.3.2 Standard practices in glossary development and deployment

From an Information Governance standpoint, terms for a Business Glossary should adhere to some of the following standard practices in development and deployment:

- ▶ Start small and focused, around 200 terms maximum.
- ▶ Learn about available relationships; establish standard methods for expressing meaning.
- ▶ Start with semantic business definitions (IT semantics and assignment to technical assets should be added later).
- ▶ Concentrate on key high-value information areas, which are driven by business outcomes.
- ▶ Take advantage of the following sources:
 - New BI or Data Warehousing initiatives as focal points
 - Master Data Management implementation and associated terms

- New Data Governance initiatives
- Heavily-regulated information areas where there are changing requirements for reporting and compliance
- ▶ Get the language of the business from existing lexicons, BI reports, business-generated IT requirements for data reporting and analysis, and personnel.
- ▶ Establish straw man definitions as starting points; perfection comes later.
- ▶ Encourage collaboration early; differences of opinion are good.
- ▶ Assign data stewardship responsibility to the most vocal contributors.
- ▶ Deploy to the interested group or groups quickly to get exposure and establish usage and best practices.
- ▶ Iterate, reiterate, and then repeat to improve content.
- ▶ Establish business benefit in focused area, including tie-back to business outcomes and enterprise adoption follows.

6.3.3 Examples of glossary categories and terms

Consider the following examples from the IBM Warehouse Pack for Customer Insight where the terms are grouped into four subcategories (as shown in Figure 6-2 on page 100): analytical requirements, business terms, dimensions, and measures.

The organization has a wanted business outcome to increase customer retention (and to decrease churn among their customers). They need analysis to understand this, which can be defined as the following term: *Individual customer churn analysis*, as shown in Figure 6-3.

The screenshot shows the 'Term Details' page for 'Individual customer churn analysis' in the InfoSphere Business Glossary. The page includes a navigation bar with 'Glossary' and 'Administration' tabs, and a search bar. The main content area displays the term name, a description, and its parent and referencing categories. Below this, there is a section for 'Associated Terms (14)' which includes a list of 'Related Terms (14)'.

Term Details

Individual customer churn analysis
This Fact table records the measures related to Individual Customers leaving the organization for a competitor or joining or returning to the organization.

Parent Category [ISW Pack Customer Insight Business Glossary](#) » [Business Terms](#)

Referencing Categories (1) [ISW Pack Customer Insight Business Glossary](#) » [Analytical Requirements](#)

Status Standard

General Information

Associated Terms (14)

Related Terms (14)

1-10 of 14
Agreement / ISW Pack Customer Insight Business Glossary » Business Terms
Calendar date / ISW Pack Customer Insight Business Glossary » Business Terms
Competitor / ISW Pack Customer Insight Business Glossary » Business Terms
Cost of acquisition / ISW Pack Customer Insight Business Glossary » Business Terms
Cost of renewal / ISW Pack Customer Insight Business Glossary » Business Terms
Customer residence area / ISW Pack Customer Insight Business Glossary » Business Terms
Individual customer profile / ISW Pack Customer Insight Business Glossary » Business Terms
Number of customers / ISW Pack Customer Insight Business Glossary » Business Terms
Number of customers returning / ISW Pack Customer Insight Business Glossary » Business Terms

Figure 6-3 A business requirements term - Individual Customer Churn analysis

Note that this term exists in two categories, a parent category and a referencing category, which allows users to search through different groupings and directly for the term. “Individual customer churn analysis” also has the following related terms:


- ▶ Agreement
- ▶ Cost of renewal
- ▶ Individual customer profile
- ▶ Number of customers returning

Some of these terms (for example, Number of customers returning, as shown in Figure 6-4) represent measures. These measures often are the Key Performance Indicators (KPIs) that are used to assess whether the business outcomes are met.

Term Details

Edit



Delete



Number of customers returning

The number of ex-customers that became active again for the Product by signing a new Agreement during the period.

Parent Category


[ISW Pack Customer Insight Business Glossary](#) »
 
[Business Terms](#)

Referencing Categories (1)




[ISW Pack Customer Insight Business Glossary](#) »
 
[Measures](#)

Figure 6-4 A measurement term - number of customers returning

6.4 Information Governance policies and rules

The Business Glossary helps establish the organizational awareness of what the business is about. To broaden or enhance the information governance perspective around these terms, it is also important to understand the business context as expressed in Information Governance policies (an enabler for the wanted business outcomes) and Information Governance rules (business expressions of the targeted Information management disciplines, whether information quality, lifecycle management, or security and privacy considerations). These Information Governance rules (and the Information Governance policies that organize them) are a way to express the following business terms:

- ▶ Ideal behavior, shape, format, and so on, of information; for example, the Social Security Number (SSN) must be formatted: XXX-XX-XXXX
- ▶ Controls that are needed around the information; for example the SSN is sensitive and private and must be masked to non-privileged users.
- ▶ Retention and storage of information; for example, customer sales transaction details should be retained for three years for tax purposes, but then should be archived.
- ▶ Which information is governed by the rule; for example, the SSN column in the Customer Database.
- ▶ Which technical asset implements the rule; for example, the US SSN Formats data quality rule.

6.4.1 Definition and management of information policies

Information Governance policies and rules are created and managed in the Business Glossary along with the terms. Similar to terms, they can be searched, browsed, and viewed. They can be labeled, given or assigned to stewards, and extended by using custom attributes that are unique to an organization. They are subject to workflow processes and associated rules that facilitate their lifecycle management.

For example, an Information Governance policy indicates that customer data is subject to Data Privacy laws, as shown in Figure 6-5.

The screenshot shows a web interface for managing information policies. At the top is a navigation bar with links: Search, Create, Terms, Categories, Rules, Policies, Labels, and Browse. A search box labeled 'Quick Term Finder' is on the right. Below the navigation bar is the title 'Information Governance Policy Details' with icons for print, email, and help. A toolbar contains 'Edit' and 'Delete' buttons. The main content area features a large heading 'Customer Data Privacy Policy' with a brief description: 'Information pertaining to customers, including their Personal Identification Information (PII), is specifically regulated by Data Privacy laws and the terms and conditions of their customer agreement.' Below this is a 'Long Description' section with a more detailed paragraph and a 'Show More' link. A 'Parent Policy' section links to 'Customer Information'. A 'General Information' section displays a table with metadata: Created By (isadmin), Created On (Oct 23, 2012 1:58:42 PM), Last Modified By (isadmin), and Last Modified On (Nov 7, 2012 1:42:58 PM). Below this are sections for 'Subpolicies' and 'Rules (1)', with the first rule titled 'Data Access restrictions for Customer Profiles'.

General Information	
Created By	isadmin
Created On	Oct 23, 2012 1:58:42 PM
Last Modified By	isadmin
Last Modified On	Nov 7, 2012 1:42:58 PM

Rules (1)	
Icon	Data Access restrictions for Customer Profiles

Figure 6-5 An Information Governance Policy: Customer data privacy

The policies provide context for Information Governance. For example, all customer information that is contained in the Information Warehouse (among the many sources in the organization) falls under the Customer Data Privacy Policy, which is a sub-policy to another policy called Customer Information. Each policy can be elaborated in its description, or linked via a custom attribute (such as for a Reference URL, as shown in Figure 6-6 on page 105) to extended material (for example, government regulations and human resource policies).

Information Governance Policy Details

Edit

Delete

Customer Data Privacy Policy

Information pertaining to customers, including their Personal Information, and the terms and conditions of their customer relationship.

Long Description

Information pertaining to customers, including their Personal Information, and the terms and conditions of their customer relationship.

Parent Policy

[Customer Information](#)

General Information

Reference URL

<http://publib.boulder.ibm.com/infocenter/isinfsv/v8r7/topic>

Created By

isadmin

Figure 6-6 Custom attribute for an external reference link

Information Governance policies also can reference associated Information Governance rules that are serving in that capacity, such as a folder or category (an Information Governance rule can be referenced by multiple policies), as shown in Figure 6-7.

Search

Create

Terms

Categories

Rules

Policies

Labels

Browse

Quick Term Finder

Information Governance Rule Details

Edit

Delete

Data Access restrictions for Customer Profiles

All customer profiles must be secured and access limited to named, privileged users.

Long Description

All customer profiles must be secured and access limited to named, privileged users.

Referencing Policies (1)

[Customer Information](#)
[Customer Data Privacy Policy](#)

General Information

Created By

isadmin

Created On

Oct 23, 2012 2:00:03 PM

Last Modified By

isadmin

Last Modified On

Nov 7, 2012 1:50:49 PM

Related Rules

Implemented By (1)

Figure 6-7 An Information Governance Rule: Data Access restrictions for customer profiles

Where the Customer Data Privacy Policy indicated that regulations apply to customer data, this Information Governance rule (Data Access restrictions for Customer Profiles) describes a specific state that must exist: “Customer profiles must be secured and access limited....”.

6.4.2 Definition and management of Information Governance rules

Information Governance rules are implemented by specific types of rules. These rules can be information quality rules from IBM InfoSphere Information Analyzer or data masking rules from IBM InfoSphere Optim™. The Information Governance rules also can indicate the assets that they govern, whether internally known to IBM Information Server, such as an Information Warehouse model or database, or external, such as an application, as shown in Figure 6-8.

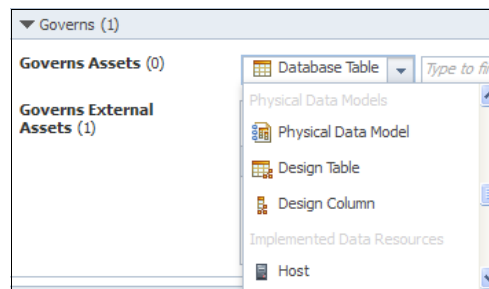


Figure 6-8 Adding governed assets to an Information Governance rule

6.4.3 Standard Practices in Information Governance policy and rule development

The development and deployment practices around the Information Governance policies and rules adhere to the following guidelines that are similar to those of terms:

- ▶ Start small and focused on relevant Information Governance policies and rules.
- ▶ Start with semantic business definitions (IT semantics and assignment to technical assets should be added later).
- ▶ Learn about available relationships. Establish standard approaches to use available resources; for example, URLs or documents that contain detailed information.
- ▶ Concentrate on key high-value information areas that are driven by business outcomes; for example, risk or compliance.
- ▶ Use new initiatives to drive an expanding coverage.

- ▶ Follow the language of the business from terms, if available; otherwise, from existing written policies and regulations and other sources.
- ▶ Establish straw man definitions as starting points; perfection comes later.
- ▶ Encourage collaboration early. In general, policies and rules are less subject to debate on definition than terms, but they can have considerably more details to address and capture.
- ▶ Assign data stewardship responsibility to those with a stake in defining, maintaining, and enforcing the policies.
- ▶ Deploy to the interested group or groups quickly to get exposure and establish usage and best practices.
- ▶ Iterate, reiterate, and then repeat the process to improve content.
- ▶ Establish business benefits in focused areas, including tie-back to business outcomes; enterprise adoption follows.

As terms and Information Governance policies and rules are established, they provide a framework to connect the Information Warehouse (and data flows into and out of the warehouse) with those terms, policies, and rules. Through the assignment and linkage of technical assets such as columns, tables, modeled attributes, and entities, the business (including IT) gains a view of the business outcomes and the stored information. Attention also can be focused on the key governance disciplines (for example, information quality or information security) around the most critical or sensitive pieces of information.


6.5 Information stewardship

This is where stewardship becomes critical to the Information Governance process because it is a quality control discipline that ensures custodial care of data for asset enhancement, risk mitigation, and organizational control. Information Stewards become focal points or references for the organization as a whole for the management of the information in relation to the wanted business outcomes. Information Stewards also are identified in the Business Glossary, and then linked to the terms, assets, and Information Governance policies and rules that they support, as shown in Figure 6-9 on page 108.

As with governance, there are stewards of business functions and Information Stewards. The term *steward* always applies to the Information Steward context.

Steward Details


Edit





Example User

First Name	Example
Last Name	User
Job Title	Data Steward
Email Address	example@sampleconsolidationco.com
Office Phone Number	555-1212
Created By	isadmin
Created On	Nov 9, 2012 5:00:23 PM
Last Modified By	isadmin
Last Modified On	Nov 9, 2012 6:06:44 PM

▼ Managed Assets (4)


[Customer Information](#)


[Customer Name must be Complete and Valid](#)


[FirstNameValid](#)




[NameExists](#)

Figure 6-9 An example Information steward with managed assets

As people in the organization browse and review the available business terms or Information Governance policies and rules, they can quickly see who is managing and governing this information. They can then contact the Information Stewards with questions and bring them into discussions concerning the appropriate use of the information, as shown in Figure 6-10 on page 109.

Information Governance Rule Details

EditDelete




Customer Name must be Complete and Valid

Customer Name must be Complete and Valid.


Long Description

Customer Name must be Complete and Valid. Completeness indicates that data must be present in the field. Valid hyphens, can contain numbers for generations only, but cannot contain other special characters or numbers in

Referencing Policies (1)

 [Customer Information](#)

Steward

 [Example User](#)

▶ General Information

▶ Related Rules

▼ Implemented By (4)

Implemented by Assets (4)





 [FirstNameNoTestData](#)
 [NameExists](#)
 [PrimaryNameNoTestData](#)
 [PrimaryNameValid](#)

Figure 6-10 Finding the Information Steward for an Information Governance Rule

6.6 Information Governance for the data warehouse

An Information Governance program for the Information Warehouse needs the following elements:⁷

- ▶ Principles (goals, values, and policies)

Principles are the core statements that guide the operation and development of the information landscape, including the information warehouse. They capture the goals and values of the Information Governance program and precisely define the implications of them to the organization, typically as a set of policies that the organization implements. This precision makes the goals and values more real to many people and clarifies for what they are signing up. Visibility through the Business Glossary into these policies facilitates that communication.

⁷ IBM Information Governance Model, version 10, IBM white paper by Mandy Chessell, pg.5-6, IBM, ©2010.

► Metrics

Metrics are the set of measurements and targets that are used to assess the ongoing effectiveness of the information supply chain and associated Information Governance program. They can measure the characteristics of the information assets themselves (for example, the number of Customers with incomplete Mailing Addresses), or the activity around them (for example, the number of processes extracting Customer Email Addresses). These can be established as terms in the Business Glossary, and can link to Information Governance rules or more specific Information quality rules, and so on.

► Organizational design and people

The organizational design defines the roles and teams that are required to govern the information supply chains. This can be thought of broadly in two segments. First, there are the people that maintain the Information Governance program and set policies that define the way information is to be governed (for example, the Owners and Stewards of the information). Then, there are the operational units of the organization that are working with the information or the infrastructure that manages the information. The Information Governance program can be set up to centralize the definition of how the policies are implemented, or decentralize these choices into the operational units.

► Capabilities

Capabilities are the tools, skills, techniques and processes required to implement Information Governance. Typically, the processes are defined by the organization and define the guidance and steps to follow on how to plan, understand, design and manage the information landscape. They combine existing capabilities together to provide new capabilities. The tools used in a process, and the skills and techniques employed by the people involved, provide the means to automate and record what is occurring, who is involved, and what needs to happen next.

To set up such a program, an organization likely will start by identifying the information assets that are most important to it and then define the principles for the Information Governance program. This provides the organization with a scope and a definition of the effects that the Information Governance program is likely to have on the organization.

For example, an organization might follow the levels of Information Governance maturity for its Information Warehouse by identifying the following milestones and benchmarks:⁸

- Level 1

Policies around regulatory and legal controls of information that are stored in the warehouse are put into place. Data that is considered critical to those policies is identified. Risk assessments also can be done around the protection of critical data.

- Level 2

More data-related regulatory controls are documented and published to the whole organization. There is a more proactive approach to problem resolution with team-based approach and repeatable processes, particularly in information coming from the warehouse. Metadata becomes an important part of documenting critical data elements.

- Level 3

Data-related policies become more unambiguous and clear and reflect the organization's data principles. Data integration opportunities into or out of the warehouse are better recognized and used. Risk assessment for data integrity, quality, and a single version of the truth becomes part of the organization's project methodology.

- Level 4

The organization further defines the value of data for more data elements in the warehouse and sets value-based policies around those decisions. Information Governance structures are enterprise-wide. Information Governance methodology is introduced during the planning stages of new projects. Enterprise data models are documented and published.

- Level 5

Information Governance is second nature. Return on investment (ROI) for data-related projects is consistently tracked and innovations are encouraged. Business value of data management is recognized and cost of data management is easier to manage. Costs are reduced as processes become more automated and streamlined.

Information Governance incorporates three core Disciplines, which are of particular relevance for the Information Warehouse: Information quality management, Information Lifecycle management, and Information Security and Privacy.

⁸ *The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance*, pg.12,
https://www-935.ibm.com/services/uk/cio/pdf/leverage_wp_data_gov_council_maturity_model.pdf, IBM, ©2007.

► Information quality management

Information quality management contains the capabilities to actively manage the quality of information in the landscape and the information warehouse, including: information profiling and analysis, cleansing and consolidation of information, and ongoing validation and monitoring against specific information quality rules. For more information, see Chapter 7, “Establishing trust by ensuring quality” on page 115.

► Information Lifecycle management

Information Lifecycle management provides the background mechanisms to manage information from the time it is created through to when it is no longer needed. Information Lifecycle management capabilities include the following examples:

– Archive

Archiving information moves it from a front-line operational system to cheaper auxiliary storage. This occurs when the information is unlikely to be needed again but must be retained to satisfy a regulation or to investigate an incident that involved the archived information. Successful archiving of information requires a knowledge of how the information is structured as it stored. Archiving is typically triggered when the information reaches a particular state; for example, it has not been accessed for a time, or the expected processing on that information item completed.

– Test Data Generation

Software project teams that are making changes to an information asset, or information supply chain, must perform extensive testing to ensure information is processed correctly. For security reasons, the team often cannot be given production data to test with. Test Data that was created manually often is missing many of the patterns that are found in the real information. What is needed is a mechanism to accurately obfuscate the values in a representative subset of the production data that preserves the patterns and profile within it. This way, the privacy and security issues are removed and the software project team has appropriate test data.

– Backup/Restore

Backup/Restore is a critical capability to enable information to be recovered if it is deleted in error, the technology storing the information fails, or facility where the information is kept suffers a disaster that requires the information and its related processing to be moved to a new location. Backup/Restore is critical for ensuring continued availability of information. The Information Governance program must ensure that the information it is responsible for is backed up frequently and the ability to restore it is tested on a regular schedule.

► Information Security and Privacy

Information security and privacy ensures that the organization can identify who has access to information and that they are the appropriate people to be given that privilege. There are two aspects to security and privacy: authentication and authorization.

– Authentication

The process of uniquely identifying a person, or system. The most typical approach is the use of a user ID and a password, although biometric authentication mechanisms (such as fingerprint readers) also are available. One of the most basic requirements of many regulations is to get to the point where every person who is accessing a system is uniquely identifiable; that is, no sharing of user IDs to access a system and control on the strength and confidentiality of passwords. Without this, it is not possible to trace the activity in the IT Systems to the individual that initiated it.

– Authorization

The mechanism to control what information and actions are available to a particular individual, and under which circumstances. This works by classifying information and then defining policies on which groups of people are enabled to perform which activities on which pieces of information and the format in which the information is displayed. The classifications can be on the type of information, sensitivity of information, state of the information, or origin or ownership of the information. A complete collection of information can have a different classification to the individual records because it enables more insight. The display of information might be read-only, editable, or obscured in some way to allow someone to validate or match the value, but not see its value. Legal issues, particularly those related to privacy, can affect whether information can be accessed outside of the country in which it was created.

6.7 Conclusion

When an organization runs a strong Information Governance program, it helps ensure that information used for critical decisions in the organization is trusted, particularly from such a central hub as the information warehouse. The information must come from an authoritative source and is known to be complete, timely, and relevant to the people and systems that are involved in making the decision. It must be managed by an Information Steward who can communicate to others about its purpose, usage, and quality. Through communication of Information Governance policy and rules, business terms, and their relationship to the information assets, the information can be clearly understood across the organization.



Establishing trust by ensuring quality

Companies today are continually moving toward greater integration that is driven by corporate acquisitions and customer-vendor linkages. As companies try to become more customer-centric, management realizes that data must be treated as a corporate asset and not a division or business unit tool.

Organizations require a detailed knowledge and understanding of the strengths and weaknesses of their data and its inherent quality both in Information Governance initiatives and as a core business management practice that helps to use the information that exists across multiple business initiatives. Their ability to gain this knowledge and apply it to their various data-related initiatives can directly affect the cost and benefits of those initiatives, including the level of trust that users or consumers have with the information. Business and technical stakeholders, from Business Users and Stewards to ETL and BI Developers to Information Governors, all have requirements for trusted information, as shown in Figure 7-1 on page 116.

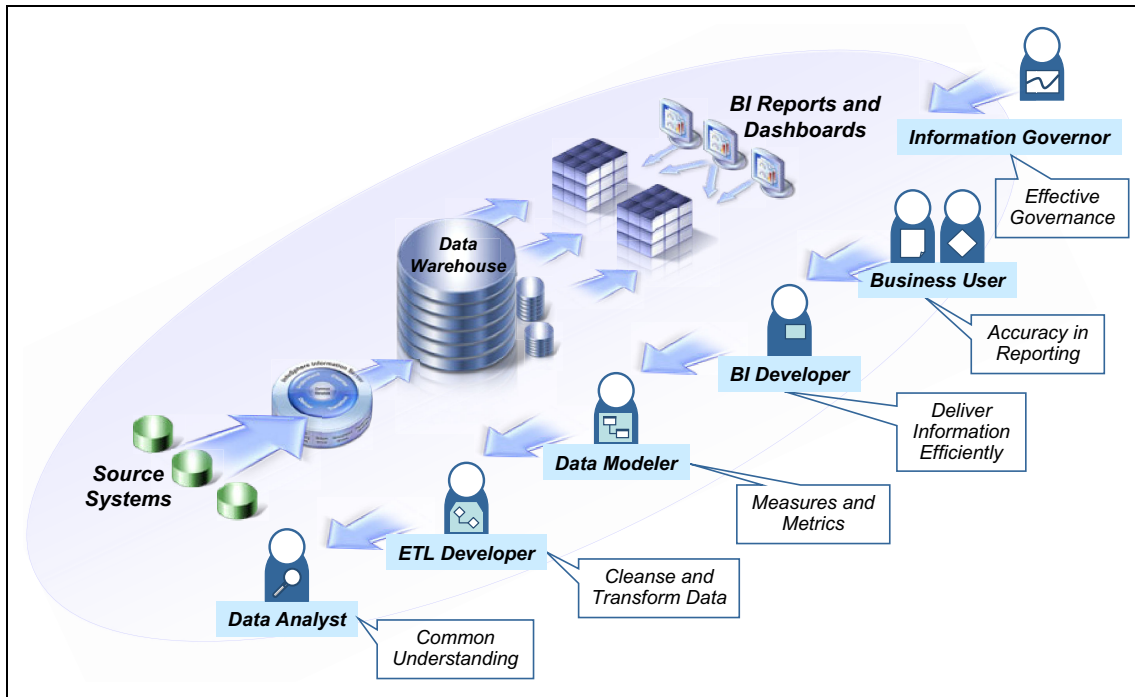


Figure 7-1 : Requirements for Trusted Information through the Information Supply Chain

This chapter includes the following topics:

- ▶ Moving to trusted information
- ▶ Mission of information quality
- ▶ Understanding information quality
- ▶ Validating data with rules for information quality
- ▶ Measuring and monitoring information quality
- ▶ Information Quality Management
- ▶ Conclusion

7.1 Moving to trusted information

Many organizations have information issues surrounding trust and control of their data. Unfortunately, many information sources are in the right form or have the right metadata or even documentation to allow a quick integration for other uses. In many established systems and enterprise applications, metadata, field usage, and general knowledge changed over time. The data might be perfectly acceptable for whatever purpose it was designed, but after it is loaded into the information warehouse, the organization discovers how inappropriate it is for what they want to do. Issues that are found include: different or inconsistent standards, missing data or default values, spelling errors, data in wrong fields, buried information, and data anomalies. In many well-publicized cases, strategic data-related projects exceeded planned cost and schedule while delivering less than expected return, or failed completely because of data quality defects that were underestimated or unknown until the implementation stage of the project.

Figure 7-2 shows some of these common information challenges in enterprise systems.

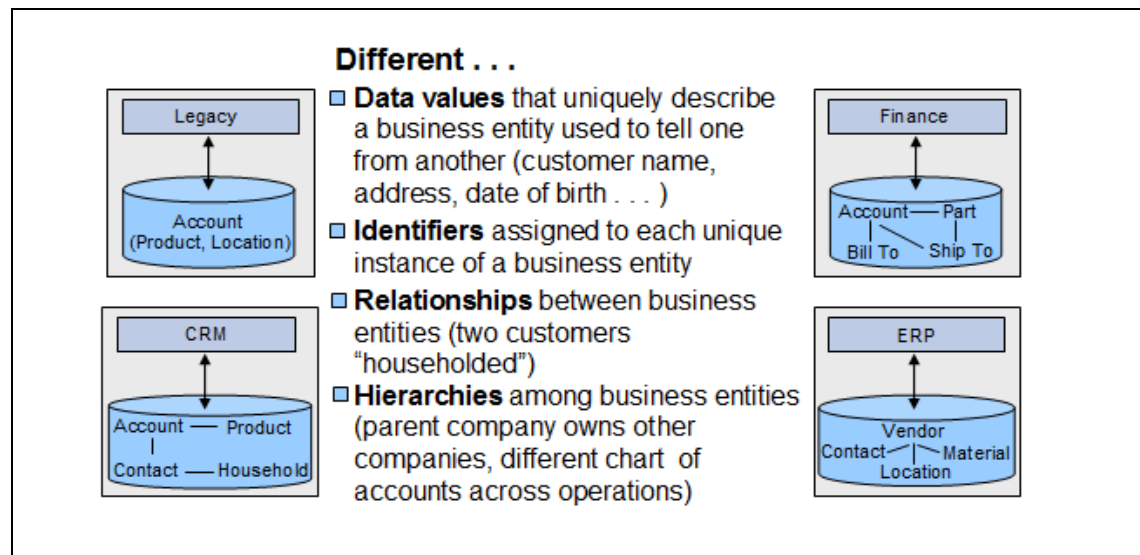


Figure 7-2 Common information challenges

7.1.1 Challenges to trusted information

Today, clients deal with information complexity every day. Most enterprises are running distinct sales, services, marketing, manufacturing, and financial applications, each with its own master reference data. There is no single system that is the universally agreed-to system of record. In data integration efforts, old data must be repurposed for new systems. Enterprise application vendors typically do not guarantee a complete and accurate integrated view; they point to their dependence on the quality of the raw input data. However, it 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, lack data quality, and have difficulty 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 adding up 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.

7.1.2 Impact of information issues

Organizations do not have trust in their information because the quality cannot be assured, the usage is unclear, and the source of the information is often uncertain. This is not an issue to address after implementation, but at the beginning and then through the lifecycle to avoid untimely and expensive fixes later. These information challenges have the following real and direct consequences to the business:

- Inability to cope with compliance rules and regulations
 - Cannot report accurately on regulations such as SOX, Basel II, customer privacy laws, and Health Insurance Portability and Accountability Act (HIPAA)

- ▶ Missed revenue opportunities
- ▶ Runaway costs:
 - Running an operation based on inaccurate information or lack of information
 - Labor cost of repairing a problem such as rejected claims (delayed payment, more human resources) can be high
- ▶ Poor customer satisfaction:
 - “Brand equity” damage
 - Loss of customer loyalty
- ▶ Inability to rescue existing investments
- ▶ Challenges to be productive and efficient when dealing with inaccurate data

So, how can an organization ensure that everyone in the business is more informed, confident, and aligned to ensure better business outcomes?

7.2 Mission of information quality

Information quality is a mission that encompasses the entire information lifecycle. From source to the ultimate destination, data goes through a long sequence of processes that convert it from one form to another and move it from one container to the next. Along this chain of processes are numerous opportunities for things to break, go astray, and produce wrong and unreliable data. Information Governance is put in place to minimize such occurrences and optimize resources to achieve the best possible results toward building trust in the information.

Trusted information is at the core of many business initiatives. It is one of the foundations for decision making processes and initiatives to optimize revenue opportunities. It enables the creation and nurturing of collaborative business processes and empowers the risk and compliance initiatives of an organization. As noted previously, organizations need a consistent and maturing information governance practice that helps to use information that exists across multiple systems and assures quality. A primary drive of information governance is the ability to establish the framework of organization, technology, and process that promotes the creation and maintenance of trusted information.

7.2.1 Key information quality steps

To support Information Quality, an organization must go beyond a reactive state through the following series of core steps:

- ▶ Perform Data Quality Assessments for all projects that identify problems up-front.
- ▶ Define business-driven Data Quality rules for consistent assessment of information.
- ▶ Define business-driven Information Quality measures with controlled monitoring by data stewards to manage and organize information and exceptions.
- ▶ Perform Information Quality Management with results and trends along key performance indicators that are deployed through standard information delivery channels (for example, dashboards and reports).
- ▶ Ongoing process improvements to consistently deliver high-quality data through managed data cleansing initiatives.

Information Quality is about addressing these issues through a consistent practice that brings together people, process, and tools, and subsequently reduces project costs and risk by discovering problems early in the data integration lifecycle, as shown in Figure 7-3.

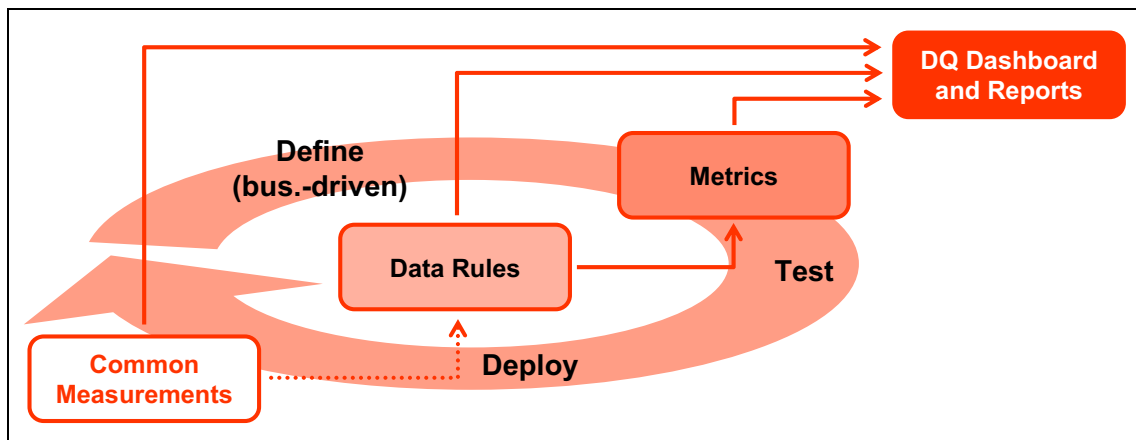


Figure 7-3 Increasing maturity in Information Quality Management

7.3 Understanding information quality

A core challenge in many enterprises is gaining an understanding of their data, particularly as systems and applications are adapted to changing business requirements. Mergers and acquisitions expand upon the existing sets of data, new sources of information are added, and consumers look across different sets of data with similar references but different meanings. Most organizations have an inconsistent understanding of information across their enterprise. Different semantics are used by different business units and IT. Different applications record data in different ways with different identifiers (for example, customer and account IDs), different formats (for example, dates stored in a date format in a database, but a string format in a file), or different values (for example, gender recorded as “M/F” in one system, but “0/1” in another).

This level of understanding is complicated further by changes in domain expertise. Often, information about the data is in people’s heads based on their specific work or it might be recorded in old documentation that is kept up-to-date as new business processes are added or system changes are made. When individuals move within or out of an organization, much of that domain knowledge is lost or fragmented.

InfoSphere Information Server through InfoSphere Information Analyzer supports an evolutionary process of knowledge creation at all levels, as shown in Figure 7-4 on page 122. The base for understanding is learning the facts and the use of profiling to find all that is to be known about the data, one column at a time. Knowledge evolves by extracting patterns and developing data rules to monitor the quality of data. Values are attached to patterns and presented as metrics for management to act upon, to evaluate the cost benefit of decisions and actions, and to measure the progress of their plan’s implementation.

7.3.1 Data Quality Assessment

Whether embarking on new business initiatives or responding to or addressing problems in data governance and risk mitigation, a core entry point to information quality is the Data Quality Assessment.

A Data Quality Assessment is intended to provide insight into this complicated issue. It establishes a foundational practice for subsequent work and a knowledge base (within a shared metadata repository) that can be used and reused across multiple projects and initiatives. A Data Quality Assessment focuses on, and shows, a business problem that is based on the underlying data.

It is important to understand that a Data Quality Assessment is not magic. Such an effort helps bring data quality issues to light, but you still require a data or business analyst to review results and draw conclusions, particularly to define the affect on business.

As shown in Figure 7-4, the data analyst is at the center of any Data Quality Assessment. Analysts must understand the scope and objectives of the assessment and what they must deliver.

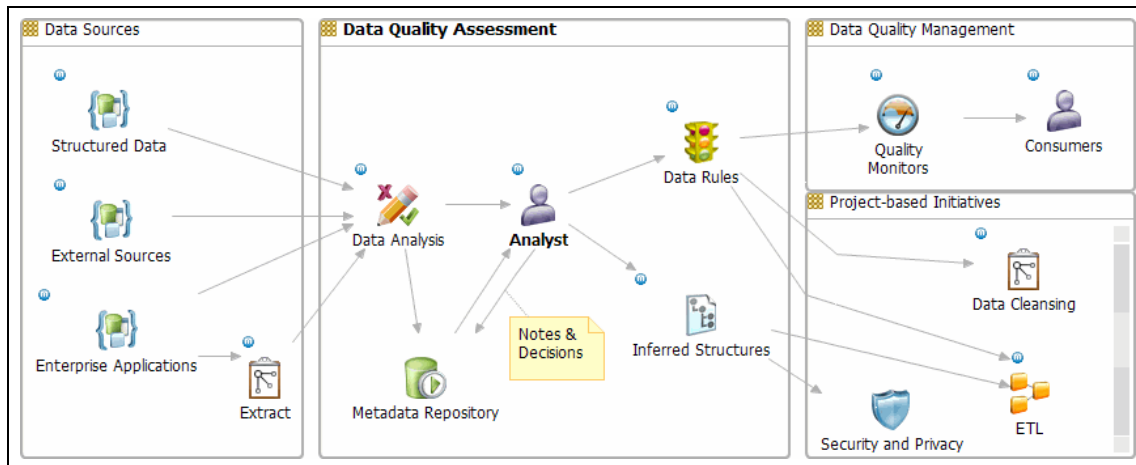


Figure 7-4 Landscape for a Data Quality Assessment

If you do not know what you are trying to achieve, how do you identify a true anomaly or problem? This question helps to shape the correct approach to data analysis: focusing on the business outcomes, the data critical to support those outcomes, and the cost of issues within that data.

Consider a healthcare example in which incomplete or invalid data on a claim results in the rejection of the claim form by an insurance company. The following costs are associated with rejection of these claims:

- ▶ Revenue is not received in a timely manner, potentially impacting business reporting and budget forecasting.
- ▶ Claims must be reviewed, corrected, and resubmitted, which results in more labor cost and lost opportunity cost.
- ▶ If claims are rejected again or pass an acceptable time limit, the patient must be billed directly, which involves further delay and potential hardship for the patient.
- ▶ If services and claims are linked to the wrong patient, there are many impacts to the patient and the facility.

Further, incorrect or invalid data also is propagated through the different applications downstream from the initial operational systems. The data warehouse and subsequent reporting systems receive incorrect, duplicate, or conflicting service transactions that also affect business decisions and analysis there.

The data analyst must focus on the following core analysis steps and best practices in the Data Quality Assessment:

- ▶ Identification of, and approach to, the data sources in question:
 - What data sources and fields are relevant?
 - Are there issues in accessibility or availability?
 - How should you address differences in data volume (for example, evaluate all data, a sample, or a targeted segment)?
- ▶ Advantages of automated data content-driven functions:
 - Capture of insights across the full volume of data
 - Ability to look at results for each and every column; primary, natural, and foreign keys; and cross-domain and cross-table overlaps
- ▶ Usage of data classifications to focus analysis
 - Can the business language (terminology) be connected to the data?
 - What issues are most common to a specific class of data?
- ▶ Validation of data formats and domains:
 - What is valid, defaulted, or invalid based on the inferred or known classification?
 - Which pieces of information can be captured as references for data cleansing or extract, transform, and load (ETL) initiatives?
 - What should be annotated for further review?
- ▶ Reporting and delivery of findings and results:
 - What inferences should be made available to other users and in what form?
 - What reports are relevant to the Data Quality Assessment and how should they be delivered?
- ▶ Retention of analysis results over time:
 - What are the baseline measurements to assess against after initial results are established?
 - How long are analysis results retained?

The flow of information is shown in Figure 7-5.

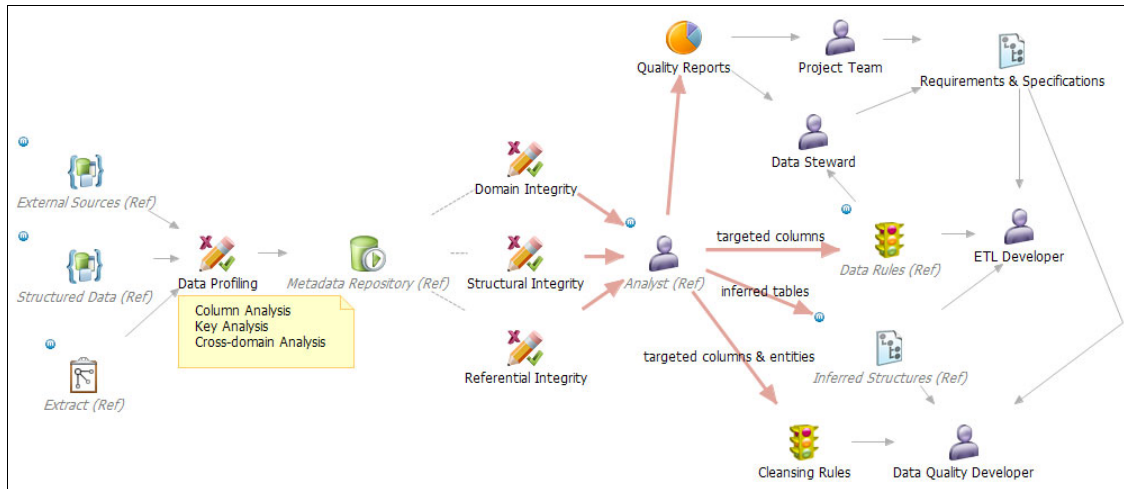


Figure 7-5 Flow of Information Quality Analysis

7.3.2 Expanding on the initial assessment

Core analysis in a Data Quality Assessment shows a broad range of issues with many domains (such as defaulted data, missing values, non-unique or duplicated keys). However, other techniques might be required to focus on certain types of conditions (usually on specific targeted columns or entities) that often are expressed as data validation rules. The analyst can use these rules to test for valid value combinations, correct formulas and aggregations, complex format requirements, or more comprehensive assessment of entire records or tables. These other tests also can be reported, retained, and trended over time.

With this foundation in core data analysis practices, the analyst can extend or expand upon a Data Quality Assessment to make it a core practice in broader enterprise initiatives, whether it is focused data cleansing efforts, information integration projects, or Information Governance initiatives. The analysis processes, data validation rules, and reports can be scheduled to run on a regular basis to provide ongoing Data Quality Monitoring.

The metadata from analysis is directly available to users of other products within the IBM InfoSphere Information Server. Data modelers and DBAs can use the inferred structures and identified classifications to establish staging areas with the correct structure or to refine privacy and governance policies. Developers in IBM InfoSphere DataStage and IBM InfoSphere QualityStage can use the statistics and annotations to ensure that appropriate cleansing routines are applied to the data, or incorporate reference tables generated from the analysis.

Further, the data validation rules from InfoSphere Information Analyzer can be plugged directly into data cleansing and ETL processes by developers via an integrated Rule Stage. This stage applies such rules in-flight, which ensures that problematic and invalid data conditions are addressed before they are loaded into the target environments.

From a basic Data Quality Assessment, InfoSphere Information Analyzer allows an enterprise to start small and continuously expand the knowledge of their data domains. This knowledge base supports and addresses the challenges that are inherent in the continuous expansion and acquisition of data, systems, and applications that are the foundation of every organization's business.

Data quality assessments of critical data at the start of a project help to identify and measure existing data defects. By performing this assessment early, the organization can take any necessary corrective action on the data, or circumvent any data problems that must be avoided.

7.4 Validating data with rules for information quality

A key factor in establishing ongoing Information Quality in an organization is moving beyond the Data Quality Assessment to validate information and measure information quality in a systemic fashion. This entails specifying consistent and re-usable data rules (also called *validation rules*), which are driven by the requirements and knowledge of the business.

Rules that are established during a Data Quality Assessment can be used as a starting point to measure data quality throughout the project lifecycle by allowing developers to test the accuracy of their code or jobs in delivering correct and expected results, by assisting quality assurance of functional and system accuracy, and by allowing business users to gauge the success of system load processes. This specification should be done within the context of a project, whether one with a larger focus (for example, expanding the data warehouse) or a narrow data quality improvement project.

7.4.1 Incorporating business value and objectives

To incorporate business value and objectives, consider the following questions:

- ▶ What data is relevant (start with the end goal in mind) and is there any of the following factors:
 - A clear focal area for evaluation?
 - A clear understanding of what data requires a data quality evaluation?

- ▶ What key performance indicators (KPIs) or metrics should be applied? You cannot manage data quality if you cannot measure it; this is about aligning goal and outcome to business impact.
 - What are the key performance indicators around data quality?
 - What metrics tie to these KPIs and to what data are the metrics applied?
 - Do these KPIs and metrics link to other business measures?
- ▶ How are KPIs and metrics communicated? No one is satisfied if they cannot find it or understand it.
 - What are the channels of communication?
 - Who receives the KPIs and metrics and what do they do with that information?
 - Is there a clear understanding of how to respond to this information?
- ▶ How will these data quality measures be tracked over time? As your business changes, so must your measurements.
 - Is there a lifecycle to the information?
 - What is the stewardship process to identify trends or handle exceptions?
 - Will these measures drive quality improvement cycles or identify when appropriate quality levels are achieved?

7.4.2 Defining the primary requirements

Based on the insights from the Data Quality Assessment, the analyst must define the requirements to address critical or key data quality issues to meet the project objectives.

The analyst addresses the following questions:

- ▶ What conditions generate constraints and exceptions, and how are those issues handled?
- ▶ What processes are used to monitor the overall data quality in the sources, targets, or during processing?
- ▶ What processes are used to review, manage, and report exception conditions?
- ▶ What processes are used to remediate data quality conditions in the data sources, targets, or implement changes to the rules based on subsequent changing business conditions?
- ▶ Where the data cleansing or exception or remediation processes change and modify the data, what data or information must be retained and what is the retention policy?

7.4.3 Designing the data rules

The following questions are reviewed in relation to defining the data validation rules:

- ▶ What policies or rules are required and what policies or rules are optional?
If there are time constraints for deployment and implementation of the project, which policies are most critical for delivery?
Identify the policies that must be enforced on the data.
- ▶ What data must be processed regardless of policies/rules?
Identify any exceptions to the policies. For example, all customers must be accounted for in the warehouse and MDM systems, thus a customer record cannot be dropped from the database because certain policies are not met.
 - When are constraints to the data applied?
 - Do these occur during process flow or after?
 - Are these applied after specific data cleansing steps are taken, such as Standardization, Matching, Consolidation, or Enrichment or only at the conclusion of all steps?Identify the points in or after specific process flows when policies are enforced on the data.
- ▶ What happens to the data when constraints are implemented?
Identify whether the data is flagged, but continue processing, or whether data is treated as exceptions or rejections from the processing for further review.
- ▶ If exception conditions occur or data is rejected, how is that data remediated or reprocessed?
 - If exception conditions are because of problems with the rules or the process flow, how is that data be reprocessed?
 - Identify whether remediation can be automated or must be handled manually. Also, steps for reprocessing or correcting the data (or rules) must be identified.

The decisions that are made in defining the constraints influence subsequent definition of Information Quality Management routines, particularly exception handling and remediation processes.

Data Validation Rules (Data Rules) evaluate the compliance or quality of data in terms of specific conditions that involve single or multiple data fields within or across records (or rows) that are logically related. In most cases, the type of data rules that are needed for analysis are not documented or even explicitly known before the Data Quality Assessment, though Information Governance policies and business processes can suggest likely conditions.

Therefore, data validation rules must be developed, or at least refined, for use in the project. The most common sources for developing explicit validation rules that are applicable to data quality work are knowledgeable people (subject matter experts), system documentation, and occasionally metadata repositories. Results from the Data Quality Assessment also can suggest other validation rules, as shown in Figure 7-6.

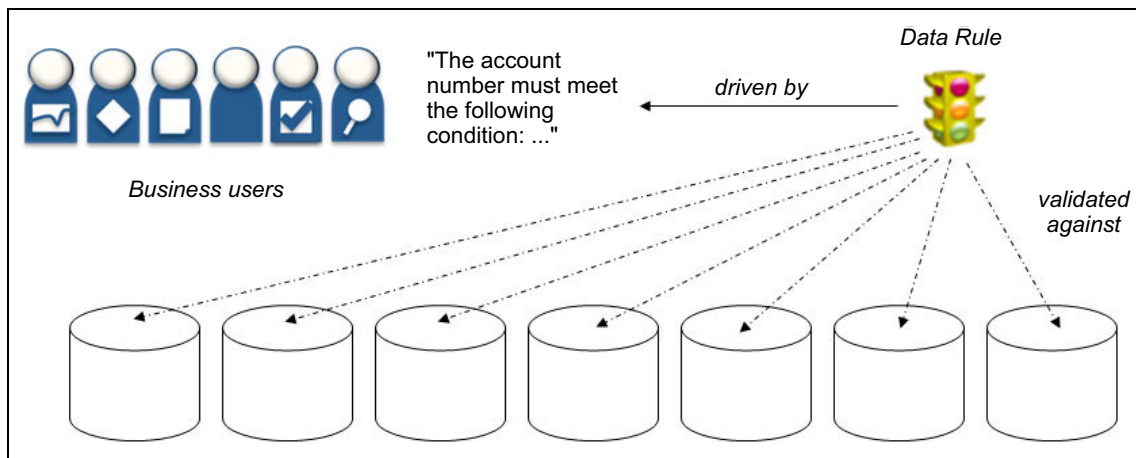


Figure 7-6 Definition of data rules

7.4.4 Example of data rule analysis

To illustrate an example of data rule analysis, consider an information warehouse for Customer Insight. One data element is a code that represents the various channels through which customers can purchase goods. Valid sites include an online website, a store, a third-party retailer, specialized installers, and so forth. Another data element is a code that represents the various products that are offered. Valid products include clothes, hardware, appliances, and so forth. The data validation rule states the legitimate relationships between site codes and product codes. A particular customer transaction includes the site code for the online website and the product code for microwave installation. Each of these codes taken separately passes a validity test. When taken together, however, the resulting information is suspect in that microwave installation is not usually purchased through the online website without an associated purchase of a microwave appliance.

7.4.5 Setting priorities and refining conditions

There are virtually endless possibilities for designing data rules that are applicable to data. Therefore, when this level of analysis is performed, consider the following suggestions when validation rules are developed and used:

- ▶ Although many potential validation rules can be developed in Information Analyzer for their effectiveness in determining data quality, a limited (but prioritized) set of validation rules should be selected for final use in analysis. This selection is based on their effect on the business or their ability to measure the integrity and reliability of key information that is taken from the data environment.
- ▶ A validation rule must be specified, reviewed, tested by using Information Analyzer, and refined with the realities of the actual data before the validation rule should be considered accurate and precise enough for use in ongoing Information Quality Management results.

7.4.6 Types of Data Rules

Data Rules can be implemented to evaluate the following integrity constraints:

- ▶ **Completeness**
The existence or presence of actual data (that is, not null, not spaces).
- ▶ **Structural Integrity:**
 - Data type checks (date, numeric)
 - Data format validation (by character or regular expression)

- ▶ Validity:
 - Data equality or difference (=, not =, >, <, >=, <=)
 - Data string containment; for example, contains “TEST”, which is a search for strings that contain this specific word
 - Data within a valid range
 - Reference lookup (by list or to a reference table)
- ▶ Uniqueness (or frequency of occurrence, usually of a key or identifying field)
- ▶ Referential Integrity

A key to one record exists in a corresponding list of keys in another table.
- ▶ Valid-Value Combinations

Multiple validity conditions often are paired with AND or OR clauses (for example, CustomerType = “G” AND CustomerStatus = “A”)
- ▶ Computational rules

Arithmetic expressions (fieldA + fieldB = fieldC) or Aggregations (for example, Sum, Count)
- ▶ Time rules

Ordered sequencing of dates
- ▶ If...Then...Else rules

Conditional expressions that might combine any of these constraints, but only in relation to a segment or subset of the data (for example, only IF CustomerStatus = “A” ...)

7.4.7 Examples of rules

The following example rules are built into Information Analyzer:

- ▶ The Gender field must be populated and in the list of accepted values.
- ▶ The Social Security Number must be numeric and in the format: xxx-xx-xxxx.
- ▶ If Date of Birth Exists AND Date of Birth > 1900-01-01 and Date of Birth < TODAY Then Customer Type Equals “P”
- ▶ The Bank Account Branch ID is valid if it is in the Branch Reference master list.

More examples of rules are shown in Figure 7-7.

Select Quality Control to Work With		
Quality Controls		
Name	Type	Description
▼ All		Global category for all QualityComponents
◊ ActiveIndividualCustomerClassMissing	Data Rule Definition	Customer Class Missing for Active Individual Customers
◊ ChannelTypeValidRef	Data Rule Definition	Channel Type in reference list
◊ CostAmtInvalidRange	Data Rule Definition	Cost Amount in valid range; applied to numeric data
◊ CustAgeInRangeNumeric	Data Rule Definition	Age: Age >= 15 and < 100; applied to numeric age data
◊ CustomerExists	Data Rule Definition	CustomerID Exists; null check only
◊ CustomerTransactionStatusValid	Data Rule Definition	Customer Transaction Status in reference list
◊ GrossIncomeInvalidRange	Data Rule Definition	Gross Income in valid range; applied to numeric data
◊ NameExists	Data Rule Definition	Name Exists; null and blank value check
◊ NetIncomeInvalidRange	Data Rule Definition	Net Income in valid range; applied to numeric data
◊ PaymentMethodValidRef	Data Rule Definition	Payment Method in reference list
◊ ProfitAmtValid	Data Rule Definition	Profit Amount computation is valid (arithmetic); applied to numeric data

Figure 7-7 Rule definitions built in Information Analyzer

7.4.8 Considerations in Data Rule design

A core aspect to the development of data validation rules is an understanding of the business problems to address. The business wants to build rules for specific conditions that are key drivers for addressing business outcomes. The following questions should be considered:

- ▶ What are the key business elements to address?
- ▶ What are the key criteria for evaluating these business elements?
- ▶ What data is involved or included in these business elements?
- ▶ Are there multiple parts or conditions to the validation?
- ▶ Are there known qualities about the data to consider?
- ▶ What are the sources of the data (for example, external files)?
- ▶ Are there specific data classes to focus on (for example, Dates, Quantities, etc.)?
- ▶ Are there aspects to the rule that involve the statistics from the validation?
- ▶ Are there aspects to the rule that involve understanding what happened previously?

Many or most of these questions might be answered from an earlier Data Quality Assessment. However, even where a Data Quality Assessment was not conducted, rule building and evaluation still can occur.

7.4.9 Breaking requirements into building blocks for Data Rules

Data validation rules often are presented in a complex, compound manner or in a business-specific vocabulary. When rules are constructed, you can look for building blocks for the rule as a starting point, such as Terms or common expressions.

Consider the following example:

To improve Customer Loyalty Tracking, the business needs a rule that identifies missing Customer Classification data but only for customers whose type is listed as “Individual” and only where the customer status is “Active”, all other customer types should be ignored for this rule. The business notes that sources variously enter the expressions “Individual” and “Active” in upper, lower, or mixed case.

To break this down into a usable format, the following building blocks can be found:

- ▶ Terms: Customer Classification, Customer Type, Customer Status
- ▶ Conditional situations: IF...THEN
Example: IF Customer Type = 'Individual'
- ▶ Alternative conditions
Example: Ignore all others
- ▶ Specific types of tests to include: Exists, Equals (=)
 - Example: 'Missing data' is a signal for the Exists type of test
 - Example: 'Is' or 'must be' indicate an Equals type of test

At this point, an initial data rule can be constructed, as shown in the following example:

```
IF Customer Type = 'Individual' AND Customer Status = 'Active'  
THEN Customer Classification EXISTS
```

After rule definitions are created, they must be bound (linked) to the relevant data sources for subsequent running. These can be run ad hoc or on a scheduled basis, but the initial focus should be to test and debug the rules against identified sources (typically using a data sample), assess the results and look for specific data conditions, incrementally add conditions as needed, and eliminate any extraneous conditions.

In our example data rule, the following initial tests indicate some conditions to address, much as the business users noted:

- ▶ “Individual”, “Active”: Could be upper, lower, or mixed case

- ▶ “Individual”: Could have leading or trailing spaces
- ▶ Customer Classification data: Might exist but be blank (for example, just contain spaces)

Functions might be needed to address these actual data conditions (which can vary from source-to-source). An updated version of the data rule definition might look similar to the following example:

```
IF trim (ucase (Customer Type)) = 'INDIVIDUAL' AND ucase(Customer
Status) = 'ACTIVE'
THEN Customer Classification EXISTS AND len(trim(Customer
Classification)) <> 0
```

This alternative version applies three functions: **ucase** (to change any alphabetic characters to uppercase), **trim** (to remove extraneous space characters), and **len** (to calculate the length of the value in number of alphanumeric characters). This rule definition now can be applied across a number of data sources and address permutations that are not of concern to the business.

However, if the Business Intelligence reports that are written for Customer Loyalty Tracking expect a standard value for Customer Type of “individual” and the data is received in differing case formats, there are other issues with that report. More data rules likely are required to ensure compliance with business expectations and the ability of the information to meet the wanted business outcomes.

With Information Analyzer, multiple data rules can be combined into larger rule sets to evaluate every condition that is needed to support the business objectives. As results or subsequent information are available, the focus can be expanded around new or more targeted rules, rule sets, and metrics to track these rules over time.

Data validation rules can be developed through several approaches, primarily dependent on whether they are to be run completely post-process (after target load), as ETL validation (after ETL, before load) or in process (within ETL as part of job flow). If the policies are post-process or ETL validation, they might be run directly in Information Analyzer or via the Information Analyzer APIs.

For more information about developing data rules, see “Managing a data quality rules environment” in the IBM InfoSphere Information Center at this website:

http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp?topic=%2Fcom.ibm.swg.im.iis.productization.iisinfsv.home.doc%2Ftopics%2Fic_homepage_IS.html

7.4.10 Evaluating Data Rule results

In evaluating the Customer Classification rule, the organization evaluates a sample set of data in the Customer Insight warehouse. The results shown are shown in Figure 7-8.

CSTINSIGHT_CST_...

View Output

View Output

Overview

Result

Benchmark Status:

Fail

Benchmark:

% Not Met <= 0.0000 %

Variance :

93.2353 %

Total Records:

340

Met #:

23

Met %:

6.76470588 %

Not Met #:

317

Not Met %:

93.23529412 %

Output: Do not meet rule conditions

1 - 50 of 317

Page 1

CUSTOMERCLASSIFICATION	CUSTOMERTYPE	CUSTOMERSTATUS	CST_ID	CST_NUM	FULL_NM	
[NULL]	Individual	Active	101	A101	Fred Smith	
[NULL]	Individual	Active	102	A102	Mary Jones	
[NULL]	Individual	Active	111	A111	Joe James	
[NULL]	Individual	Active	1061	A1061	Brendan Clark	
[NULL]	Individual	Active	1062	A1062	Anna Rodriguez	
[NULL]	Individual	Active	1063	A1063	Jim Lewis	
[NULL]	Individual	Active	1064	A1064	Pat Lee	
[NULL]	Individual	Active	1065	A1065	Anne Walker	
[NULL]	Individual	Active	1066	A1066	John Hall	
[NULL]	Individual	Active	1067	A1067	Peter Allen	
[NULL]	Individual	Active	1068	A1068	Jane Young	
[NULL]	Individual	Active	1069	A1069	P Hernandez	
[NULL]	Individual	Active	1070	A1070	Eric King	
[NULL]	Individual	Active	112	A112	Emma Field	
[NULL]	Individual	Active	1071	A1071	Joe Wright	
[NULL]	Individual	Active	1072	A1072	Emma Lopez	

Figure 7-8 Detail output results for the Customer Classification rule

The results from the small sample of 340 records indicate that 317 records or over 93% of the records failed to meet the business criteria. To achieve better outcomes for the Customer Loyalty Tracking initiative, the business must review how they can improve their quality measures.

As the individual types of defects are reviewed with subject matter experts and assessed for their impact on the business, the organization must develop appropriate action plans that lead directly to material improvements. They also must make recommendations concerning the cleaning of existing data and preventing the proliferation of the defects in the future. The first action in these plans often can require a root cause analysis to determine an appropriate corrective solution. These proposed action plans are presented to management for approval, funding, and staffing with the goal of elevating the data quality condition of the data environment.

Based on the recommendations that are accepted, establish the Quality Measures and Monitoring that are needed to track and achieve the wanted level of data quality over time. As that process is established, also make the transition to Information Quality Management to ensure appropriate governance of the environment. An information quality management environment is in Figure 7-9.

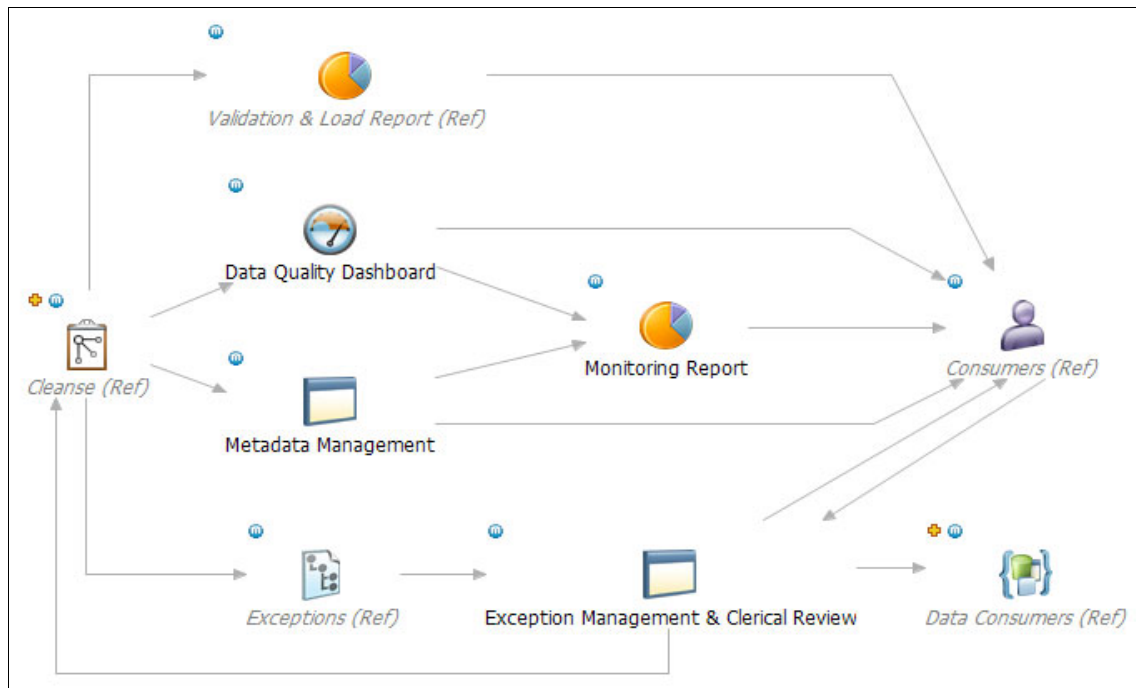


Figure 7-9 An Information Quality Management environment

To conduct initial Quality Monitoring, complete the following steps:

1. Select, define, and prioritize issues.
2. Set data quality objectives in the form of measures or metrics.
3. Develop the Quality Monitoring process.
4. Implement the Quality Monitoring process.

5. Conduct continuous improvement reviews.
6. Transition to ongoing Information Quality Management.

7.5 Measuring and monitoring information quality

Establishing Information Quality Measures and Monitoring supports a focused effort on quality metrics that repeatedly retest the data environment for the critical defects that are identified in the Data Quality Assessment or through defined Data Rules. This evaluation occurs regularly (for example, monthly) and is driven by data stewards. During this phase, you should expect a reduction in the number of critical defects to zero or to some other acceptable level and assume a level of issue remediation or resolution. This reduction can be shown as a monitoring trend. As the required quality goals of the data environment are eventually satisfied, the monitoring of improvements in the total data environment can continue as is or be replaced by quality monitoring based on a statistically valid sample of the data.

7.5.1 Establishing priorities

The first step in preparing to perform focused data quality monitoring is to select, define, and prioritize the data quality issues that are identified in the Data Quality Assessment report or through the definition and evaluation of Data Validation rules. Some or all of this step might be done by management in response to previously reported results.

Usually these issues involve the implementation of corrective action for the data defects that have follow-up action plans. For reporting purposes, group the individual data defect types into broad issue categories that include a number of defective data conditions that are combined into one metric measurement. For example, consider consolidating street, PO Box, city/town, state, country, and postal code defects into a single address rule set that can generate specific statistics across the records-in-error.

Consolidation includes grouping across levels of evaluation and by grouping data elements into related subjects. Consolidation helps to focus the data quality improvement and monitoring by limiting or reducing the rules or issues that must be tracked and monitored (that is, it is easier to monitor a single rule set that looks consolidates 15 rules than to track and monitor 15 distinct rules, especially when expanded across multiple areas of focus).

7.5.2 Setting objectives and benchmarks

Set data quality objectives, or benchmarks, that reflect the minimally acceptable level of data quality. This is a critical step in Data Quality Monitoring. The benchmark is expressed as the percentage of records in the data environment that must be defect-free, and should align to the Governance Policies and Rules for those data elements. Give each condition its own benchmark. For example, the Customer Insight warehouse has the categories and benchmarks that are shown in Table 7-1.

Table 7-1 Example Data Quality benchmarks

Key Terms (Categories)	Benchmark	Business Object
Customer Name	100%	The organization must know who its customers are. This allows the organization to target customers with new products or to target potential new customers with existing products.
Customer Address	100%	The organization must know where its customers are located. This allows the organization to deliver information to target customers.
Customer Age	98%	Most organizations like to capture customers at a young age and retain them. If an organization finds that its customer age profile is increasing, it could indicate a potential decline in business.
Customer Preferred Contact Method	95%	Customers want organizations to deliver information through specific channels (email, flyers, and so forth)
Customer Do Not Solicit Indicator	100%	Customers who indicated this status do not want any solicitations and might withdraw their business, if solicited.

Important: A 100% defect-free benchmark might not be the optimal benchmark for a rule or measure. The expense of becoming defect-free must be weighed against the effect on the business of the data defect. Some industries and data environments are more intolerant of defects than others. These factors must be considered when objectives are set.

7.5.3 Developing the monitoring process

Use these requirements to develop a monitoring system in Information Analyzer that can be run periodically as an automated procedure. Data Rules can be brought forward from earlier Data Quality Assessments or other projects to support the monitoring initiative. Rule sets can consolidate multiple rules into focused statistics. Metrics can be put in place to establish the appropriate weighting or criticality to the measurements and benchmarks.

Consider the following example:

The business implemented the rule that identifies missing Customer Classification data for the Customer Loyalty Tracking initiative. They expect the data to be perfectly valid from day 1 to ensure a valid marketing campaign. However, working with the targeted data sample, they quickly find that over 93% of the data is missing, which is a significant issue for the business process.

An effort is undertaken to remediate the issues by working through different sources and work units. Over a two-month period, corrections are made and tracked until they reduce errors to about 25%. While the business decided that they should adjust their target to a 5% error level, they still have some work to do, as shown in Figure 7-10.

CSTINSIGHT_CST_...									
View Output									
<input type="checkbox"/> Include Tests									
History									
Type	Timestamp	Status	Total Records	Statistics				Validity	
				# Met	% Met	# Not Met	% Not Met	Benchmark	Variance
Baseline	9/23/2012 7:41 PM	✖	340	23	6.76470588 %	317	93.23529412 %	% Not Met <= 0.0000	93.2353 %
Run	10/5/2012 10:07 AM	✖	340	117	34.41176470 %	223	65.58823529 %	% Not Met <= 0.0000	65.5882 %
Run	10/16/2012 11:59 PM	✖	340	128	37.64705882 %	212	62.35294118 %	% Not Met <= 0.0000	62.3529 %
Run	10/28/2012 1:49 PM	✖	340	158	46.47058824 %	182	53.52941176 %	% Not Met <= 0.0000	53.5294 %
Run	11/9/2012 3:38 AM	✖	340	137	40.29411765 %	203	59.70588235 %	% Not Met <= 0.0000	59.7059 %
Run	11/20/2012 5:36 PM	✖	340	256	75.29411765 %	84	24.70588235 %	% Not Met <= 5.0000	19.7059 %

Figure 7-10 Execution of Customer Classification rule over six points in time

Since monitoring requires at least two data points for trend analysis, run a first cycle that represents the current condition (baseline) of the data. Thereafter, run the Information Quality Monitoring process on a schedule as a fully automated procedure. Each subsequent cycle of the monitoring process creates more data points, which reflects the changes in data quality over time, as seen in Figure 7-11 on page 139.

You can view all information from one dashboard. If you want to view full results of an analysis or data-rule runs, you can perform the following tasks:

- ▶ Export or extract results for more analysis.
- ▶ Report results and then deliver those results to others on an as-needed or scheduled basis and through various alternative formats.

You can develop a report by using standard report templates (various report templates are available). Some templates are for the data quality analysis functions and some are associated with data rules, rule sets, and metrics. You also can output data rule results to named tables for use by other reporting tools or extract data from the InfoSphere Information Analyzer repository into external tools to generate reports or dynamic dashboards. For an example of an implemented report through IBM Cognos, see Figure 7-13.

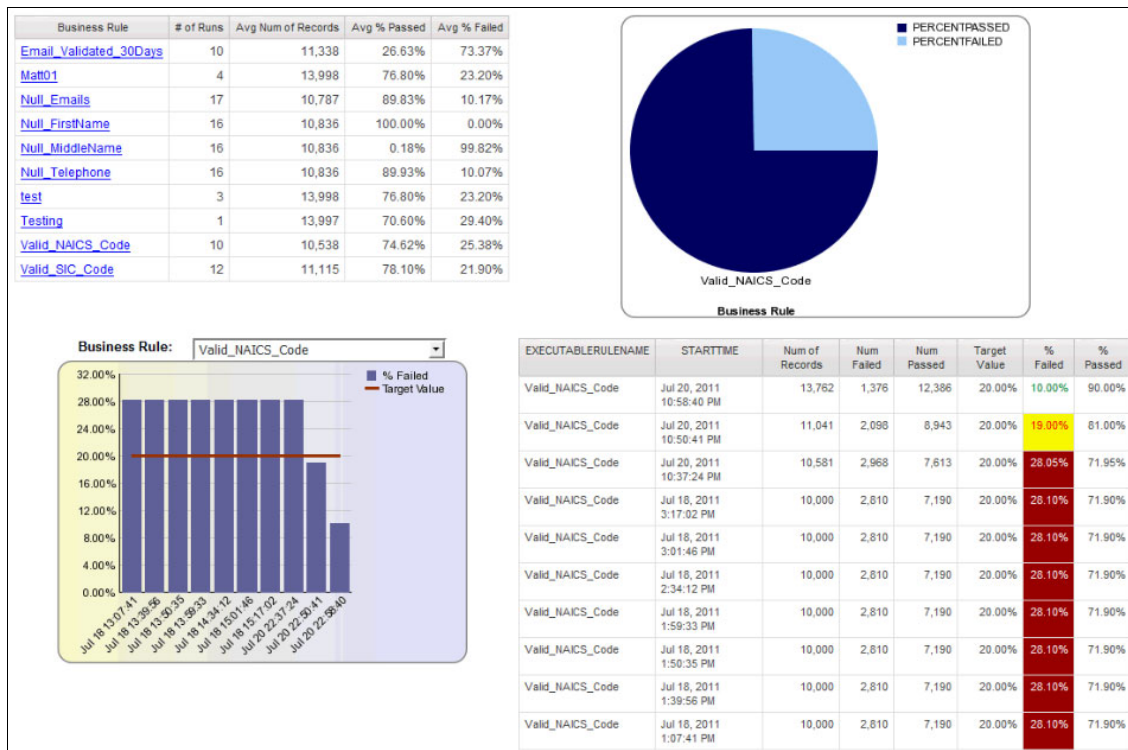


Figure 7-13 Integrated Data Quality dashboard through IBM Cognos reporting

Typically, rules results tables are stored in the Information Analyzer database repository with unique system-generated tables names. Those tables are immediately available for use in “viewing results” for rules only. However, Information Analyzer's data rules also can be configured to write to user-named rule output tables. This approach provides for the following results:

- ▶ Collecting data in targeted output tables that are readily accessible to external applications.
- ▶ Appending data to an existing user-defined table.
- ▶ Sharing of tables between rules (so multiple rules can update the same table).
- ▶ Delivering output of one rule as the input to another (for example, filtering the Customer Insight Warehouse Customer table to only include and evaluate Active Customers or exception data in subsequent rules).
- ▶ Enhanced custom reporting through IBM Cognos or other reporting tools.

When the data environment reaches the agreed upon goals for acceptable data quality, make the transition to the final (and permanent) phase of Information Quality Management, which represents a high level of maturity in the Information Governance Framework.

7.6 Information Quality Management

Information Quality Management is the practice that data quality requirements and transforms the actions to meet them into an effective Information Governance discipline.

7.6.1 Lifecycle and deploying Data Rules

One core part of Information Quality Management is managing the lifecycle and the deployment of Data Rules. *Lifecycle management* is the practice and process of determining when a rule (or a new version of a rule) has reached an acceptable level for ongoing usage, or when it is ready to be deprecated or discontinued. *Deployment* is the practice and process of moving one or more data rules from a development to a production environment.

InfoSphere Information Analyzer operates within the confines of a project. Every analysis task or every rule definition testing and run task is done within the context of an InfoSphere Information Analyzer project. The project consists of a set of data assets, tables, columns, files, and fields that are selected from the available metadata that were imported into the repository. It includes a group of authorized users with their designated roles within InfoSphere Information Analyzer. Roles for users are defined in the InfoSphere Information Server administrative console with specific roles for analysis and separate roles for data rules.

Any number of projects can exist simultaneously. They can have the same or different data sources and the same or different users and roles. An authorized user in a project can perform various tasks, running analysis tasks, data quality-related tasks, or both.

Rules that are created in one project are visible only to users who are authorized in that project. Rules can be published to make them visible and available for use in all other projects. Project staff can manage its rules within a multilevel structure of folders that it can create to suit its needs, distribution of subjects, responsibilities, or deployment schedules, as shown in Figure 7-14.

Quality Controls				
Name	Type	Description	Status	Data
▼ All		Global category for all QualityComponents		
◊ ActiveIndividualCustomerClassMissing	Data Rule Definition	Customer Class Missing for Active Individual Customers	Candidate	1
◊ ChannelTypeValidRef	Data Rule Definition	Channel Type in reference list	Candidate	1
◊ CostAmtInValidRange	Data Rule Definition	Cost Amount in valid range; applied to numeric data	Candidate	0
◊ CustAgeInRangeNumeric	Data Rule Definition	Age: Age >= 15 and < 100; applied to numeric age data	Candidate	0
◊ CustomerExists	Data Rule Definition	CustomerID Exists; null check only	Candidate	0
◊ CustomerTransactionStatusValid	Data Rule Definition	Customer Transaction Status in reference list	Candidate	1
◊ GrossIncomeInValidRange	Data Rule Definition	Gross Income in valid range; applied to numeric data	Candidate	0
◊ NameExists	Data Rule Definition	Name Exists; null and blank value check	Accepted	0
◊ NetIncomeInValidRange	Data Rule Definition	Net Income in valid range; applied to numeric data	Candidate	0
◊ PaymentMethodValidRef	Data Rule Definition	Payment Method in reference list	Candidate	1
◊ ProfitAmtValid	Data Rule Definition	Profit Amount computation is valid (arithmetic); applied to numeric data	Candidate	0
◊ QtyInValidRange	Data Rule Definition	Quantity in valid range; applied to numeric data; remove one condition if range is unbounded	Accepted	0
◊ TransactionTypeValidRef	Data Rule Definition	Transaction Type in reference list	Candidate	0
◊ TxnQtyInValidRange	Data Rule Definition	Transaction Quantity in valid range; applied to numeric data	Candidate	0
◊ CST_TXN_CustomerTransactionStatusValid	Data Rule Definition	CST_TXN_CustomerTransactionStatusValid	Draft	0
◊ CST_TXN_PaymentMethodValidRef	Data Rule	CST_TXN_PaymentMethodValidRef	Draft	0
◊ CSTINSIGHT_CST_ActiveIndividualCustomerClassMissing	Data Rule	CSTINSIGHT_CST_ActiveIndividualCustomerClassMissing	Draft	0
◊ CSTINSIGHT_TXN_ChannelTypeValidRef	Data Rule	CSTINSIGHT_TXN_ChannelTypeValidRef	Draft	0
▼ Customer Insight				
▼ Customer				
◊ CustAgeInRangeNumeric	Data Rule Definition	Age: Age >= 15 and < 100; applied to numeric age data	Candidate	0
◊ CustomerExists	Data Rule Definition	CustomerID Exists; null check only	Candidate	0
▼ Transaction				
◊ ChannelTypeValidRef	Data Rule Definition	Channel Type in reference list	Candidate	1

Figure 7-14 Management of Data Quality rules

A user with the role of rule manager can decide if a rule is ready to be promoted from Draft and Candidate to Approved or Standard status. It is up to the manager of the organization (through policies and process) to decide the meaning of the status and how to use it. Promotion from draft to candidate might be that the rule moved from the stage of formulation to evaluation where possibly more tests are required to determine that it serves the intended purpose.

Some rules can be basic and simple enough that they can be promoted directly to approved or standard, but others might require more testing and evaluation. After changing the status to approved or standard, the rule is locked from further editing. A rule cannot be changed intentionally or erroneously unless its status is changed back to draft or candidate by the rule manager. The status of rules is displayed in the Status column in the rule overview pane. The change to the rule status is applied in the rule edit pane.

7.6.2 Publishing data rules for reuse

Rules also can be published from a project for broader reuse across the organization. Publication provides a method for other data analysts in other projects or developers who are adding validation rules into in-process job flows to consistently use the same criteria. This helps to achieve greater trust across the organization that the right rules are in place.

For instance, to share a rule definition with others you should perform the following tasks:

- ▶ Provide a standard, consistent rule form.
- ▶ Use knowledge that is already established.
- ▶ Use a typical definition pattern as a base for new definitions.

As shown in Figure 7-15, a set of common rule definitions were published for general usage.

◊ CSTINSIGHT_TXN_ChannelTypeV	Data Rule	CSTINSIGHT_TXN_ChannelTypeValidRef	Draft
▶ Customer Insight			
▶ Published Rules			
▶ AdultInrangeCalc	Data Rule Definition	Derived Age Adult: Age >= 18 and < 125; applied to derived age calculated as the absolute va	Accepted
▶ AdultInrangeNumeric	Data Rule Definition	Adult: Age >= 18 and < 125; applied to numeric age data	Accepted
▶ AdultInrangeString	Data Rule Definition	String data Adult: Age >= 18 and < 125; applied to string age data	Accepted
▶ AgeInrangeCalc	Data Rule Definition	Derived Age: Age >= 0 and < 125; applied to derived age calculated as the absolute value of t	Accepted
▶ AgeInrangeNumeric	Data Rule Definition	Age: Age >= 0 and < 125; applied to numeric age data	Accepted
▶ AgeInrangeString	Data Rule Definition	String data Age: Age >= 0 and < 125; applied to string age data	Accepted
▶ AlphaFormatValid	Data Rule Definition	Example Alphabetic Format; excludes null values	Accepted

Figure 7-15 Published rules for cross-project usage

7.6.3 Deploying data rules to production

Rule Deployment supports a typical lifecycle approach that moves rules through a standard cycle of development, test, and production. The developed and approved validation, exception management, and remediation processes are migrated and deployed together as a package across defined deployment paths to the target run environment. For one-time, source system cleansing processes, this deployment can represent a limited or special exception to standard production procedures. Typically for this pattern and for all other Managed Data Cleansing patterns, the deployment follows standard production procedures established for ETL and application environments.

After a rule, rule sets, or metrics were defined, tested, and approved in the development environment, they must be moved and deployed in the production environment. This task is performed by the rule administrator, who exports the rules from the development environment and imports them into the production environment.

With Information Analyzer, an export package is created by a rule administrator from one environment for delivery, and import and deployment in another. For more information, see “Deploying data quality components in the IBM InfoSphere Information Center at this website:

<http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp?topic=%2Fcom.ibm.swg.im.iis.dataclick.doc%2Ftopics%2Fclickplugin.html>

The export package can include: data rule definitions, rule set definitions, data rules, global variables, rule sets, and metrics. After the package is imported in the second environment, the data rules and rule sets must be checked to ensure that all variables in the rule logic, join conditions, and output are properly bound to data sources that are available to that environment. This is handled through standard Information Analyzer processes for setting bindings, setting join keys, and setting output conditions.

Alternately, deployment can use the available HTTP APIs to retrieve and export sets of rule information and subsequently deploy and import them into another environment. This approach supports the use of scripting tools to configure standard deployment and run processes.

All import and export actions produce an audit event that can be traced by administrators.

As a preferred practice, the approach for rule definition, testing, and deployment must be clearly established. In such an environment, naming standards become important in moving from the initial design and testing to the deployed rules, rule sets, or metrics. Rules are no longer under the control of a single user, but people in other roles must now schedule, run, and monitor them. The ability to correctly identify the rule, rule set, or metric is paramount.

7.6.4 Running Data Rules in production

Running rules, particularly in a standard production environment, should be handled in a formalized manner to ensure that the right sets of rules are run in a timely and consistent manner. The management and ongoing monitoring of deployed components for running and tracking the Information Integrity policies are performed by following defined approaches for evaluating the effectiveness of the policies in maintaining or improving the defined data quality and cleansing goals.

For one-time, source system cleansing processes, this management evaluation can represent single review of the run processes and policies to verify whether the defined Managed Data Cleansing objective was met. For all other ongoing Data Quality processes, the evaluation and management tracks the results and trends of the run processes and policies to verify whether the defined Managed Data Cleansing objectives are met.

A Data Operator can be assigned to run Rules and Reports via a standard scheduler (external or internal) or through command-line scripts. While ad hoc runs can still occur, the use of a scheduled approach helps ensure that Data Stewards understand when to review and track the data rules for which they are responsible.

7.6.5 Delivering and managing data quality results

The results of each Quality Monitoring cycle should be disseminated to the appropriate managers, data stewards, or data owners. Publication can be through face-to-face meetings or through distribution of reports to the appropriate managers or data owners. However, as an ongoing practice, regular result distribution channels are recommended that include the following tasks:

- ▶ Review results via Information Analyzer UI or Data Quality Console.
- ▶ Generate and deliver reports.
- ▶ Extract and publish results.
- ▶ Manage results tables over time (for example, archive or purge).

As shown in Figure 7-10 on page 138 and Figure 7-11 on page 139, results can be viewed directly in Information Analyzer.

Exceptions that are generated from the data rules also can be processed through the Data Quality Console interface. This web-based UI supports data stewards who must bring reviewers in to handle exception processing. Exceptions can be tracked by status and priority, as shown in Figure 7-16.

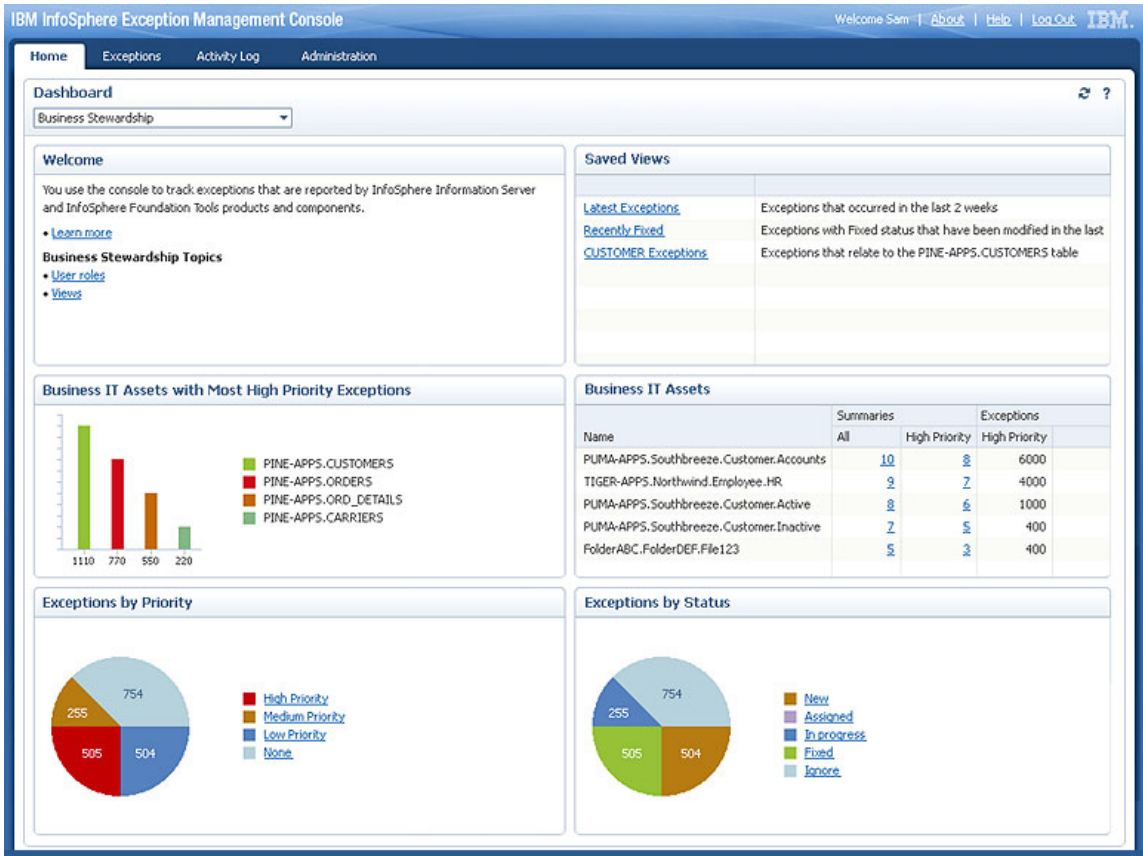


Figure 7-16 Data Quality Console UI

Further, data stewards can assign exceptions to specific individuals for review and analysis. This is a key process in the broader handling of Information Quality Management tasks, as shown in Figure 7-17.

Status	Priority	Exceptions	Objects Processed	Business IT Asset	Owner
		2000	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		1000	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		500	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		500	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		500	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		250	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		250	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		400	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		400	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None
		200	5000	PUMA-APPS.Southbreeze.Customer.Accounts	None

Figure 7-17 Assignment of exceptions to analysts for review

7.6.6 Developing Information Quality reports

You can develop a report by using standard report templates. Information Analyzer provides over 80 ready-to-use reports, for Data Quality Assessment and Quality Monitoring. Some of these focus on run history, others on ongoing trends in data quality or the current list of data quality rules within a project. Reports can be run ad hoc or scheduled and support delivery in various output formats including HTML, XML, and PDF that can be viewed by a wide audience.

A list of some available reports is shown in Figure 7-18.

▼ Data Rule Exception	
◊ Data Rule Exception Report	Provides an output of rows associated with a specific Data Rule exception
◊ Data Rule Exception Table Chart Report - Summary	Provides a graphical output of rows associated with a specific Data Rule Exception Table and Column
◊ Rule Set Exception Report	Provides an output of rows associated with a specific Data Rule exception
▼ Data Rule Summary	
◊ Data Rule Execution History Summary	Execution History summary for selected data rules.
◊ Data Rule Summary Project List	List of Data Rule Definitions and/or Data Rules.
◊ Project Folder Data Rule Validation List	List of Validation Details of Data Rules
▶ Domain Quality Summary	
▼ DO Dashboard	
◊ Dashboard Report	Dashboard Report
▶ Foreign Key Analysis	
▶ Metadata Summary	
▼ Metric Summary	
◊ Metric Execution History Summary	Execution History summary for selected data metric.
◊ Metric Summary Project List	List of metrics.
◊ Project Folder Metric List	List of Metrics and its details for selected folders
▶ Project Status	
▶ Project Summary	
▼ Rule Set Summary	
◊ Project Folder Rule Set Validation List - Summary	Rule Set Validation Summary for selected Rule Sets
◊ Rule Set Execution History Statistics	Execution History Summary for selected Rule Sets
◊ Rule Set Summary Project List	List of Rule Set Definitions and/or Rule Sets.

Figure 7-18 A list of some selected Data Quality reports in Information Analyzer

You can use a command-line interface (CLI) or HTTP to request the run of various InfoSphere Information Analyzer commands or to extract data from the InfoSphere Information Analyzer results repository (IADB). This way, you can create custom reports and extract data to populate data marts that feed dashboards.

The following types of functions are covered by the HTTP and CLI API:

- ▶ Creation and modification of InfoSphere Information Analyzer projects
- ▶ Registration of sources in the project
- ▶ Creation of virtual columns and virtual tables
- ▶ Creation, modification, and deletion of rules and rule sets
- ▶ Creation and modification of global variables
- ▶ Run column analysis (base profile)
- ▶ Retrieval of column analysis results (base profile)
- ▶ Retrieval of frequency distributions
- ▶ Running of rules and rule sets
- ▶ Retrieval of rules and rule sets run history
- ▶ Retrieval of rules and rule sets output tables

With this functionality, you can integrate InfoSphere Information Analyzer in third-party environments without the use of the rich client. You can create custom reports by extracting and combining results of several requests in one report and use XSLT to format, as needed.

7.6.7 Retaining and archiving old results

Management of data quality results includes the ability to archive or purge detail results. Over time, organizations that often use data quality buildup large numbers of output tables with detailed content. As with most data, there is a lifespan to this information and organizations must decide what to retain online, archive, and purge. This can vary based on the criticality or risk that is related to the specific business element as defined in Information Governance policies.

The Information Server metadata repository (XMETA) contains the summarized information quality results while the Information Analyzer analysis results repository (IADB) contains detail-level quality results (for example, Frequency Distributions, Rule Exceptions). Both of these repositories should be backed up regularly by using standard database backup procedures. Archiving strategies should focus on frequency distribution tables from Data Quality Assessments or user-named output tables from data rules because these tables are the easiest to identify and apply archiving rules to.

All data quality output tables can be manually purged from a project by using the IBM InfoSphere Information Analyzer user interface or APIs. Output tables also can be purged automatically through settings in the project's properties to establish consistent global or project level practices. Output Tables can be purged automatically based on age or number of runs, or can be purged on a per rule basis.

7.6.8 Improving ongoing processes

Ongoing Process Improvement encompasses the maturing of Information Quality Management practices, particularly those practices that are focused on making the processes more efficient.

After changes are implemented to the warehouse, process improvement is useful for the following reasons:

- ▶ Provides continuous feedback on the data quality condition of the warehouse.
- ▶ Detects unexpected changes in source data feeds to the data warehouse early.
- ▶ Reinforces users' confidence in the usability of the data warehouse on a continual basis through monitoring reports.

Where changes to policies are required, this activity assesses what policies (whether deployed as rules or processes) to alter and how those alterations should occur.

Stewards and information owners take information about policy change, which is typically received via content-based documents (such as email and Word documents), and translates that information into associated business concepts.

Using a Business Glossary facilitates greater understanding. Where a Business Glossary is mapped to associated data assets (including data sources and rules), the stewards and owners begin the evaluation of potential change.

Impact Analysis and Data Lineage reports that are defined in Impact Analysis and Manage Data Lineage tasks facilitate the work of tracking potential change and enforcement.

Rules that are developed through IBM InfoSphere Information Analyzer can use the Usage section for individual rule definitions to identify associated and deployed rules.

Stewards and information owners must determine whether the policy changes are local (specific to single or sets of rules or processes) or global (applicable to all related rules and processes).

Changes to existing policies can alter the following processes and procedures:

- ▶ Validation rules
For example, adding new valid values to a rule, adding more rule conditions to data elements, and so forth.
- ▶ Timing of validation processes
For example, validation must occur in-stream after data consolidation rather than after data load.
- ▶ Data flows based on validations
For example, runs must go to an exception process rather than continuing to the target system.
- ▶ Exception management processes.
For example, changes to data are now made in the originating source rather than the target system. Data Stewards must monitor and report on conditions to downstream users.

After the impact of policy change is evaluated, changes to policies can be identified, scheduled, and deployed.

Throughout the lifecycle of the data quality management environment, conduct a complete audit of the data quality practices and processing periodically; for example, annually or biannually. Frequency is determined by the volatility of the data. A full data quality environment audit is conducted for the following reasons:

- ▶ To demonstrate the effectiveness of the rules used in Quality Monitoring.
- ▶ To identify any significant changes in the data, such as the introduction of new valid values.

The audit consists of running or rerunning the Data Quality Assessment with particular focus on new valid values that were accepted into the data environment or new Governance Policies and Rules that must be considered. (You might need to update validity tables or range tables if this is the case.)

You also might need to modify Data Rules to support new data situations. You also must verify that the conditions that are evaluated are still the conditions that are critical to the business and update accordingly rules might need to be deprecated and added to the Quality Monitoring system.

7.7 Conclusion

When an organization focuses on improving Information Quality, it initiates a process that is designed to establish trust in the information that serves its business. Information Quality is not a matter of running a tool on a set of data and producing a set of statistics on values. Information Quality is an active process that involves people and process that use tools to facilitate initial assessment, validate data according to data rules that are driven by business requirements, establish metrics and benchmarks to track improvements (or degradation), and integrate into a broader Information Governance strategy.



Data standardization and matching

In general, populating a data warehouse is the process of integrating, consolidating, and aggregating data from various disparate data sources. These sources are rarely designed by a single authority. More often, they are designed through individual, siloed projects or introduced through mergers and acquisitions where each one uses its own types of identifiers and maintain its own version of the truth. Trying to integrate such fragmented data into a single system presents a number of data quality (DQ) problems that must be addressed before loading the data into the Data Warehouse to ensure accurate, complete, and consistent information.

The success and value of the data warehouse is determined by the adoption of business users. Only if they find the information reliable and trustworthy will they use it for critical business decisions.

8.1 Use cases

In section 7.3.1, “Data Quality Assessment” on page 121, we described the importance of Data Quality Assessments (DQA) to gain an understanding of the data and to provide insights about existing or potential data quality issues. We also explained that evaluating data quality results, drawing conclusions, and establishing an action plan always must be made in connection with a specific business case and its business impact. This relationship can be shown when comparing the following use cases:

- Use Case A: Stand-alone data quality assessment of an operational data source
- Use Case B: Assessing the source data quality for a project to build and load a customer data warehouse

For Use Case A, Assessing the data quality: An assessment of an operational data source is made to identify data entry issues (for example, incomplete or missing data validation). As a consequence, an action plan should include an improvement of the data entry process.

For Use Case B, Loading a data warehouse: Cross-data source profiling can identify conflicting and inconsistent information even though the data quality for each individual source was determined to be satisfactory. Such problems might not be resolved by changing the data entry process, but must be reconciled through a data cleansing process before the data is loaded into the warehouse. The data cleansing process is typically run as a workflow of various data cleansing tasks by which the data is parsed, standardized, verified, de-duplicated, consolidated, and enriched, as shown in Figure 8-1.

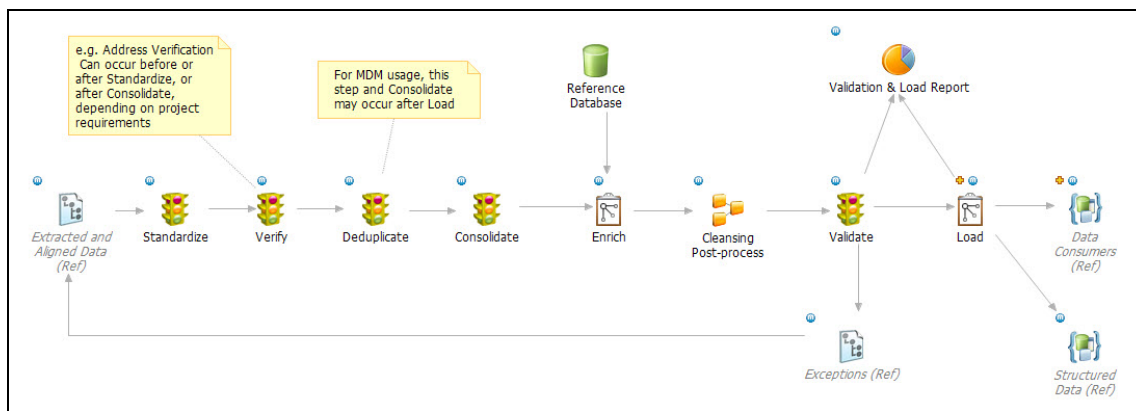


Figure 8-1 Stages in a Data Cleansing process and Initial Data Warehouse load

IBM InfoSphere QualityStage, an integral component of IBM Information Server, provides data cleansing functions on an easy-to-use, design-as-you-think flow diagram. The tight coupling between InfoSphere DataStage and QualityStage through a unified design and runtime environment allows for easy embedding of data cleansing tasks into any information integration process, as shown in Figure 8-2.

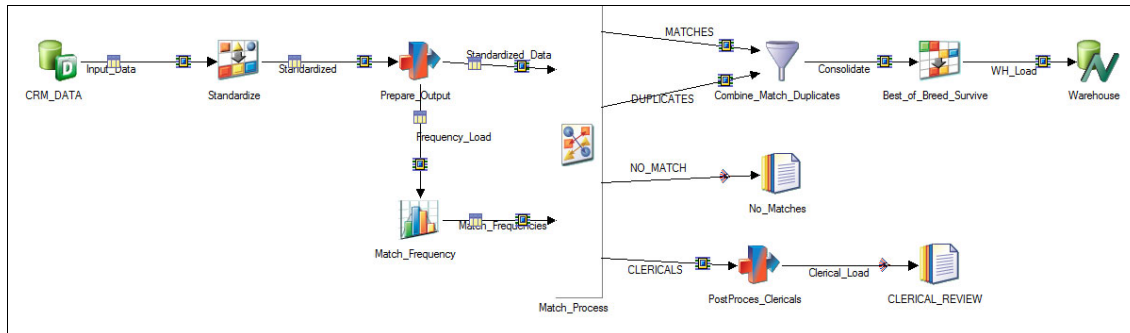


Figure 8-2 End to end extract, cleanse and load process

In a data warehouse loading project context, an organization can include any or all of the following methods to ensure information is complete, consistent, and accurate before such information is loaded into the warehouse:

- ▶ Source access or extraction
- ▶ Conditioning
- ▶ Standardization
- ▶ Address verification
- ▶ Matching
- ▶ Group association
- ▶ Survivorship
- ▶ Data enrichment
- ▶ Output formatting
- ▶ Auditing the load process

The methods do not prescribe that each one must be applied for all use cases, nor is it a prescription that they must be run in the given sequence. The business case, DQA results, and the data domain are the fundamental drivers of which data quality method should be applied. For example, loading product information into a warehouse or Product Information Management system does not require an address verification phase. Similarly, an initial load phase from a single source might not necessitate a reference matching phase.

In this chapter, we describe the following data quality methods:

- ▶ Conditioning and standardization
- ▶ Address verification
- ▶ Matching and de-duplication
- ▶ Consolidation and enrichment

8.1.1 Conditioning and standardization

Standardization is a process that is used to normalize the data to defined standards. This incorporates the ability to parse free-form data into single-domain data elements to create a consistent representation of the input data and to ensure data values conform to an organization's standard representation.

The standardization process can be logically divided into a conditioning or preparation and a standardization phase. Conditioning decomposes the input data to its lowest common denominators, based on specific data value occurrences. It then identifies and classifies the component data properly in terms of its business meaning and value. Following the conditioning of the data, standardization removes anomalies and standardizes spellings, abbreviations, punctuation, and logical structures (domains).

Conditioning and standardization of data is the first step you must take after DQA identified the data quality issues. The following DQ issues might require standardization:

- ▶ Lack of information standards

Identical information is entered differently across different information systems (for example, various phone, identifier, or date formats or address information). This makes the information look different and presents challenges when trying to match and consolidate such information.

- ▶ Unexpected data in individual fields

This describes a problem where data is misplaced into an incorrect data field or certain data fields are used for multiple purposes. For further data cleansing, the system must prepare the data to classify individual data entries into their specific data domains.

Table 8-1 on page 157 shows an example where address information is spread out into various Address_XX fields that contain a mixture of data domains within each field.

Table 8-1 Address information

Address_1	Address_2	Address_3	Address_4
IBM	555 Bailey Ave	San Jose, CA	95141
425 Market Street	San Francisco	CA	94111
4400 N 1st ST	#100	San Jose	95134

Information is buried in free-form text fields. Free-form text fields often carry valuable information or can be the only information source. To take advantage of such data for classification, enrichment, or analysis, it must be standardized first.

► Lack of consistent identifiers across different data systems

Disparate data sources often use their own proprietary identifiers. In addition, these sources can apply different data standards to their textual data fields and make it impossible to get a complete view across these systems.

Table 8-2 shows three products from three different systems. At first glance, they look different but they actually represent identical products.

Table 8-2 Product information

Identifier	Product
19-84-103	RS232 Cable 6' M-F CandS
CS-89641	6 ft. Cable Male-F, RS232 #87951
C&SUCH6	Male/Female 25 PIN 6 Foot Cable

Redundancy within different information systems can exist.

This issue is a side effect of the lack of information standards which causes the same information (for example, customer record) to be entered into an information system multiple times by using different spelling and formatting variations.

The QualityStage standardization process uses standardization rule sets to condition and standardize data based on specific domain requirements. Similar to data validation rules that were defined in Chapter 7, “Establishing trust by ensuring quality” on page 115, QualityStage provides a range of built-in rules that can be used as-is or as the starting point to customize and fine-tune the standardization to specific requirements.

The following built-in Standardization Rule Sets by Category are available:

► Country name and address standardization:

More than 20 countries are supported.

► Product Data:

- Pharmaceutical
- Generic Product data
- Outdoor product data

► Validation/Standardization:

- Phone numbers
- Dates
- Tax Identifiers
- Email
- Country
- Expanded Company Names

These built-in rule sets cover the most common data patterns within their specific domain. However, the standardization process is fully configurable and extendable to accommodate deviations in customer's data pattern and to ensure that input data is standardized as wanted.

If and how much customization you need to apply depends on the following factors:

- Data domain: Certain data domains (for example, location) are more standardized than others.
- Structure of, or the lack of, input data.

Based on the initial situation, such as input data domain and complexity, you might differentiate between a Data Analyst-driven process to fine-tune and enhance existing rule sets and a Developer driven process to create new rule sets. The following processes can be used:

- Data Analyst-driven process to fine-tune and enhance existing rule sets on an ongoing basis.

Because input data can vary over time (for example, introduction of new products, brands, or more attributes), the standardization rules must be adjusted to handle those changing situations without affecting the overall data cleansing process. Standardization rules can define override conditions that take precedence over the rules and conditions that are defined in the actual rule set.

In QualityStage, such fine-tuning of the standardization process is performed collaboratively by the Data Analyst who uses the QualityStage Standardization Rules Designer.

By using a web-based, drag-and-drop GUI, data analysts can see pattern rules and individual input records that require adjustment to the current standardization process.

To calibrate a standardization rule to changing input data conditions, a data analyst follows a reviewing and conditioning process, as shown in Figure 8-3.

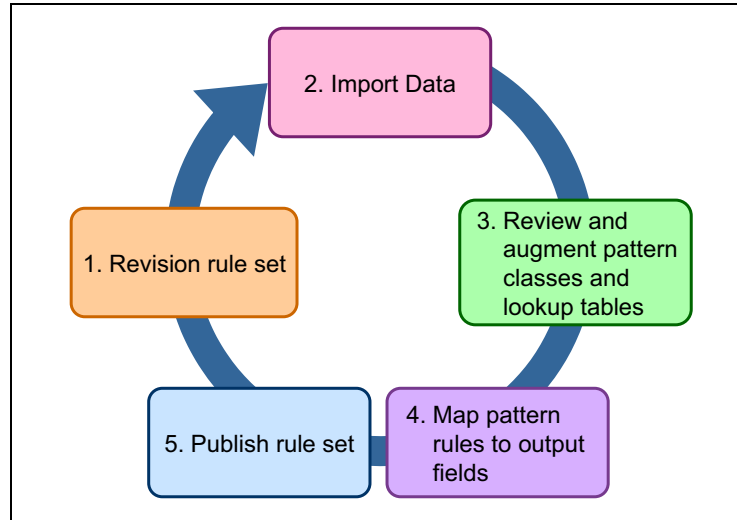


Figure 8-3 Rule set customization process

The process includes the following steps:

a. Create a rule set revision.

For a data analyst to make changes to an existing standardization rule set, a rule set must be placed in revision state (a form of checkout), which is easily done by opening a rule set in the QualityStage Standardization Rules Designer.

b. Import “un-handled” sample data.

Those data records that are not correctly standardized by this rule set must be imported by using the “Import Sample Data” link, as shown in Figure 8-4 on page 160. The import process parses the records and for each record determines the pattern rule by using the classification information that is contained in the rule set. For example, by using the Outdoor Retail rule set, the record EPOCH Infinity Men's Yellow Watch has the pattern B+SCT, as shown in Figure 8-4 on page 160.

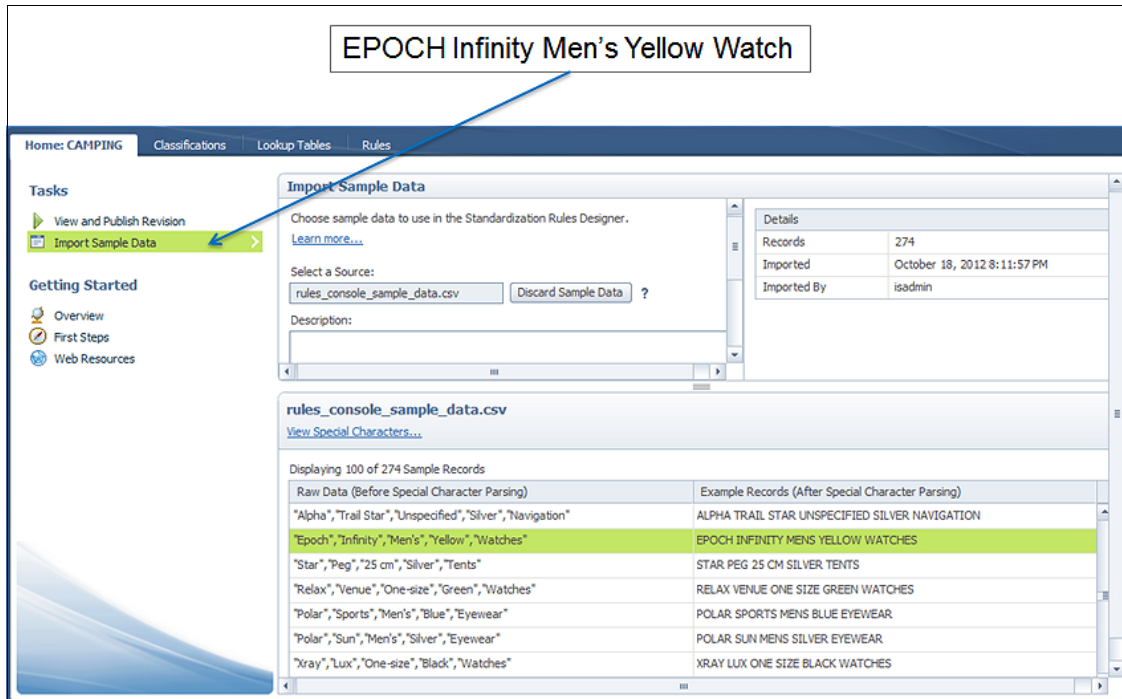


Figure 8-4 Import "un-handled" sample data

c. Review and augment pattern classes and lookup tables.

Changing data conditions is one reason that data records are not correctly standardized by the existing rule set. Changing conditions might include the introduction of new categories of data or new class values. By using the Outdoor Retail rule set, an example for requirements to augment class values is the inclusion of products from a new manufacturer into the retailer's portfolio. In such a case, the new product brand names, types, or other specific attributes are not recognized by the rule set and therefore must be added.

By using the Classification and Lookup Tables panes on the Standardization Rules Designer, a Data Analyst can easily review and extend the classifications and lookup tables, as shown in Figure 8-5 on page 161 and Figure 8-6 on page 162.

Class Description	Product Brand	Product Name	Product Size	Product Color	Product Type
Pattern	B	+	S	C	T

The screenshot shows the 'Classifications' tab in the CAMPING software. The 'Browse Classes' panel on the left lists various classes, with 'B Product Brand' selected. The 'B Product Brand' panel on the right shows a list of brands, with 'EPOCH' highlighted. The 'EPOCH' panel at the bottom shows the 'Define Value' tab with 'EPOCH' as the value and 'B' as the class.

Value	Standard Value	Class	Frequency	Count	Patterns
ALPHA	ALPHA	B	5.47%	15	3
BUGSHIELD	BUGSHIELD	B	1.82%	5	3
CANYON	CANYON	B	2.18%	6	3
COURSE	COURSE	B	1.82%	5	4
EDGE	EDGE	B	1.82%	9	4
EPOCH	EPOCH	B	12.04%	33	1
EVERGLOW	EVERGLOW	B	1.82%	5	2
EXTREME	EXTREME	B	4.01%	11	10
FIREFLY	FIREFLY	B	3.64%	10	8

EPOCH

Define Value Notes

If the **Apply** button is disabled, the selected definition is the active definition. To change the active definition, select a different definition and click **Apply**. ☐ Deactivate definitions

Value: Standard Value: Class: Similarity Threshold:

⊙ EPOCH EPOCH B 900 (Exact)

[Add Definition](#)

Figure 8-5 Add class values

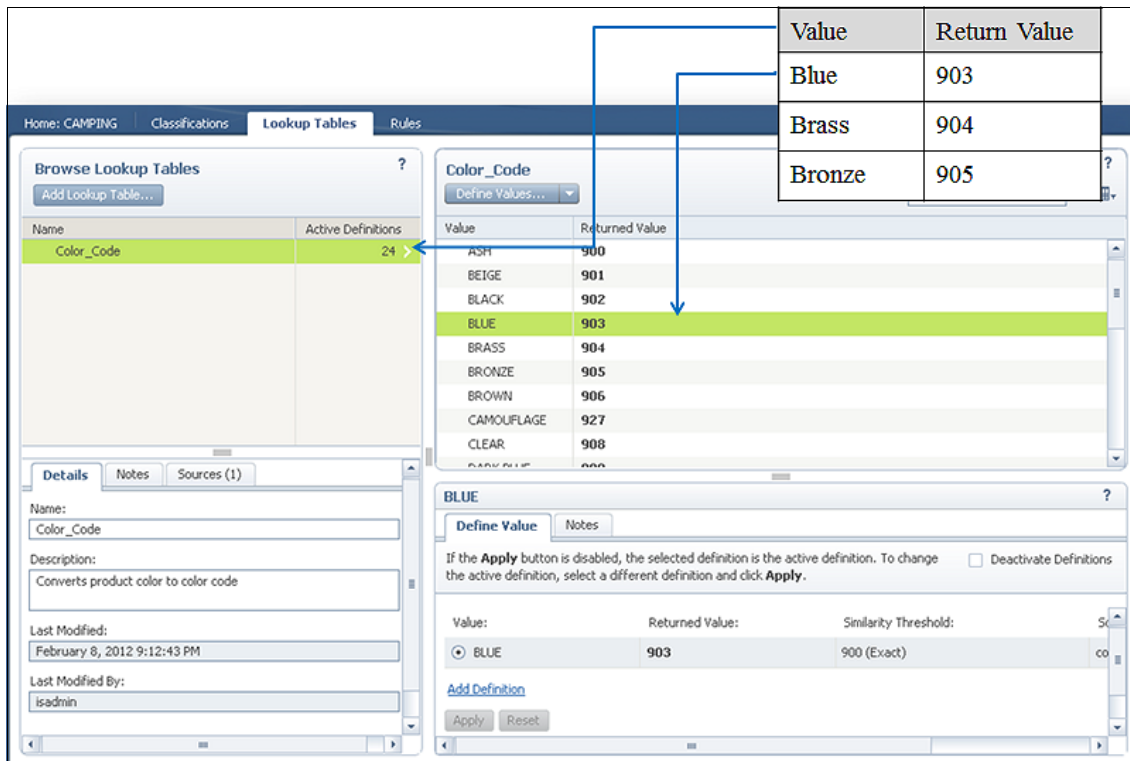


Figure 8-6 Add Lookup table and values

d. Map pattern rules to output fields.

Another form of changing data conditions is structural differences in the input data set to what is currently accepted for standardization by the rule set. Because of these differences, such records are ignored.

By using the un-handled records for individual input pattern rules, a data analyst can review un-handled patterns and through drag-and-drop paradigm, control how they should be mapped to output fields (see Figure 8-7 on page 163) and what corrective actions should be applied (Figure 8-8 on page 164) to ensure the entire data sets is properly standardized.

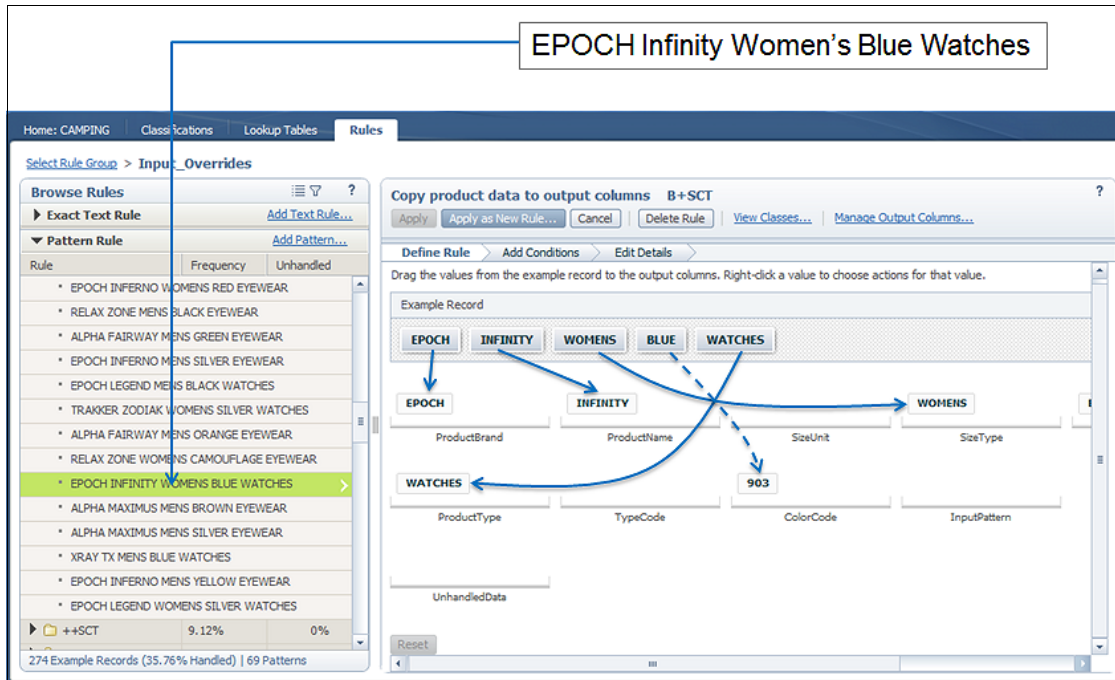


Figure 8-7 Mapping input records to output fields

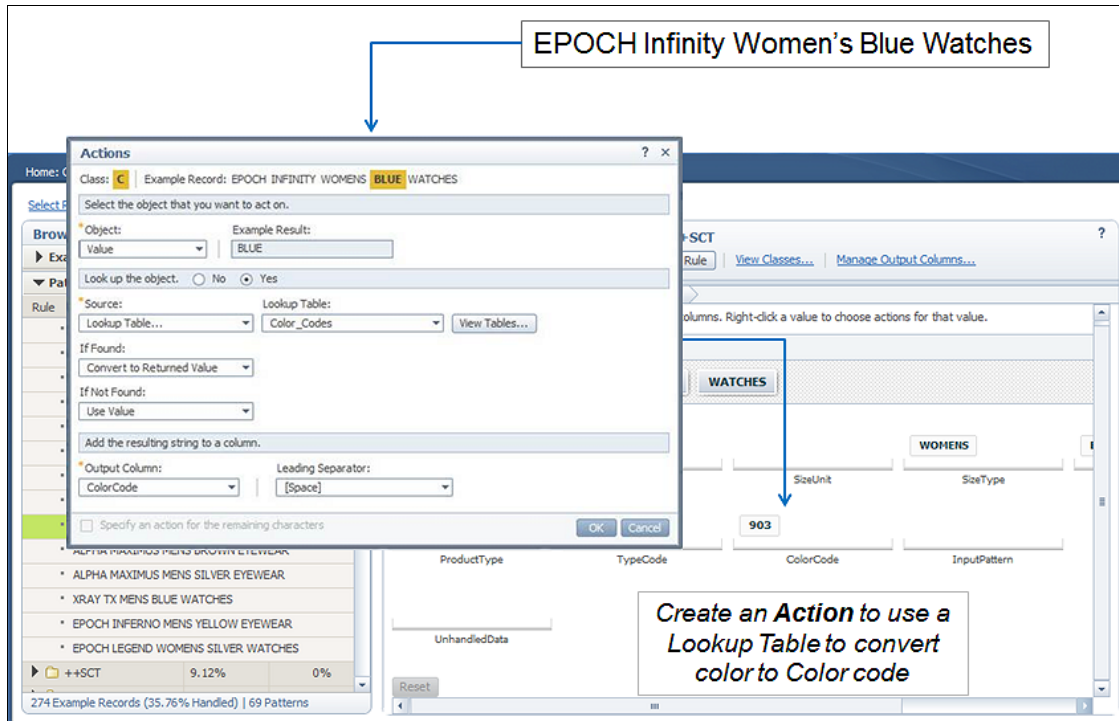


Figure 8-8 Applying a lookup action

e. Publish a rule set.

The final step is to publish the rule set to apply the changes back to the rule set that is used by a particular QualityStage project.

- Developer driven process: Create new standardization rule sets to handle custom domain data.

Even an extensive library of built-in rule sets might not always provide base standardization solutions for every data condition. However, a base standardization solution can be the prerequisite to apply a Data Analyst enhancement process.

To ensure that such data is appropriately standardized, QualityStage equips users with a development environment to design new rule sets that use the declarative Pattern-Action (PAL) language.

For more information about how to design new rule sets by using PAL, see the Pattern Action Reference guide that is included in the Information Server documentation, which is available at this website:

http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp?topic=%2Fcom.ibm.swg.im.iis.qs.patguide.doc%2Ftopics%2Fcontainer_topic.html

8.1.2 Address verification

Address verification is the process that is used to validate an address against a postal reference file that greatly increases the accuracy of your address data. There are many direct and indirect benefits of accurate address information, including increased response rate, lower shipping cost, increased effectiveness of direct marketing initiatives, and ultimately higher customer satisfaction.

Unlike standardization, which ensures only that an address format conforms to specific lexical rules, an address verification process determines if such address also is a valid physical address. For example, 555 Main Street, Springfield, CA 95444 US appears to be a valid US address, but address verification shows that it is, in fact, a fabricated address.

Worldwide address verification within QualityStage is provided through the QualityStage Address Verification Interface (AVI). AVI supports address verification for more than 240 countries and territories and includes the following features:

► Address parsing:

Performs a country-specific lexicon analysis and decomposes the address into its dedicated address fields without validation against a postal reference file, as shown in Table 8-3.

Table 8-3 Parsing table

Input	Output	
555 Bailey avenue, San Jose, CA, US	House Number	555
	Premise	Bailey Ave
	Locality	San Jose
	Admin Area	CA
	Country	US

► Address verification

Corrects and enhances an address by using country-specific postal reference files, as shown in Table 8-4.

Table 8-4 Verification table

Input	Output	
555 Bailey avenue, San Jose, CA, US	House Number	555
	Premise	Bailey Ave
	Locality	San Jose
	Primary Postcode	95141
	Secondary Postcode	1111
	Sub Admin Area	Santa Clara
	Admin Area	CA
	Country	US

► Address suggestion

If an address verification is ambiguous because of missing input information (for example, missing house number), the “suggestion” feature can be used to retrieve the various address alternatives, as shown in Table 8-5.

Table 8-5 Address Suggestion table

Input	Output	
Bailey avenue, San Jose, CA, US	1	400 Bailey Ave, San Jose, CA 95141
	2	500 Bailey Ave, San Jose, CA 95141
	3	555 Bailey Ave, San Jose, CA 95141

► Bidirectional Transliteration

For countries with non-Latin character sets, it supports the transliteration of an address between its native character sets and Latin, as shown in Figure 8-9 on page 167:

- Native → Latin → Native
- For example, Cyrillic → Latin or Latin → Simplified Chinese

Input	Output
БЕЛОВЕЖСКАЯ УЛ 39 МОСКВА 121353 RUS	Belovezskaja Ulica 39 Mozaiskii Moskva 121353, RUS

Figure 8-9 Bidirectional Transliteration table

You might question why you do use or should use standardization and address verification against your location domain data. The answer is that the processes complement each other. The complementation of AVI is that it verifies if and to what degree an address is recognized as a valid mailing address. But, AVI also can complement the standardization process for countries without built-in rule sets. In such a case, by using the stand-alone country-specific address parsing capability in AVI, you can standardize location data globally without the need to expand the country standardization rule sets.

8.1.3 Matching and de-duplication

After conditioning and standardization are completed, the match processing starts.

In a data warehouse scenario, data from multiple sources is consolidated into a single target. To avoid creating duplicate data in the warehouse, the source data must be matched against the data in the warehouse.

The objective of match processing is the establishment of entity-level relationships (client, household, vendor, product, or parts) across all input records, which generate each record's appropriate logical relationships. Unique entity keys are created to allow the organization to create entity-oriented views in addition to their existing operational views. The matching process can be set up to run independently for each defined entity relationship and can be run in multiple parallel streams, depending on business rules.

One of InfoSphere QualityStage's core strengths is its ability to precisely match data, even when it appears to be different. To do so, QualityStage uses a statistical matching technique called *Probabilistic Record Linkage*. This method evaluates each match field, taking into account frequency distribution, discriminating value, and data reliability, and produces a score (or match weight), which precisely measures the content of the matching fields. This measurement decisively gauges the probability of a match.

Probabilistic matching systems are the preferable choice in situations where data set sizes and numbers of attributes are large, and high levels of accuracy and low total cost of ownership are important.

InfoSphere QualityStage performs matching through one of its two match stages: One-source and Two-source match stages. Before matching can take place, a data analyst must configure the specific match conditions through the Match Designer user interface, as shown in Figure 8-10.

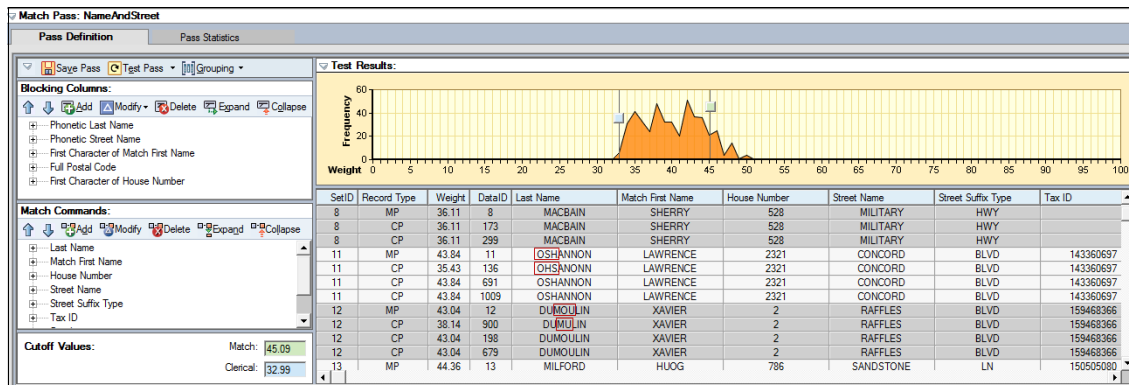


Figure 8-10 Match Designer Match Pass Configuration

A Match Specification includes the following fundamental principles:

- ▶ Match passes

Match passes are the method that is used to define specific match conditions. Each match specification can define any number of match passes to implement complementary or independent business rules to compensate for data errors, missing values in blocking columns, or reduce the complexity when you are dealing with large data volumes.

- ▶ Blocking columns

Blocking provides a method for limiting the number of record pairs to examine if it is infeasible to compare all record pairs for sources of reasonable size. Blocking partitions the sources into mutually exclusive and exhaustive subsets, and the matching process searches for matches only within a subset. If the subsets are designed to bring together pairs that have a higher likelihood of being matches and ignore those that are less likely matching pairs, successful matching becomes computationally feasible for large data volumes.

- ▶ Matching commands

Match commands are the method that is used to specify matching columns, match comparison types, agreement and disagreement weights and weight overrides.

► Cutoff Values

Match and clerical cutoffs are thresholds that determine how to categorize scored record pairs.

Cutoff values are based on the composite weight, which is assigned to each record pair. All record pairs with composite weight equal or above the match cutoff value are considered duplicates, record pairs below the clerical cutoff are considered non-matches and records with composite weight between the two values are considered clerical records.

Because the cutoff values have direct influence whether a record pair is considered a match, the actual business purpose should determine how aggressive or conservative those values might be defined. For example, if de-duplication is performed to create a mailing list for shopping catalogs, it might be acceptable to set more aggressive (lower) match cutoff value than you do with patient records.

8.1.4 Consolidation and enrichment

The data enrichment process (survivorship) creates a single representation of an entity across business lines with the “best of breed” data.

This process can be performed at the following levels:

- Record level
- Logical domain level (Name, Address, Product, and so forth)
- Field level
- Any combination of these levels

By storing survivorship records in the target data store for each match process, a common institutional representation of data can be viewed across the organization and ensures that the entire organization is using most complete and highest quality data.

Figure 8-11 shows various ready-to-use, “best of breed” survive rules. Users also can construct their own survive rules by using the Survive Stage expression builder.

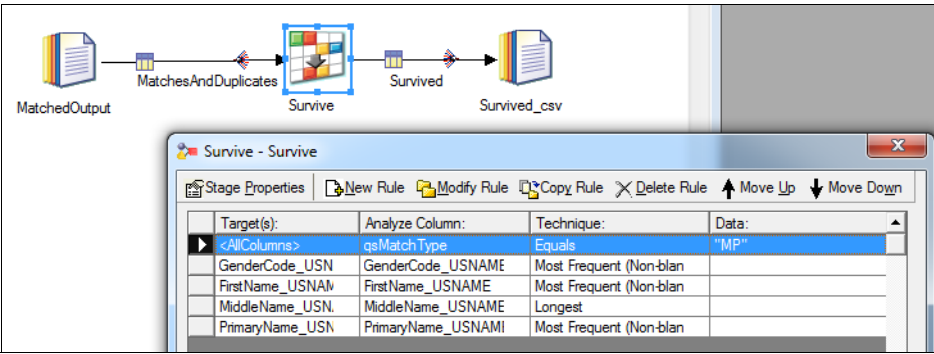


Figure 8-11 Defining Survivorship rules

8.1.5 Summary

For many critical business decisions, an Enterprise Data Warehouse is the central hub from which analytical data is sourced. As such a central nervous system, organizations must have an appropriate data quality process in place that ensures only first class data is fed into this hub.

In this chapter, we provided an introduction into the core set of data cleansing methods that often are applied when data is consolidated into a data warehouse. We also outlined that the sequence and emphasis towards individual methods is determined by the actual business case, the data domain, and data quality assessment results.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics that are covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide more information about the topic in this document. Note that some publications that are referenced in this list might be available in softcopy only:

- ▶ *Metadata Management with IBM Information Server*, SG24-7939
- ▶ *IBM WebSphere QualityStage Methodologies, Standardization, and Matching*, SG24-7546
- ▶ *IBM WebSphere Information Analyzer and Data Quality Assessment*, SG24-7508
- ▶ *IBM InfoSphere Information Server Installation and Configuration Guide*, SG24-4596
- ▶ *IBM InfoSphere Information Server Deployment Architectures*, SG24-8028
- ▶ *Smarter Business: Dynamic Information with IBM InfoSphere Data Replication CDC*, SG24-7941
- ▶ *IBM InfoSphere DataStage Data Flow and Job Design*, SG24-7576
- ▶ *InfoSphere DataStage Parallel Framework Standard Practices*, SG24-7830
- ▶ *Deploying a Grid Solution with the IBM InfoSphere Information Server*, SG24-7625

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft, and other materials, at this website:

<http://www.ibm.com/redbooks>

Other publications

The following publications also are relevant as further information sources:

- ▶ IBM InfoSphere Channel on YouTube:
<http://www.youtube.com/user/IBMinfosphere>
- ▶ IBM InfoSphere Information Server Integration Scenario Guide:
<http://publibfp.boulder.ibm.com/epubs/pdf/c1938050.pdf>
- ▶ IBM InfoSphere Information Server v9.1 Information Center,
<http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp>
- ▶ IBM developerWorks community: IBM InfoSphere Information Server Best Practices:
<https://www.ibm.com/developerworks/community/groups/service/html/communityview?communityUuid=23355691-5fbb-4d69-bcd9-1dd5358daa45>

Online resources

The following website also is relevant as another information source:

- ▶ IBM Information Server v9.1 Information Center, IBM, ©2011:
<http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp?topic=%2Fcom.ibm.svg.im.iis.dataclick.doc%2Ftopics%2Fclickplugin.html>

Help from IBM

IBM Support and downloads:

<https://ibm.com/support>

IBM Global Services:

<https://ibm.com/services>

IBM Information Server: Integration and Governance for Emerging

(0.2" spine)
0.17" <-> 0.473"
90 <-> 249 pages



IBM Information Server

Integration and Governance for Emerging Data Warehouse Demands



Information Quality and Governance

Self-service Data Integration

Big Data Integration

This IBM Redbooks publication is intended for business leaders and IT architects who are responsible for building and extending their data warehouse and Business Intelligence infrastructure. It provides an overview of powerful new capabilities of Information Server in the areas of big data, statistical models, data governance and data quality. The book also provides key technical details that IT professionals can use in solution planning, design, and implementation.

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