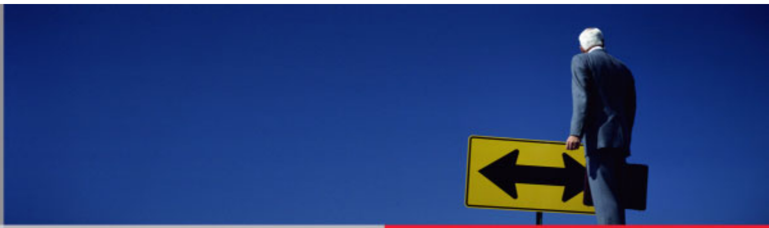
Governing and Managing Big Data for Analytics and Decision Makers



Redguides

for Business Leaders

Mandy Chessell
Ferd Scheepers
Nhan Nguyen
Ruud van Kessel
Ron van der Starre

- Gain insight into why a data reservoir adds value to analytic projects
- Understand the architectural details of a data reservoir
- Learn ways to deliver the value of information governance

Introduction

It is estimated that a staggering 70% of the time spent on analytics projects is concerned with identifying, cleansing, and integrating data, because of the following issues:

- Data is often difficult to locate because it is scattered among many business applications and business systems.
- Frequently the data needs reengineering and reformatting in order to make it easier to analyze.
- The data must be refreshed regularly to keep it up-to-date when it is in use by analytics.

Acquiring data for analytics in an ad hoc manner creates a huge burden on the teams that own the systems supplying data. Often the same type of data is repeatedly requested and the original information owner finds it hard to keep track of who has copies of which data.

As a result, many organizations are considering implementing a *data lake* solution. A data lake is a set of one or more data repositories that have been created to support data discovery, analytics, ad hoc investigations, and reporting. The data lake contains data from many different sources. People in the organization are free to add data to the data lake and access any updates as necessary.

However, without proper management and governance, such a data lake can quickly become a *data swamp*. A data swamp is overwhelming and unsafe to use because no-one is sure where data came from, how reliable it is, and how it should be protected. IBM® proposes an enhanced data lake solution that is built with management, affordability, and governance at its core. This solution is known as a *data reservoir*. A data reservoir provides the right information to people so they can perform the following activities:

- Investigate and understand a particular situation or type of activity.
- Build analytical models of the activity.
- Assess the success of an analytic solution in production in order to improve it.

A data reservoir has capabilities that ensure the data is properly cataloged and protected so subject matter experts have access to the data they need for their work. This design point is critical because subject matter experts play a crucial role in ensuring that analytics provides worthwhile and valuable insights at appropriate points in the organization's operation. With a data reservoir, line-of-business teams can take advantage of the data in the data reservoir to make decisions with confidence.

This IBM Redguide™ publication discusses the value of a data reservoir, discusses how it fits into the existing business IT environment, and identifies sources of data for the data reservoir. It also provides a high-level architecture of a data reservoir and discusses key components of that architecture. It identifies key roles essential to creating, supporting, and maintaining the data reservoir and how information integration and governance play a pivotal role in supporting the data reservoir.

A view from ING

Before diving into the details of the data reservoir, it is worth pausing to consider the business value of the data reservoir solution. This solution was the result of a partnership between ING and IBM.

Ferd Scheepers of ING stated "Having both the perspective of a large international bank and one of the biggest IT vendors in the world, and challenging each other at every turn really was a great example of what a partnership can be. The data reservoir architecture was developed to work independent of technology choices. But having IBM at the table helped in making sure that it didn't become a great architecture on paper only, but an architecture that can be realized by technology that is available today, whilst being open enough to embrace new technologies in the future."

ING is a large international bank. Banks have much data about their customers. This data includes how much a person earns, what they spend their money on, where they live, even where they travel or eat. Similar types of information may be shared on a social media site. However, people who willingly share their information on a social media site know that this data will become more or less public. When people share their data with their bank, they trust that the bank will use this data responsibly, for the purposes that the data was shared, and this responsibility goes further than of just abiding by the law.

Take a customer's payment transactions as an example. Many customers would be unhappy if they felt the bank was monitoring how they spent their money. However, they would probably also expect the bank to detect fraudulent use of their debit card.

Both of these use cases involve the same data but the first example seems to be prying into a person's privacy and the second is an aspect of fraud prevention. The difference between the cases is in the purpose of the analytics. So as the bank makes data more widely available to its employees for the purpose of analytics, it must monitor both the access to data and the types of analytics it is being used for.

Bringing data together in a data reservoir makes it easier to enforce standards, but having all the data in one place creates a much higher risk of loss or misuse of information if the security of the system is compromised. The data reservoir addresses this dilemma with its information governance and security capabilities.

No data can enter the data reservoir without first being described in the data reservoir's catalog. The data owner classifies their information sources that will feed the data reservoir to determine how the data reservoir should manage the data, including access control, quality control, masking of sensitive data, and data retention periods.

The classification assigned to data created different management actions in the data reservoir. For example, when data is classified as highly sensitive, the data reservoir can enforce masking of the data on ingestion into the data reservoir. Less sensitive data, that is nevertheless personal to the bank's customers, may be stored in secured repositories in the data reservoir, so it can be used for production analytics. However, when it is copied into sandboxes for analytical discovery, it will be masked to prevent data scientists from seeing the values, without loosing the referential integrity of the data. Behind the scenes, the data reservoir is auditing access to data to detect if employees are accessing more data than is reasonable for their role. Thus the data reservoir is opening access to the bank's data, but only for legitimate and approved purposes.

A second challenge banks face is adopting real-time processing. In the past, banking was mostly batch-oriented. Payments were processed over a few days and banking was done in office hours. Nowadays people buy online 24 hours a day, every day of the week, from many countries around the world, and they expect their bank to support this. People interact with banks more and more on their mobile devices, creating more contact moments, but they are much shorter. This situation has a huge impact on how the bank operates which impacts the supporting IT systems.

In a real-time environment, the actuality of data becomes the most important factor. The bank must focus on real-time processing, both for fraud detection and also for customer service. Offering a loan in real time requires real-time analytics work with the most up-to-date risk profile for the requesting customer. Banking customers also expect real-time insights into their spending patterns, including the transaction they just did a few seconds before. So data must be continuously consolidated and reconciled for real-time processing, while also supporting the traditional, batch-oriented processes such as financial reporting that relies on daily reconciled results. The architecture is open in supporting a combination of real-time and batch ingestion. Even in a real-time world, some processes will stay batch-oriented for years to come.

A third challenge that most companies face is making data accessible to the business users in the organization. Data for business users must be formatted to support simple visualization tools, labeled with relevant business terminology and in step with the data in the systems or record. At the same time, broad access to all types of raw data is needed by data scientists to develop advanced analytics algorithms. The open architecture supports different formats of data for different types of users, based on the same data values.

The close cooperation between ING and IBM is a win-win. For IBM, it is an opportunity to explore improvements to their product portfolio and for ING, we believe the partnership will support our journey to becoming a real-time predictive bank, offering great service and trust to our customers.

What is a data reservoir

Organizations that want to improve their decision making with analytics face various challenges:

- Getting enough data points to understand key situations and building the analytic algorithms that capture the appropriate decision logic.
- Deploying analytical algorithms so they can perform the following activities:
  - Detect appropriate and related situations.
  - Retrieve the stored information needed to understand the broader context.
  - Use historical experience to predict what could happen.
- Ensuring the insights from the analytics is acted upon by the organization in a timely manner.
- Collecting evidence on the effectiveness of the analytical decisions to enable a process of continuous improvement.

Success requires a triad of trusted information, a suite of analytical algorithms and tools, along with human engagement. The data reservoir provides the following capabilities:

- Flexibility in supplying data to analysts, data scientists, and business teams.
- Efficiency in extracting and maintaining data.
- Dependability in the protection and governance of data.
- Choice in the analytics deployment environment for performance and impact.

The high-level structure of the data reservoir is shown in Figure 1.

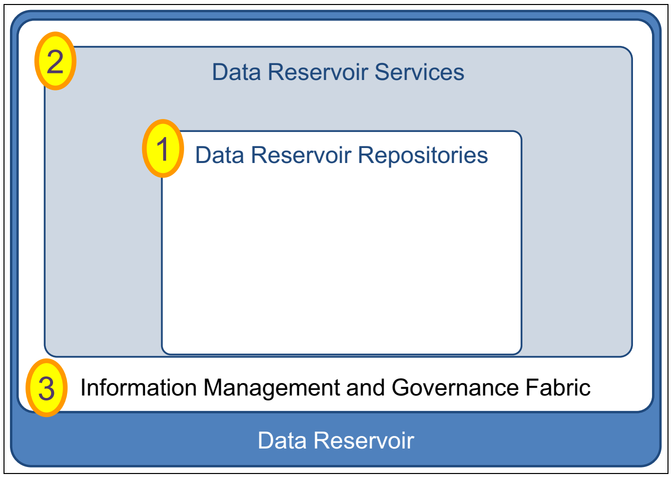

2
Data Reservoir Services
1
Data Reservoir Repositories
3
Information Management and Governance Fabric
Data Reservoir

Figure 1 Data reservoir overview

There are three main parts to a data reservoir described as follows:

- The data reservoir repositories (Figure 1, item 1) provide platforms both for storing data and running analytics as close to the data as possible.
- The data reservoir services (Figure 1, item 2) provide the ability to locate, access, prepare, transform, process, and move data in and out of the data reservoir repositories.
- The information management and governance fabric (Figure 1 item 3) provides the engines and libraries to govern and manage the data in the data reservoir. This set of capabilities includes validating and enhancing the quality of the data, protecting the data from misuse, and ensuring it is refreshed, retained, and eventually removed at appropriate points in its lifecycle.

The data reservoir is designed to offer simple and flexible access to data because people are key to making analytics successful.

Analytics in the business world

In order to understand how data and people are key to analytics, let us step back and look at a real world situation and understand how a subject matter expert makes a decision.

Consider an experienced sales assistant working with a customer buying a gift for a friend. The sales assistant recognizes the customer and remembers that this person typically buys expensive gifts and frequently at the last minute. The sales assistant understands that the gift must be immediately available.

The sales assistant asks about the friend's preferences and the type of gift the customer wants. Using their experience, the sales assistant suggests some products that sell well in these circumstances.

In this example, the sales assistant is making a decision on what to offer this customer. When people make decisions it is rare that all of the information they need is readily available. Effective sales assistants combine key information, such as:

- Details about the current situation (this is a repeat customer wanting to buy a gift that is available immediately)
- Knowledge of the broader context (the customer’s previous buying habits)
- Experience in similar situations with successful outcomes (knowledge of products popular with similar people within the same price bracket)

Throughout the process, the assistant is continually testing the results of their suggestions in order to tune their understanding and improve their decisions. This additional understanding not only improves this particular decision, but impacts future decisions as well.

Many analytic algorithms work in a similar way to the human decision making process. These algorithms draw information from the present situation and combine it with stored knowledge bases in order to determine a recommendation or classify a situation. It is important to understand, however, that analytics are not as adaptive to subtle changes and differences in each new situation as humans are.

Even the most sophisticated analytics available today is not as good as a human subject matter expert at creating new knowledge and improving the decision-making process. Human expertise is almost always a part of the process of creating new analytics, improving existing analytics, and making ad hoc decisions. Experts must teach the analytic tools how to make decisions in the same way that the experienced sales assistant trains a new assistant. However, once the analytic algorithm is configured properly, the expertise it represents can be used over and over again.

Working with a data reservoir

The objective of the data reservoir is to give people access to the data they need to operate aspects of the business and build analytic solutions. Another objective is to ensure that much of the creation and maintenance of the data reservoir is accomplished with little to no assistance and additional effort from the IT teams.

Figure 2 on page 6 shows the overall usage flow for a data reservoir.

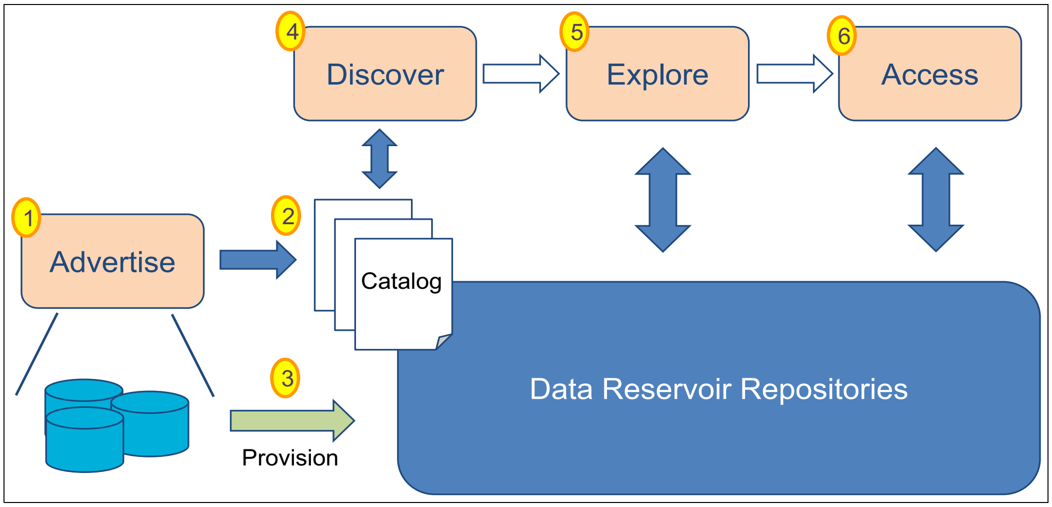


Figure 2   How does the data reservoir operate?

The activities identified in Figure 2 are described as follows:

- Advertise (Figure 2, item 1)

  Whenever there is a new source of data to add to the data reservoir, it is advertised in the data reservoir's catalog.

- Catalog (Figure 2, item 2)

  The catalog described the data in the data reservoir and how it will be managed and governed. It enables people to locate and manage the data they need. This catalog is organized so that data can be classified in various ways making it easy for different teams to find what they need.

- Provision (Figure 2, item 3)

  Provisioning is when the data is incorporated into the data reservoir. Typically all future changes made to the original source of data are synchronized with the copies in the data reservoir. This approach provides for a regular flow of data into the data reservoir.

- Discover (Figure 2, item 4)

  The catalog is used to discover where data of a particular type is located.

- Explore (Figure 2, item 5)

  Once located, the data values can be explored to verify that this is the right type of data.

- Access (Figure 2, item 6)

  Data can then be accessed directly or copied into a sandbox for use by an analysis tool.

With this basic set of capabilities identified, the key activities that can be performed on a data reservoir are covered in more detail in the following sections.

Discovering data

An information source documented in the catalog has many types of descriptive fields that explain where it came from, who owns it, last time the data was refreshed, and structural and profile information. It also includes links to resources that use the information source in order to make it possible to understand both the lineage of an information source and the impact of any change to it. The catalog can be extended to include additional classifications, links, and resource types. By having all this information in the catalog, it creates a wealth of knowledge about the content of the data reservoir. Examples of attributes that can be used to organize the information in the data reservoir are shown in Table 1.

Table 1 Example attributes

Attribute	Description
Subject area	The subject area describes what topic the information is about. For example, it could be customer data, payment data, or product data. A subject area has an associated glossary of terms that can be linked to the individual fields in an information source to pinpoint the exact meaning of the information values it contains.
Zone	A zone provides a course-grained grouping of information sources for a particular type of usage. An information source can be located in multiple zones.
Location	Location defines where the information source is located. Is it within this data reservoir, a related reservoir, or a particular external source that can be provisioned into the data reservoir on request?
Level of confidence	Level of confidence indicates the known level of quality of the information and consequently how much confidence to have in the information.

From the catalog it is possible to select a small collection of related information and either access the data in its original location or have it copied into a sandbox for private use by analytics and visualization tools.

Creating simple reports and views of information

A catalog query shows the information sources that contain information of interest. These sources are copied into simple files for use by a spreadsheet or reporting tool assuming the requester has sufficient security access. The process of copying information into the simple files is initiated from a selection menu to choose which values you are interested in, how frequently you want the information refreshed, and how much information you want returned. The files are managed by the data reservoir.

Investigating unusual activity

Investigating unusual activity often requires a wide range of information sources. The catalog for the data reservoir provides search capability to identify which sources of information are potentially of interest to the investigation.

Creating new analytics

Data scientists can also use the catalog to locate useful data for building new analytics routines. Once the wanted data is located, it can easily be copied into a sandbox so the data scientist can start preparing and analyzing the data as part of the analytics definition process.

Capturing new data and insight

The data reservoir is an interactive data environment with the ability for all types of users to add new data. When someone wants to add a new source of data, they perform the following steps:

- Advertise that the data is available in the catalog
- Arrange for the data to be provisioned into one or more of the data reservoir’s repositories
- Enable others to find, explore, and access the data

Validating the authenticity and protecting information

A key to information governance is the classification of data, data repositories, types of processing, and the people consuming the data. These classifications are linked to governance policies and rules that are applied to the data as it is processed in the data reservoir.

The data reservoir's catalog contains the definitions of the policies, rules, data classifications, data repositories, data types, and processing descriptions. It also includes links to compliance reports. It is possible to search for specific items and navigate between these definitions to understand how data is being protected and managed in the data reservoir.

The catalog is also recording the movement of data into, out of, and around the data reservoir. With this approach, it is possible to generate a report that traces where a particular set of data values came from and how they have been processed. This type of report is called a *lineage report*. It can be displayed in varying levels of detail from a high-level overview of the data reservoir repositories and information sources involved, down to details of each step in the processing that was performed along the way.

Data reservoir in the business environment

Figure 3 shows how the data reservoir fits into the business environment and is used by various teams/roles and resources within the enterprise.

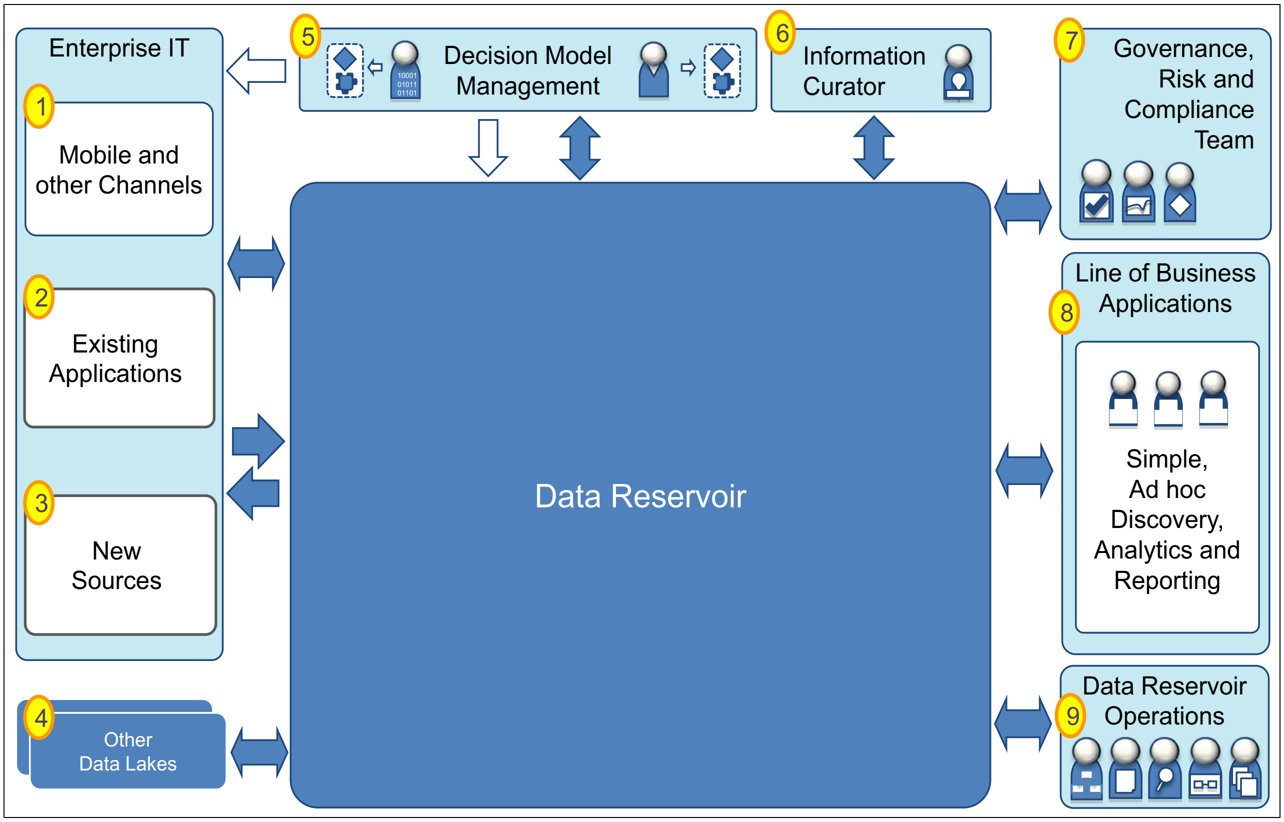

Enterprise IT
1
Mobile and other Channels
2
Existing Applications
3
New Sources
4
Other Data Lakes
5
Decision Model Management
10001 01011 01101
6
Information Curator
7
Governance, Risk and Compliance Team
Line of Business Applications
8
Simple, Ad hoc Discovery, Analytics and Reporting
Data Reservoir Operations
9
Data Reservoir

Figure 3 Data reservoir users and connected systems

The details of the various data reservoir users and the connected systems are described as follows:

- Mobile and other channels (Figure 3 on page 8, item 1)

  These channels are operational applications that support the interaction with people such as, customers, suppliers, and employees. The data reservoir could supply key data values and analytical insight to a high-speed cache in order for these applications to improve the performance of simple lookups. The data reservoir is able to refresh this cache after an outage.

- Existing applications (Figure 3 on page 8, item 2)

  Included in this group are operational applications that drive an organization's daily business. They supply information to the data reservoir that describes this daily operation and its associated master data. These applications also receive analytical insights and other derived information such as, micro-segmentation and alerts.

- New sources (Figure 3 on page 8, item 3)

  New sources describe information outside of the business data managed by the system of record applications. This information could be log files from customer interactions or information from third parties such as, social media services and data providers.

- Other data lakes (Figure 3 on page 8, item 4)

  In this case, the data reservoir could be exchanging information with other data lakes, swamps, or reservoirs. These sources of data can be owned by this organization, part of a cloud deployment, or owned by an external party.

- Decision model management (Figure 3 on page 8, item 5)

  Decision model management describes the systems used by data scientists and business analysts as they configure analytics models and rules to execute inside the data reservoir. Advanced analytics and data mining is managed from decision model management. These teams need access to samples of the data. This data must be formatted for analysis tools and the environment must have sufficient performance capacity to handle intense, lumpy workloads from the mining and testing processing.

- Information curator (Figure 3 on page 8, item 6)

  An individual or group of people in the organization that are maintaining the catalog information. This activity includes advertising new sources, enhancing and categorizing existing data descriptions, and helping people locate the data they need.

- Governance, risk and compliance team (Figure 3 on page 8, item 7)

  The information governance, risk, and compliance team use a reporting environment to demonstrate compliance in industry regulations and business policies. Some of these policies apply to the management of information. These policies are defined and managed in the data reservoir's catalog. The information governance roles involved are shown from left to right (Figure 3 on page 8 item 7): Information Governor, Information Owner, and Auditor.

- Line of business applications (Figure 3 on page 8, item 8)

  These are applications designed to provide reports, search, and simple analytics capabilities that are under the control of the lines of business. The interfaces that they use are designed to be self-service with simple configuration to define new insight.

- Data reservoir operations (Figure 3 on page 8, item 9)

  The data reservoir operations team maintains the data reservoir behind the scenes. In item 9 from left to right the roles are enterprise architect, information steward, data quality analyst, integration developer, and infrastructure operator.

Architecture of the data reservoir

With the broader view of how the data reservoir fits into the enterprise environment in place, Figure 4 shows the major components inside the data reservoir.

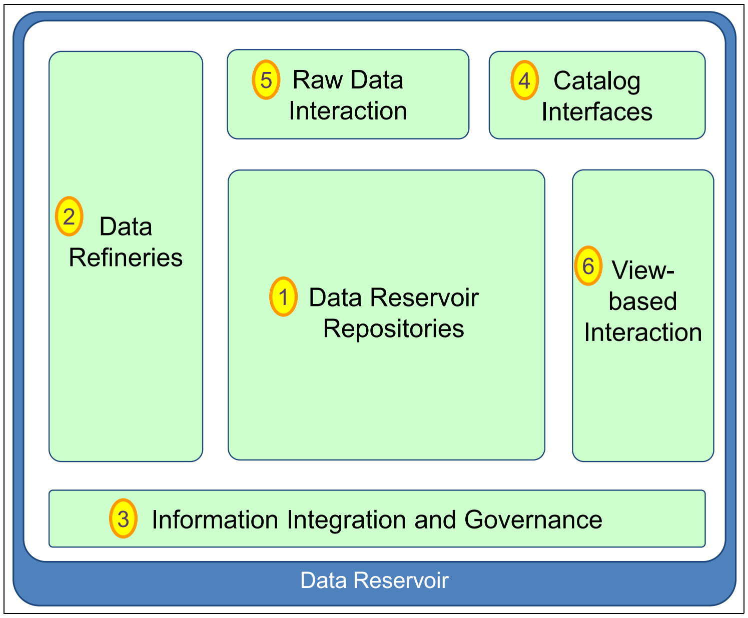


Figure 4 Data reservoir logical architecture

The key components of the data reservoir (Figure 4) are defined as follows:

- Data reservoir repositories (Figure 4, item 1)

  These repositories provide the storage for data and are able to host analytics to execute against the data it stores.

- Data refineries (Figure 4, item 2)

  The refineries provide the ability to move and transform data in, out, and between the data reservoir repositories. The data refineries use the information integration and governance fabric capabilities to efficiently process the data and enforce the governance policies.

- Information integration and governance (Figure 4, item 3)

  Information integration and governance provide the libraries to support different provisioning, transformation, and governance capabilities that are used in the data reservoir, particularly the data refineries.

- Catalog interfaces (Figure 4, item 4)

  The catalog interfaces provide information about the data in the data reservoir. The catalog includes details of the information collections (both repositories and views), the meaning and types of information stored, and the profile of the information values within each information collection.

- Raw data interaction (Figure 4 on page 10, item 5)

  Raw data interaction provides access to most of the data (security permitting) in the data reservoir for advanced analytics. It is responsible for masking sensitive personal information where appropriate.

- View-based interaction (Figure 4 on page 10, item 6)

  View-based interaction provides access to data in the data reservoir (subject to security permissions) for line-of-business teams that want to perform ad hoc queries, search, simple analytics, and data exploration. The structure of this information has been simplified and it is labeled using business relevant terminology.

The following sections go into more detail about each of the data reservoir components.

The data reservoir repositories

Organizations tend to create and collect a large amount of data that is very diverse. Examples of this data range from the following data types:

- Data about people, organizations, and assets that are the essential resources of the organization.
- Data about the day-to-day activities such as, orders for new business, payments, and marketing campaigns.
- Log data that is generated when a system or device runs. Examples are the logs generated by websites or call switchboards.
- Derived or external data such as, reference data, classifications, key performance indicators (KPIs), or news feeds.

The data reservoir needs at least one repository in which to accumulate data. Typically it will have various repositories, each focused on supporting particular types of workload. Figure 5 illustrates the types of repositories that might be found in a data reservoir.

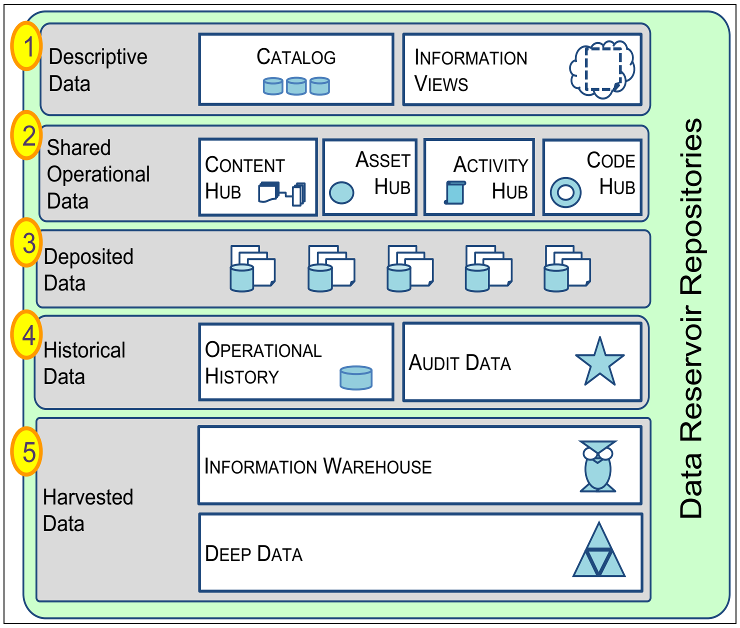


Figure 5   Data reservoir repositories

The types of repositories found in a data reservoir (Figure 5 on page 11) are defined as follows:

- Descriptive data (Figure 5 on page 11, item 1)

  The descriptive data is sometimes referred to as *metadata* because it describes the data in the data reservoir. The descriptive data repositories include:

  - Catalog: A repository and applications for managing the catalog of information stored in the data reservoir.
  - Information views: Definitions of simplified subsets of information stored in the data reservoir repositories. These views are created with the information consumer in mind.

- Shared operational data (Figure 5 on page 11, item 2)

  Shared operational data contains consolidated operational data that is being shared by multiple systems. The data reservoir may host the master copy of this data, or a reference copy of this data supplied from one or more operational systems. This shared operational data includes the following components:

  - Asset hub: A repository for slowly changing operational master data (information assets) such as, customer profiles, product definitions, and contracts. This repository provides authoritative operational master data for the real-time interfaces, real-time analytics and for data validation in data ingestion. If it is a reference repository of the operational master data management (MDM) systems it might also be extended with new attributes that are maintained by the reservoir. When this hub is taking data from more than one operational system, there may also be additional quality and de-duplication processes running that improve the data. These changes are published from the asset hub for distribution both inside and outside the reservoir.
  - Activity hub: A repository for storing recent activity related to a master entity. This repository is needed to support the real-time interfaces and real-time analytics. It may be loaded through the data ingestion process and through the real-time interfaces. However, many of its values were derived from analytics running inside the data reservoir.
  - Code hub: A repository of common code tables and mappings that are available through the data reservoir's application programming interfaces (APIs).
  - Content hub: A repository of documents, media files, and other content that has been managed under a content management repository and is classified with relevant metadata to understand its content and status.

- Deposited data (Figure 5 on page 11, item 3)

  Deposited data are information collections that have been stored by the data reservoir information users. These information collections may contain new types of information, analysis results, or notes.

- Historical data (Figure 5 on page 11, item 4):

  Historical data contains read-only records of historical activity such as:

  - Operational history: A repository providing a historical record of the data from an operational application. This data is stored in the same format and with the same level of quality as is found in the application itself. This approach makes it possible to use this data to investigate activity around a specific application.
  - Audit data: A repository used to keep a record of the activity in the data reservoir. It is used for auditing the use of data and who is accessing it, when, and for what purpose.

- Harvested data (Figure 5 on page 11, item 5)

  Harvested data includes data from other systems that has been extracted and is being transformed, consolidated, aggregated, and analyzed to create new insight. Harvested data includes the following types of data store:

  – Information warehouse: A repository optimized for high-speed analytics. This data is structured and contains a correlated and consolidated collection of information.

  – Deep data: A repository holding a copy of most of the data in the data reservoir. It provides a place where raw data can be located for analysis. The data may be annotated, linked, and consolidated in deep data. Data may be mapped to data structures after it is stored, so effort is spend as needed rather than at the time of storing. This repository is designed for flexibility, supporting both high volumes and variety of data.

Feeding the data reservoir

Much of the data in the data reservoir comes from the organization's IT systems. These sources for data might be the systems operating the business or other sources that are monitoring activity. Another source might be a log of the usage of the organization's website.

The data refineries (Figure 4 on page 10, item 2) provide a number of functions in the data reservoir. The functions include the following:

- Moving data in, out, and between the data repositories
- Restructuring data to make it easier to work with
- Verifying the quality of data and reporting problems
- Standardizing and enriching data
- Masking and protecting data
- Archiving data

Data refineries can be used independently of a data reservoir to provide data to new applications to move data between clouds and on-premises application, and for synchronizing data between data reservoirs. In addition, they provide much of the dynamic behavior of the data reservoir. Figure 6 on page 14 illustrates the data refineries in the data reservoir.

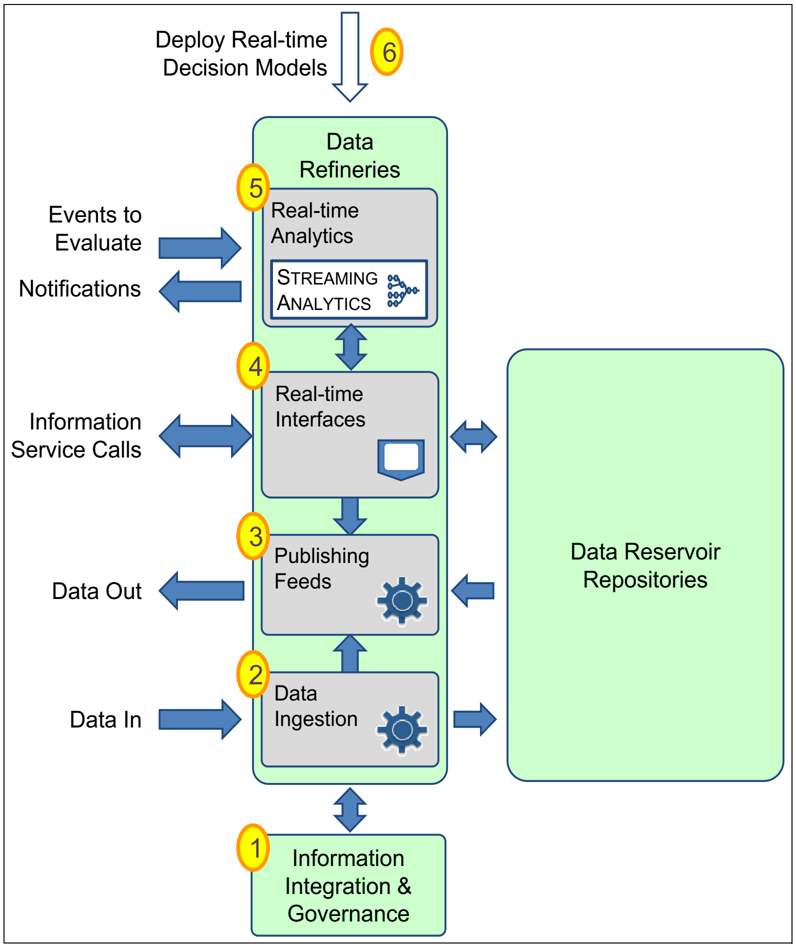


Figure 6   Data refineries in the data reservoir

The data refineries (Figure 6) are described in more detail as follows:

- The data refineries make extensive use of the capabilities in the information integration and governance fabric (Figure 6, item 1).
- As new data is imported into the data reservoir, the data ingestion processes (Figure 6, item 2) ensure the data is transformed, logged, and copied into the appropriate data repositories.
- Periodically, data and analytics insight can be published (Figure 6, item 3) by the data reservoir for use by systems outside of the data reservoir.
- These systems could also make API calls (Figure 6, item 4) to the data reservoir to access/update data or insight. These APIs might interact with a single data repository or may federate data from multiple repositories.
- The data reservoir could process events in real time (Figure 6, item 5) and produce insight that could be stored in the data reservoir's repositories, and published externally for other systems to act upon.
- Any of the data refineries might execute analytics (decision models) (Figure 6, item 6) as part of their execution.

Information governance

The requirements for information governance are documented in the catalog as policies, rules, and classifications. The data reservoir's information integration and governance fabric is responsible for ensuring these requirements are met by all operations of the data reservoir. Figure 7 shows the components of the information integration and governance fabric.

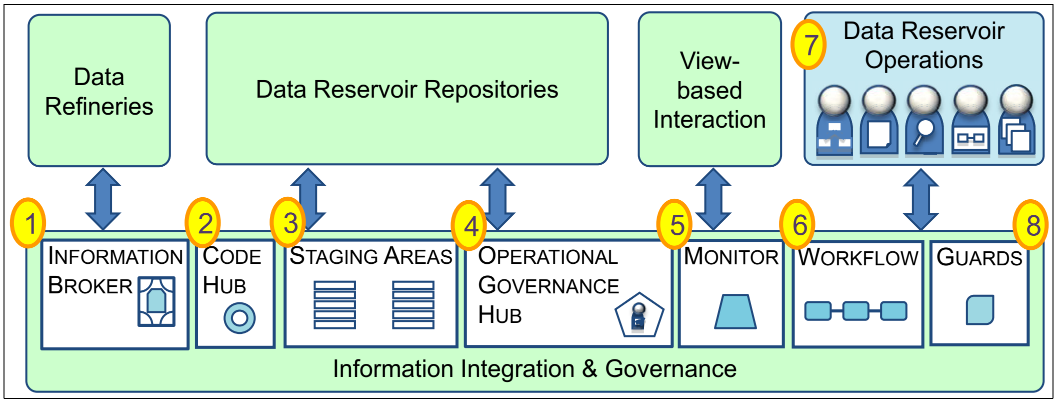

Data Refineries
Data Reservoir Repositories
View-based Interaction
7
Data Reservoir Operations
1
2
3
4
5
6
8
INFORMATION BROKER
CODE HUB
STAGING AREAS
OPERATIONAL GOVERNANCE HUB
MONITOR
WORKFLOW
GUARDS
Information Integration & Governance

Figure 7 Operationalizing information governance

Information integration and governance consist of the following key components:

- Information broker (Figure 7, item 1)

  The information broker is a runtime server environment for executing the integration processes (such as the information deployment process). These processes move data in and out of the data reservoir and among the components within the reservoir.

- Code hub (Figure 7, item 2)

  The code hub is a repository of common code tables and mappings used for joining information sources to create information views.

- Staging areas (Figure 7, item 3)

  A server supporting the staging areas that are used to move information around the data reservoir.

- Operational governance hub (Figure 7, item 4)

  A repository and applications for managing the information flow and information governance within the data reservoir. This information node supports the metadata services.

- Monitor (Figure 7, item 5)

  The monitor watches over the overall function and responsiveness of the data reservoir to ensure a consistent working environment.

- Workflow (Figure 7, item 6)

  Workflow consists of a server running stewardship processes that coordinate the work of the individuals responsible for fixing any problems with the data in the data reservoir.

- Data reservoir operations (Figure 7, item 7)

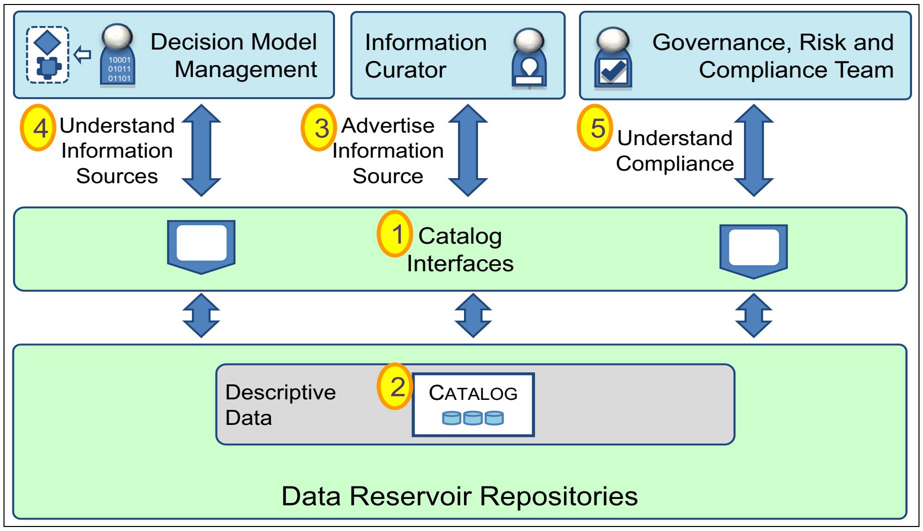
  The data reservoir operations team that manages the operations of the data reservoir. Their roles are shown in Figure 3 on page 8, item 9.

- Guards (Figure 7, item 8)

  The guards provide capability to control access to information.

Maintaining the data reservoir catalog

The catalog in the data reservoir is key to ensuring data can be located and is properly managed. Figure 8 shows how the catalog is used and by what teams and roles.



10001
01011
01101
Decision Model Management
Information Curator
Governance, Risk and Compliance Team
4 Understand Information Sources
3 Advertise Information Source
5 Understand Compliance
1 Catalog Interfaces
Descriptive Data
2
CATALOG
Data Reservoir Repositories

Figure 8   Data reservoir catalog

Figure 8 shows how various resources use the catalog:

- Catalog interfaces (Figure 8, item 1)

  The catalog interfaces provide the user interfaces and APIs to interact with the catalog.

- Catalog (Figure 8, item 2)

  The catalog is a repository and a set of applications used to maintain details about the data in the data reservoir.

- Advertise information source (Figure 8, item 3)

  An information source can be described in the information catalog to indicate that new information is available. When this data is copied into one or more of the data reservoir's repositories, the descriptive data about these repositories is updated to include this new data. The linage information is updated to show that these repositories are also fed by this new information source.

- Understand information source (Figure 8, item 4)

  The catalog provides multiple mechanisms for searching, querying, and browsing details about the data in the data reservoir.

- Understand compliance (Figure 8, item 5)

  The catalog records policies, rules, and classifications that define how data is to be managed and used in the data reservoir. The data refineries call the information integration and governance fabric to implement these policies. The governance team is able to query which polices, rules, and classifications are attached to each type of data to verify that they are being managed correctly. They can also see the statistics for the quality exceptions that have been reported.

Providing access to the data reservoir

Teams retrieve information from the data reservoir either in the format that it is stored or they use simplified views of the data that has been created to make it easier to use with business reporting and analysis tools.

Figure 9 shows the process for working with the data in the format it is stored in. This process is known as *raw data interaction*.

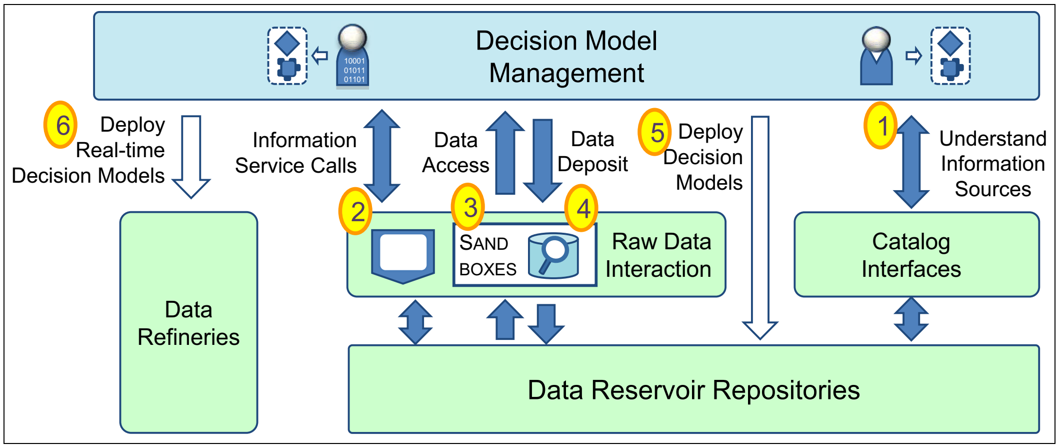

Decision Model Management
10001 01011 01101
6
Deploy Real-time Decision Models
Information Service Calls
Data Access
Data Deposit
5
Deploy Decision Models
1
Understand Information Sources
2
3
4
SAND BOXES
Raw Data Interaction
Catalog Interfaces
Data Refineries
Data Reservoir Repositories

Figure 9 Data reservoir raw data interaction

The data reservoir raw data interaction is described as follows:

- Understand information sources (Figure 9, item 1)

  The catalog provides the means to locate the data reservoir repository that has the data required.

- Information service calls (Figure 9, item 2)

  Information service calls provide APIs to directly access the data in the repository.

- Data access (Figure 9, item 3)

  By using a simple wizard, it is possible to export either a sample or the complete collection of data in to a sandbox.

- Data deposit (Figure 9, item 4)

  It is also possible to import data from the sandbox back into the data reservoir repositories.

- Deploy decision models (Figure 9, item 5)

  After analyzing the data and building a decision model that contains rules and analytics, it is possible to deploy the decision model into the data reservoir's repositories. This decision model can then be run regularly.

- Deploy real-time decision models (Figure 9, item 6)

  It is also possible to deploy the decision models into one or more data refineries to generate insight from the data as this data is being processed.

Figure 10 on page 18 shows the use of the simplified views to interact with the data. These views perform two functions as follows:

- Create flattened structures that are easier to use in a spreadsheet of visualization tool.
- Use appropriate business vocabulary to label the data so it is easier to understand what the values mean.

The use of simplified views is referred to as *view-based interaction*.

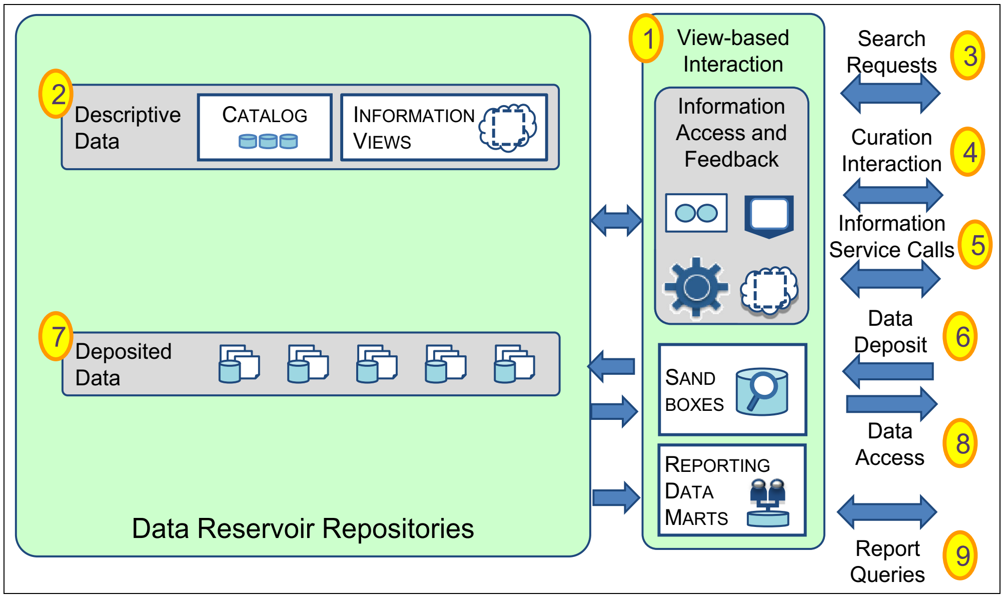


Figure 10 Data reservoir view-based interaction

Data reservoir view-based interaction (Figure 10) consists of the following elements:

- View-based interaction (Figure 10, item 1)

  View-based interaction provides access to data in the data reservoir (subject to security permissions) for line-of-business teams. This access enables them to perform ad hoc queries, searches, simple analytics, and data exploration. The structure of this information has been simplified and it is labeled using business relevant terminology.

- Descriptive data (Figure 10, item 2)

  Descriptive data provides the definitions used to create information views and access the data in the data reservoir using the information view.

- Search requests (Figure 10, item 3)

  Search requests help locate the data that a person is interested in. A search is able to look at the catalog and the data values themselves.

- Curation interaction (Figure 10, item 4)

  Curation interaction provides the capabilities to define new information views using business vocabulary.

- Information service calls (Figure 10, item 5)

  Information service calls are APIs to access data through the information views.

- Data deposit (Figure 10, item 6)

  Data deposit enables a business user to store new data into the reservoir. Similar to data export, data import also calls an appropriate data refinery to validate, transform, and store the data as appropriate.

- Deposited data (Figure 10, item 7)

  A repository where any data deposited through the view-based interaction is typically stored.

- Data access (Figure 10 on page 18, item 8)

  Data access allows a business user to extract a collection of data from the data reservoir. It is stored in the format described by the information view, rather than the format that it appears in the data reservoir's repositories. It calls an appropriate data refinery to validate, transform, and store the data as appropriate.

- Reporting data marts (Figure 10 on page 18, item 9)

  The reporting data marts provide departmental/subject-oriented sources for line-of-business reports.

Bringing it all together

Figure 11 is an illustration of the data reservoir, showing all of the pieces fitted together.

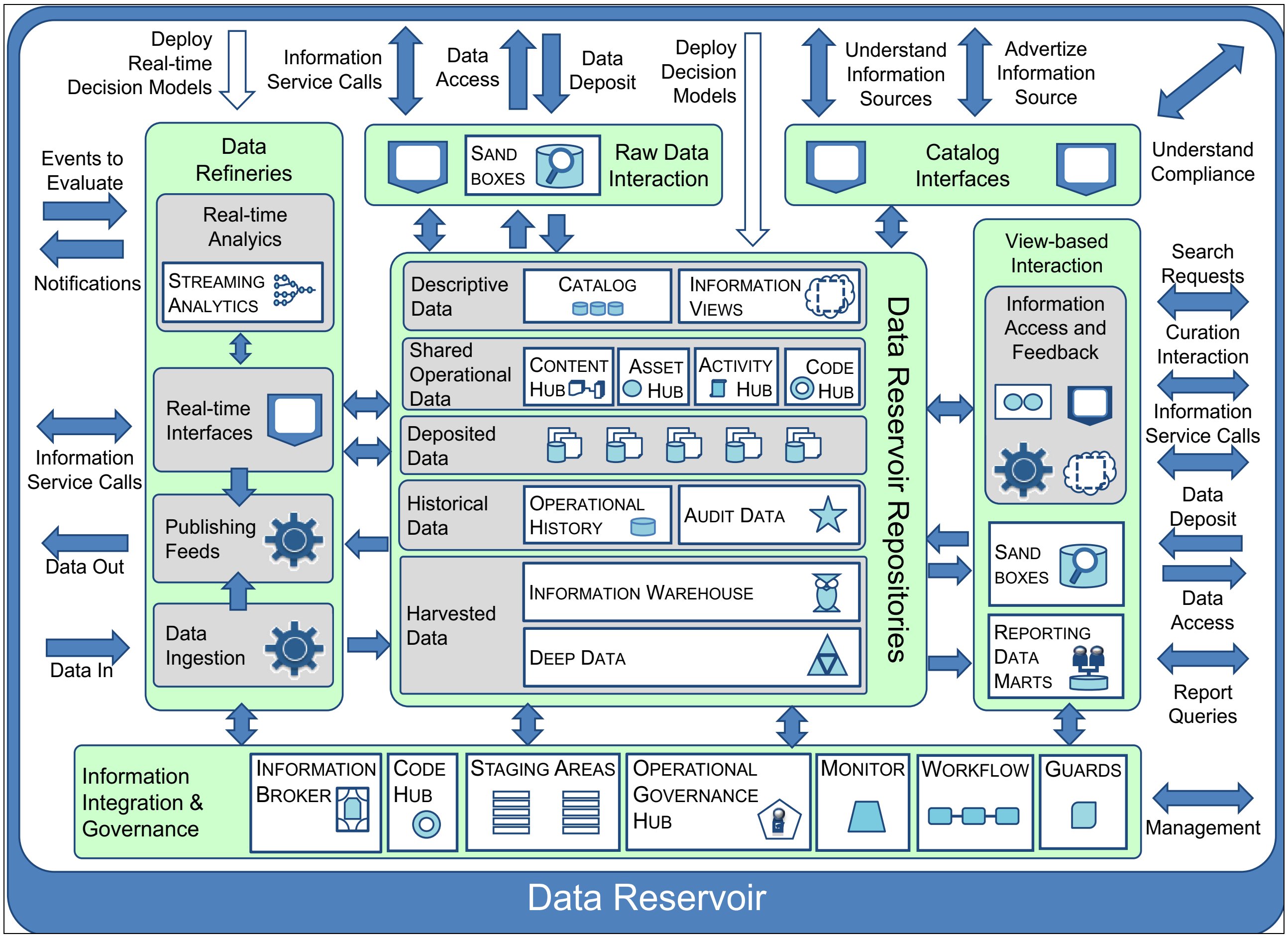


Figure 11 Data reservoir summary

Organizational change and impact

The data reservoir starts to deliver deep value as the organization takes on these responsibilities:

- Willing to share and exchange information.

- Willing to remove unnecessary complexity and inconsistency in the way different parts of the business operates, the vocabulary they use, and the data they collect.
- Willing to use data and insight in decision making.
- Willing to provide feedback and improve data and any resulting insight that comes from the data reservoir.

For many organizations a focused effort is required to make this transition. The most effective approach is to focus on a few specific use cases. Initially, just load the data needed for these use cases into the reservoir and only train the teams that are involved in the use cases. This way, the processes that are used to set up and enhance the data reservoir are tested and streamlined in advance of a more widespread deployment.

As your experience grows, more teams can be given access to the data reservoir services, broadening both the supply and the use of the shared data.

Rolling out a data reservoir

A data reservoir is a dynamic, agile environment for the business teams to control and make use of in an interactive, self-service mode. As such, there are some initial bootstrap activities to lay down the governance and management framework, which includes the following:

- Installation of the information integration and governance platform and at least one data repository.
- Definition of the governance policies and related implementations for managing data for each of the subject areas that will be initially stored in the data reservoir.

Once this framework is in place, the following teams continuously enhance the richness of the data reservoir:

- Governance team

  The governance team makes it possible for the data reservoir to accept data on new subject areas by defining the governance policies and related definitions for this subject area data.
- IT teams play a key role by enhancing the data reservoir in the following ways:
  - Adding new types of repositories in the data reservoir to support specialist data storage and analytics.
  - Adding new data refineries to exchange data between the operational systems and the data reservoir. This approach ensures the data reservoir has the latest operational information and that the operational systems benefit from the insight generated in the data reservoir.
  - Adding feeds from non-traditional sources of information such as, a log data and social media.
- Information curators

  These curators define new sources of information that can be used to extend the insight in the data reservoir.
- Business teams

  Business teams add their knowledge and departmental data into the reservoir to bring an additional perspective to the data from the operational systems.

The result of this commitment and the effort by these teams is an adaptable data reservoir that grows with your organization's data needs.

How IBM helps make it happen

IBM has years of experience in supporting business transformation projects, particularly those involving analytics. IBM itself continuously reinvents its business operations to make better use of information and analytics. The IBM service organizations are able to bring you the benefit of this experience both from the IBM transformation experience and from assisting IBM clients in their transformation activities. The offerings includes the following items:

- Proof of Technology (POT)

  A POT is a short session run in a single day to enable you to experience new technology hands on.
- Proof of Concept (POC)

  A POC is a longer lasting engagement of a few weeks to a few months where the IBM team builds out a specific scenario to demonstrate new technology. This activity is done often onsite at a client's venue.
- Lab and other implementation support services

  Ongoing mentoring and subject matter expertise to support a client who is implementing new technology. This type of engagement is focused on skills transfer and accelerating the project.
- Delivery services

  Onsite support of a project providing the right mix of skills to manage and implement the project in collaboration with the client's teams.

IBM also has a world-class software portfolio that provides the tools, services, and industry content to support clients taking on their own transformation. IBM can help clients with the following activities:

- The governance and protection of data in the data reservoir.
- The transformation, quality management, and self-service delivery of information to the business.
- The management of the infrastructure that supports the data reservoir.
- Advanced analytics tools and algorithms to support greater insight in your organization.

IBM has workload optimized platforms that enable effective execution of analytics.

Whether you are considering a completely new implementation of a data reservoir or want to incorporate existing data warehouses and analytics repositories into a data reservoir, IBM has the expertise and technology to help you deliver this capability.

Summary

Success with analytics requires determining the decision areas you want to improve, identifying the insights that could drive positive activity, and defining the information that underpins the related processes.

The process of investigating and analyzing a situation or identifying a pattern of activity can require access to a wide variety of information. There must be a way to review past activity in order to understand the factors that determine either a successful outcome or that it is time to try something different.

The data reservoir manages data for analytics, by providing:

- Flexibility in supplying data to analysts, data scientists, and business teams.
- Efficiency in extracting and maintaining big data.
- Dependability in the protection and governance of data.
- Options in the analytics deployment environment to increase performance and impact.

This Redguide publication described why a data reservoir is valuable and how the organization must change to take real advantage of a data reservoir. The guide also discussed the architectural details of a data reservoir.

Other resources for more information

For additional information, review the following documents:

- *Smarter Analytics: Information Architecture for a New Era of Computing*, REDP-5012
- *Patterns of Information Management*

  http://www.informit.com/store/patterns-of-information-management-9780133155501?WT.mc_id=Author_Chessell_PoIM

Authors

This guide was produced by a team of specialists from around the world working with the International Technical Support Organization (ITSO).

Mandy Chessell is an IBM Distinguished Engineer and Master Inventor and is currently the Chief Architect of IBM InfoSphere® Solutions, working in the Chief Technology Officer (CTO) Office of SWG Information Management. She has expertise in designing information supply chains for information intensive solutions. Mandy joined IBM in 1987 and has held roles for developing new features for various IBM products such as IBM CICS®, IBM TxSeries, Encina, Component Broker, and IBM WebSphere® Application Server. She is a Fellow of the Royal Academy of Engineering (FREng), a Chartered Engineer (CEng), and a Fellow of the British Computer Society (FBCS).

Ferd Scheepers is the Chief Information Architect for ING. He has worked for ING since 1995 and in that time has held many different roles, ranging from Lead Architect for Development Environments, Business Intelligence and Middleware to Enterprise Architect for Payments, Transaction Services, and Customer Centricity. In his current role, Ferd is responsible for the Data and Information Architecture for ING Global.

Nhan Nguyen is a Building Block Architect for Marketing and Customer Analytics at ING Netherlands. He has over 15 years of experience in IT, with many of those years in the field of Information Management. In previous roles, he was responsible for the architecture of the International Network of ING.

Ruud van Kessel is Solution Architect for Marketing and Customer Analytics at ING Netherlands. Ruud has over 10 years in IT and is one of the technology experts in the field of Information Management, with extensive knowledge on Hadoop, Data Warehouses and extract, transform, and load (ETL).

Ron van der Starre is an Information Management Architect for the IBM Software Group in the Netherlands. He has over 20 years of experience in IT with the last 15 years within IBM. He started within the services organization as an IT Specialist for packaged solutions and after a couple of years switched to work as a consultant in the area of business process redesign.

Thanks to the following people for their contributions to this project:

LindaMay Patterson
IBM International Technical Support Organization, Rochester

Henrik van Bruggen
ING, Netherlands

Patrick van der Drift
IBM, Netherlands

Frans Nieuwerth
IBM, Netherlands

Now you can become a published author, too!

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

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

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  https://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

- Explore new IBM 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

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 Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites 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.

This document, REDP-5120-00, was created or updated on August 26, 2014.

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:

CICS®
IBM®
InfoSphere®
Redbooks®
Redguide™
Redbooks (logo) ®
WebSphere®

The following terms are trademarks of other companies:

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