Redbooks

ibm.com/redbooks

Designing and Operating a Data Reservoir

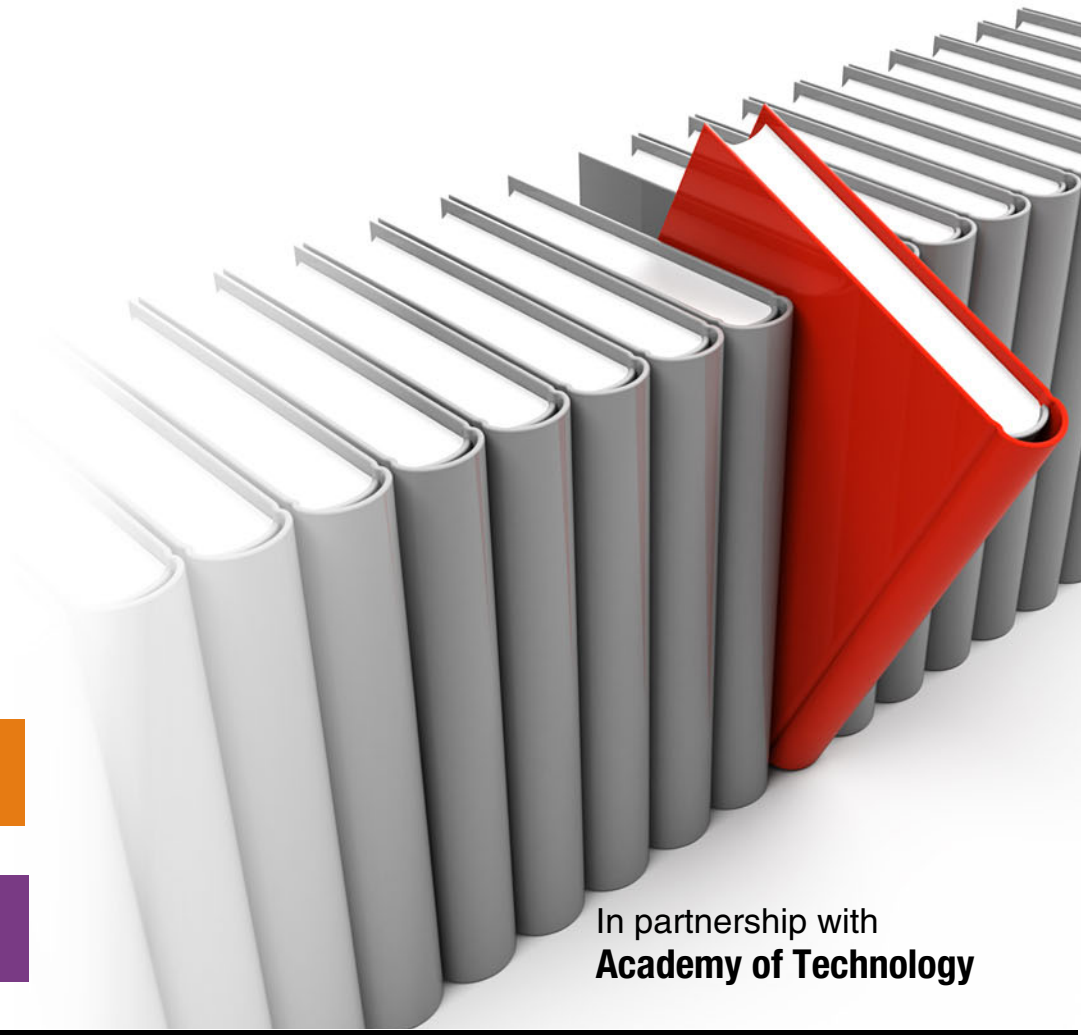
Mandy Chessell

Nigel L Jones

Jay Limburn

David Radley

Kevin Shan



Analytics

Big Data

In partnership with
Academy of Technology

IBM

Redbooks

International Technical Support Organization

Designing and Operating a Data Reservoir

May 2015

SG24-8274-00

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

First Edition (May 2015)

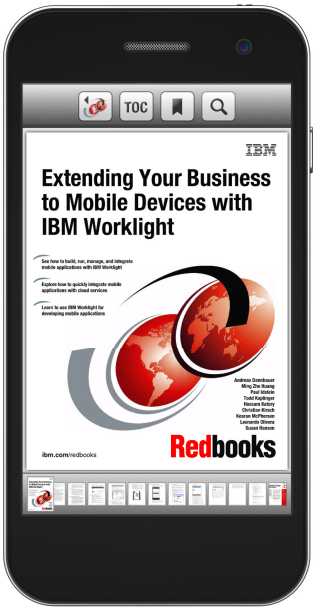
Contents

Find and read thousands of IBM Redbooks publications

- Search, bookmark, save and organize favorites
- Get up-to-the-minute Redbooks news and announcements
- Link to the latest Redbooks blogs and videos

Get the latest version of the Redbooks Mobile App

iOS **Download Now** Android



Promote your business in an IBM Redbooks publication

Place a Sponsorship Promotion in an IBM® Redbooks® publication, featuring your business or solution with a link to your web site.

Qualified IBM Business Partners may place a full page promotion in the most popular Redbooks publications. Imagine the power of being seen by users who download millions of Redbooks publications each year!

ibm.com/Redbooks
About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK

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.

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™
CICS®
Cloudant®
DataStage®
DB2®
developerWorks®
Guardium®
IBM PureData™
IBM®
IMS™
InfoSphere®
Insight™
Optim™
PureData®
Redbooks®
Redbooks (logo) ®
WebSphere®

The following terms are trademarks of other companies:

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

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

Preface

Together, big data and analytics have tremendous potential to improve the way we use precious resources, to provide more personalized services, and to protect ourselves from unexpected and ill-intentioned activities. To fully use big data and analytics, an organization needs a system of insight. This is an ecosystem where individuals can locate and access data, and build visualizations and new analytical models that can be deployed into the IT systems to improve the operations of the organization. The data that is most valuable for analytics is also valuable in its own right and typically contains personal and private information about key people in the organization such as customers, employees, and suppliers.

Although universal access to data is desirable, safeguards are necessary to protect people's privacy, prevent data leakage, and detect suspicious activity.

The data reservoir is a reference architecture that balances the desire for easy access to data with information governance and security. The data reservoir reference architecture describes the technical capabilities necessary for a system of insight, while being independent of specific technologies. Being technology independent is important, because most organizations already have investments in data platforms that they want to incorporate in their solution. In addition, technology is continually improving, and the choice of technology is often dictated by the volume, variety, and velocity of the data being managed.

A system of insight needs more than technology to succeed. The data reservoir reference architecture includes description of governance and management processes and definitions to ensure the human and business systems around the technology support a collaborative, self-service, and safe environment for data use.

The data reservoir reference architecture was first introduced in *Governing and Managing Big Data for Analytics and Decision Makers*, REDP-5120, which is available at:

http://www.redbooks.ibm.com/redpieces/abstracts/redp5120.html

This IBM® Redbook publication, *Designing and Operating a Data Reservoir,* builds on that material to provide more detail on the capabilities and internal workings of a data reservoir.

Authors

This book was produced by a team of specialists from around the world working at the International Technical Support Organization, Poughkeepsie Center.

Mandy Chessell CBE FREng CEng FBCS is an IBM Distinguished Engineer, Master Inventor, and member of the IBM Academy of Technology. Her current role is the Chief Architect for Information Solutions in the IBM Analytics Group CTO office. She leads the design of common information management patterns for different industries and solutions. This includes the Data Reservoir, Next Best Action solution, and the strategy for Information Governance.

In earlier roles Mandy led the development of new features for IBM CICS®, Encina, TxSeries, WebSphere® and InfoSphere® products. She has over 50 issued patents worldwide in the fields of transaction processing, event management, business process management, model driven development, and information management.

Outside of IBM, Mandy is a Fellow of the Royal Academy of Engineering and a visiting professor at the University of Sheffield, UK. Identified in 2000 as one of MIT Technology Review's hundred young people most likely to make significant 21st Century technical innovation, she is distinguished as the first woman to win a Royal Academy of Engineering Silver Medal. Mandy has an honorary fellowship of the Institution for Engineering Designers (IED) and a honorary doctorate of science from the University of Plymouth. In 2015, Mandy was awarded a CBE in the Queen's New Years Honours List for services to engineering.

Mandy's numerous published titles include *Patterns of Information Management*, a book on design patterns for better information architecture and management. These patterns form the basic language and component model for the data reservoir architecture. More information about Mandy's publications can be found here:

http://www.linkedin.com/pub/mandy-chessell/22/897/a49

Nigel L Jones is an Information Solutions Architect based at IBM Hursley within the IBM Analytics Group. His current role is working on data reservoir capabilities and information governance end to end including Hadoop, and has a focus on open source technologies. He has been working with IBM master data products for the last 10 years supporting the development of new capabilities and integration into solutions. Before that, he worked extensively with IBM voice solutions. Nigel has a degree in Physics and Computer Science from the University of Manchester.

Jay Limburn MBCS CITP, is an IBM Senior Technical Staff Member at the IBM Software Development Laboratory in Hursley UK. Jay works within the IBM Analytics organization and is the lead architect for Master Data Governance and consumption. In this role, he is responsible for ensuring that master data can be delivered efficiently to business users and for accelerate IBM clients' ROI. Jay is a recognized expert on data governance and strategies that allow organizations to extract value from their data quality projects. He has presented at conferences worldwide on these topics, and as a UK Senior Inventor holds 12 patents in these areas.

David Radley is an IBM Analytics Information Solution Architect in the IBM UK Hursley lab. He has over 25 years of experience in IT, with the last 10 years in Information Management. In his role, David promotes and develops information architecture to underpin analytics and EMM solutions, by putting master data at the heart. David has published best practices in these areas. David is keen to promote innovation and has filed two patents with IBM. David holds a degree in Physics from Birmingham University.

Kevin Shank is a Senior Technical Staff Member (STSM) and the Chief Architect for Metadata, Governance, and Semantic Technologies for the IBM InfoSphere product group. He has recently been a principal contributor to the efforts to develop the DataWorks cloud products. Kevin was the original chief architect and team lead for the InfoSphere Metadata Server, and one of the principal architects for the InfoSphere Information Server platform itself. Kevin's technical background includes expertise in Object-Oriented Design, Distributed Systems, Software Modeling, and structures for representing Metadata and Information (EMF/RDF/etc). Kevin has numerous patents and publications, and has a BSEE from The University of Tennessee, as well as MS and PhD degrees in Computer Engineering from Syracuse University.

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

Philip Monson, Redbooks® brand Software lead
International Technical Support Organization, Poughkeepsie NY

Martin Borrett, Director of the IBM Institute for Advanced Security Europe
IBM Hursley

Chris Nott, CTO Analytics, UKI
IBM London

Chris Grote, Big Data Architect for Banking & Financial Markets
IBM London

Now you can become a published author, too!

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

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

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

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

- Stay current on recent Redbooks publications with RSS Feeds:

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

1

Introduction to big data and analytics

This chapter presents a high-level overview of the data reservoir.

This chapter includes the following sections:

- Data is key to success
- About this publication
- Case study: Eightbar Pharmaceuticals
- Summary and next steps

1.1 Data is key to success

In the modern world, data is a key element of business success. Over the last few years, IBM has worked with different businesses around the world, from banks to telecommunication companies, retailers to utility companies. These businesses are requesting help in building a sustainable and adaptable big data and analytics environment.

You might ask *Why are big data and analytics important to business?*

Many of the companies that IBM has worked with have been identified as some of the best in their industry. They have developed advanced processes for dealing with complex situations. But this is not enough. They also need to gather and analyze data from a wide range of sources to understand and act in a changing world.

Big data and analytics shines a light on what is really happening in an organization. It can identify what has changed and identify unexpected obstacles that might cause you or the business to stumble even on a known path.

Expectations are changing.

We are in a world of rapid change and ubiquitous communications. The younger generation does not remember a world without the Internet and mobile phones. Social media such as Twitter and Facebook create an expectation that as an individual, they have a voice and a right to be heard. These people expect service to be personalized and they can mobilize support for change with ease.

Consider the digital trail.

The act of using mobile phones and other mobile devices leaves a digital trail, illuminating where you go and what you do. Supporting this connectivity is an infrastructure so complex that it is too complex for the individual to understand or to control as a system without feedback.

Change has occurred because automation is affordable.

Sensors are becoming so cheap they can be placed anywhere. There is more data and more affordable processing power to collect, correlate, and act on this data.

Consider as an example that an electricity company's revenue stops if the power fails. Scheduled, routine maintenance reduces the chance of failure by replacing potentially worn out components. However, these actions do not eliminate failures. It is only when the whole grid is modeled and monitored does the company identify where components are being overloaded and which components need maintenance sooner. With this information, the company can reduce the number of failures further.

When weather predictions of storms are overlaid on this grid, new vulnerabilities are uncovered. Moving maintenance crews to targeted locations in advance of the storm minimizes outages, and reduces the cost of overtime for the crews.

Collecting, correlating, interpreting, and acting on a wide variety of data through analytics improves the resilience and quality of services. It is not easy to do and takes a disciplined, agile, and open approach. Without it, an organization is stumbling in the dark, unable to see the most obvious obstacles in its path.

1.2 About this publication

This publication describes how to manage information for an analytics and data-driven organization. It expands on the information available in *Governing and Managing Big Data for Analytics and Decision Makers*, REDP-5120, which introduced the concept of a data reservoir. This publication provides enterprise architects and information architects with more detail on how to create a data reservoir for their organization. It also presents a high-level overview of the data reservoir architecture.

The data reservoir architecture acts as a blueprint for a generic system of data repositories that are managed under a single information governance program. These repositories together offer a data distribution and self-service data access capability for analytics and other big data use cases.

This publication provides more details about the data reservoir architecture and how to customize it for your organization.

Before delving into the details of how to do this, it is worth exploring the impact that big data and analytics will have on an organization (both in terms of the new capabilities that are enabled and the cultural shift it takes to be data-driven) through a simple case study.

1.3 Case study: Eightbar Pharmaceuticals

Eightbar Pharmaceuticals (EbP) is a fictitious pharmaceutical company.

The company has grown from a small group of researchers working together in a spirit of open communication, collaboration, and trust to a medium-sized successful pharmaceutical company that has a small range of successful drugs in the market and many others in development (three of which look very promising).

Up to this point, their investment in IT has been focused on the automation of the manufacturing process driven by the growth in demand for their most successful drugs. However, the market is shifting to personalized medicine and the company realizes they need a greater investment in data and analytics to embrace this new market.

In recent weeks, they also uncovered some fraudulent activity relating to their manufacturing supply chain and realized that they needed to improve operations security.

The owners of EbP ask Erin Overview to develop an investment proposal for their IT systems to support personalized medicine and help the company become more data-driven.

1.3.1 Introducing Erin Overview

Erin Overview is an enterprise architect working at EbP. In fact, she is the only architect working at EbP. Her speciality is information architecture and she developed the organization's data warehouse. This data warehouse takes feeds from the key manufacturing, sales, and finance systems to create reports on the operational part of the business. She has had minimal interaction with the research team up to this point because they have their own tools and systems.

Erin has a small team that includes:

- Peter Profile is an information analyst who monitors the quality of the data flowing into the data warehouse.
- Gary Geeke is their IT infrastructure administrator who ensures the IT systems are up and running. He is able to perform simple installations and upgrades, and relishes the chance to work with new technologies.

This team (Figure 1-1) is responsible for the support of the core IT systems used by the company. They are not the only people involved, but they each have a potential leadership role in changing the way the company's information is managed.

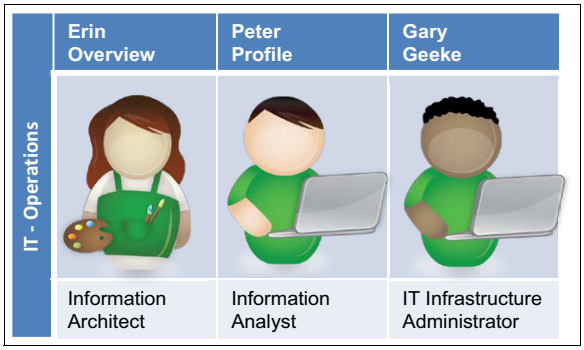


Figure 1-1 The IT Operations team

1.3.2 Perspectives from the business users at EbP

Erin realizes that enabling EbP to become data-driven and particularly to change their business towards personalized medicine is an enormous challenge that requires sizing and prioritizing if it is ever going to be delivered.

Her first task is to get some clarity on the requirements from various stakeholders in the business. This activity took a number of attempts because the various stakeholders needed help to understand how to develop realistic use cases.

Fraud investigation

Tom Tally is the accounts manager in the EbP's Finance team. He explained how hard it was during the recent fraud investigation to locate, extract, and then piece together the data from various systems to uncover the cause of the fraud.

To become more data-driven, he believes that they need a catalog of all of the data they have, better data consistency between the systems, and a simple way to access the data (Figure 1-2). He was also keen on developing fraud detection rules that can be deployed into the systems.

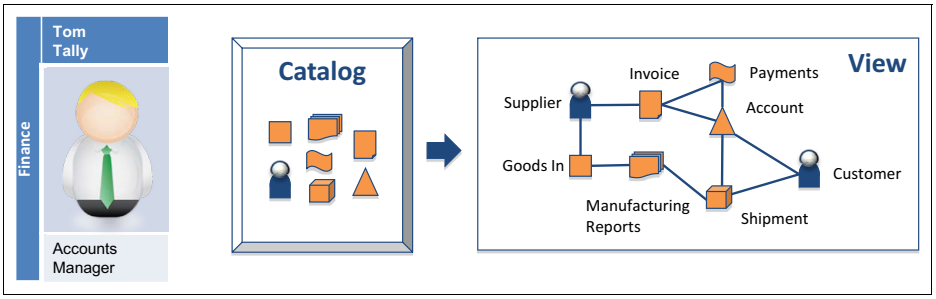


Figure 1-2 Use case for fraud investigation

Sales campaign

Harry Hopeful is a sales specialist for EbP. He has many contacts in various hospitals, and maintains a number of spreadsheets that record the sales visits he makes, who he sees, and the outcome (Figure 1-3). He would like to be able to refresh his spreadsheets with details of the latest customer and product information from EbP's core systems. This would save him time in planning his sales campaigns.

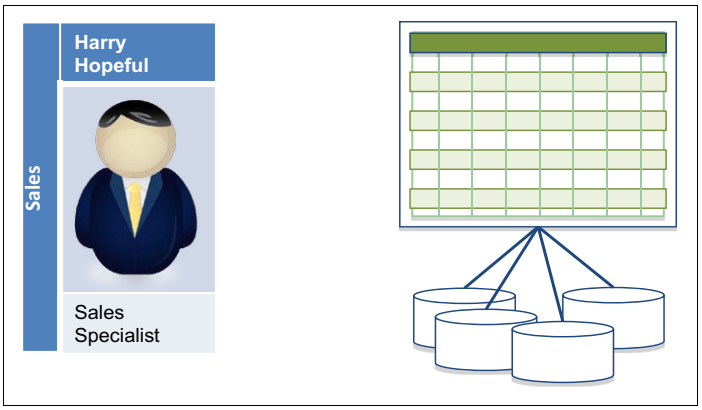


Figure 1-3 Analysis for sales campaigns

Erin asks if Harry would like to receive recommendations on who to visit and what to offer to his clients. He is interested, but skeptical that an IT system could do this.

Erin shared details of different types of customer analytics and the next best action solution, and Harry agreed that this could be useful to work towards. This process is particularly important because he has long-term experience in the industry and it would be useful to capture this knowledge before he retires.

Harry is asking for access to up-to-date operational data about their customers and products that can be used to dynamically update his spreadsheets. This data could also be used for customer analytics that makes recommendations on what to sell to each of Harry's customers.

Clinical trials

Erin then moves on to talk to the research teams. Tessa Tube is the lead researcher for their new range of personalized medicine. She has grand plans for how their personalized

medicine will work, creating a close relationship between the patients and medical staff working with the treatments. Tessa describes a system that uses mobile devices to connect medical staff and patients to EbP through a cloud. This solution would capture measurements and dispense information about the drugs being tested and potential side effects.

As a starting point Tessa would like to build a simple version of this solution for her clinical trials (Figure 1-4). This initial solution would collect measurements and notes from the people involved in the clinical trial. This data would then be routed to her team for analysis and further development.

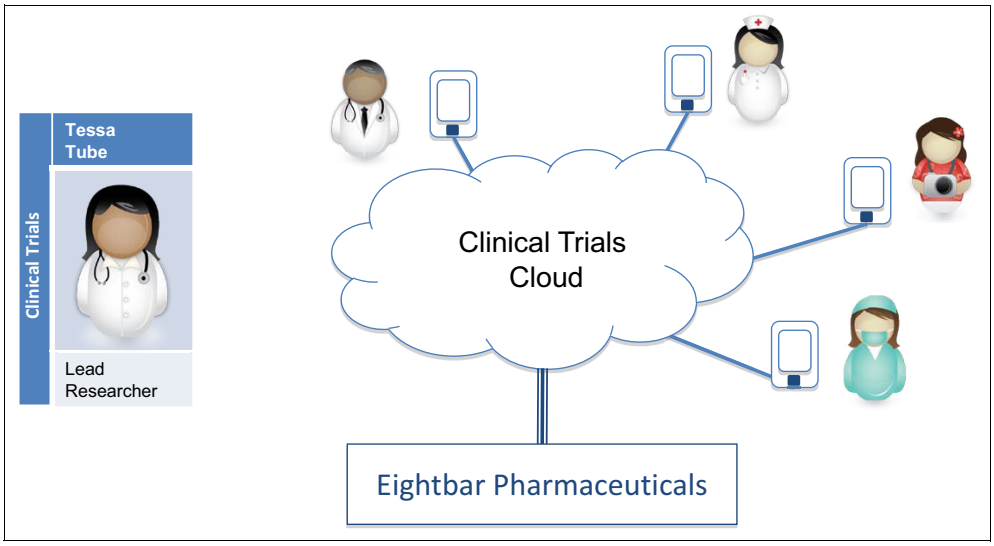


Figure 1-4 Clinical trials mobile-based solution

As Erin sketches out the solution that Tessa is asking for, she concludes that this clinical trials solution is going to require these components:

- Real-time access to information about the drugs being tested.
- Real-time execution of analytics to calculate the optimum mix and dosage for each patient.
- Initiation of requests to the manufacturing supply chain to manufacture the required treatment for each patient.
- The ongoing capture of data from medical staff and patients that describes the activity during a course of treatment.

As the personalized medicine part of this business grows, this solution is going to need to scale to support the needs of all their patients and medical staff. It will be the portal used by medical staff to understand the details of the products that EbP offers. It might well also become the principal sales channel for EbP, supporting advertising for new clinical trials and recommendations on treatments to use.

Clinical record management

Erin is keen to understand what type of information is collected in a clinical trial. She talks to Tanya Tidie, the clinical records clerk who is responsible for managing the clinical trial registrants and the results. Much of her work today is manual, dealing with paper-based records.

Initially the clinical trials are set up. The sales team (including Harry Hopeful) sign up consultants who are interested in participating in the trial. The consultants work with Tessa's team to determine which patients would benefit most from the trial. Next, they need to get the

patient's consent. Tanya must ensure that the correct documentation of this phase of the trial is captured, including the patient's consent, plus personal and clinical details of the patient to provide a basis for analyzing the results.

After the clinical trial is running, there is an ongoing collection of information about the treatments given to the patient, observations made by the medical staff, and measurements of the patient made during the treatment.

When the clinical trial is complete, Tanya must assemble the results of the clinical trial for regulatory approval.

Tanya assembles all of the evidence for a clinical trial on her personal computer using a local application. When the clinical trials solution is in place, although she will still need to manage the participant sign-up and permission through the paper-based process, the ongoing collection of evidence, such as treatments given and measurement of the patient's condition, throughout the clinical trial should be much simpler (Figure 1-5).

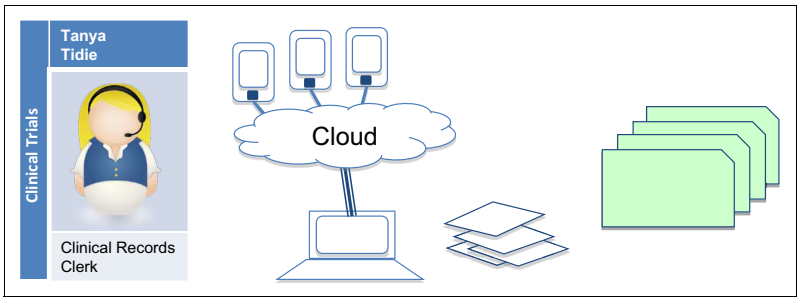


Figure 1-5 Clinical records management

Advanced analytics processing

Callie Quartile, one of Tessa's data scientists, also provides more details about the data management required to support the Treatment Advice Cloud for personalized medicine (Figure 1-6 on page 8).

Data feeds from the cloud are processed in real time to create a treatment plan based on information about the patient and their condition. This solution also feeds the historical data stores with both the measurements and calculations, and sends instructions to manufacturing to ensure that the correct course of treatment is available for this patient.

An object cache is fed details of the treatment plan to the patient and relevant medical staff from the historical data store. This approach provides a continuously available source of data for the mobile devices connected to the cloud.

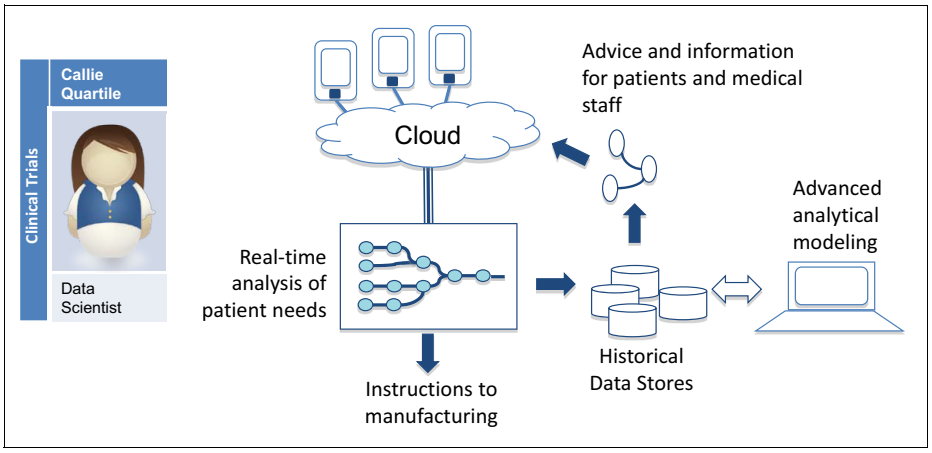


Figure 1-6 Analytics process for personalized medicine

Callie demonstrated her advanced analytics tools to Erin, showing how she creates samples of data, transforms each one for the analytics processing, and then creates candidate models and tests them on the different samples. These tools need to be able to connect to the historical data stores to extract samples of data to discover patterns in the data that will be used to configure the analytical models.

1.3.3 Signs of deep change

Erin is thoughtful after her discussion with Callie. Each of the business stakeholders is imagining that they will continue to operate as before, aided by more data. With her experience in information architecture, Erin can see that the real-time analytics driving both patient care and manufacturing is going to change the way that the whole organization operates.

Real-time analytics often causes the boundaries between the traditional silos (such as between research, manufacturing, and sales) to disappear and the way that drugs are sold and paid for is going to become more fine-grained.

Realizing that this is more than an IT project, Erin returns to the owners of EbP with her findings. She summarizes their data management requirements:

- The ability to drive and audit their operations to ensure that the research and manufacturing units are operating effectively together and are free of fraudulent activity.
- The ability to drive their research agenda through on-demand access to a wide variety of data sources. Data from these sources is combined and analyzed from different perspectives to create understanding of the conditions they are treating and the differences between individuals.
- The ability to meet the regulations associated with drug development, manufacturing, and personalized patient care.
- The ability to offer a more social interaction, with real-time insight and data gathering between both medical staff and patients during clinical trials and normal treatment schedules.

Figure 1-7 is a sketch of the systems she is envisioning.

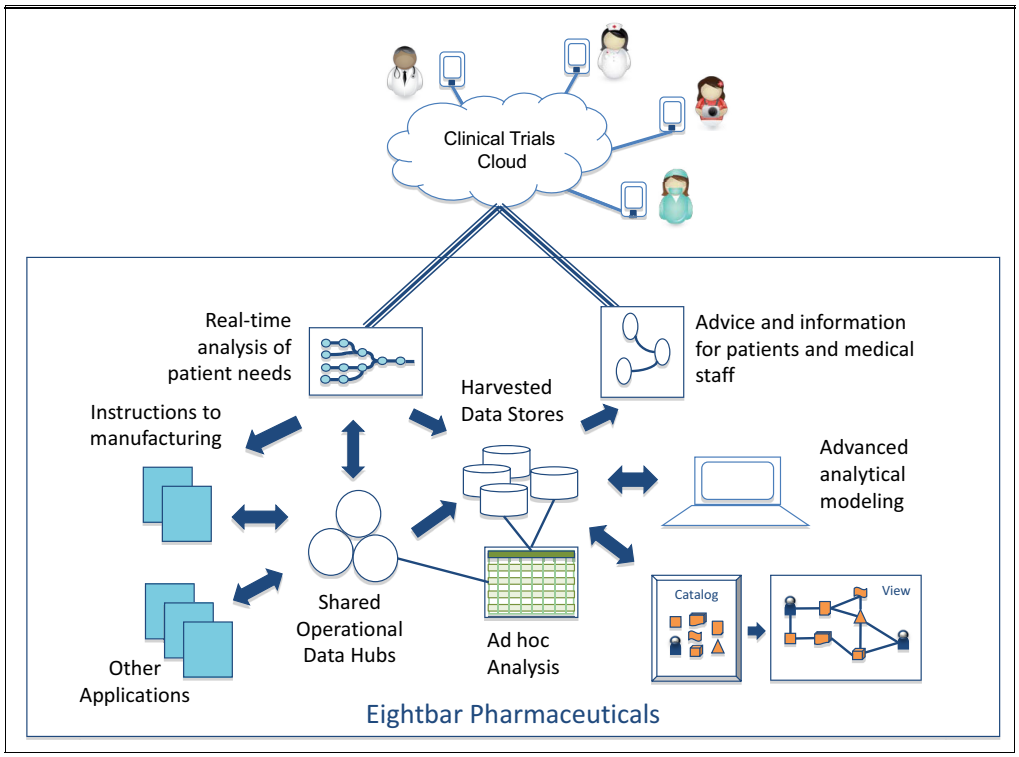


Figure 1-7 The new world?

At the heart of the solution are data stores of harvested data and a set of operational shared operational data hubs. The harvested data stores are fed from the operational systems, the treatment advice cloud, and other external sources that the research teams are using. Between them, they provide a coherent view of the business for daily operation and analytical processing. New data sources can be brought into the harvested data stores to extend the analytics in use by the medical teams.

In addition, she highlights that the culture and skills of the organization need to change:

- Consulting data and analytical results should become a standard practice in all types of decision making. Today, they rely on personal experience and best guess approaches.
- Information should be treated as a key company asset and managed with the same care as their drug development and manufacturing.
- The IT tools that the teams use need to be upgraded so that they are communicating and collaborating more effectively, both within the company and with external parties, while protecting their intellectual property.

These are all essential attributes of a data-driven organization.

The owners of EbP are understandably concerned at the complexity in Erin's sketch and the potential impact of this change. There are two outcomes from the meeting:

- They ask for some external validation of the architecture and a framework/roadmap to enable them to roll out the new capabilities in an iterative manner. This task is the role of the data reservoir architecture from IBM. This architecture was developed using IBM experience in building big data and analytics solutions for organizations that want to be data-driven.

- It covers the integration of both historical and operational data necessary to support real-time analytics and is highly componentized to allow for an incremental rollout.
- It provides self-service access to data for business and analytics teams while protecting and managing the data that the organization depends on.
- They appoint a Chief Data Officer (CDO) named Jules Keeper to manage the business transformation to a data-driven organization.

1.3.4 Governance and compliance perspectives

Jules Keeper is an experienced CDO who has led a number of successful information governance programs in other companies. He sees the role of information governance as an enabler of a data-driven organization. It should deliver these advantages:

- Understanding of the information that an organization has
- Confidence to share and reuse information
- Protection from unauthorized use of information
- Monitoring of activity around the information
- Implementation of key business processes that manage information
- Tracking the provenance of information
- Management of the growth and distribution of their information

Each of these aspects needs an element of skills and training for the people who use the information. The system also needs special procedures when people need to collaborate and agree on a particular course of action, and technology to automate much of the daily management and monitoring of information.

Information governance is important in the personalized medicine business because they must demonstrate proper information management and use at all stages of a product (drug) lifecycle, not just during the clinical trials. This change affects most people in the organization, and requires the systems to gather the evidence that information is being properly managed.

Jules is keen to work with Erin on the changes to the IT systems. He also introduces Erin to two of his colleagues who are involved in the compliance and governance of EbP:

- Ivor Padlock is the security officer. His background is in physical security, but a couple of years ago he branched out into IT security. He wants to understand how they can ensure that the personal data and research IP can be properly protected by the new systems.
- Faith Broker is an auditor focused on EbP's compliance with pharmaceutical industry regulations. She established the current set of standards that the teams work to today, and is an expert in the regulations in each of the countries they sell their treatments to. Personalized medicine is an emerging field in their industry and Faith has been working with the regulators on the safety and compliance requirements that they need to implement.

Together Jules, Ivor, and Faith make up the governance team that Erin will collaborate with during the project (Figure 1-8).

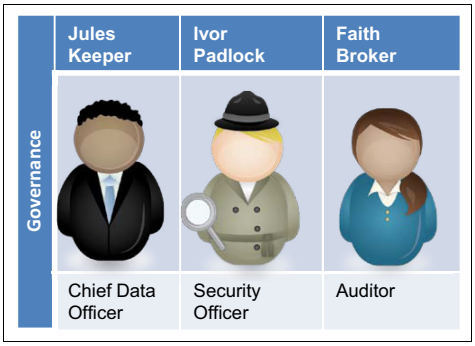


Figure 1-8 The governance team

1.3.5 Positioning the data reservoir in the enterprise architecture

Mindful of the request by the owners of EbP that she should follow established practices in the architecture and build out of the new data driven ecosystem, Erin starts to document and classify these items:

- The existing systems that will integrate with the new ecosystem
- The existing systems that will need major upgrade or change
- The new systems and capabilities that will be required

In particular, Erin wants to document the scope and position of the data reservoir in their existing IT landscape.

Most of the existing IT systems at EbP are referred to as systems of record (SoR). *Systems of record* are the operational systems that support the day-to-day running of the business. At EbP, these are the manufacturing, finance, sales, and administration systems. They are transactional systems focusing on the efficient operation of the business.

Systems of engagement (SoE) are systems that provide personalized and highly responsive support to individuals. At EbP, the new clinical trials solution is an example of a system of engagement.

The data reservoir is called a *system of insight* (SoI). The role of a system of insight is to consolidate data from systems of record and systems of engagement to support analytics. The system of insight supports both the creating of analytics and the execution of analytics to generate insight that can be used by both the systems of record and systems of engagement.

Together, the systems of record, systems of engagement and systems of insight provide a complete ecosystem for the modern business enabling responsive mobile applications, with the trusted reliability of the systems of record and added intelligence of the systems of insight.

Figure 1-9 shows that these three groups of systems are linked together and exchange data.

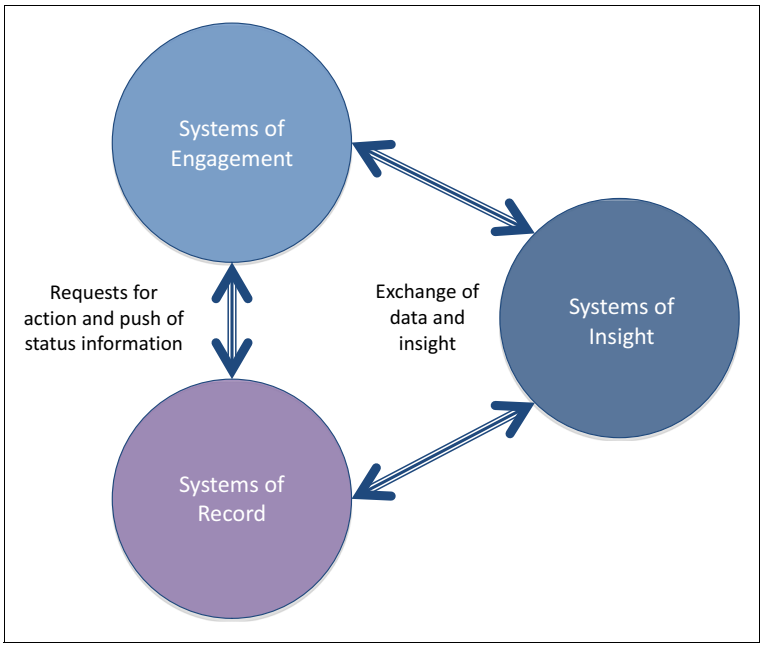


Figure 1-9 Connecting systems of record, systems of engagement and systems of insight

The systems of insight typically receive data from the systems of record and systems of engagement and then distribute both data and analytical insight in return. Between the systems of record and systems of engagement are requests for action and push messages that update status related to the activities in either group of systems.

Systems of record, systems of engagement, and systems of insight each have different architecture and governance requirements. These types of systems have the following differences:

- Systems of engagement support individuals.

  They combine functions supported by multiple parts of the business to create a unique and complete experience for the users they support. Their function tends to evolve rapidly and they must to be available whenever the individual needs them. The systems of engagement use caches and queues to provide their services even when the systems of record and systems of insight are not be available.

- Systems of record support the business and focus on efficient processing of transactions.

  They are typically long lived and each maintains their own database of information. This information is typically organized around business activity or transactions. This information from the systems of record must be fed into the systems of insight to enable the systems of insight to keep up-to-date with the latest activity in the business.

- Systems of insight need to support a wide variety of data types, sources, and usage patterns.

  They must be highly adaptive to the increasing requirements for data and analytics.

From a governance perspective, all systems need to be secure. The systems of insight provide much of the archiving and retention support for the other systems. Data quality is managed in the systems of record and systems of engagement at the point where data enters the systems. The only exception to this is the quality work performed on the feeds of data from external sources. However, this is limited because subsequent updates from the external source will overwrite any data that has been corrected.

Erin overlays the EbP systems on top of the ecosystem (Figure 1-10). The clinical trials application is shown in the systems of engagement. The manufacturing, sales, and other administrative systems are systems of record and the new analytics capability, with its associated data, is sitting in the data reservoir as a system of insight. This gives her a scope for the data reservoir.

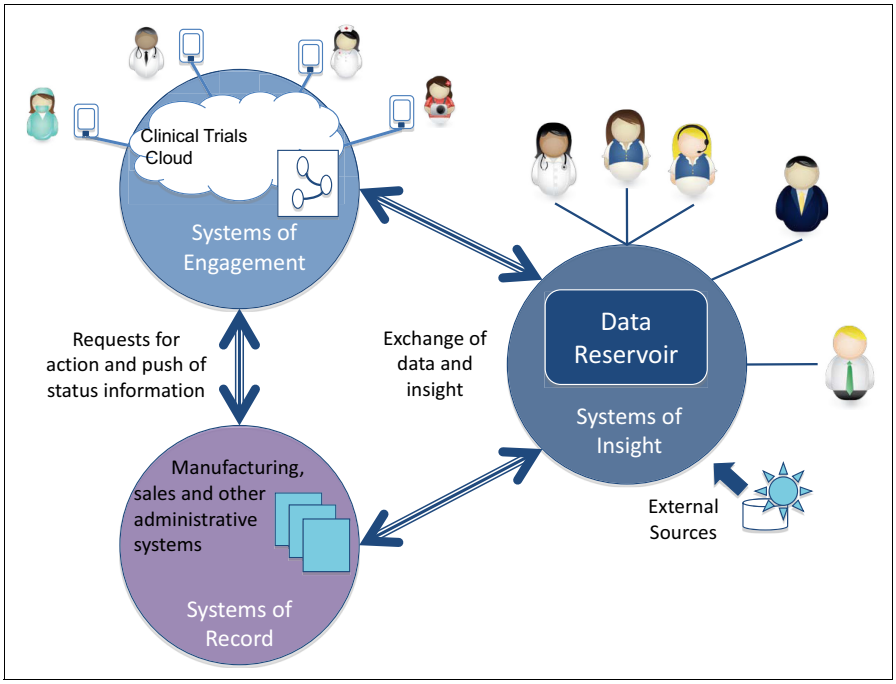


Figure 1-10 EbP systems overlaid on systems or record, systems of engagement, and systems of insight ecosystem

1.3.6 The data reservoir

As a system of insight, the data reservoir is the data and analytics intensive part of the ecosystem that supports all these teams:

- The research team in their daily work
- Providing sales with the latest information about customers and products
- Supports the finance team with ad hoc data queries

Much of the operation of the data reservoir is centered on a catalog of the information that the data reservoir is aware of. The catalog is populated using a process that is called curation (Figure 1-11).

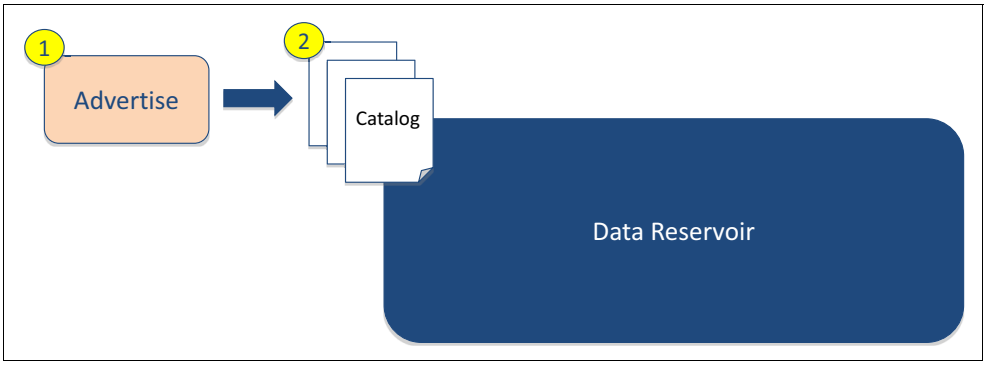


Figure 1-11 Advertising data for the data reservoir

The first type of curation is the advertisement of existing data sources in the catalog (Item 1 in Figure 1-11 on page 13). A description of a data source is created either by the owner of the data source or a trusted curator, and stored in the catalog (Item 2 in Figure 1-11 on page 13). For more information about information curators, see “Role Classifications” on page 41.

The description of the data source includes these details:

- Data source name, short description, and long description
- The type of data that is stored in the data source and its classification
- The structure of the data (if present)
- The location of the data, in terms of its physical location and electronic address

This description of the data source helps someone looking for data to discover and assess the appropriate data sources (Item 3 in Figure 1-12).

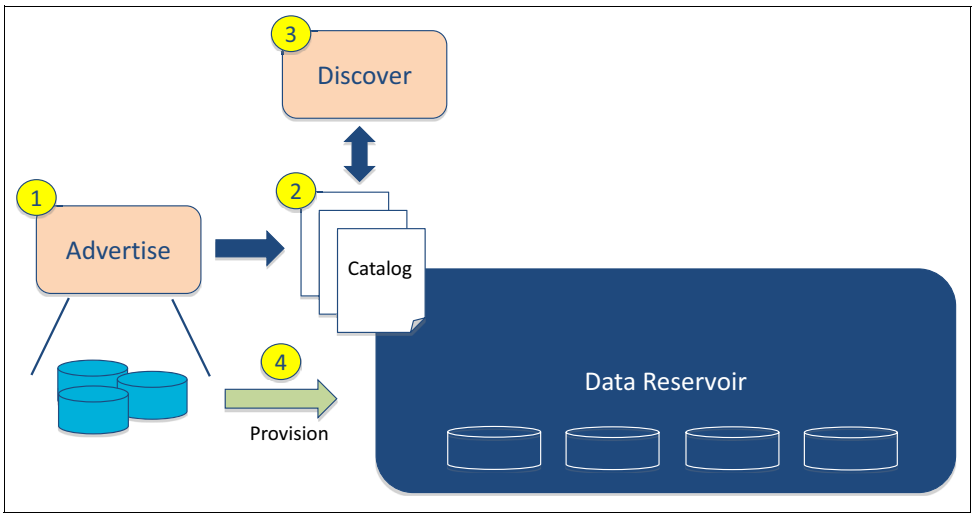


Figure 1-12 Data discovery in the data reservoir

After it is clear that data from a particular data source is needed in the data reservoir, it is provisioned into the data reservoir. The provisioning process typically includes an initial copy of data into one or more of the data reservoir repositories, followed by incremental updates to the data as it changes in the original sources.

In special circumstances, a real-time or federated approach can be used to provision the data source for the reservoir.

Real-time interfaces retrieve data from the data source on demand. Federation enables data to be returned in real time from multiple data sources in a single response. These approaches have the advantage that a copy of the data source does not need to be maintained in the data reservoir repositories. However, the data source must have the capacity to handle the unpredictable query load from the data reservoir users.

Real-time provisioning could be an initial approach for some new data sources that can be changed to a copy style of provisioning as the number of users (and hence the load on the original data source) increases.

Whichever way the data source is provisioned into the data reservoir, after the wanted data is located, it is copied into a sandbox for exploration (Item 5 in Figure 1-13)

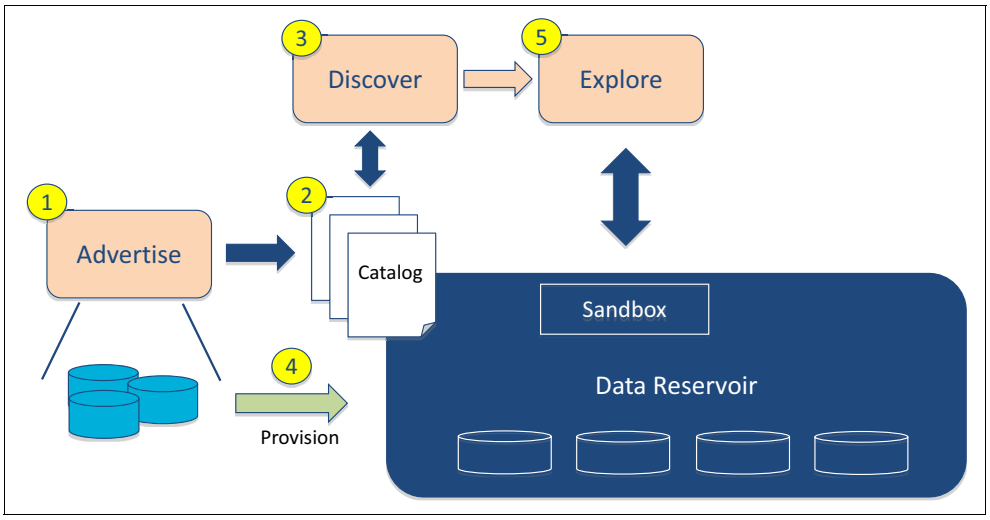


Figure 1-13 Data exploration and the data reservoir

Data exploration is the process of experimenting with and visualizing data to understand the patterns and trends in that data. It often involves reformatting data and combining values from multiple data sources. This is one of the reasons why data is copied into a sandbox for this work. The other reason is to limit the direct access to the data reservoir repositories to, ideally, just the data reservoir services, so that the use of data can be properly logged and audited.

Data exploration might result in the development of new analytics models and business rules. These new functions might be deployed in the data reservoir or in the connected systems of record or systems of engagement (Item 6 in Figure 1-14).

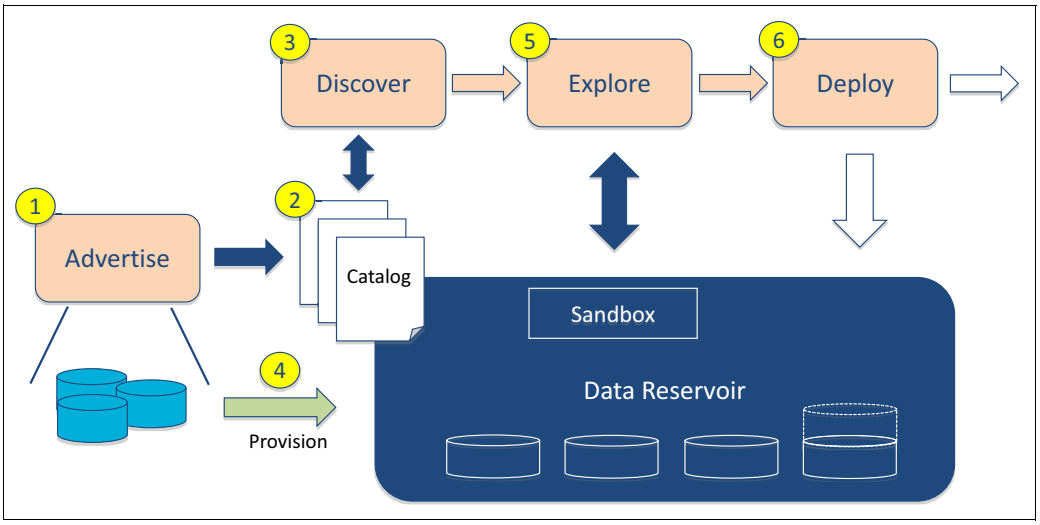


Figure 1-14 Deploying analytics into production

The deployment of these functions involves some quality checks, potentially some reformatting of the data exchange that the function expects, and then integration into the deployment environment.

Another common role for the data reservoir is to act as a data distribution broker between different systems that are connected to it (Figure 1-15).

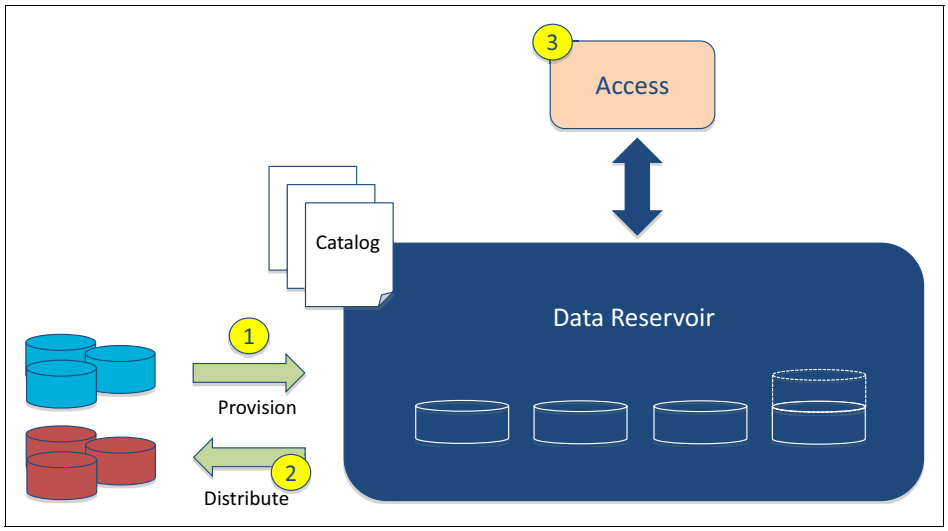


Figure 1-15 Data distribution from the data reservoir

The latest values from original data sources are continuously entering the data reservoir (Item 1 in Figure 1-15). Selected values can be distributed to other systems (Item 2 in Figure 1-15). These values can also be accessed on demand through real-time interfaces, providing the latest values when they are needed.

1.3.7 Inside the data reservoir

From the outside, the data reservoir seems to be a simple collection of data sources. Inside the data reservoir is a complex set of components that are actively governing, protecting, and managing the data. The internals of the data reservoir are presented as a series of levels of subsystem and component diagrams of increasing levels of detail.

Figure 1-16 shows level 1, the high-level structure of the data reservoir.

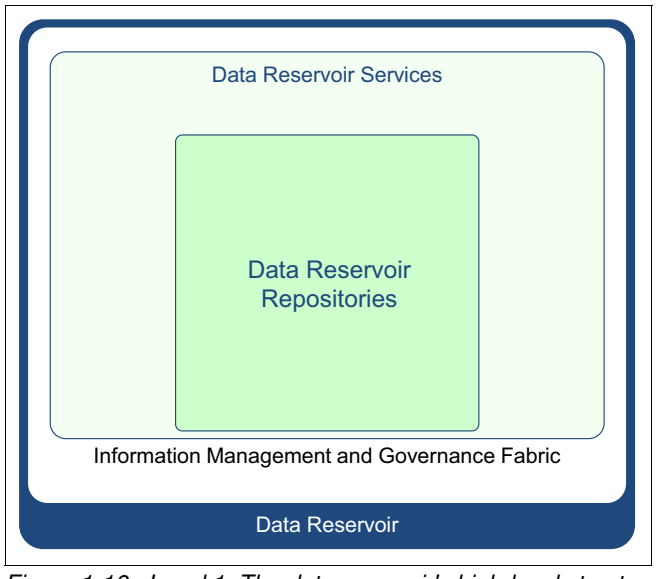


Figure 1-16 Level 1: The data reservoir's high-level structure

Data reservoir repositories

In the center are the data repositories. These repositories provide shared data to the organization. Each repository either supports unique workload capabilities or offers a unique perspective on a collection of data. New repositories can be added and obsolete ones removed during the lifetime of the data reservoir.

The same kind of data can be present in multiple repositories.

Data reservoir services

It is the responsibility of the data reservoir services to keep these copies in synchronization, and to control and support access to the data reservoir repositories. The data reservoir services include a catalog to enable people to locate the data they need and verify that it is suitable for their work.

Information management and governance fabric

Underpinning the data reservoir services is specialist middleware that provides the information management and governance fabric. This middleware includes provisioning engines for moving and transforming data, a workflow engine to enable collaboration between individuals working with the data, and monitoring, access control, and auditing functions.

1.3.8 Initial mapping of the data reservoir architecture

When Erin overlays her initial architecture sketch (Figure 1-7 on page 9) on the data reservoir structure (Figure 1-17), she begins to separate the different concerns of the data-driven ecosystem. This approach clarifies the boundary of the shared data and its management process from the users and feeds interacting with it.

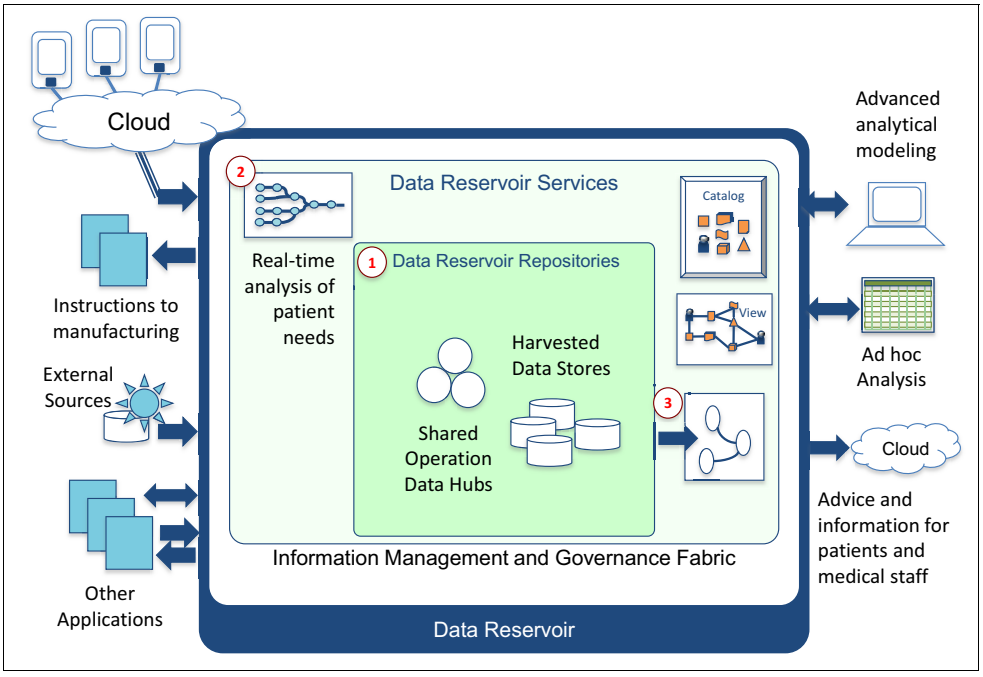


Figure 1-17 EbP's requirements overlaid on the data reservoir's high-level structure

Erin classifies the shared operational data hubs and harvested data stores as data reservoir repositories (Item 1 in Figure 1-17). The real-time analysis of patient needs becomes a data reservoir service (Item 2 in Figure 1-17).

The object cache for the clinical trials cloud is also a data reservoir service rather than a repository, even though it stores data (Item 3 in Figure 1-17). This is because its contents are derived from the harvested data stores, so it is a type of materialized view for the data reservoir. This cache is in the clinical trials cloud. However, logically it sits inside the data reservoir because it is managed and governed by the data reservoir processes.

The next level of detail of the data reservoir architecture gives some structure to the data reservoir services (Figure 1-18).

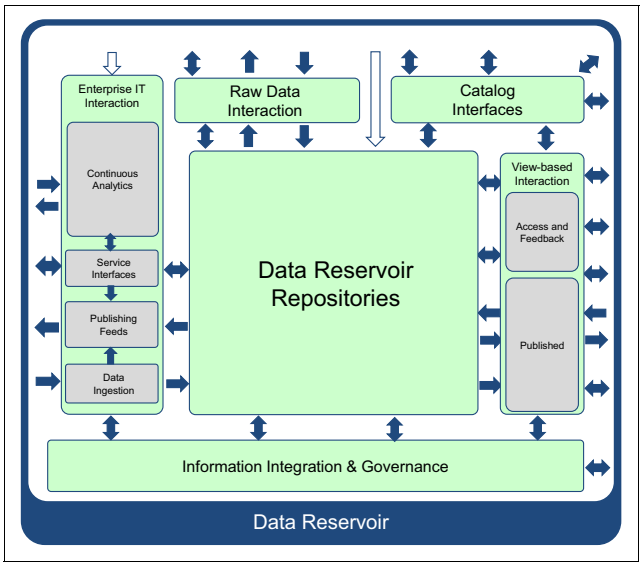


Figure 1-18 Level 2: The data reservoir's internal subsystems

Level 2 of the data reservoir architecture helps Erin identify that some users, such as Callie Quartile the data scientist, are able to work with raw data. However, others, such as the accounts manager Tom Tally, need simpler views of the data (Figure 1-19).

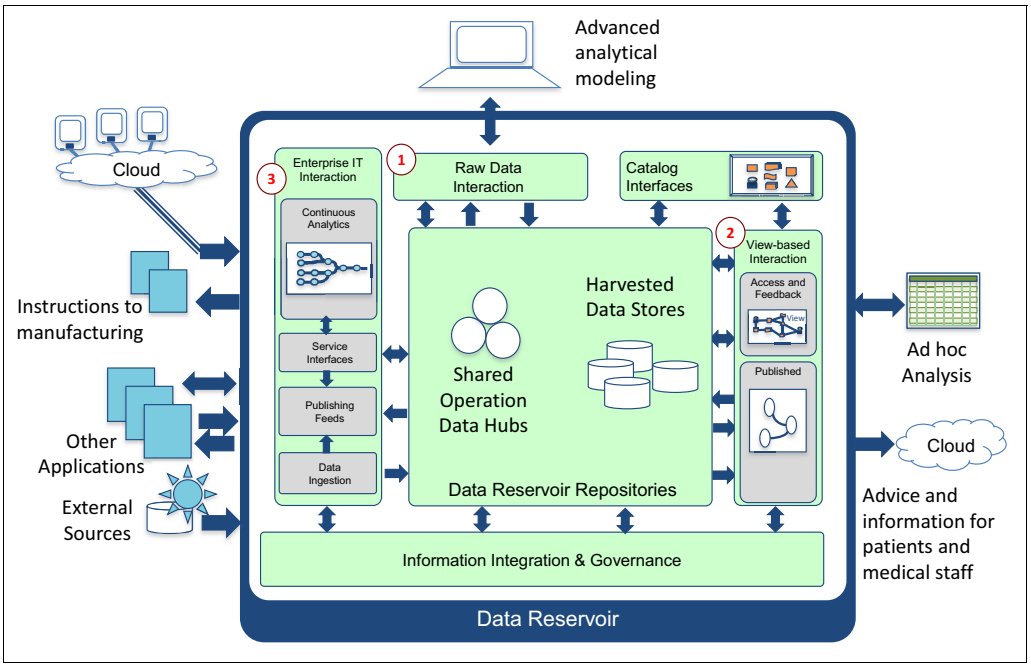


Figure 1-19 Internal subsystems in EbP's data reservoir

Erin connects the advanced analytical modeling to raw data interaction (Item 1 in Figure 1-19) and the ad hoc analysis to view-based interaction (Item 2 in Figure 1-19). She also added the different enterprise IT interaction subsystems because she knows she needs them to connect to their existing applications and new data feeds from eternal sources (Item 3 in Figure 1-19).

The fine detail architecture picture (Figure 1-20) adds pattern-based components that act as a checklist for any additional detail that Erin wants to highlight at this stage. Chapter 3, “Logical Architecture” on page 53 covers these components in detail. Much of the additional detail describes the kinds of data repositories that could be supported in a data reservoir (an organization rarely deploys all possible types of repositories).

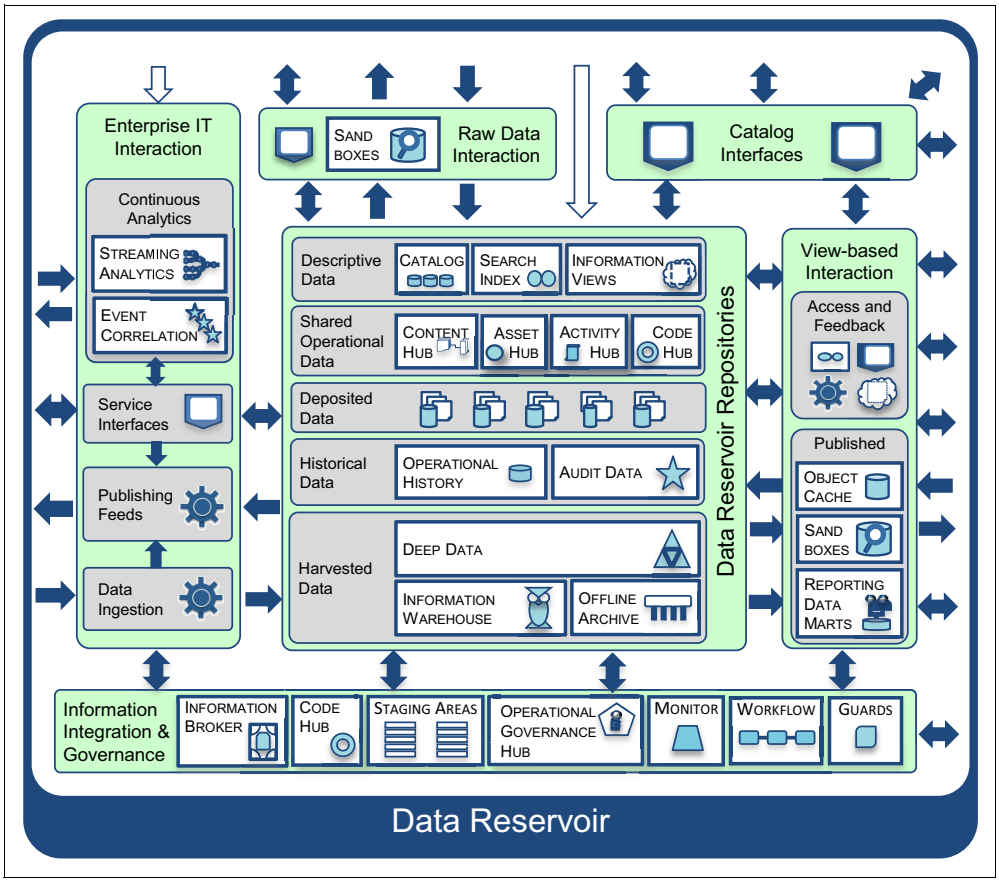


Figure 1-20 Level 3: The data reservoir's internal detail

In the data reservoir architecture, each type of repository characterizes a different kind of information collection, or data set, in terms of the disposition of its content (that is structure, current or historical values, read/write, or read-only). For example, the Operational History repository stores historical data from a single operational system. Except for time stamps, this data is stored in the same format as the original system so it can be used for operational reporting, quality analysis, and archiving for that operational system. There is a different instance of the Operational History repository for each operational system that needs this type of store. Deep Data, however, holds raw detailed data from many sources, along with intermediate and final results from analytics.

The infrastructure that hosts the data reservoir repositories is selected based on the amount of data and the anticipated workload. It is possible that a general-purpose data platform, such as Apache Hadoop, can support all of the selected repositories. However, it is more common to see a mix of infrastructure (including data warehouse and analytics appliances) making up the infrastructure platform of the data reservoir.

Erin decides to focus on selecting the data repositories and usage paths at this stage. She will work with Gary Geeke, their infrastructure expert, after they have a clear view of the volume, variety, and velocity of data for the data reservoir (Figure 1-21).

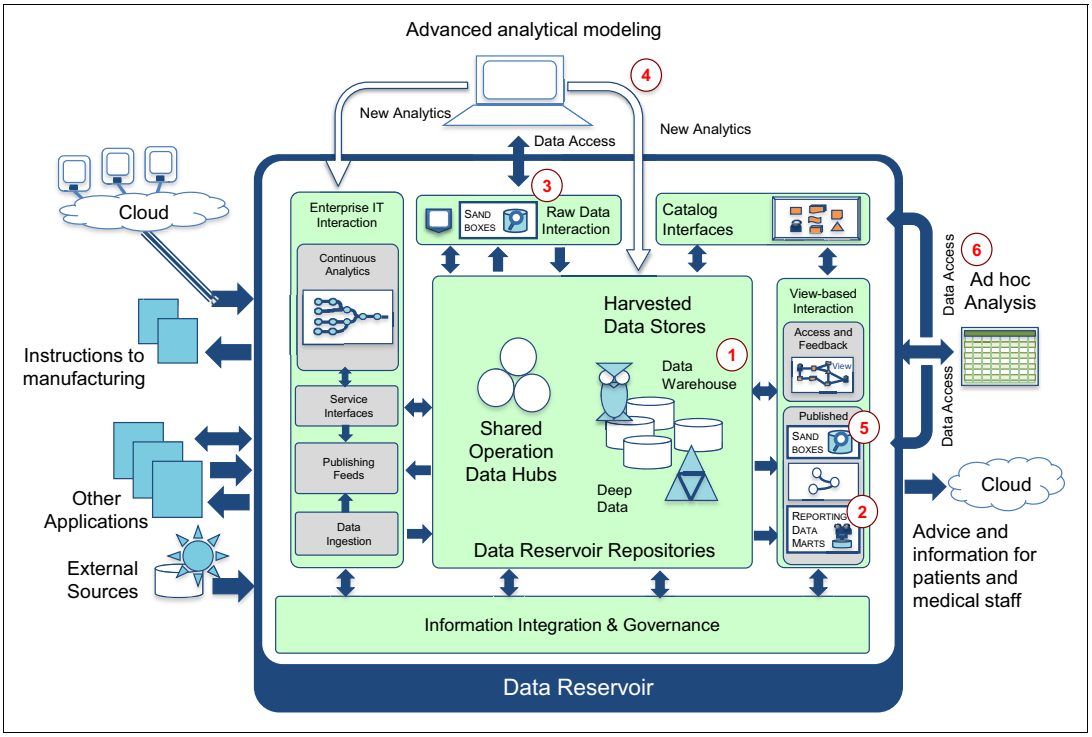


Figure 1-21 Internal detail of EbP's data reservoir

Erin notices that there is an information warehouse in the data reservoir architecture and decides to include their existing data warehouse (Item 1 in Figure 1-21) in the data reservoir. It contains much of the data that will be needed for the data reservoir.

Erin also adds the data marts populated from the data warehouse (Item 2 in Figure 1-21). They are added to the view-based Interaction because they are derived from the data warehouse and are entirely recreatable. They also offer simplified, specialized views of different subsets of the data reservoir, which further confirms their place in the view-based interaction subsystem.

Erin adds a sandbox and data access interface to the raw data interaction (Item 3 in Figure 1-21). The data scientist will not access the data reservoir repositories directly. The raw data interaction subsystem allows them to access and copy any data that they need, ensuring they have access rights to the data they are requesting and monitoring their data use.

Erin adds detail around the advanced analytical modeling, showing it creating analytical models that run both in the data reservoir repositories and the continuous analytics service (Item 4 in Figure 1-21). They must work out a quality assurance process for the deployment of these models because both hosting destinations will be operational production systems when they are supporting the personalized medicine business. The research teams are not used to the discipline associated with working with operational systems.

Erin also added sandboxes to the view-based interaction (Item 5 in Figure 1-21) to allow the business teams to create copies of data they want to manipulate. This means they can view the data descriptions in the catalog, access the data through an API, and copy it into a sandbox (Item 6 in Figure 1-21).

At this stage, Erin's confidence in the lifecycle of the data in the data reservoir and how it will be accessed increases.

The data reservoir architecture helps to classify and organize the different data repositories, processing engines, and services required for big data and analytics.

1.3.9 Additional use cases enabled by a data reservoir

Looking into the future, the data reservoir can act as the analytics source for the customer care organization in these ways:

- The customer marketing team can request subsets of data for selected marketing campaigns.
- The manufacturing team might want to analyze the demand for different types of treatments so they can plan for improvements to the manufacturing process.

After data is organized, trusted, and accessible, it helps people to understand how their part of the organization is working and improve it.

1.3.10 Security for the data reservoir

As you can imagine, the data reservoir has the potential to hold a copy of all kinds of the valuable data for an organization. In EbP's case, it will contain their valuable pharmaceutical IP that they have built their business on, plus details of their customers, manufacturing activity, and suppliers. In the wrong hands, this data could be leaked to external parties or be used to exploit weakness in their operations, exposing them to further fraud. Security is a key concern.

Security requirements come from various sources:

- An assessment of the key security threats that need to be guarded against, including:
  - Deliberate (theft, denial of service, corruption, or removal of information)
  - Accidental
  - Failures
- Legal, statutory, regulatory, and contractual requirements
- Business requirements and objectives

This section considers how security should be implemented in the data reservoir.

Figure 1-22 shows the stakeholders for data reservoir security and the effect it should achieve.

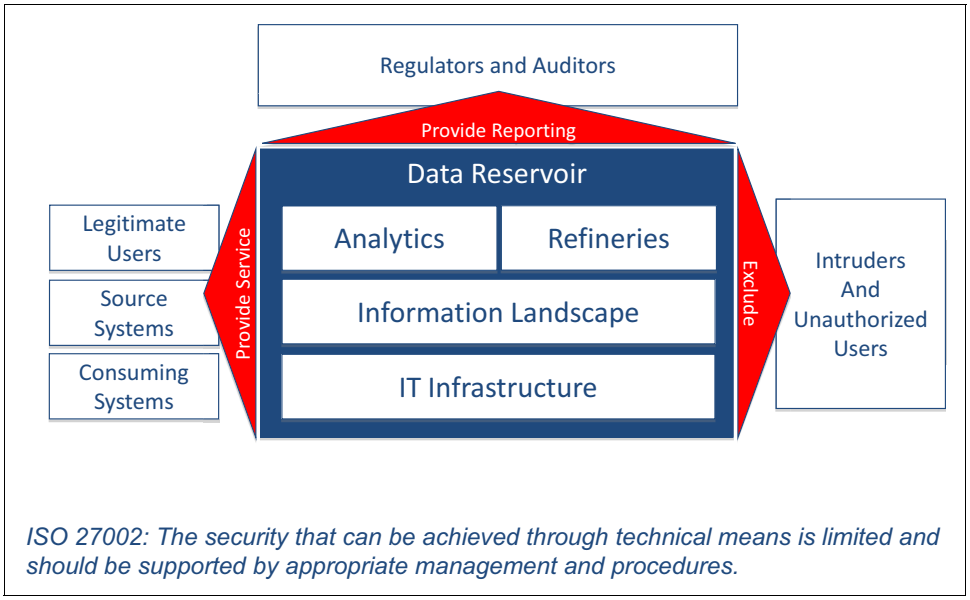


Figure 1-22 Stakeholders for data reservoir security

The data reservoir must provide a useful service to legitimate users and both source and consuming systems. It must provide evidence to regulators and auditors that information is being properly managed. Finally, it must effectively exclude intruders and unauthorized users.

This protection must extend over the IT infrastructure, data analytics, and the data refineries that manage the data in the data reservoir.

Erin and Ivor Padlock, EbP's security officer, defined these as EbP's core security principles:

- An individual's privacy will be maintained.
- Information is only made available on a need to know basis.
- The organization's information will be protected from misuse and harm.
- Only approved users, devices, and systems can access the organization's data.

These principles should be thought of as the high-level goals of the data reservoir's security. The implementations of these goals are delivered through security controls.

Building the security manual for the data reservoir

The following security control categories list the different aspects of security that need to be considered. These are based on the guidance from the ISO27000-5 set of security standards.

These control categories are the chapters in the data reservoir's security manual. Each focuses on how a particular facet of security will operate.

- Vitality of security governance

  Vitality of security ensures that the information security of the data reservoir provides adequate protection despite the changing risk landscape.

  Vitality includes processes for raising concerns and new requirements for security of the data reservoir. There should be clear responsibilities on who should review and update the security processes.

- Information classification and handling principles

  The information classification and handling principles clarify kinds of information, and the requirements for accessing and using information.

  The classification schemes are the key to ensuring that appropriate security is applied to data based on its value, confidentiality, and sensitivity. The classifications are stored in the catalog and linked to the descriptions of the data in the data reservoir. These classifications are used by the data reservoir services to ensure that data is properly protected.

  Many organizations have security classification schemes defined already and, if possible, these should be adopted for the data reservoir so they are familiar to its users. Otherwise, there is a default set of classification schemes defined on IBM developerWorks®.

- Information curation

  Information curation identifies catalogs, and classifies and describes collections of information to define the scope and ownership of information in the data reservoir and the appropriate protection responsibilities.

  Related to classification, curation ensures that the way data is described in the catalog is accurate. A high-quality catalog means that data can be found, understood, and used appropriately. There are three control points where curation occurs and there is a potential to ensure the protection of data or the people who use it:

  – When a new data source is advertised in the catalog for the data reservoir. This is a key point where data is classified. Has the correct classification been used?

  – When a new project is started and the project leader establishes a list of the data sources that the team will use. Have the most appropriate data sources been selected?

  – When a person is searching for data and leaves feedback or ratings. Is this feedback highlighting a valid problem either in the data or in the description of the data?

- Provision of information

  Provision of information is used to accomplish these objectives:

  – To ensure that information in the data reservoir is appropriately hosted, classified, and protected, while still being made available to authorized users.

  – To ensure that users of the reservoir are confident that information will be available whenever it is needed.

  Assuming that data has been correctly curated and classified, the data reservoir infrastructure and services should provide appropriate custodianship of the data.

- Maintenance of information trust

  Maintenance of information trust ensures that users of information know the pedigree of the information they are using

The data reservoir should maintain lineage for the data that it manages, and that lineage data should be trusted and protected. The data reservoir must ensure that data flowing into the data reservoir comes from verified sources (Figure 1-23).

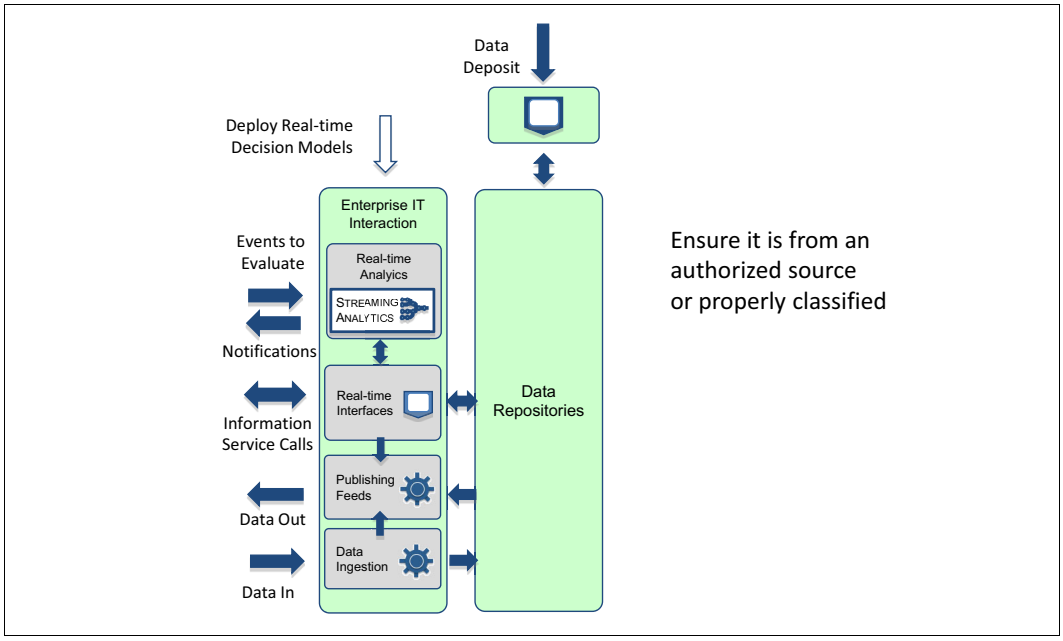


Figure 1-23 Trusting data flowing into the data reservoir

- User access of information

  User access of information ensures only authorized users access information and they use it correctly. An organization must answer several questions around the way that access control is managed. Here is a set of candidate questions to consider:

  - Who are the people who have a right to access the information in the reservoir?
    - All employees?
    - Contractors?
    - Business partners?
  - Who vouches for (and validates) a user's access rights?
  - What are the terms and conditions you want users of the data reservoir to sign up to?
  - What are the risks you need to guard against?
  - How granular should security access be controlled (collection, entity, attribute)?
  - What are the guarantees that the data reservoir offers to information providers and users?
  - How deep does the specific user identification penetrate?
    - Does the repository know who the user is?
    - Are repositories accessed directly?
  - What is valid processing in the data reservoir?
    - What is the lifecycle that information in the reservoir goes through?
    - Does its security requirements change at different stage of its lifecycle?
    - What is the analytics lifecycle and what are the security implications?
    - What are the software lifecycle processes and how is that process assured?

– How does the context of a request affect access rights?

Typically a data reservoir is set up to ensure that no user accesses the data reservoir repositories directly. Authorized processes access the data reservoir's repositories and people access the data from the data reservoir through the services. In addition, there are the following items to consider:

– Individuals are identified through a common authentication mechanism (for example Lightweight Directory Access Protocol (LDAP)).
– Data is classified in the catalog.
– Access granted by business owners.
– Access controlled by data reservoir services.
– All activity is monitored by probes that store log information in the audit data zone.

After these questions have been answered, a number of end-access services that must be covered by access management. These are covered by Figures 24-27.

Which data is discoverable in the catalog (Figure 1-24)?

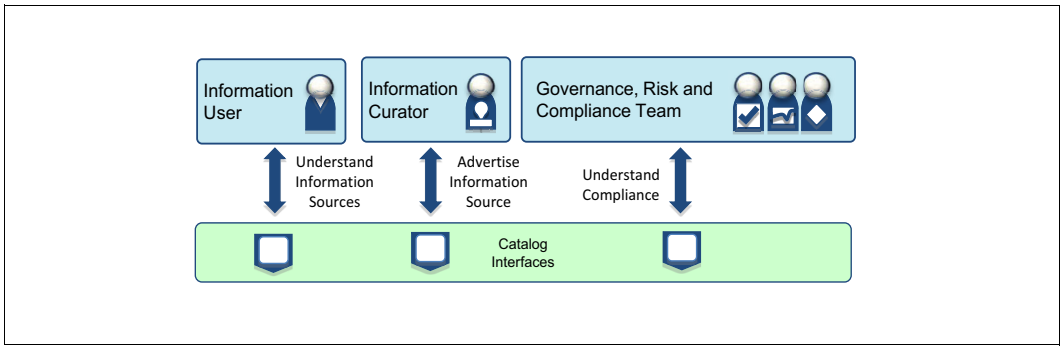


Figure 1-24 What data is discoverable in the catalog?

What data can be extracted from the sandboxes and who can access the sandboxes (Figure 1-25)?

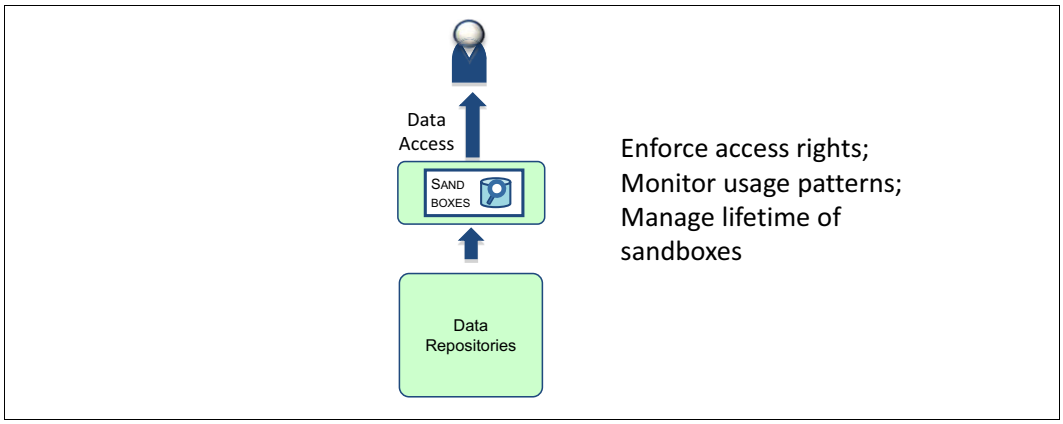


Figure 1-25 What data can be extracted from the data reservoir repositories?

What data values can be located with search technology or accessed by the real-time interfaces (Figure 1-26)?

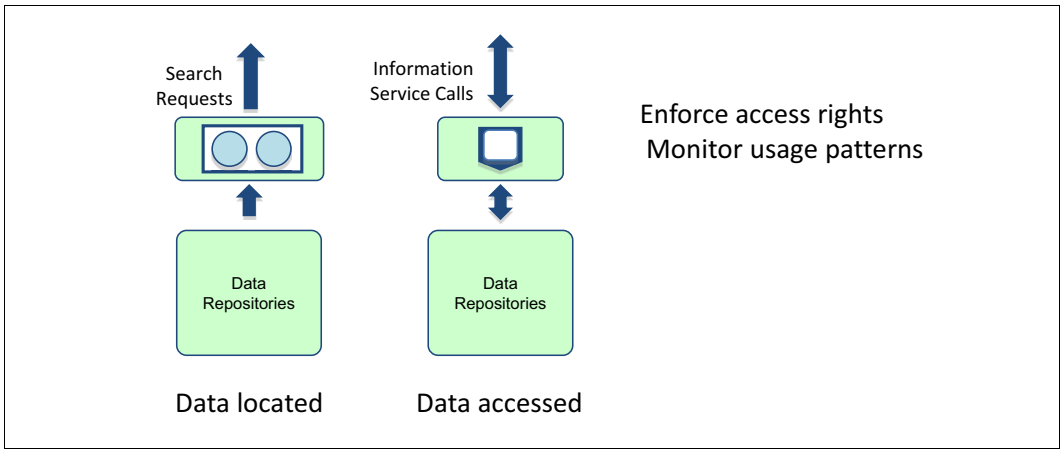


Figure 1-26 What data values can be located and accessed?

Finally, what access to data is given to the teams managing the data reservoir infrastructure (Figure 1-27)? Can they see the data values?

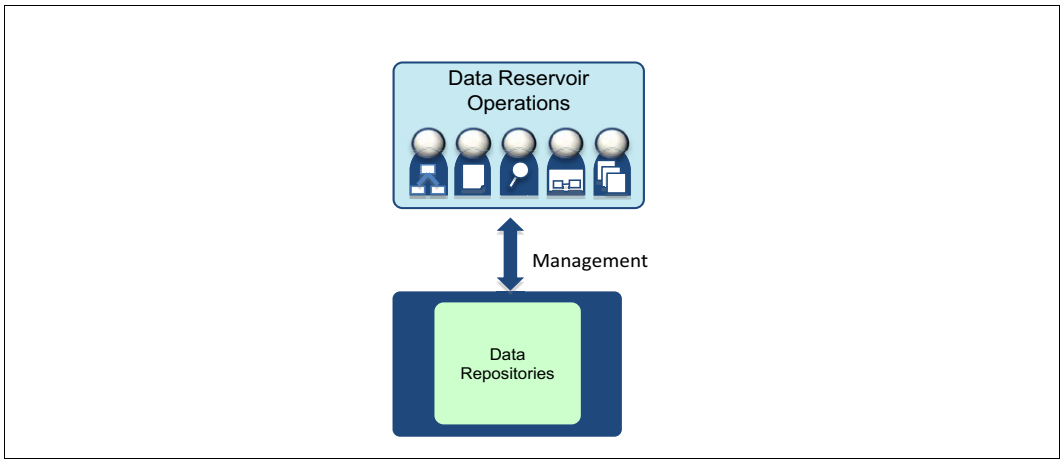


Figure 1-27 What access to data do the data reservoir operations and maintenance teams have?

- Protection of information subject

  Protection of information subject ensures the privacy and appropriate use of information about an individual.

  Most data reservoirs store personal data. There are laws on how this data must be managed in most countries. The data reservoir must respect these requirements.

- Analytics validation

  Analytics validation ensures that the insight generated is meaningful and ethical.

  The consolidated and linked data in the data reservoir enables the analytics teams to create detailed insights into individuals. How does the organization decide which analytics are appropriate for its business? The organization needs a framework for deciding on the appropriateness of the analytics they develop and ensuring proper safe-guards are in place.

- Data refinery development

  Data refinery development ensures software that is transforming information is properly reviewed, tested, and approved.

  The development team building data refinery services need proper checks to ensure erroneous or malicious code cannot be introduced into the data reservoir that could corrupt or leak data.

- Return of information

  Return of information to ensure that information is returned or deleted when no longer needed.

  Sensitive data represents a risk, so there might be occasions where data is only retained in the data reservoir for short periods to minimize this risk. How do you ensure that all copies of this data are eliminated? Similarly, there are regulations that enable a person to request that information about them is deleted from a company's data stores. How does the data reservoir show that this has happened?

- Maintenance of enterprise memory

  Maintenance of enterprise memory ensures that records are correctly managed and retained.

  The data reservoir can be used as an online archive for many systems, in which case it must take on the responsibility for the proper retention of this data.

- Incident reporting and management

  Incident reporting and management ensures that potential and actual security breaches are reported and resolved in a timely and effective manner.

  Incident reporting and management should be a closed loop, ensuring that all incidents are raised, reviewed, and the appropriate action is taken. There should be reporting to highlight the kinds, levels, and severity of incidents that are occurring over time.

- Integrity of information security

  Integrity of information security ensures that the information is used to guard the data reservoir is correctly managed.

  This might seem obvious, but if anyone can access and change audit information, then it is not possible to trust that it is an accurate record of the activity in the data reservoir.

1.3.11 What does IBM security technology do?

Technology's role in protecting the data reservoir is to provide automation at key points in the processes and services that surrounds the data, including the following items:

- Centralized management of identity authentication of users connecting to the data reservoir
- Classification of data and detection of incorrectly classified data
- Implementation of rules associated with classification
- Auditing, monitoring, and report generation of all activity in the data reservoir
- Encryption, masking, and redaction of data both at rest and in motion
- Authorization of access to data reservoir repositories and services
- Automated choreography of actions to grant/revoke access, resolve issues, investigate suspicious activity, and audit access

Each of these seeks to support the education, manual procedures, and controls that create a culture where the protection of data is a priority.

1.4 Summary and next steps

In the past, analytics has been a back-office function that reported on the past activities of the organization. It enabled a business to understand trends in their operation or market so they could plan for the future.

Real-time analytics and general access to a broader range of data is disruptive, often causing the boundaries between the traditional silos in an organization to break down as people start to see a broader view of the organization and desire more responsive decisions. Mobile applications provide more data and enable real-time insight to be delivered to individuals, further accelerating the demand for a more responsive organization.

Together, the systems of record, systems of engagement, and systems of insight provide a complete ecosystem for the modern business. This system enables responsive mobile applications, with the trusted reliability of the systems of record and added intelligence of the systems of insight.

Systems of record, systems of engagement, and systems of insight each have different architecture and governance requirements. The data reservoir provides a generic architecture for a governed and managed system of insight.

Data Reservoir = Efficient Management, Governance, Protection, and Access.

The data reservoir architecture is presented as a number of levels of increasing detail that helps to classify and organize the different data repositories, processing engines, and services required for big data and analytics. It also acts as a checklist to identify all of the capabilities that are needed to manage a system of insight.

2

Defining the data reservoir ecosystem

The data reservoir solution defines the types of technology components necessary to manage and share a wide variety of information. In addition, it considers the ecosystem of processes and teams that interact with these technical components to ensure that the data in the data reservoir is available, findable, useful, and properly protected.

This chapter takes a closer look at the ecosystem around the data reservoir.

This chapter includes the following sections:

- How does the data reservoir support the business?
- Process tools and lifecycles
- Defining the information governance program
- Creating a culture that gets value from a data reservoir
- Setting limits on the use of information
- Conclusions

2.1 How does the data reservoir support the business?

After sketching the overall architecture of the data reservoir, Erin begins working with the users of the reservoir across the business. Her goal is twofold. She wants to get to the next level of detail on their use cases, but also wants to get their buy-in on the role and use of the reservoir. She is confident that by starting with specific, narrow initial proof of concepts (POCs) for each of her constituent groups, she can demonstrate value. And she intends that this effort will also shake out the system for the larger use cases.

2.1.1 Extended data warehouse

EightBar Pharmaceuticals (EBP) has had a data warehouse as part of their IT infrastructure for a number of years. At the time it was originally conceived, they hoped that the infrastructure would satisfy the ever-growing information consumption needs of EbP. However, as time has gone on, there have been some notable issues:

- Keeping the warehouse up to date with the changing needs of the business has proven the biggest challenge. The business information, and the needs of the business are changing faster than the IT staff can evolve the data warehouse itself.
- Although the reports created from the warehouses do provide a great deal of value, they could be more valuable if newer types of information, such as the unstructured data sources that are cropping up frequently, could also be incorporated in them.
- The costs associated with the warehouse continue to grow. The payment to their warehouse vendor gets larger every year, as they must pay for more storage and more compute cycles.
- The frequency of access of the data is very different, although the storage costs are the same for all of it. With nowhere else to put some of this data, the warehouse starts to contain increasingly obsolete historical data, which is used less and less in the more contemporary business reports.

A data reservoir can help address many of these issues:

- The data reservoir provides support for a wider variety of storage mechanisms than are found in the traditional data warehousing environment. Many of these allow for storage of information in formats where prior knowledge of the information structure is not needed. This allows for the storage of the information to evolve with the information itself as it changes over time. Using tools such as IBM Big SQL over open source components such as HBase, you can continue to allow business intelligence (BI) report builders to use the SQL that they are familiar with, but over an ever more diverse set of data.
- The overall storage costs of the different repositories in the data reservoir can be dramatically lowered compared to those of the data warehouse alone. This lower price point means that storage decisions get easier, with complementary information going to cheaper repositories, while the warehouse remains dedicated to the core data for the traditional BI reports that it runs.
- Older warehouse data can also be moved to cheaper storage. Using federated interfaces, the existing BI reports can run as is, but against data sets that are now a combination of those that are in the warehouse and those in the reservoir.
- New styles of repositories can enable a broader range of analytical processing. This is particularly true on unstructured or semi-structured data. In particular, text analytics and graph-based analytics become cheap and effective.

After further discussions with Erin, the warehouse team is excited about using the data reservoir as a way to extend their data warehouse. The lower storage costs will finally allow

them to bring some other data sources online that have been a challenge for them in the past. The flexible storage mechanisms will also allow them to keep up with the changing information itself.

2.1.2 Self-service information library

Erin next spends some time with Callie Quartile, the data scientist. Today in EbP, Callie finds much of her time goes to just tracking down the information that she needs. The IT staff are great, but they are very busy. Having to interrupt them all the time to ask where this particular piece of data is, or where a particular report from last year is becomes tiresome to say the least. After she has found a particular piece of information, it is difficult to know when it was last updated or by whom. It can even be hard to figure out who is responsible for the information if she wants more detail about how the information is structured or the meaning of certain elements within it.

Erin explains to Callie how the data reservoir will be associated with a catalog. That catalog will track the assets within the reservoir and be everyone's window into the content there. An important aspect of a catalog (such as the IBM InfoSphere Information Governance Catalog) is the ability to provide a self-service experience for people such as Callie. This includes the ability to use simple phrases or terms to search for associated assets. After an asset is found, it should be clear who the curator and information owner are, when it was added to the catalog, and in many cases what the lineage of the asset is. Many elements of the data sources that Callie finds can be linked back to business terms that define the precise semantics of the values themselves. This allows Callie to use the numerical information with much more confidence.

This can all be done without requiring IT involvement, allowing Callie to find, recognize, and use the information much more quickly.

2.1.3 Shared analytics

In talking with Erin, Callie shares one of her other long standing frustrations: Sharing analytical results with colleagues. She and her other data scientists all end up with file after file of SAS data sets or R results, and no effective way of sharing results. Callie explains that the sharing issue is not just about getting access to the files, but also understanding the sources that contributed to the analytic results themselves. She makes it clear that for the analytics to be reusable, the data scientists need to be able to understand what data contributed to a particular result set.

Erin again explains the role of the reservoir and its catalog. Analytic results can be published back to the reservoir, but when this is done, not just the results themselves should be published. Metadata about the sources and in many cases the analytic operations themselves that were performed to obtain these results should be included. As tooling around the reservoir evolves, this metadata publication along with the result sets will become automatic.

With the analytic result sets published to the reservoir, and details about the analytic operations that were used to create the result set, the collaboration between Callie and her colleagues can change dramatically. They will be able to *shop for* analytics, in the same way that they can use the catalog to s*hop for* data within the reservoir.

2.1.4 Tailored consumption

For a data reservoir to serve the diverse community of users that need information, it must have a certain amount of contextual awareness.

Predefined vocabularies and business object models can be used by the data reservoir catalog to match against the data itself. This allows users such as Harry Hopeful, the sales specialist for EbP, to query the repository using terminology that is familiar to them. The catalog uses that language to locate the assets that best match Harry's search and objectives. In this way, the same physical assets within the reservoir might be arrived at through different paths by Harry or Tessa Tube, the lead researcher. Each of them operates in a different part of the business with a different vocabulary, but in some cases will need to find the same assets.

This tailored navigation is also supported by a corresponding tailored consumption. Harry needs to find certain data sets, but after they are found, he wants to see, manipulate, and use those data sets using the terms that he is familiar with, not whatever database table or column names where assigned by a contract extract, transform, and load (ETL) developer several years ago.

2.1.5 Confident use

Erin and Jules Keeper, the new Chief Data Officer (CDO) at EbP, are also clear that enabling self-service, storing analytics, and allowing easy consumption will all be for naught if the information that is found cannot be used with confidence.

Confident use of information requires many factors:

- The information itself enters the reservoir as part of a well-defined governance program.
- The meaning of the data is captured and easily understood. This starts with the simple metadata about a particular asset such as a database table, such as the column names and their types. This information will often be enhanced with an association to a corresponding business term or element within an industry vocabulary.
- It is easy to identify the sources of the data. Any legal or usage restrictions, if those exist, should be clear.

2.2 Process tools and lifecycles

The data reservoir is more than a set of passive repositories. The data within the data reservoir is being actively managed to ensure that people can find and access the data that they need.

2.2.1 The need for self-service

The initial objectives of a data reservoir are typically aimed at a small subset of the information available within an enterprise. It is unlikely that a *big bang approach*, where all data is fed into the reservoir and then *switched on* will be acceptable to the business stakeholders or feasible to manage from a project perspective. Initially most data reservoirs will demonstrate immediate value to pockets of information users. Over time, as the value for this small subset of users is demonstrated, further phases of the project will increase the volume of data being fed into the reservoir and increase the breadth of the users within the enterprise that will operate within it. Quick wins in those initial phases where tangible business value can be demonstrated is key to ensuring business buy-in from stakeholders across the enterprise.

While the data reservoir implementation is in its infancy, management of the operational aspects of the reservoir can be relatively lightweight. Typically, users of the reservoir will have

a similar level of skills, and importantly, given the limited volume of data, will likely be familiar with the type of information that can be discovered within the reservoir. However, over time as the value proposition is proven, as increasing numbers of users want access to increasing volumes of data, the management of the reservoir requires a more automated approach.

2.2.2 Facets of self-service

Self-service can be considered from various perspectives:

- A user being provided with all the capabilities they need to be able to search and find the data that they need without support
- A data owner being able to ensure that the correct level of governance is enforced on their data
- A system administrator being notified of impending resource allocations being exceeded on a file system
- A fraud investigation officer being provided with the tools they need to investigate suspicious activity within the reservoir

In all of these situations, self-service allows these individuals to perform their tasks without support and effectively.

2.2.3 Enablers of self-service

Initial stages of a data reservoir might not be focused on the self-service requirements of the future state of the reservoir. However, it is important to understand upfront and plan for how self-service can be used to ensure that the operation of the reservoir is as efficient as possible. Two of the key aspects to self-service are workflow and catalog management.

2.2.4 Workflow for self-service

There are various items to consider when defining a workflow for self-service:

- What is a business process

  A business process is a series of actionable steps, tied together into a logical sequence to achieve a business result. The actionable steps can be human-centric or system-centric. Human-centric steps typically require an individual to perform a unit of work, often through a user interface. System-centric steps define automatic steps within the business process such as running a business rule, reading/writing from a system, or sending out notifications. A business process will consist of many human and system steps tied together to provide a definition of how individuals and systems should coordinate their interaction with data.

- Workflow versus Business Process

  Often, the terms *business process* and *workflow* are used interchangeably. However, a business process can be described as the abstract definition of the work that must be done, whereas the workflow is the concrete implementation of the business process. Workflows can be defined in many ways, but when a workflow is defined as an implementation of a business process, specialized software, known as Business Process Management software can be used to model, build, and define complex workflows. IBM Business Process Manager is an example of this specialist software. As the number and variety of people using the data reservoir increases, workflow becomes an enabler of efficient and effective operation of the data reservoir, connecting people and ensuring action is taken to address missing, lost, or incorrect data.

For more information about Workflow, see:

http://www-03.ibm.com/software/products/en/category/BPM-SOFTWARE

- Types of Workflow for self-service

  When considering workflow as an enabler to self-service, workflow can be classified into the areas shown in Table 2-1.

Table 2-1 Forms of workflow

Classification	Description
Data quality management	A collection of workflows within the system that enforce the accuracy of the data and the governance policies associated with the data
Data curation	A collection of workflows that support the curation of the data within the reservoir. These workflows provide mechanisms for Curators to massage the metadata and respond to notifications that inaccurate metadata exists.
Data protection	Typical workflows here manage the granting of access to data and allow the system to notify security and compliance officers of security alerts and suspicious activities, and allow for auditing of system usage.
Data lifecycle management	These workflows allow for the management of information assets within the reservoir or for artifacts of the governance program. Typical examples here include managing the lifecycle of a piece of reference data and managing the lifecycle of changes in policy and rules that need to be applied to the data within the reservoir.
Data movement and orchestration	Workflows classified in this area define the flow of data into and through the reservoir. Workflows in this category define the steps that must be followed for new sources to be shared within the reservoir. They can also define the masking events that should occur before the data is shared with a specific group of users

For more information about each type of workflow, see Chapter 5, “Operating the data reservoir” on page 105.

2.2.5 Catalog management for self-service

The heart of the data reservoir is the catalog, which contains the descriptions of the data in the data reservoir and related detail that defines how this data is being managed.

The catalog is the first access point for most people using the data reservoir. The accuracy and the usefulness of its content will create the first impression of the data reservoir's usefulness and quality. Therefore, its design and the effort to populate it with useful and relevant content is a key success factor for the data reservoir. Human curators maintain some of this content, incorporating feedback from users along with automated processes that are surveying and monitoring the activity in the data reservoir.

2.3 Defining the information governance program

Establishing a governance program for the data reservoir is not just a necessary step, but an essential one for the reservoir effort to succeed. Without a governance program, the data reservoir can easily turn into a place where any and all manner of information is dumped. At first glance, this might seem to be part of the goal of the reservoir: A place to put and share

the information. Indeed that is the goal, but for it to be information, not just raw bits and bytes, it needs governance to ensure that the data that is shared is properly characterized and curated. This section details how to set up a governance program.

2.3.1 Core elements of the governance program

There are three core elements to establishing a governance program that can operate at scale:

- Classification

  Data in the data reservoir is classified using simple classifications schemes that express the business's perspective on the value, sensitivity, confidentiality, and integrity. Processes and infrastructure are also classified according to their capability along with the different types of activity. These classifications become the vocabulary to express how the data reservoir manages data.

- Policies and rules

  For each classification of data, corresponding policies and rules should dictate how that data will be handled in different situations. The rules are described using the classifications in Table 2-1 on page 34. For example, sensitive information must be stored in a secure data repository. The rules support one or more policies that define the wanted end state for the data reservoir. Together the policies and rules make up the requirements of the information governance program.

- Policy implementation and enforcement

  The rules are implemented in the data reservoir processes. Some rules test whether a wanted state is true or not. These are called verification rules. If the wanted state is not true, an exception is raised. The exception can be corrected, logged, and ignored if it is not worth fixing or an exception granted for a period of time. This is a common pattern for data quality because the errors are already present in the data. Enforcement rules, in contrast, are able to force the wanted state. They are more common in the protection of data in the reservoir. If an enforcement rule fails, it is due to an infrastructure failure or a set up error.

These core elements support the information governance principles.

2.3.2 Information governance principles

The information governance principles are the top-level information governance policies. They are typically the first set of definitions that the governance leader, Jules Keeper, establishes because they underpin all other information governance decisions.

The information governance principles define the scope of the information governance program. This is a fairly comprehensive set that covers the responsibilities of individuals in their use and management of information.

These first three principles outline the scope of the information governance program. Define it to cover all information. At a minimum, all information is potentially shareable and must be classified to determine how it is managed. Some classifications indicate that the information is of sufficiently low value and sensitivity that it does not require any special management. As the classification levels rise, the requirements increase.

1. Information is a company asset. It will be managed according to the prescribed governance policies.
2. Information is identified and classified. All information that is stored by the company will be identified and classified according to its sensitivity and content. This classification will determine how the information will be governed.
3. Information is a sharable resource and should be made available for all legitimate needs of the company.

These four principles define the roles that people assume when they work with information:

1. Information is owned. There is an individual responsible for the appropriate management and governance of each information collection.
2. Information users are identified. An individual will be identified and be accountable for each access and change they make to information.
3. Information users are responsible. Individuals are responsible for safeguarding the information that they own, access, and use.
4. Decision makers are responsible for ensuring they use information of appropriate integrity for their work.

These three principles establish the three main disciplines related to information governance: Information lifecycle management, information protection, and information quality:

1. Information is kept as long as it is needed. Information that is no longer needed will be disposed of correctly.
2. Information is protected. Information is secured from unauthorized access and use.
3. Information quality is everyone's responsibility. Information is validated and where necessary it is corrected and made complete.

This principle establishes the need for information architecture to ensure that information is being managed in the most efficient and cost effective manner:

- Information is managed in a cost effective manner. This is achieved through a well-defined information architecture that follows standards and best practices.

This last principle establishes the point that because something is technically possible, and legal, it does not mean that it is appropriate to do. Thought must be given to the consequences and impact on customer trust and related brand image:

- Information and analytics will only be used for approved, ethical purposes.

The information governance principles are then supported by obligations, delegations, and standards.

Obligations

The information governance obligations are policies that have been delegated from other governance focus areas, governance domains, and business teams. They can be thought of as subpolicies under the policies defined by the originating teams.

The following are typical examples:

- Risk Management: The risk management team might handle requirements for classification, data quality, and lineage.
- Intellectual Property Protection: Policies that define the requirements for protecting the organization's intellectual property.
- Export Controls: Defining the restrictions when moving information between countries.
- Financial Reporting: Providing policies related to the collection, management, and retention of financial information.
- Privacy: Defining the requirements for safeguarding the privacy of individuals and organizations.
- Human Resources: Definition of policies around the management and retention of data about employees.
- Marketing: Definition of policies on how customer information should be used.

Delegations

Delegations are policies where the responsibility for implementation has been passed to another governance team. These policies are documented in the catalog to explain where responsibility lies.

- Information Security: The information security team typically takes responsibility for managing access control and setting standards for infrastructure security. The information governance classifications provide a business perspective on where these controls need to be applied.
- IT Governance: Information governance needs a reliable IT infrastructure to operate successfully. IT infrastructure failures can lead to loss or corruption of data. If data is not available when people need it, they tend to take private copies of the data that they manage themselves.

Approaches

The information governance approaches define the approved standards, best practices, and architecture that underpins the information governance program. Enforcement of standards typically reduces cost, avoids errors and speeds up the development of new IT capability. The following are examples of typical standards:

- Standards for data structures and definitions
- Standards for reference data
- Standards for authoritative sources
- Standards for protection of information
- Standards for retention and disposal of information
- Standards for managing information quality

2.3.3 Classification schemes

Classification is at the heart of information governance. It characterizes the type, value, and cost of information or the mechanism that manages it. The design of the classification schemes is key to controlling the cost and effectiveness of the information governance program.

A classification scheme consists of a discrete set of values that are used to describe one facet of an asset's character. When an asset is classified with a particular classification scheme, it is assigned one or more of these classification values.

The information governance policies and rules are then expressed in terms of the classification values to provide explicit guidance on how information of that classification should be managed and used.

Every classification scheme has a definition for what it means if information is unclassified.

Each classification scheme must be memorable and meaningful to individuals creating and using information. This is why classification schemes can be applied to the information resources at multiple levels of granularity to provide the correct level of behavior at an appropriate cost.

The classification schemes shown below are suggestions for an information governance program. These classification schemes are implemented in the InfoSphere Information Governance Catalog. Business classifications, role classifications, resource classifications, and semantic classifications are implemented as terms in the glossary. Technical data classes are implemented as data classes.

There are five main groups of classification schemes:

- Business Classifications: Business classifications characterize information from a business perspective. This captures its value, how it is used, and the impact to the business if it is misused.
- Role Classifications: Role classifications characterize the relationship that an individual has to a particular type of data.
- Resource Classifications: Resource classifications characterize the capability of the IT infrastructure that supports the management of information. A resource's capability is partly due to its innate functions and partly controlled by the way it has been configured.
- Activity Classifications: Activity classifications help to characterize procedures, actions, and automated processes.
- Semantic Classification: Semantic classification identifies the meaning of an information element. The classification scheme is a glossary of concepts from relevant subject areas. These glossaries are industry-specific and are included with industry models. The semantic classifications are defined at two levels:
  - Subject area classification
  - Business term classification

Business Classifications

Business classifications characterize information from a business perspective. This captures its value, how it is used, and the impact to the business if it is misused.

Confidentiality

Confidentiality is used to classify the impact of disclosing information to unauthorized individuals:

- Unclassified: Unclassified information is information that is publicly known. Open data is an example of unclassified data.
- Internal Use: This information is used to drive the everyday functions of the organization. It is widely known within the organization and should not be shared with external parties. However, if it leaks, the impact on the organization is minimal.

- Confidential: This information is only for people with a need to know. It is key information that, if disclosed beyond this group, could have a localized negative impact.
  - Business Confidential: This information provides an organization with a competitive advantage.
  - Partner Confidential: This information is about a partner organization (such as customer or supplier) that has requested that this information be kept in confidence.
  - Personal Information: This information is about an individual. Disclosure of this information could harm or expose the individual.
- Sensitive: This information is only for people with a need to know. This information can have lasting damage if it is disclosed beyond this group.
  - Sensitive Personal: This information is about an individual. Disclosure of this information could cause lasting harm or exposure to the individual.
  - Sensitive Financial: This information relates to the financial health of the business and disclosure could have a lasting impact on the financial health of the organization.
  - Sensitive Operational: This information relates to how a business is operating. For example, this might involve high value proprietary procedures and practices. Disclosure of this information beyond the trusted group of people could expose the organization to fraud, threat, or loss of competitive advantage in the long term.
- Restricted: This type of information is restricted to a small group of trusted people. Inappropriate disclosure could be illegal or seriously damage the organization's business. Every copy of this information must be tracked and accounted for. There are three subtypes:
  - Restricted Financial: Details of the financial health of the organization.
  - Restricted Operational: Details of the operational strategy, approaches, and health of the organization.
  - Trade Secret: Core ideas and intellectual property that underpins the business.

Retention

Retention is used to characterize the length of time this information is likely to be relevant or needed by the organization.

- Unclassified: This type of information is useful but not critical to the organization. It will be retained for the default retention period of two years.
- Temporary: This information is a copy of information that is held elsewhere and will be removed shortly.
- Project lifetime: This information is needed for the lifetime of the associated project. It can be archived after the project completes.
- Team lifetime: This information is need for the lifetime of the associated team. It can be deleted or archived after the team disbands.
- Managed lifetime: Managed lifetime determines how long information should be kept before it is archived. This value can be set by the business because there is no regulatory requirement that controls the retention period. The subtypes defined for this classification are Six Months Retention, One Year Retention, Two Years Retention, Five Years Retention, Ten Years Retention, and Fifty Years Retention.
- Controlled lifetime: This information's retention is controlled by regulation or legal action. The length of time is typically dependent on the type of information and will be captured in the related rules.

- Permanent: This information is likely to be needed for an extended period unless the business that the organization is changed dramatically, at which time the retention of this information should be revisited

Confidence

Confidence indicates the known level of quality of the information, and consequently how much confidence to have in the information. The following are suggested confidence levels:

- Unclassified: New data that has not had any analysis applied to it. Its level of confidence is unknown. This data might turn out to be high quality, but until it has been assessed, it should be treated with caution.
- Obsolete: This information collection is out of date and has been superseded by another information collection. It should only be used to investigate past decisions that were made using this information.
- Archived: This information has been archived and is available for historical analysis.
- Original: Original information comes from operational applications. It has not been enhanced and so contains the information values that were used by the teams during normal operations. Often this information has a localized perspective and can be narrow in perspective.
- Authoritative: Authoritative information is the best knowledge that the organization has on the subject area. This information is continually managed and improved. This information has the highest level of confidence.

Severity

Severity is used to classify the impact of a particular failure of the information technology infrastructure or issue with the data values stored in an information collection. The following are example severity levels:

- Severity 1: Critical Situation/System Down/Information Unusable. Business critical software component is inoperable or a critical interface has failed. This indicates that you are unable to use the program, resulting in a critical impact on operations. This condition requires an immediate solution.
- Severity 2: Severe impact. A software component is severely restricted in its use, causing significant business impact. This indicates that the program is usable, but is severely limited.
- Severity 3: Moderate impact. A noncritical software component is malfunctioning, causing moderate business impact. This indicates that the program is usable with less significant features.
- Unclassified: Minimal impact. A noncritical software component is malfunctioning, causing minimal impact, or a nontechnical request is made.

Business impact

The business impact classification defines how critical an information collection is to the organization ability to do business. This classification is typically associated with business continuity and disaster recovery planning. However, it is also an indication of the value of the information in the information collection.

- Unclassified: Information that is used by an individual, so only that individual is impacted.
- Marginal: Occasional or background, non-essential work is impacted.
- Important: Parts of the business are unable to function properly.
- Critical: The business is not able to function until capability is restored.
- Catastrophic: The business is lost and restoration is unlikely.

Role Classifications

Role classifications are used to control the types of data that an individual can see. They are typically relative to the data itself. Some role classifications are to a subject area or entity/attribute type, whereas others are related to instances, particularly when it comes to personal data. As such, some role classifications can be calculated dynamically rather than manually assigned. Here are some examples of more static user roles:

- Information Owner: A person who is accountable for the correct classification and management of the information within a system or store.
- Information Curator: A person who is responsible for creating, maintaining, and correcting any errors in the description of the information store in the governance catalog.
- Information Steward: A person who is responsible for correcting any errors in the actual information in the information store.

Examples of classifications that are more instance-based might be labels that show the relationship between a user of data and the data subject:

- Close Neighbor
- Relative
- Colleague
- Manager
- Spouse

These types of classification are specific to the instance and need to be dynamically calculated.

Resource classifications

Resource classifications characterize the capability of the IT infrastructure that supports the management of information. A resource's capability is partly due to its innate functions and partly controlled by the way it has been configured.

Governance zone

The governance zone provides a course-grained grouping of information systems and information collections for a particular type of usage. The governance zones are overlapping so an information system or information collection can be in multiple zones.

These zones are commonly found in a data reservoir:

- Traditional IT zones
  - Landing area zone

    The landing area zone contains raw data just received through the Data Ingestion component from system of record applications and other sources. This data has had minimal verification and reformatting performed on it. Processes inside the data reservoir called *data refineries* take this data and process it to improve its quality, simplify its structure, add new insight, and link related information together.
  - Integrated warehouse and marts zone

    The integrated warehouse and marts zone contains consolidated and summarized historical information that is managed for reporting and analytics.
  - Shared operational information zone

    The shared operational information zone has information sources that contain consolidated operational information that is being shared by multiple systems. This zone includes the master data hubs, content hubs, reference data hubs, and activity data hubs. They support most of the service interfaces of the data reservoir.

- Audit data zone

  This is where log information about the usage of data in the data reservoir is kept. Analytics models run in this zone to detect suspicious activity. It is also used by security experts for investigating suspicious activity and for auditing of the data reservoir operations.

- Archive data zone

  Archive data is no longer needed for production, but has potential value for investigations, audit, and understanding historical trends.

- Self-service zones
  - Descriptive data zone

    The descriptive data zone contains the metadata that describes and drives the management of the data in the data reservoir. This zone starts out as a simple metadata catalog, but as the business gains self-service and governance capability, the descriptive data zone grows in sophistication.

  - Information delivery zone

    The information delivery zone contains information that has been prepared for use by the lines of business. Typically this zone contains a simplified view of information that can be easily understood and used by spreadsheets and visualization tools. Business users access this zone through the View-based Interaction subsystem.

  - Deposited data zone

    The deposited data zone is an area where the users of the data reservoir can store their own files either for safety or for sharing. The inclusion of deposited data zone helps to reduce the data leakage from the data reservoir because business and analytics teams are not required to set up their own local files store.

  - Test data zone

    The test data zone contains obfuscated data for testing. This is nested in the deep data zone, and is used by developers and analysts when testing new function and analytical models.

- Analytic zones
  - Discovery Zone

    The discovery zone contains data that is potentially useful for exploring for new analytics. Experienced analysts from the line of business typically use this zone for these purposes:

    - Browse catalog to locate the data they want to work with
    - Understand the characteristics of the data from the catalog description
    - Populate a sandbox with interesting data (this sandbox is typically in the exploration zone.)

  - Exploration zone

    The exploration zone contains the data that the analysts and data scientists work with to analyze a situation or create analytics. Users of this zone reformat and summarize the data to understand how a process works, locate unusual values (outliers), and identify interesting patterns of data for use with a new analytical algorithm.

  - Analytics production zone

    The analytics production zone contains detailed information that is used by production analytics to create new insight and summaries for the business. This data is kept for some time after the analytics processing is complete to enable detailed investigation of

the original facts if the analytics processing discovers unexpected values. There is a large overlap in the data elements found in the analytics production zone and the exploration zone because the production analytics models are typically developed in the exploration zone. Repositories in this zone will have production service level agreements (SLAs) applied to them, particularly as the organization becomes more data and analytics driven.

– Derived insight zone

  The derived insight zone identifies data that has been created as a result of production analytics. This data is unique to the data reservoir and might need additional procedures for backup and archive.

Transport Security Classification

The transport security classification describes the ability of an information provisioning technology to protect information from interception and tampering while it is being provisioned between systems. It typically uses these classifications:

- Unclassified: Unsecured
- Secured: Access to the technology is controlled so that only approved processes can access it.
- Encrypted: Information is encrypted so even if it is accessed, no one can steal or alter the values.

Information store location

Location classifies where an information store is located. The definitions below are examples from an information governance scheme centered on a data reservoir. The classification helps to identify which sources are part of the reservoir and which are connected. The catalog includes sources that are outside of the reservoir to enable lineage to be captured. Each information store can only have one location:

- Unclassified

  A potential source of information for the data reservoir. It is present in the information governance catalog to advertise that it exists. However, no attempt has yet been made to integrate it with the data reservoir.

- Internal system

  A system that is owned by the organization, but sits outside of a data reservoir that is exchanging data with the data reservoir repositories.

- Third-party source

  A system that is operated by a third party.

- Adjacent reservoir

  An information source that is managed by a different data reservoir. This adjacent reservoir is either sending or receiving information.

- Data reservoir repository

  A core repository of the data reservoir.

- Data reservoir service store

  A sandbox or data mart that contains information for the business to use for analytics.

- Data refinery store

  A private information store that is used internally in the data reservoir to transform raw data into useful information.

Activity classifications

Activity classifications help to characterize procedures, actions, and automated processes.

Business process type

The information governance program must be seen to support the business strategy directly and be flexible to adapt to changing business needs. For example, the information governance program should cover five types of business process:

- Communication: Ensuring each individual employee is aware of his/her roles and responsibilities related to their use and management of information.
- Compliance: Ensuring requirements are met and incidents of non-compliance are reported.
- Exemption: Handling special cases in an effective and timely manner.
- Feedback: Measuring the effectiveness of the program and handling suggestions for improvement and complaints.
- Vitality: Evolving the program to support new requirements and reach deeper into the organization.

These business processes can be manual procedures, tasks, or automated processes, and together they keep the governance program grounded in the needs of the organization.

Control point decision classification

Control points are decisions made in business process that determine the response to a governance requirement. The governance program should provide descriptions on how to proceed after each possible choice is made. It has these classifications:

- Correct to comply: The data or processing environment will be changed to bring it into compliance.
- Request exemption: The current situation will not be changed and an exemption requested. Typically, exemptions are for a set time period. Many permanent exemption requests suggest that the governance program is not meeting the needs of the business.
- Ignore: This is used when what is causing the situation has low business impact and a conscious decision is made to ignore it, at least for the short term. For example, there might be quality errors detected in the contact details of the main customer database. A control point decision can be made not to correct contact details for customers that have been inactive for more than two years.
- Request clarification: Governance requirements are unclear and more information is needed.

Enforcement point classification

The enforcement point classification is used to characterize the behavior of an automated rule or component implementation. Typically it is one of these components:

- Verification rules: Testing that a particular governance requirement has been met. For example, verifying that an attribute contains valid values. If the rule fails, an exception is raised.
- Enforcement rules: Ensuring that a particular governance requirement is met. For example, masking sensitive data as it is saved in an unsecured repository.

Examples of enforcement point classifications include copy, delete, mask, validate values, validate completeness, derive values, enrich, standardize, archive, back up, link, merge, collapse, and raise exception.

Semantic classification

Semantic classification identifies the meaning of an information element. The classification scheme is a glossary of concepts from relevant subject areas. These glossaries are industry-specific and are included with IBM industry models.

The semantic classifications are defined at two levels:

- Subject area classification

  Provides a course-grained classification of the source and use of information. This classification is typically used to provide a context or scope to the business classifications. For example, to define retention periods for Controlled Longevity information collections, or to specify the subject area that an authoritative information collection supports.

- Business term classification

  This uses the business terms that are defined in the business glossary to provide a fine-grained semantic classification for information elements. This helps people find the data that they need and helps prevent integration errors as information is copied and consolidated between information collections.

Data classifications

Business classifications typically define the type of governance that is required. Data classifications characterize the way that data is typed and supported technically, which is key information when automating the actions associated with information governance.

Data classes

Data classes define the fine-grained logical data types. They are used to determine which implementation of a governance action to run. The following are examples of data classes:

- Personal information such as first and surname, gender, age, Passport Number, Personal Identification Number, personal income, date of birth, country-related identification number, and drivers license number
- Company name
- Location information such as address, city, postal code, province/state, and country
- Financial information such as Credit Card Number, Credit Card Verification Number, and Account number
- Contact information such as phone number, email address, Internet Protocol Address, Uniform Resource Locator, and computer host name

The concept of a data class is common between IBM InfoSphere and IBM InfoSphere Optim™ products.

2.3.4 Governance Rules

Information governance rules are defined to explain how the information governance policies will be implemented. Typically they apply to the activity of a system or team, but they can also apply to specific types of data. Either way, they are organized according to the business owners who are responsible for its implementation.

There is a governance rule for defined for each situation where an information policy is relevant. The rules are then expressed in terms of the business classifications. For example, the owner of a collaboration space might define a set of governance rules that define these classifications:

- Which classifications of data can be posted in the collaboration space
- Those classifications of data that are allowed, what are the restrictions on use and access that must be observed.

2.3.5 Business terminology glossary

For reservoir catalog's such as the IBM Information Governance Catalog, the classification process includes assignment of assets to business terms. *Business terms* are part of a *business glossary*, where terms are organized into folder-like structures called *business categories*. A business term captures the vocabulary used by the business and includes a textual description that defines the term itself. By associating assets in the catalog with one or more business terms, a powerful semantic is created. That association says that this information asset is classified by this particular business term. This classification can allow for searching by business terms so the users can find all the assets that are associated with that business concept. Navigation can also be facilitated in the other direction with a classification in place. After a particular information asset is found, semantically related assets can be found by seeing what other assets have also been classified by business terms associated with this asset.

For some technical users, seeing the actual term asset assignments can be interesting. For most of the others, this information will be used to drive more meaningful search and to provide *suggested assets* that are related to the one selected, all in a fully automated manner.

Creating a full *Business Glossary* from scratch can be difficult, but fortunately there are usually starting points to ease the process. For Jules Keeper and his team, they started with industry standard glossaries that facilitate interaction between different pharmaceutical companies and government regulators. Other industry's predefined vertical glossaries (and more) can be obtained from sources such the IBM Industry Models.

For more information about IBM Industry Models, see:

`http://www.ibm.com/software/data/industry-models/`

2.3.6 EbP starts its governance program

Jules Keeper sees the governance program as a key part of his role as Chief Data Officer (CDO) for EbP. As an industry veteran, he has been responsible for establishing governance programs at several previous companies. He has learned that it is vital to start with a fairly narrow vertical slice through the domain to be governed, establish early success there, and build on it.

With the information governance principles in place, Jules creates a core governance team. The primary initial goal of the team is to establish the core set of policies, rules, and procedures that will govern the information in the reservoir. As the scope of the reservoir grows and changes over time, so will the composition of the governance team itself.

In this case, the team begins with a vertical domain slice that deals with the US Federal Food and Drug Administration (FDA). Within this domain, they further focus on the handling of patient information for clinical trails. Working from a set of existing FDA vocabularies and other internal sources, the team assembles a core Business Glossary. These terms give

explicit meaning to what would seem to be self-evident terms such as *Patient*. The clear definition removes any ambiguity about when a person moves from a candidate for a particular trail, to a patient in that trail, and thus can be properly considered for inclusion in reports.

As part of this process the team also identifies personally identifiable information (PII). Compliance is required with US and other country statutes around personal privacy and the handling personally identifiable information. PII is a key business responsibility for Jules Keeper and is essential for the success of their reservoir itself. US Social Security Numbers, Patient Identifiers, and other data elements are given formal definition in terms of their structure so that they can be formally tagged by automated means when information with these fields is added to the reservoir. Rules for who can see the data, and what transformations must be applied to the data (masked, removed, blocked) are all defined. These rules can include which zones within the reservoir data with unmasked PII is allowed and which zones will require masking.

It might seem like all PII should be masked by default on entry to the reservoir, but there can be valid business reasons for adding it unmasked. Providing strong governance around those fields is the next step, so that only authorized individuals can see the unmasked data. One of those use cases can be for fraud investigations. A risk officer for instance will often need access to all the data to facilitate an investigation.

Another area that Jules directs the core governance team to look into is information lineage. Information lineage shows the primary upstream sources and the downstream users for pieces of information. This perspective can be invaluable in determining whether a particular data set is the correct one for use, if it is up to date, and so on. However, lineage needs to be governed like anything else. The team establishes rules around which assets must be published to the reservoir with lineage, and which ones do not. Further they also define policies that validate lineage on a periodic basis for certain key data sources. Because lineage is often built and reported through automated means, it can be subject to bugs and errors similar to any other software process. It is important that a selected subset of lineage flows is regularly audited for correctness. Given the large number of assets in the reservoir and the huge number of flows that are defined, it is impractical to check them all. Checking the key flows can help gain confidence that the others are proper too.

2.3.7 Automating Curation Tasks

Much of what is being described can be classified as various types of curation, which is giving and maintaining meaning for a collection of information assets. For the reservoir to scale, it is vital to use both automated and social curation mechanisms.

The following automated curation tasks have already been described:

- Assignment of terms to assets

  These can be done on a contingent basis, automatically, and then confirmed (in batch) by others, or can be fully automated. The effectiveness of automated assignment can vary with asset types, but expect this to be an area where research and technology advancements rapidly increases effectiveness.

- Data type determination and assignment to data classes

  This can discover US Social security codes in data sets, credit card numbers, and so on. These data classes can be used to determine verification rules to run, and identify potential cases where data has been misclassified.

- Assignment to other ontologies

  It turns out the business glossary is just one means of classifying the assets in the repository. Other classifications schemes are not tied directly to the business language used, but can instead reflect reusable logic. For example, a geographic classification of an asset would understand that in the US, an address exists within a town, which is a type of political jurisdiction. A town is in a county, which is another political jurisdiction, which is in a state, and so on. This geographical classification can be used by data scientists for an example query such as finding “all clinical trials and patients conducted in Orange County California in 2012”.

Social curation uses the knowledge of the users of the data reservoir. Although automated techniques should do the bulk of the work, there is no substitute for a subject matter export (SME) to tag, comment, rate, or classify a particular asset or set of assets. These users and others can also rate data sources (zero to five stars, for example). Different users can have different weights assigned to their ratings and classification depending on the zone and domain of area of the reservoir that they are operating in. All of this allows for more assets to be discovered and suggested in a self-service manner.

2.3.8 Policies for administering the reservoir

Jules and the governance team do not stop just at policies around data privacy. Other policies are defined to make sure that the reservoir itself is maintained in a manner that ensures that relevant data is easy to find.

To do this, they also define policies for data retention. Certain data sets are added to the reservoir from well-defined sources on a scheduled basis. How long should these be retained? Where do they go when they are retired? If the reservoir catalog supports it, policies can also be defined that handle unused data, archiving it as needed.

2.4 Creating a culture that gets value from a data reservoir

The data reservoir is part of a profound shift of the culture at EbP to data-centric decision making. With the proper information from the reservoir and easy to use analytic tools, the days of acting on *gut instinct* are replaced by data rich analysis based on trusted data.

2.4.1 Reservoir as a vital daily tool

To make this shift, both Erin and Jules realize that the reservoir must be a vital daily tool for knowledge workers such as Callie and Tessa. Callie, Tessa, and other knowledge workers have been involved from day one on the project and played a key voice in selecting the vendor for the catalog and collaboration tools. The knowledge workers considered these key points in their evaluation:

- Is there a notion of project or investigation that allows virtual notebooks to be created that target a particular analytic task or research item?
- Can their favorite analytic tools easily load data from the reservoir and produce results? Is lineage automatically generated for these analytic flows?
- Are quality and trust scores computed automatically for data sets within the reservoir?
- Can the knowledge workers easily flag bad data, rate good data highly, add tags/labels, and otherwise curate the data that they find there?
- Is it easy to find and use lineage for information assets?

The goal for knowledge workers like Callie or Tessa is that the data reservoir should become similar to a trusted assistant or colleague. It should always be ready to help find the data that they need, help determine whether the data is good or bad, record analytic results, and share the results and overall investigation in a form that others can find, share, and reuse.

2.4.2 Reassuring information suppliers

So far this section has focused on the consumption side of the reservoir in making it a vital component, but the supply side is also key. Information suppliers, be they other business units or other entities within the company, need to be assured that the data that they supply to the reservoir is both managed properly and eventually used. The governance program that Jules has developed helps with the concerns about the way information is managed. Individual information suppliers can have a say in the policies that are defined for their data sources. Usage statistics and other metrics for the reservoir itself can also show which data sets are being accessed and with what frequency.

2.5 Setting limits on the use of information

This chapter has touched briefly on some of the ethical and security aspects of data within the reservoir, but these are such important topics that they are worth a more detailed look.

2.5.1 Controlling information access

The goal of the reservoir is to get high quality, trusted data to the correct people in a self-service manner. This allows them to do their jobs and creates value for the company. However, the wide user community of the reservoir itself means that there is a requirement to ensure that data is both used appropriately and only by those that have authorization to do so.

The key to information security is the appropriate classification of data, either by the owner or by a trusted curator upon entry to the data reservoir. This defines the type of protection that is appropriate for the data. The classification can be applied at the information collection or attribute level depending on the variety of data in the data source.

Access is authorized by the business owner of the data using a well-defined business process that maintains a record of the people with access rights and then makes requests to the IT team to update the access control list.

2.5.2 Auditing and fraud prevention

Applications tend to limit the amount of data an individual can see and the actions they can take. The purpose of the data reservoir is to remove these restrictions by freeing the data from boundaries of the original application. However, with this freedom comes the responsibility to use data responsibly.

After information access controls are in place, it is vital to monitor patterns of usage and raise flags/alarms when access falls outside the boundaries of acceptable use. The combination of actionable policies to prevent data from being exposed to inappropriate individuals and the monitoring to look for access that might fall outside of prescribed policies can help ensure that governance policies and procedures are being adhered to by the data reservoir community.

One other area of information use in the reservoir concerns the terms and conditions around purchased data sets, and other data sets that might have precise legal restrictions on their

use. Proper information governance policies and rules must be attached to these assets so that the terms and conditions can be clearly seen, honored by users, and in some cases enforced by the reservoir itself.

A strong set of policies and procedures for information access and audit contributes to the confidence in the reservoir itself. Information suppliers can be confident that their information is being used within the terms and conditions that they have specified, and by people that are authorized to see and use the data. Information users can be confident that the information they find in the reservoir can be used without further restrictions when used within the clearly stated and easily identifiable policies and rules. Both are key components of a self-service data reservoir.

2.5.3 Ethical use

The ability to process large amounts of data from multiple sources is moving faster than the governmental and other regulatory bodies can define what is the proper use of information is. This can result in a gray area between what is currently clearly legal, and what is newly possible in terms of insights gained from perhaps what until now have been disparate data sources.

This situation leaves companies and the individuals within those companies to use their best judgment in the ethical treatment of this data. IBM has recently published a paper, *Ethics for big data and analytics*[1], discussing some of these issues. The paper highlights some of the factors to consider when deciding what is ethical use of information:

- Context

  For what purpose was the data originally surrendered? For what purpose is the data now being used? How far removed from the original context is its new use? Is this appropriate?

- Consent and choice

  What are the choices given to an affected party? Do they know that they are making a choice? Do they really understand what they are agreeing to? Do they really have an opportunity to decline? What alternatives are offered?

- Reasonable

  Is the depth and breadth of the data used and the relationships derived reasonable for the application it is used for?

- Substantiated

  Are the sources of data used appropriate, authoritative, complete, and timely for the application?

- Ownership

  Who owns the resulting insight? What are their responsibilities towards it in terms of its protection and the obligation to act?

- Fair

  How equitable are the results of the application to all parties? Is everyone properly compensated?

- Considered

  What are the consequences of the data collection and analysis?

[1] Ethics for big data and analytics, available at http://www.ibmbigdatahub.com/whitepaper/ethics-big-data-and-analytics

- Access

  What access to data is given to the data subject?

- Accountable

  How are mistakes and unintended consequences detected and repaired? Can the interested parties check the results that affect them

Each of these aspects can be used to evaluate a specific use of data in the data reservoir and to design appropriate measures into the solution to safeguard both the privacy of the data subjects and the reputation of the organization.

An important point to remember is that there is no fixed definition of what is the ethical use of information. This varies between people of different backgrounds and age groups. Collectively, our perception of what is expected and what seems creepy is changing all of the time. Therefore, it is important for the organization to have a position they are comfortable with, that they can be transparent about their use of data with any data subjects and stakeholders, and there is an appropriate process where people can raise concerns and have a situation redressed that they consider is an unethical use of data about them.

2.5.4 Crossing national and jurisdictional boundaries

If your company maintains data that crosses national boundaries, you must pay particular attention to a growing and diverse set of laws that governs the allowable use of data for individuals within a particular country and the movement of data across national boundaries.

As a starting point to get basic information about data protection laws worldwide, see the *Data Protection Handbook* at:

http://www.dlapiperdataprotection.com/#handbook/world-map-section/c1_CN/c2_DK

If you are operating across national or other jurisdictional boundaries, you must consider regional requirements for data handling and establish polices and rules that respect those requirements. This can result in particular reservoir repositories being cited in particular countries with limits on who can access the data and the activities this data can be used in.

Where regulations differ wildly from country to country, it might be easier to have a data reservoir in each country with links between them to share summarized data where needed.

2.6 Conclusions

The data reservoir is an ecosystem that enables an organization to share information both effectively and safely. The definition of how this ecosystem works is embedded in the information governance program.

This governance program must be comprehensive to ensure that the ecosystem operates effective. It must be flexible because there is a wide variety of data stored and it is not appropriate to implement a single set of standards and processes for all types of data.

3

Logical Architecture

Chapter 1, “Introduction to big data and analytics” on page 1 explained how the data reservoir can help companies meet new expectations for personalized service and in the fictitious company example saw how it can to drive radical improvements in patient care and manufacturing. Chapter 2, “Defining the data reservoir ecosystem” on page 29 described how the business might approach this, looking at culture, ethics, and overall workflow.

This chapter explains how the data reservoir is put together in more detail. It details the different functional areas of the reservoir, their roles, and responsibilities. The chapter also covers some of the key interactions with the reservoir and some considerations you might need to make when creating a data reservoir.

This chapter includes the following sections:

- The data reservoir from outside
- Overview of the data reservoir details
- Data reservoir repositories
- Information integration and governance
- Component interactions
- Summary

3.1 The data reservoir from outside

Existing processes and systems in the enterprise are already in place supporting the daily business of the organization, whether that is sales, customer service, manufacturing, support, or research and development (R&D). The business of the organization is not going to stop just because of a new data reservoir.

However, with the data reservoir in place, those processes and systems can take advantage of the easier access to data and analytics that the data reservoir offers.

Some core systems primarily provide data to the reservoir, and therefore do not get as much value in return, but the impact on them will be small. In fact, this is a key consideration. Those systems might have stringent 24 x 7 requirements, while the initial deployment of the reservoir might not require the same level of availability.

Other systems and users who are interested in analytics, reporting, and enabling new applications will be able to reap the benefits of this data that is easy to collect, but hard to manage. These systems and users can also be confident that data is not being misused. In addition, the business might decouple this analytical use from the operational systems and ensure that each remains fit for purpose.

This section describes the different kinds of external systems and users that interact with the reservoir. This is not an exhaustive list, but it is intended to provide some ideas of what might be useful to connect, and some of the considerations for doing so. These systems lie outside the control of the reservoir itself. They might be other reservoirs as well. There might also be times when you ask “should this be inside the reservoir or outside?” The answer is that it depends. Existing high quality, governed sources might be best left as is. Organization boundaries might also mean leaving data outside because the intent of the reservoir is to apply consistent governance and management throughout.

The following sections cover the systems that are shown in Figure 3-1 on page 55, starting with the lower left and working clockwise around the data reservoir.

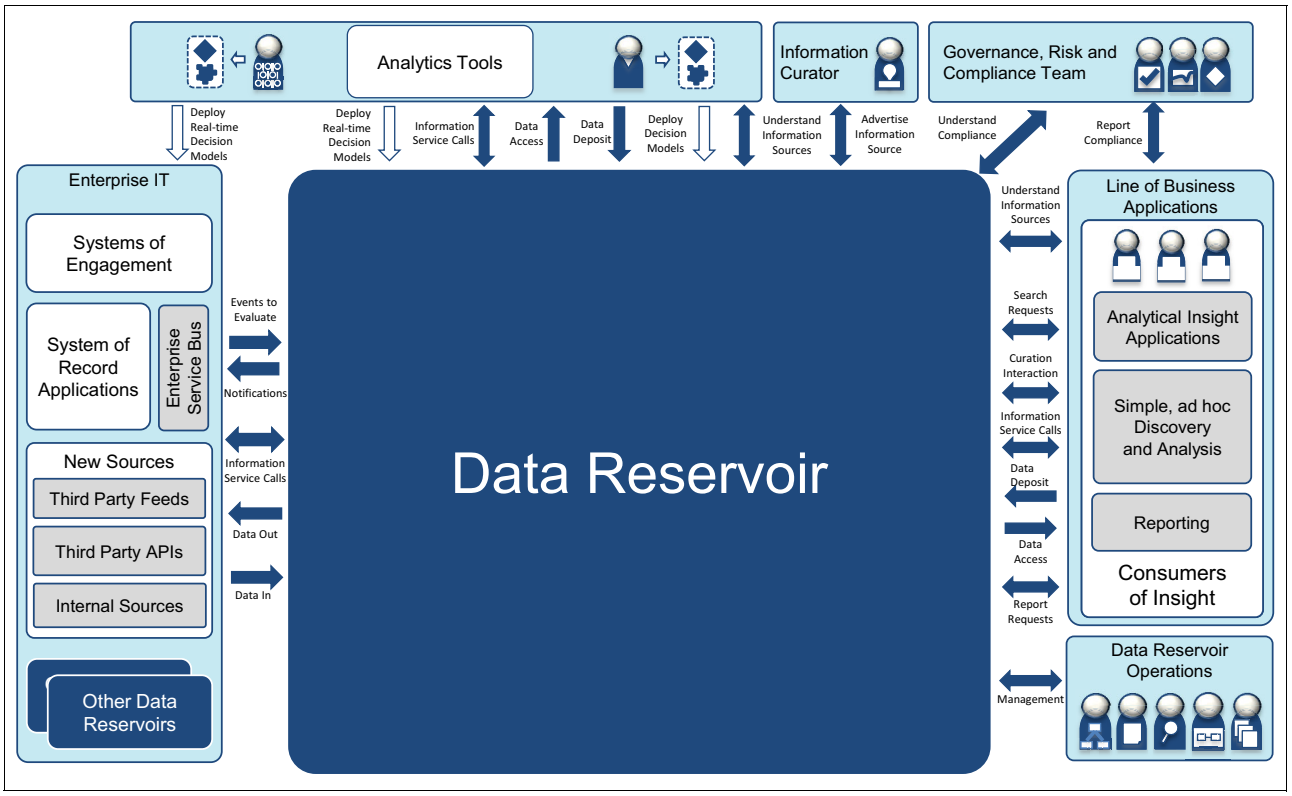


Figure 3-1 Data reservoir's system context

3.1.1 Other data reservoirs

You might be thinking about setting up your first data reservoir, but it is also worth considering how many you really need. A data reservoir has a governance program that determines how data is managed. It might be that you have distinct operating units that it does not make sense to bring together. Another situation might be that you have operations in different countries where rules around sharing data mean that you want to maintain easy access to data within the country, but have much tighter control over what can move between them to satisfy the applicable laws in each country.

A data reservoir can interact with other data reservoirs in a hierarchical, or peer-to-peer manner depending on the requirements. If for example multiple data reservoirs are needed on a country-by-country basis, you might decide to have a global reservoir that is fed a subset of data from the in-country data reservoirs.

Multiple data reservoirs can also provide a high degree of decoupling. Different rules, policies, and even infrastructure can be used by each country, even though for simplicity you aim to minimize those differences. Typically, a central center of excellence could document a suitable starting template for a reservoir, which could then be customized for each country.

Data reservoirs can supply or use data from each other.

3.1.2 Information sources

An organization such as EbP will have various existing systems responsible for running the business. These systems are already in place and critical to daily operations. As you start building a data reservoir, you will look to these systems to provide data that can then be analyzed.

In the past, the owners of these systems might have had frequent requests to provide regular extracts of data such as a database dump or a spreadsheet. This action could be done in an ad hoc manner or it could have been automated daily. Some other teams were picking up this data for their own use. However, from the point of view of the owner, there is no mechanism to control what happens to the data after it leaves the original system. The owner does not have any knowledge of these factors:

- Who is using their data?
- How many copies are there?
- What business decisions are being made based on their data?

The reservoir aims to provide governance in how data extracted from original systems is later used. The role of the information source is purely to make data available to the reservoir and to let the IT team running those systems get on with what they do best rather than having to respond to random requests for data.

Over time it is likely these sources could also become users of some of the data in the reservoir, adding nuggets of insight to those existing systems. This can be done using some summarized analytical data calculated in the reservoir.

Mobile and other channels

Customers and employees interact with the enterprise's IT systems in many different ways. A few years ago, this might have primarily been a web-based interaction, but these days an increasing proportion of that interaction is coming from mobile devices. These interactions are known as Systems of Engagement because these interactions have evolved beyond simple transactional systems to being much more user-centric and personalized. As the volume of *Internet of Things* devices takes off, users will also interact more with a myriad of devices such as within their vehicle, their heating systems, and lighting systems. The variety and volume of the data is increasing rapidly. Some data can arrive in batches, but much more can arrive as a stream of data.

Understanding this data, and deriving business value from it is one of the challenges organizations face. Therefore, all of these systems make ideal sources for the data reservoir, which needs to be able to ingest this information rapidly and reliably and ensure that it is made available for investigation and analytical use.

To engage with their users, these systems will also use data from the reservoir. For example, a mobile application might require access to a customer's segmentation or churn propensity. This might be termed summarized analytical data. Given the real-time nature of the interaction, a cache is typically used to minimize any delay to the application and maintain a high level of availability.

Systems of Record

Systems of Record manage the data around the core business of the enterprise. These systems have often been developed as stand-alone applications and can have been in existence for a long time. It is likely they are not subject to a high amount of change or redesign. However, they remain a critical source of data for the reservoir because they provide critical business information such as a list of customers, product inventory, or bank transactions.

Systems of record manage the enterprise's core data, and might include order systems, master data management, and product inventory data.

These systems are carefully managed and in general will be providers of data to the reservoir. Any use of the output from analytical processes in the reservoir by these systems is likely to be through published feeds, which are then fed back into those systems.

New sources

More data is being captured by organizations. Examples of further sources that can be explored include social media and open data.

These sources would not necessarily be seen as useful in their initial form, but by using analytics and merging with other information, can become valuable.

Many weather data sets are now made available. You might feed this information into the data reservoir to make it easy to compare user purchasing decisions with the weather at the time. Or you could look at forecasts to see not how the weather has affected a user's decision, but how what weather they think is coming next will.

Capturing social media feeds, whether by following selective accounts or sampling everything you can get your hands on (though social media) can prove useful in analyzing customer sentiment. Was that new product reviewed favorably? Does it make owners happy?

A flurry of planning applications captured from a town council could be used as an indicator of a stronger economy in a particular area. Could this be a place worth targeting for product sales?

Often it is only after experimenting with this type of data that its value can be understood.

3.1.3 Analytics Tools

Data scientists usually have a solid foundation in computer science, mathematics, and statistics, rather than substantial business knowledge. Their role is to explore different kinds of data to look at patterns so that they can understand the meaning of relationships between sets of data. They often have a thirst for new information sources to develop new ideas on how to gain value from combining that data with other sources. The tools used by the data scientists include capabilities for preparing and transforming data for analysis. They typically work with data in its original (raw) format.

Data can be accessed through the creation of a sandbox copy of the data for the data scientist to use because the data scientist will subset and transform the data as part of their work. Their activity should not affect the shared version of the data in the data reservoir. The sandbox will be accessible as an area in a file system such as Hadoop Distributed File System (HDFS) or accessible by using common mechanisms such as Java Database Connectivity (JDBC) or Open Database Connectivity (ODBC).

As the data scientists explore samples of this data, they will develop analytical models and try them out against test data. Those models can then be deployed to refinery services within the data reservoir or to operational systems outside the data reservoir.

3.1.4 Information curator

Data that is governed needs to have an owner. This person is accountable for the correct management and use of the data. When an information owner agrees to their data being copied into the reservoir, they need to ensure that the management procedures in the data

reservoir will provide proper protection for this data. The management procedures are driven by the classifications assigned to the data in the data reservoirs catalog.

The creation of the description of the data in the data reservoir's catalog is called *information curation*. The information owner is responsible for the accuracy of this definition. However, for enterprise IT systems, the information owner is often a senior person who delegates to an information curator. The information curator understands what that data is and can describe it in an entry in the catalog. Ultimately the success of the reservoir depends on having data that is useful and findable.

The information curator manages the catalog description over time, possibly adding more details as they become known.

3.1.5 Governance, risk, and compliance team

A data reservoir has governance so that data can be relied on. This data needs to be managed and validated.

The information governance, risk, and compliance team will use the catalog to define the governance program and provide a reporting environment to demonstrate compliance to this program.

3.1.6 Line-of-business applications

Individual lines-of-business have developed, and will continue to develop, a myriad of applications. These are not managed by IT and often rely on isolated sets of data for a focused need.

As the data reservoir evolves, these applications can start to have access to a far greater variety of enterprise information than was previously available. The developers of these applications can browse through the catalog of available information sources and have confidence in what the data is, where it came from, and that they can access it in a suitable form.

These applications can access information in the reservoir through service calls and can go through a sandpit environment to isolate their usage from ongoing updates in the reservoir. In some cases, these sandpits can hold raw data sets as seen with the data scientist, but with two distinct differences:

- The line-of-business user is likely to be interested in an entire set of data rather than a sample.
- The line-of-business user might prefer to find data in a more simple or flattened format approach suitable for their need. At one extreme, this might mean something as simple as a flat CSV file for import into a spreadsheet. The View-based Interaction subsystem described in 3.4.4, “View-based interaction” is responsible for providing this simpler format.

3.1.7 Data reservoir operations

The daily operation of the data reservoir is supported by people with these roles:

- Monitor the operation of the reservoir, that is, is it working efficiently?
- Manage metadata in the repository that means ensuring that appropriate classifications schemes are in place

- Deploy changes to the infrastructure components
- Other tasks similar to any IT system

The operations team set up and manage the infrastructure that forms the data reservoir. This team includes the enterprise architect, information steward, data quality analyst, integration developer, and infrastructure operator.

3.2 Overview of the data reservoir details

Previous sections have detailed the users and systems that interact with the data reservoir. This section describes how the reservoir itself is structured from an architectural standpoint and the capabilities it offers.

The data reservoir includes three major parts (Figure 3-2):

- The data reservoir repositories where the data is stored.
- The information management and governance fabric that provides the infrastructure and middleware to operate the data reservoir.
- The data reservoir services provide the mechanisms where data is exchanged between the data reservoir and the systems and people that use it.

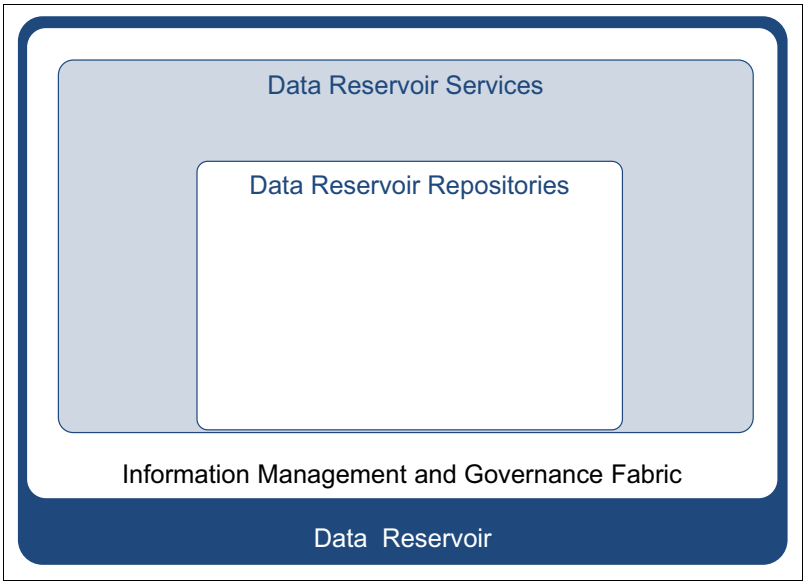


Figure 3-2 Top-level structure of the data reservoir

Figure 3-3 shows the major subsystems inside the data reservoir services and the information management and governance fabric that surrounds the data reservoir repositories.

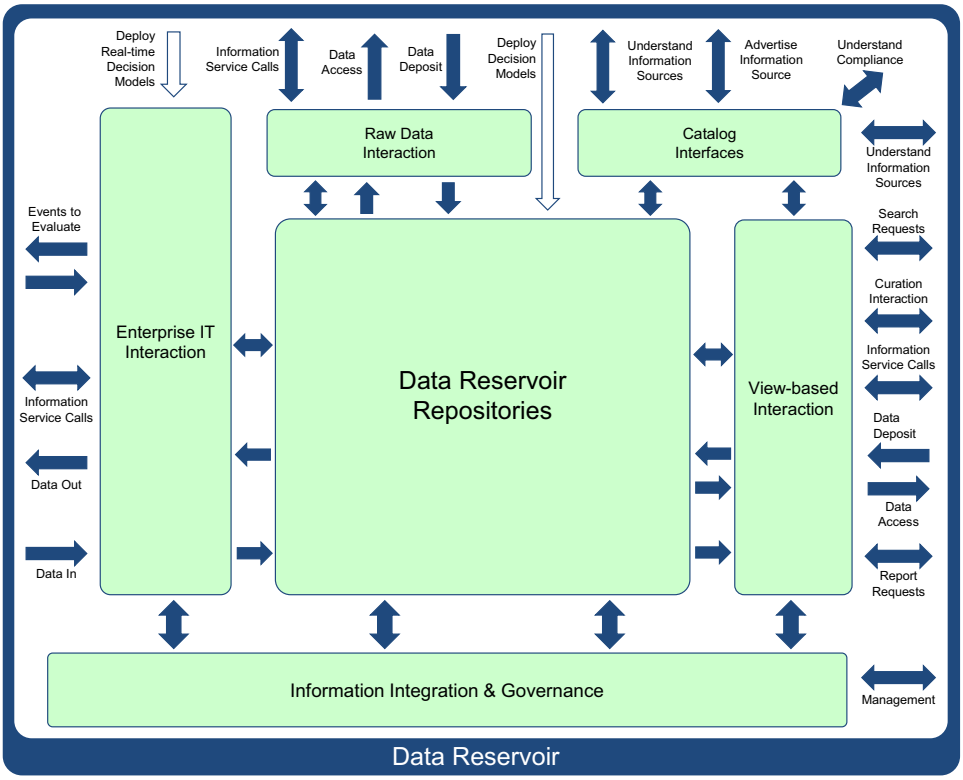


Figure 3-3 Major subsystems in the data reservoir

The data reservoir services contain four subsystems:

- Enterprise IT interaction responsible for the exchange of data with the systems managed by the enterprise IT teams.
- Raw data interaction responsible for the exchange of data with the data scientists and business analyst who are creating analytics.
- Catalog services manage the descriptions of the data in the data reservoir and the definitions of how it is governed.
- View-based interaction responsible for the exchange of data with the line-of-business users.

The Information management and governance fabric has a single subsystem in the logical architecture called Information Integration and Governance.

The sections that follow describe these subsystems in more detail.

3.3 Data reservoir repositories

The data reservoir repositories provide the storage for the data in the reservoir. There are multiple types of data repository, with each being responsible for supporting a different kind of access with the data being stored in a suitable format. A particular data reservoir will not

necessarily have every type of repository. It is common for a data reservoir to have multiple instances of a particular type of repository.

Each type of repository must support the appropriate service levels around performance, latency, and availability for the usage it is supporting as shown in Figure 3-4.

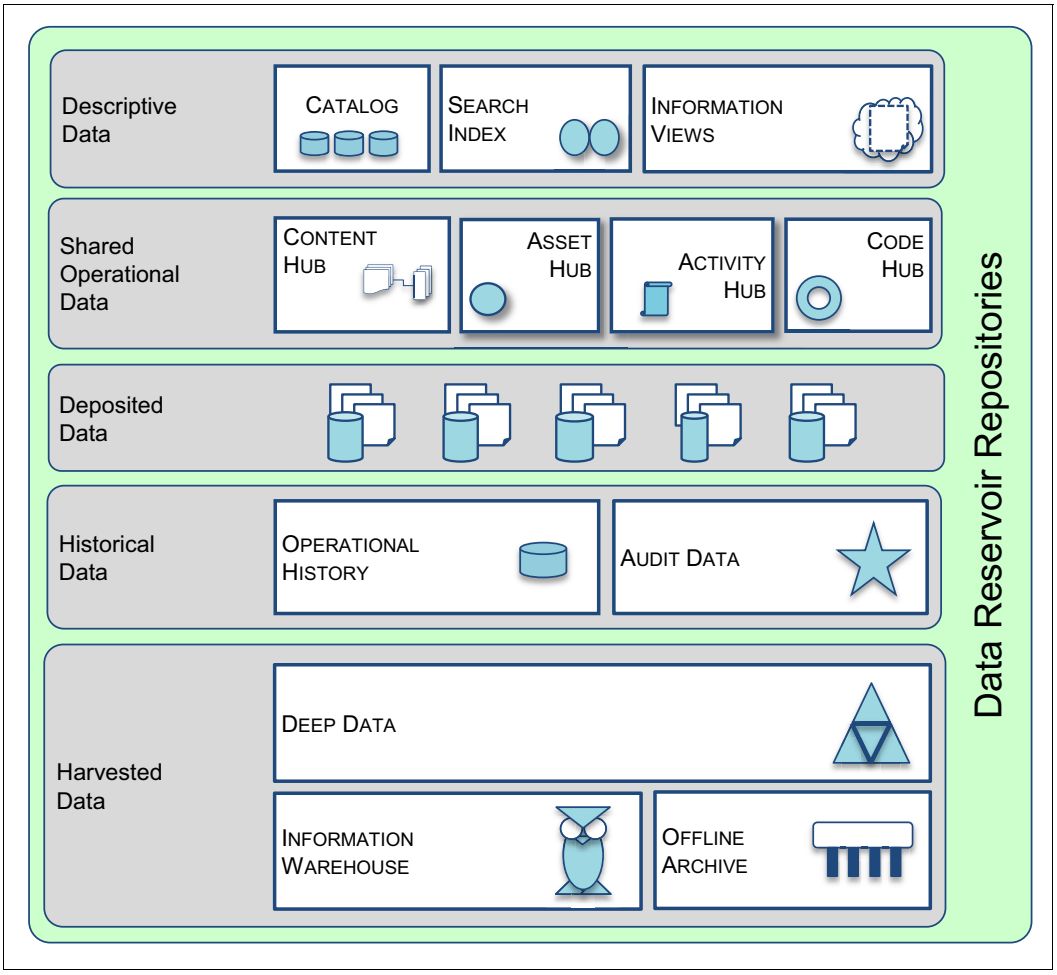


Figure 3-4 Types of data reservoir repositories

Different technologies can be used to satisfy the needs of these repositories, such as relational database or HDFS. However, the types are really related to the disposition of the data (that is structure, current or historical values, read/write, or read-only) used rather than the technology applied.

A data repository needs to store new information provided to it. Some transformation can occur to get data into the appropriate format, but this process depends on the data type. This process of storing new data is usually done during a staging to decouple the provider of the data from the repository itself.

A data repository will also support analytics processing deployed in it and calls to access and manage its information from other services. It also needs to publish insight and other relevant information to other reservoir components.

Users will not directly access the data repositories. Instead, access will be through a data reservoir service.

3.3.1 Historical data

This data provides a historical record of activity in the original format of the source. After this data is provisioned in the data reservoir, it is read only.

Operational history is a repository that stores a historical record of data from a system of record. This data will be stored in a similar format to the original operational data and is used for local reporting or as an archive for the system of record. Operational history repositories can also be used for applications that have been decommissioned and the data from which is being kept for compliance or reassurance reasons. Some analytics processes can also use this data.

Audit data contains collections of logs recording activity around who, what, and why in the reservoir, and can be used to understand how data is being used, or more importantly potentially misused. The guards and monitors of the data reservoir create this data.

3.3.2 Harvested data

This type of repository stores data from information sources outside of the data reservoir. This data is then available for processing and use by other reservoir services and external users or application. For example, it can be joined with other data, create subsets, add some analytical data, or be cleansed or otherwise changed into a form different from the original sources that better meets the needs of how it is going to be used.

An information warehouse is a repository that provides a consolidated historical view of data. Often it feeds the reporting data marts that provide dimensional or cube views, and other forms of traditional online analytical processing (OLAP). Its structure is optimized for high-speed analytics, and it contains a related and consolidated set of information.

A deep data repository is optimized to support data in various formats and at a large volume. Data is initially stored schema-less, although over time data structures can be applied, such as by defining hive tables.

Even if the data is largely schema-less, care needs to be taken in deciding how to structure the file and directory structure within HDFS to make it easier for users to use and process the data. This concept is discussed further in Chapter 4, “Developing information supply chains for the data reservoir” on page 75.

This is an ideal place to capture all kinds of ad hoc data that might be available, and offer the data sets for innovative analytical use within the organization

3.3.3 Deposited data

An important aspect for the data reservoir is that it enables its users not only to search for and use data put in the reservoir by others, but also makes it easy for those users to contribute their data back for others to use.

Deposited data consists of these information collections that have been stored in the data reservoir by users. This data can include various different types of data stored in different ways, such as analytics results or intermediate data. This data is often a snapshot and not actively managed or refreshed by other systems in the reservoir.

The depositor of the data is often the data owner and appropriate metadata will be entered by them into the catalog that describes the data to the best of their ability.

3.3.4 Shared operational data

These repositories store a copy of operational data within the organization. On occasion, these are in fact the only copy. Often they are updated from upstream system of record in near real time and the data might be minimally augmented on route to the repository.

Data is stored in the same format as the source and will be used as authoritative by other systems and processes within the reservoir. This data can be stored in one of these types of hubs, depending on the type of data:

- An Asset Hub stores slowly changing operational master data.

  This might include customer profiles, product information, and contracts. Within the reservoir and by the reservoir users, this is seen as the authoritative source of master data. Deduplication and other quality improving techniques are applied to this data. Updates are available for publishing.

- An Activity Hub stores recent activity related to a master entity.

  This is often rolled up periodically and will be used by analytical process to (for example) understand a customer's recent support calls with an organization or recent orders.

- A Code Hub is a set of code tables or mappings that are used for connecting different information sources.

  The Code Hub has a canonical definition of a reference data set such as a country code. Systems that use a country code might use their own definitions, so the code hub helps map between the canonical form and the application-specific definition. This service is made available to systems outside the reservoir.

- A Content Hub contains unstructured data and appropriate metadata in a content management repository.

  For example, this can be where you store business documents, legal agreements, and product specifications in their raw form. It might also include documents to publish externally such as to a website or rich media. The rich media can include audio and video such as educational material, announcements, and podcasts.

3.3.5 Descriptive data

This set of repositories is responsible for managing much of the metadata that is used for the reservoir itself to operate. These repositories need to be highly available or the reservoir will not function.

The catalog is one of the most crucial parts of the reservoir because it contains information about the systems and assets within the reservoir and the classification schemes used for those assets. It is used by some automated process to discover governance information such as information classification, by data owners to record details about their data, and by other users of the reservoir shopping for new data to use. It consists of a repository and applications for managing the information stored in the data reservoir. It is accessed through the catalog interfaces.

Information views provide definitions of simplified subsets of information that are geared towards the information user. These are typically done through federation, so this component stores information about the sources and relationships. In addition, the data is related to a semantic model that is also managed here. Search index is an index of data in the reservoir and associated metadata to enable applications to locate information they are after.

3.4 Information integration and governance

The Information Integration and Governance subsystem (Figure 3-5) includes the components necessary for managing and monitoring the use of information in the reservoir, including data mapping, movement, monitoring, and workflow. Its aim is to enforce standards on how information deployment processes are implemented and to provide a set of standard services for aiding these processes.

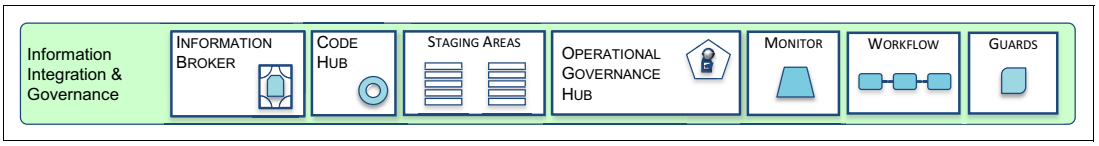


Figure 3-5 Information integration and governance components

The following are descriptions of the components shown in Figure 3-5:

- Information broker

  The Information broker is the runtime server environment for running the integration processes that move data into and out of the data reservoir and between components within the reservoir. It typically includes an extract, transform, and load (ETL) engine for moving around data.

- Code hub

  Similar to the code hub found within the shared operational data, this code hub has a canonical definition of a reference data set, such as a country code, and the specific implementation versions of this reference data set. This code hub is used primarily to facilitate transcoding of data coming into the reservoir and data feeds flowing out. Additionally, to support analytics the reference data can map the canonical forms to strings to make it easier for the analytics user and their activities.

- Staging areas

  Staging areas are used to manage movement of data into, out of, and around the data reservoir, and to provide appropriate decoupling between systems.

  The implementation can include database tables, directories within Hadoop, message queues, or similar structures.

- Operational governance hub

  The operational governance hub provides dashboards and reports for reviewing and managing the operation of the data reservoir. Typically it is used by the following groups:

  – Information owners and data stewards wanting to understand the data quality issues in the data they are responsible for that have been discovered by the data reservoir.

  – Security officers interested in the types and levels of security and data protection issues that have been raised.

  – The data reservoir operations team wanting to understand the overall usage and performance of the data reservoir.

- Monitor

  Like any piece of infrastructure, it is important to understand how the data reservoir is performing. Are there hotspots? Are you getting more usage than you expected? How are you managing your storage?

  The data reservoir has many monitor components deployed that record the activity in the data reservoir along with its availability, functionality, and performance. The management of any alerts that the monitors raise can be resolved using workflow.

- Workflow

  Successful use of a data reservoir depends on various processes involving systems, users, and administrators. For example, provisioning new data into the data reservoir might involve an information curator defining the catalog entry to describe and classify the data. An information owner must approve the classifications, and an integration developer must create the data ingestion process. Workflow coordinates the work of these people. For more information, see Chapter 5, “Operating the data reservoir” on page 105.

- Guards

  Guards are controls within the reservoir to enforce restrictions on access to data and related protection mechanisms. These guards can include ensuring the requesting user is authorized, data masking being applied, or certain rows of data being filtered out.

3.4.1 Enterprise IT interaction

Figure 3-6 shows the components of the Enterprise IT interaction subsystem.

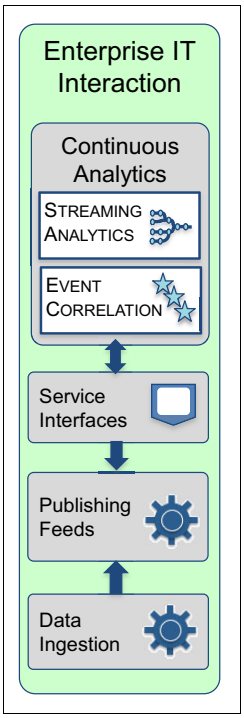


Figure 3-6 Enterprise IT interaction

This area of the reservoir offers services that are primarily targeted at other enterprise applications and the IT organization rather than business and analytics users:

- Continuous analytics

  Continuous analytics integrates and analyzes data in motion to provide real-time insight. This enables insights to be detected in high velocity data from real-time sources. For example, you might want to analyze streaming data from a user's interaction with a website in real-time to suggest a new product offering to them.

  - Streaming analytics is the component responsible for running the analytics on streaming data in real-time.
  - Event correlation provides for complex event processing based on business events that can occur over an extended period and can correlate these events to develop

additional insight. Some of these events can come from insight obtained by the streaming analytics component.

- Service interfaces

  Service interfaces provide the ability for outside systems to access data in the reservoir repositories, and for systems within the reservoir to query data from both inside and outside. These interfaces can be REST web services, SQL style through JDBC, or various other forms.

- Data ingestion

  The ability to easily import data from a multitude of sources is one of the selling points of the data reservoir. You want to be able to capture anything easily, and to allow decisions about how that data can be used later in the process.

  That being said, information needs to be incorporated into the reservoir in a controlled manner. During the process, metadata must be captured. It is also important to ensure appropriate separation between other systems and the reservoir to ensure an issue in one does not impact both. The Information Ingestion component is responsible for managing this. A staging area will often be used as part of the process to support loose coupling and ensure that delays in processing will not affect the source system.

  The information ingestion component will apply appropriate policies to the incoming data, including masking, quality validation, deduplication, and will push appropriate metadata to the reservoir catalog.

  Data with a known structure can be stored in a more structured repository such as a relational database, whereas less structured or mixed data can end up in a file system such as HDFS.

- Publishing feeds

  Publishing feeds are responsible for distributing data from within the data reservoir to other systems external to the reservoir. This can include other analytical systems, other data reservoirs, and systems of record, but all are outside the governance control of the supplying data reservoir. Lineage for the publishing process is captured, but this might be the furthest point lineage reports are available to.

  This data can be triggered by upstream changes, or run on a schedule or on demand from a user request.

  A subscription table is used to manage the list of sources and the destinations they need to be published to. The destinations can involve populating a table, creating files, or messages to be posted to a queue.

  During the publishing process, data transformation can occur, for example by resolving code values to alternate representations

3.4.2 Raw data interaction

Figure 3-7 shows the components of the Raw Data Interaction subsystem.

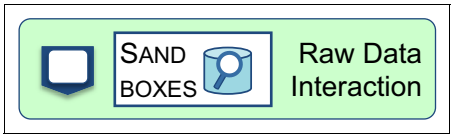


Figure 3-7 Raw data interaction components

Users of the reservoir do not access data directly. The raw data interaction component supports those who want to access data as it is stored in the reservoir repositories. They will

typically be expert users such as data scientists who accessing the data through ODBC or JDBC, or by using files such as in HDFS. They might be doing ad hoc queries, search, simple analytics, or more sophisticated data exploration.

Access control and filtering are done in this layer to ensure adherence to established governance rules, and audit logs are kept of the data accessed. For example, data might need to be masked to remove personally identifiable information. Quality of service rules can also be applied to ensure that users do not overload the repositories with excessive requests that cannot be handled without adversely affecting other workloads.

A sandbox is used to provide a user with a copy of data from selected repositories that allows for a greater level of isolation from changes to the underlying data and to limit the workload impact on the rest of the reservoir. These sandboxes will be managed through their lifecycle according to applicable governance policies.

3.4.3 Catalog interfaces

Figure 3-8 illustrates the catalog interfaces subsystem.

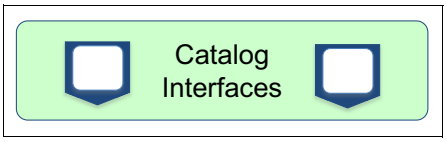


Figure 3-8 The catalog interface subsystem

The catalog interfaces provide information about the data in the data reservoir. This 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.

The catalog interfaces make it possible for the users of the reservoir to find the data they need in an appropriate format. This can be data formatted for business users to use in their visualization tools or more complex data geared towards the needs of a data scientist

The data reservoir's catalog also contains definitions of policies, rules, and classifications that govern access to data by users. The interface layer is responsible for adhering to these rules and ensures that users gain access only to the data to which they are entitled. Users are also able to request provisioning of data through the catalog interface.

The catalog interfaces also allow lineage data to be retrieved so that a user can understand where data came from and what it is being used for. This data can be viewed at various levels of detail from a high-level business perspective to a detailed operational view.

Ensuring that the catalog data is available to users of the reservoir in a timely, accessible way, easily searchable, and with appropriate categorization is crucial to the success of a data reservoir. It is a key aspect in ensuring users know what data they are using, helping ensure that the reservoir does not become a data swamp.

An information curator is responsible for ensuring the associated catalog data is accurate and up-to-date and covers concepts for the subject area, and that appropriate classification is in place.

3.4.4 View-based interaction

Figure 3-9 shows the components of the view-based interaction subsystem.

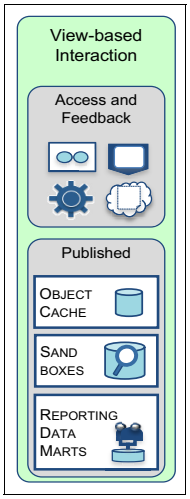


Figure 3-9 View-based interaction components

This subsystem of the data reservoir contains views or copies of the data in the reservoirs repositories that either simplify the structure or improve the labeling of the attributes so they are easy to understand for business users. It provides these benefits:

- Access and feedback

  Self-service provisioning helps to get information to business users with minimal delay, known characteristics, and without any additional IT involvement. To support this and improve the usefulness of the reservoir, users are able to request access to additional data and feedback on the data within the reservoir through comments, tagging, rating, and other collaboration. This can be done through interfaces such as the catalog as they search for data, or through other collaboration tools and social media.

  Information views provide simplified subsets of the data reservoir repositories that are labeled with business friendly terms (Figure 3-10).

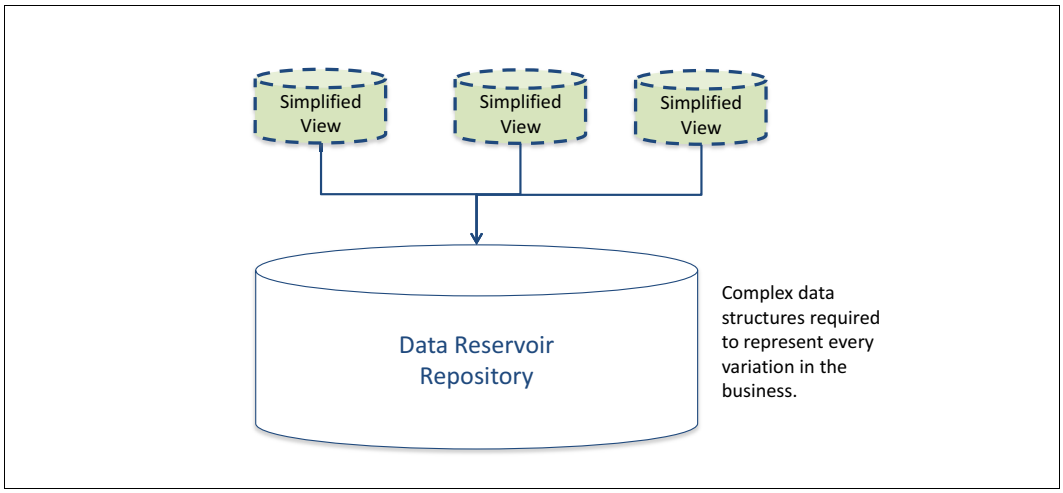


Figure 3-10 Providing views over a data reservoir repository

This component provides the ability for the business user to search for and obtain simple subsets or aggregations of data in a form more geared to them rather than to the structure of the underlying repositories. It is also labeled using business-oriented vocabulary. This makes ad hoc queries, searches, simple analytics, and data exploration much easier for these users.

The data reservoir operations team sets up useful views of data based on the data reservoir's repositories for these business users to access.

- Published

  The published subsystem contains the stored copies of data. These copies can be rebuilt as required from data stored in the data reservoir repositories.

- Sandboxes

  A sandbox is an area of personal storage where a copy of requested data is placed for the requester to use. These sandboxes are the same as the sandboxes in raw data interaction. They are populated from the simplified information collections designed for business users.

- Reporting data marts

  The reporting data marts provide departmental/subject-oriented data marts targeted at supporting frequent line of business reports. The data is often structured as a dimensional model such as star or snowflake, and are easily used by common business reporting packages.

  Data in marts will be updated incrementally as new data is made available. typically from an information warehouse.

- Object cache

  To improve performance and availability to applications, some views of data can be made available through a cache. The object cache is particularly suited for systems of engagement applications because it is document-oriented, typically using the JSON format.

3.5 Component interactions

Chapter 1, “Introduction to big data and analytics” on page 1 introduced an example pharmaceutical company, EbP, where Erin was trying to address some challenges around a few initiatives being undertaken by the business. Erin proposed a data reservoir architecture to help with these initiatives. This section uses a few aspects of those scenarios to demonstrate how the components in the reservoir interact to fulfill the needs of the business.

In each interaction, the Information Broker is the engine that runs the reference steps. Workflow is used to coordinate different aspects of the process, human or system, and the monitor tracks activity and performance metrics.

3.5.1 Feeding data into the reservoir

Figure 3-11 illustrates how the components within the reservoir collaborate during the process of importing new data into the reservoir.

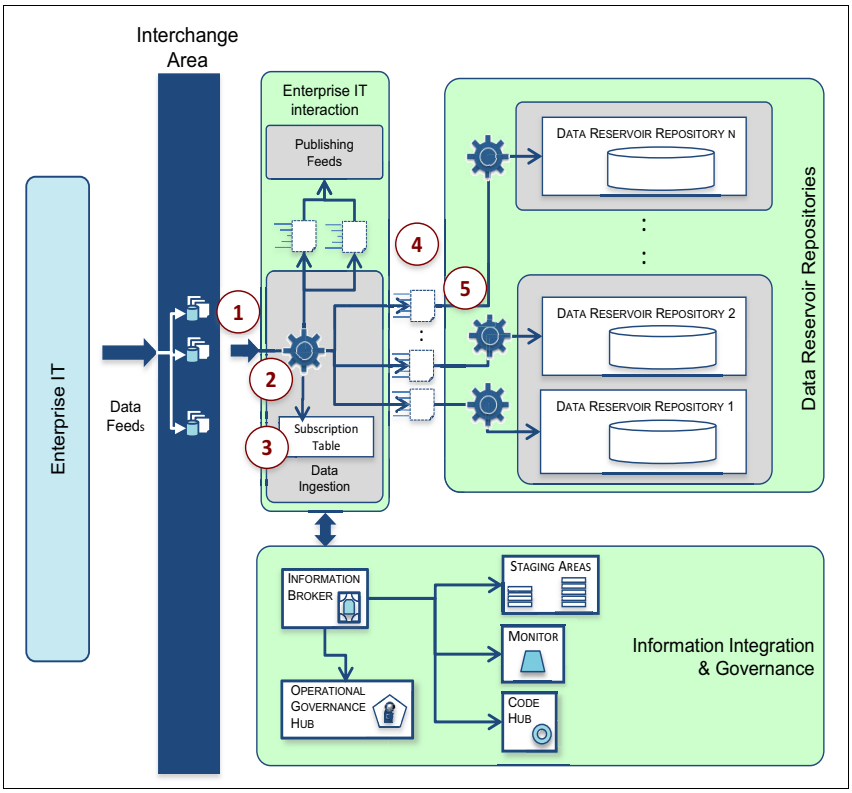


Figure 3-11 Bringing data into the data reservoir

Bringing data into the data reservoir involves the following steps (Figure 3-11):

- Step 1: Enterprise IT systems deposits data into an interchange area that is shared between the reservoir and Enterprise IT.

  This area is managed by the reservoir, and is the handoff point from existing systems. This area is only used by the Data Ingestion subsystem and will not be available to other parts of the reservoir. This interchange area can be a combination of file systems, message queues, and other technologies, and provides isolation between the source systems and the data reservoir.

- Step 2: The data ingestion process for a particular source is started by the information broker in response to an event, a manual request, or a schedule.
- Step 3: The data ingestion process looks up configuration information for the incoming data sets in the subscription table to determine the appropriate destinations.
- Step 4: The data is placed in staging areas for the publishing feeds component for destinations outside the reservoir. This data is in the same format that arrived in the interchange area.
- Step 5: The data is placed in the data repositories inbound staging area, again in the same format that arrived in the interchange area. These repositories then incorporate this new data into their internal storage, performing any necessary transformations.

Operational lineage information from the data ingestion processes is updated in the operational governance hub. This lineage data records the source and destinations of the data along with other pertinent information about the data transformations performed on it.

3.5.2 Publishing feeds from the reservoir

Figure 3-12 looks at the other end of the process and how data within the reservoir is published back to other systems.

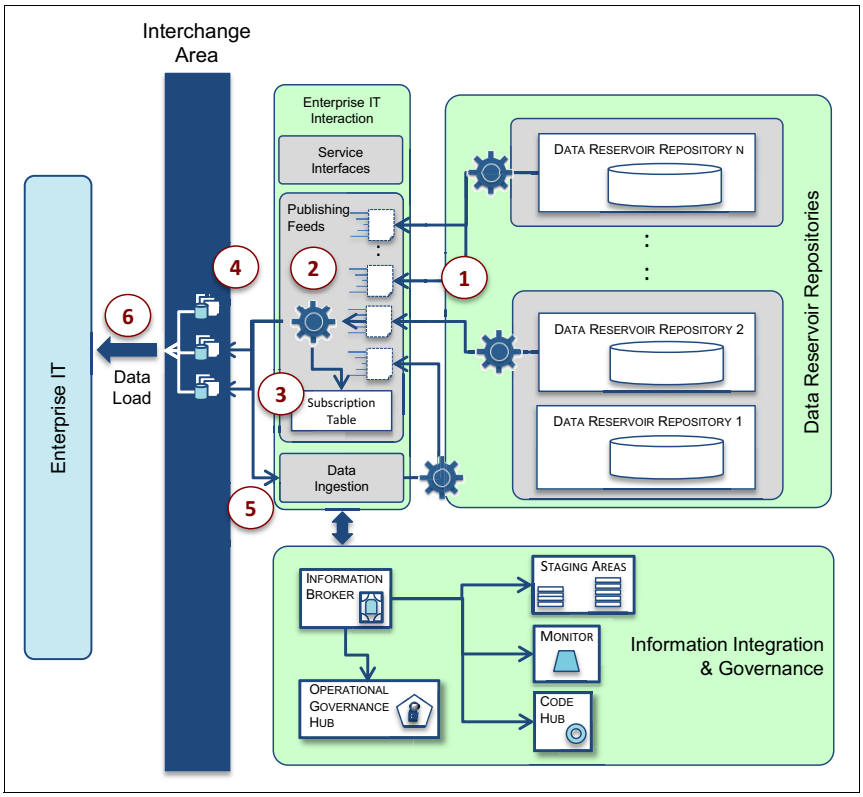


Figure 3-12 Publishing data from the data reservoir

Publishing data from the data reservoir involves these steps (Figure 3-12):

- Step 1: The data reservoir repositories publish interesting information to transient information collections. There is one transient information collection for each topic area.
- Step 2: An information deployment process is triggered (manual, scheduled, or on data arrival) to process the data in a transient information collection.
- Step 3: This process uses the subscription table to determine where information is to be distributed to.
- Step 4: The information deployment process posts a copy of the information to each relevant distribution mechanism.
- Step 5: The distribution mechanism can push data to the data ingestion subsystem to push to the data reservoir repositories. This action is how new insights generated by analytics are distributed between the data reservoir repositories.
- Step 6: Users pick up the data from the distribution mechanisms.

Each component also records data that allows operation lineage to be seen through the process.

3.5.3 Information integration and governance

The information integration and governance components collaborate to manage the movement of data within the data reservoir (Figure 3-13).

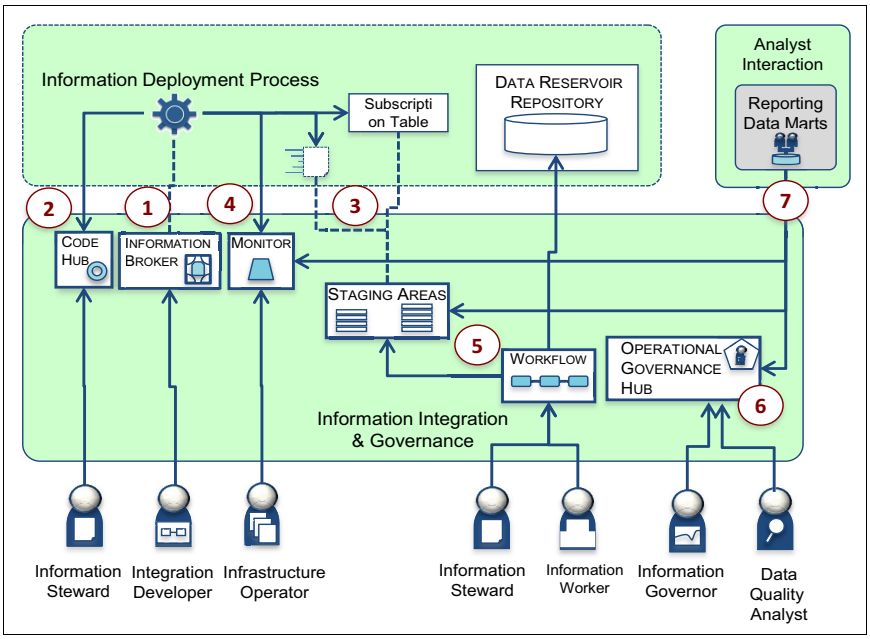


Figure 3-13 Information integration and governance interactions

The components from information integration and governance interact use the following steps (Figure 3-13).

- Step 1: Movement of information around the data reservoir is managed by an information broker. It hosts the integration processes (typically Information Deployment Processes) that transform and copy information in and out of the data reservoir repositories.
- Step 2: The integration processes use a code hub to look up code table mappings.
- Step 3: An integration process locates where to deliver data using a subscription table that lists the destination transient information collections for each relevant destination. The transient information collections are hosted in the staging areas database server.
- Step 4: The movement of information is logged by the monitor.
- Step 5: Errors discovered in the information by an integration process are recorded in a special transient information collection and are processed through stewardship workflows.
- Step 6: The policies that control the management of the data reservoir are managed in the operational governance hub.
- Step 7: Information owners and other users can see reports on the operation of the data reservoir.

3.6 Summary

This chapter provided information about the components that make up the data reservoir architecture. You should now have an understanding of each component, its role, and key interactions. In subsequent chapters, you can explore more about how data, processes, and people interact with the reservoir. Figure 3-14 shows a summary of the components of a data reservoir.

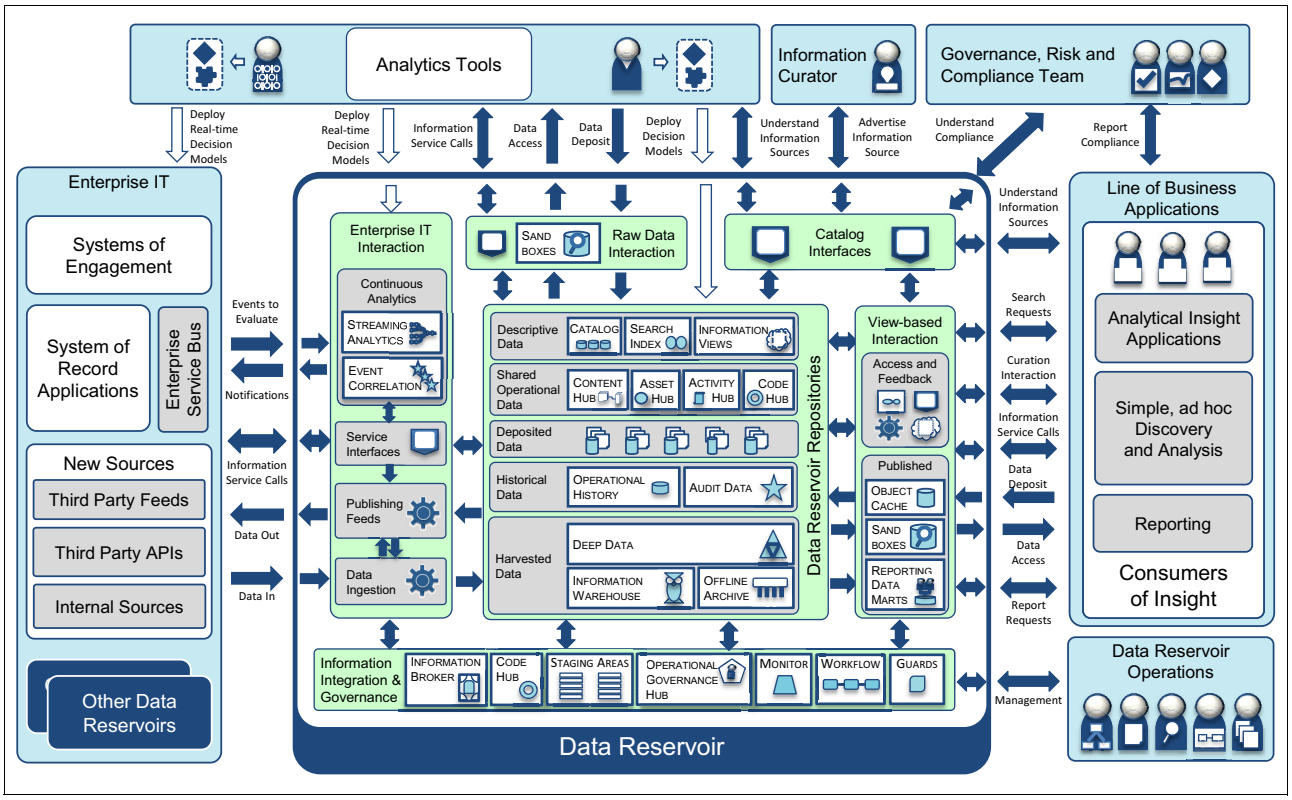


Figure 3-14 Summary of the components of a data reservoir

4

Developing information supply chains for the data reservoir

The thinking is that data has mass and therefore attracts more data through gravity. A successful data reservoir will attract increasing amounts of data. The information supply chain documents data flows so data does not disappear. Instead, it is visible and accessible for data reservoir users.

This chapter describes how to design the flow of data through the data reservoir by using information supply chains. An information supply chain describes the flow of a particular type of data from its original source through various systems to the point where it is used. It maps the movement of data from one or more information sources through the reservoir repositories and other stores to one or more users.

This chapter also describes the information supply chains designed for the data reservoir, and the considerations and best practices around implementing them for a particular organization. There are a number of stages to this effort:

- Identifying the subject areas that need to be represented in the data reservoir.
- Determining the potential sources for the data.
- Working with the information zones to identify which repositories need data from each subject area.
- Design the data reservoir refineries to link the sources with the data reservoir repositories and the users.

Data reservoir: The data reservoir manages data with different levels of quality and trustworthiness that potentially contains sensitive information. Well-designed information supply chains ensure that the origin of data and the processing that it has received are well understood so that users can decide whether the data they discover in the data reservoir is suitable for their needs.

This chapter includes the following sections:

- The information supply chain pattern
- Standard information supply chains in the data reservoir
- Implementing information supply chains in the data reservoir
- Summary

4.1 The information supply chain pattern

The IBM developerWorks article *Patterns of Information Management*[1] discusses the information supply chain as a pattern to address the problem “An organization needs to process information to fulfill its purpose. How is the flow of information coordinated throughout the organization's people and systems?” The solution is to design and manage well-defined flows of information that start from the points where information is collected for the organization and links them to the places where key users receive the information they need.

An information supply chain typically begins where people or monitoring devices enter data into an IT system. This system processes and stores the data. The data is then copied to other systems for processing and over time it spreads to a number of different systems. The path that the data flows is the information supply chain.

Figure 4-1 shows a simple diagram of the information supply chain.

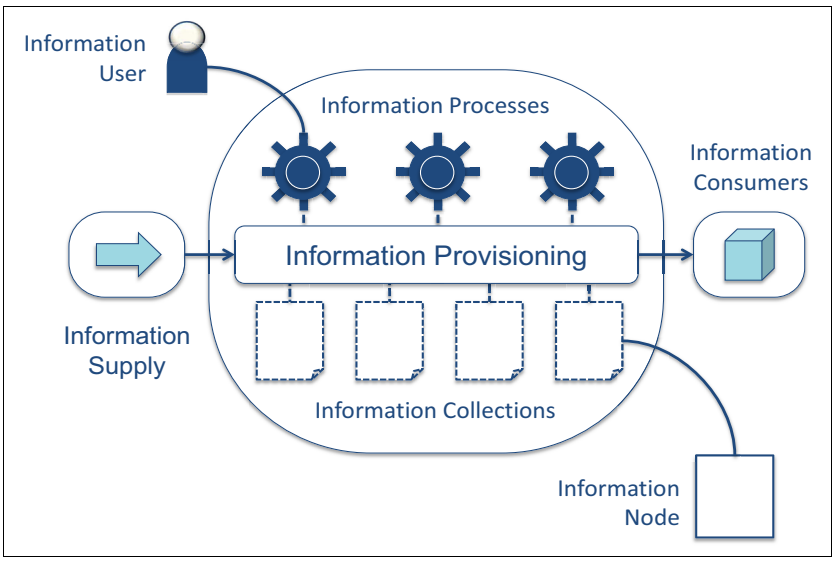


Figure 4-1 Schematic diagram of an information supply chain

The data entering the information supply chain comes from the information user and the information supply. This source can be another information supply chain outside of the control of the organization, such as from sensors or other devices that generate data.

The information collections are the places where this data is stored. In the data reservoir context, these are the data reservoirs, sources, repositories, sandboxes, and published data stores.

The information processes transform and derive new data for the information supply chain, and copy the data between the information collections. This processing is called information provisioning. In the data reservoir context, the following are the information processes:

- The data refinery processes that run in the enterprise IT interaction subsystem
- The analytics that run in the data reservoir

[1] Patterns of Information Management, Mandy Chessell, https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/W4108ee665aa0_4201_8931_923a96c3653a/page/Information%20Supply%20Chain

The information nodes are the infrastructure services that support the information collections and processes.

An information supply chain is the flow of a particular type of data through an organization's systems, from its original introduction, typically either through a user interface (UI) or external feed to its various users.

Another way of looking at an information supply chain is that it is the stories that you tell to transform the raw data into something useful. The information supply chain takes data of a particular kind with potentially different levels of quality and trustworthiness and levels of sensitivity and delivers it in various forms to different users.

Theoretically, the scope of an information supply chain is from the moment data is created through to every user of that data. However, the scope here is defined as an information supply chain to the systems and users that are within a particular sphere of influence. For this publication, the scope is to the data reservoir itself and the systems and people that connect to it.

4.2 Standard information supply chains in the data reservoir

Typically there is a different information supply chain for each subject area that the organization has. The simplicity of the data reservoir architecture means that many information supply chains will follow the same path through the data reservoir components. This section describes these standard data flows.

4.2.1 Information supply chains for data from enterprise IT systems

Enterprise IT provides the data reservoir with data that records how the organization is operating. Its inclusion in the data reservoir is a critical step for supporting the analytics that drive the business. The data scientists and analytics can copy this data into the data reservoir by using deposited data. However, this just adds a snapshot of this data. Ideally, this data is brought in through reliable, ongoing, automated processes so that the data reservoir becomes an authoritative source of a wide range of enterprise data.

This section describes the information supply chains that automate the importing of data into the data reservoir and the publication of data from the data reservoir. These automated processes are maintained by enterprise IT.

Enterprise data delivery information supply chain

Figure 4-2 shows the flow of data from enterprise IT systems into the data reservoir repositories where it can be used in various ways.

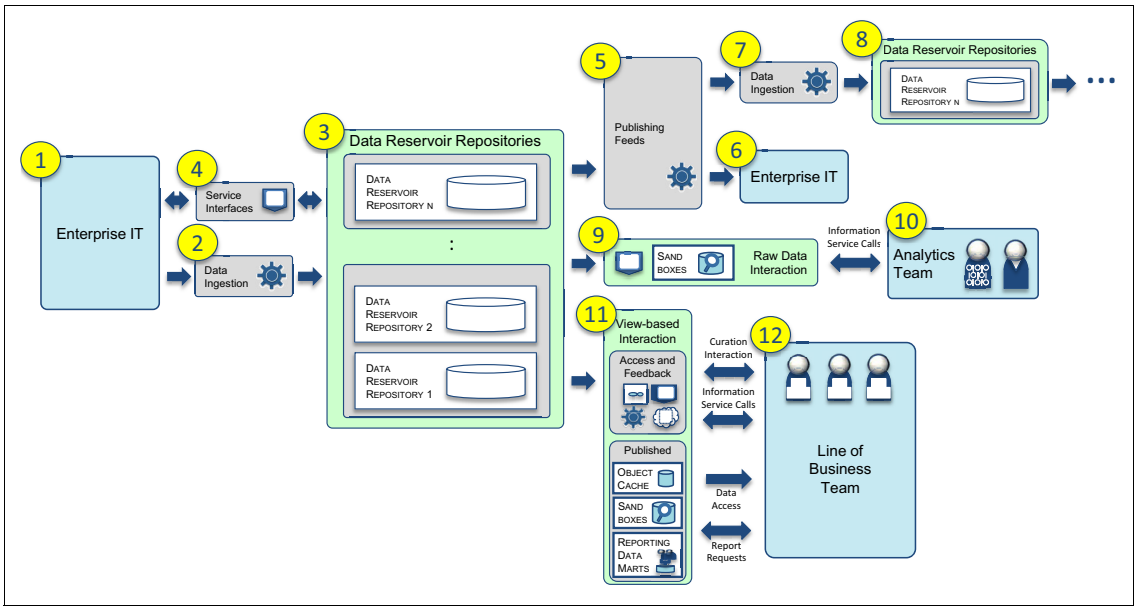


Figure 4-2 Enterprise data delivery information supply chain

The following are the steps in the enterprise data delivery information supply chain flow (Figure 4-2):

- Item 1: Data is generated through activity in the enterprise IT systems and external sources.
- Item 2: Data is copied into the data reservoir repositories through the data refinery processes in the data ingestion subsystem.
- Item 3: Data is stored in one or more data reservoir repositories.
- Item 4: Data may be accessed and updated through service interfaces.
- Item 5: Data from the data reservoir repositories, potentially augmented by analytics, is pushed into the publishing feeds subsystem for distribution.
- Item 6: The data refinery processes in publishing feeds can push the data back into the enterprise IT systems or to external feeds.
- Item 7: Alternatively, publishing feeds can push the data into data ingestion so that it is stored in other data reservoir repositories.
- Item 8: Typically, this is a flow of data from the shared operational data to either the historical data or harvested data repositories.
- Item 9: Enterprise IT data can be extracted from the data reservoir repositories through the Raw Data Interaction subsystem.
- Item 10: This Raw Data Interaction subsystem enables the analytics team to explore the data and create analytics from it. The analytics team benefit from the enterprise data by being able to explore, find, and use it in the raw data sandboxes.
- Item 11: Enterprise IT data from the data reservoir repositories can be accessed through the View-based Interaction subsystem. This action can be done either directly through information views or through the published data stores that have been populated by using data refinery processes from the data reservoir repositories.

- Item 12: The line-of-business (LOB) teams can then view simplified versions of the data. The LOB team is not as technical as the analytics team, and use the view-based interactions to work with simplified subsets of the data that are meaningful to them.

Continuous analytics information supply chain

Events and messages from enterprise IT can also be processed in real time by the data reservoir, rather than stored for later processing. This real-time processing occurs in the Continuous Analytics subsystem and the results of its processing are distributed both inside and outside of the data reservoir (Figure 4-3).

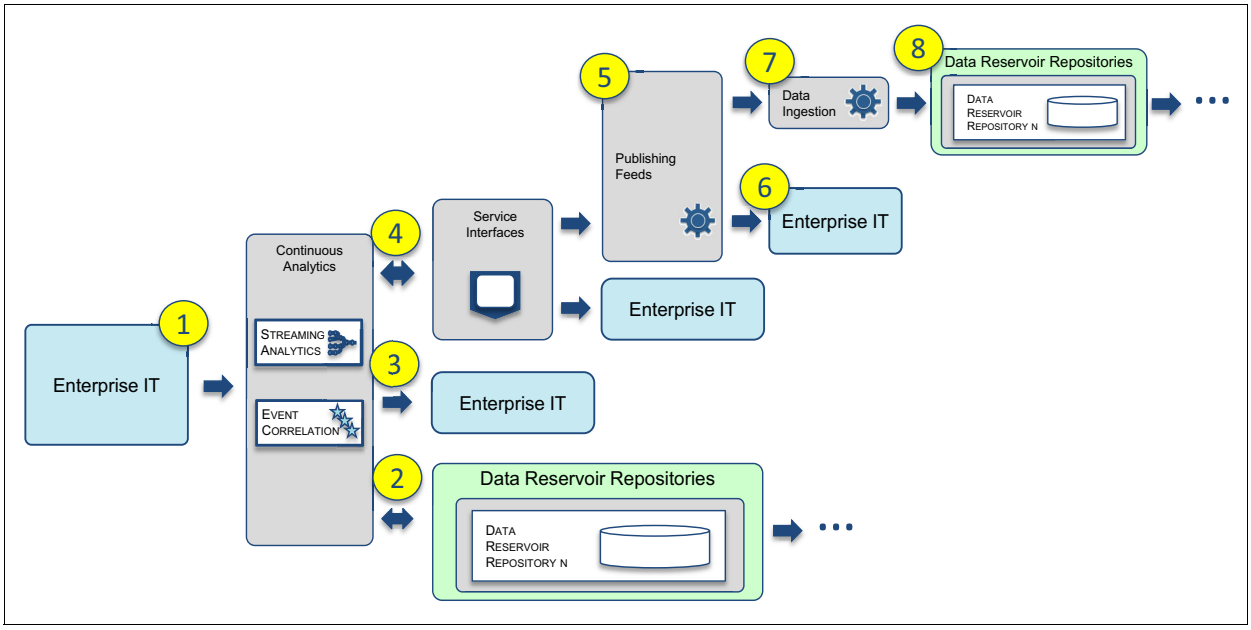


Figure 4-3 Continuous Analytics information supply chain

The continuous analytics information supply chain consists of these items (Figure 4-3)

- Item 1: New events or streaming data from enterprise IT are fed to the data reservoir and picked up by Continuous Analytics.
- Item 2: Continuous Analytics can look up data in the data reservoir repositories to correlate and interpret the incoming data.
- Item 3: The resulting events can be sent back to enterprise IT directly.
- Item 4: Results can also be pushed through the services interfaces for broader delivery. The service interfaces provide a layer of isolation between the continuous analytics processes and the delivery destination.
- Item 5: The results of the continuous analytics processes can be sent as a publishing feed.
- Item 6: The publishing feed can push data to enterprise IT.
- Item 7: The publishing feed can push data to the Data Ingestion subsystem.
- Item 8: The data ingestion processes store the new insight in the data reservoir repositories.

The choices in the different paths to publish new insights provide options on how frequently the data is published and the mechanisms used.

4.2.2 Information supply chain for descriptive data

The data reservoir is heavily dependent on the catalog and related descriptive data to ensure that the data it is managing is properly governed and understood.

This descriptive data itself requires proper management to perform its job correctly.

Figure 4-4 shows the information supply chain for descriptive data.

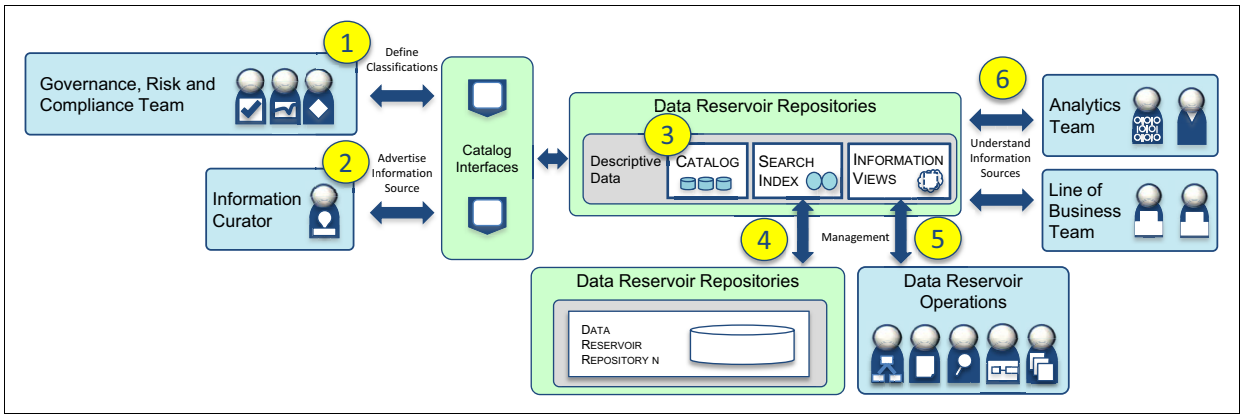


Figure 4-4 Metadata information supply chain

The metadata information supply chain consists of these items (Figure 4-4):

- Item 1: The governance risk and compliance team set up the policies, rules, and classifications using the catalog interfaces.
- Item 2: The information curators describe and classify the data that they own.
- Item 3: Both the governance definitions and the descriptions of information sources are stored in the catalog.
- Item 4: The search index is continuously updated with the results of indexing scans on the data reservoir repositories (including the catalog).
- Item 5: The data reservoir operations team manually maintains the information views.
- Item 6: The analytics team and LOB teams can then understand the data with the understanding that it meets company policies.

4.2.3 Information supply chain for auditing the data reservoir

Internally, the data reservoir continuously monitors the activity in the reservoir and the requests for data from people and systems external to the data reservoir. This monitoring activity generates a lot of data that can be analyzed and audited. Figure 4-5 shows the flow of data for auditing the data reservoir.

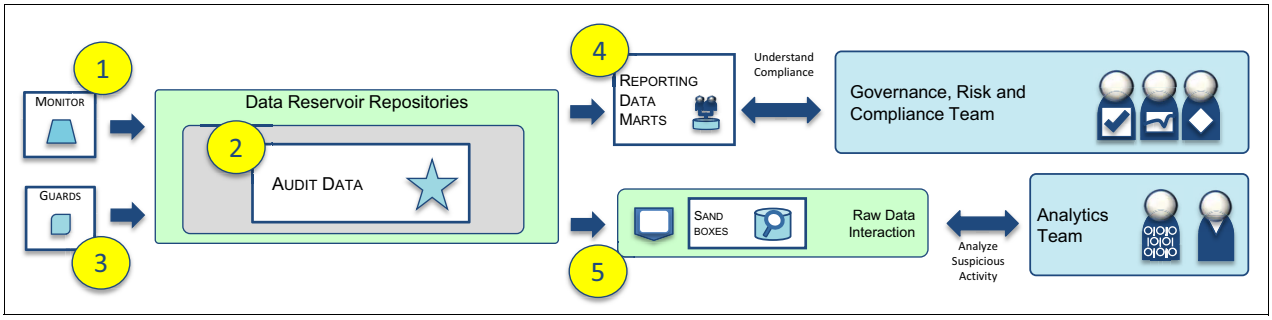


Figure 4-5 Data reservoir auditing information supply chain

The data reservoir auditing Information Supply Chain consists of the following items (Figure 4-5):

- Item 1: Much of the audit data is captured by monitoring processes. These processes watch the access of data in the data reservoir repositories, and monitor steps in the data refinery processes that create the operational lineage information for the data reservoir data.
- Item 2: This data is stored in an Audit Data repository.
- Item 3: The guards that are enforcing access control can also record both successful and unsuccessful access attempts in an Audit Data repository, depending on the repository being accessed.
- Item 4: Audit data is regularly consolidated into reports for the governance, risk, and compliance team.
- Item 5: The analytics team can generate additional analytics by exploring the values stored in the audit data. These analytics can be deployed into the audit data repositories to detect suspicious activity.

4.2.4 Information supply chain for deposited data

An important aspect of the data reservoir is that it offers a place for teams to store their own data sources and the results of their analysis in the data reservoir without needing help from the Data Reservoir Operations team. This data is stored in deposited data. So the last information supply chain is one that is a simple service-oriented style of access to the deposited data repositories (Figure 4-6).

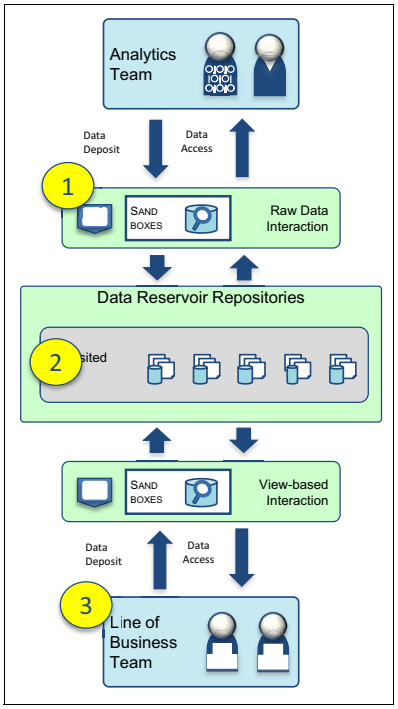


Figure 4-6 Deposited Data Information Supply Chain

The Deposited Data Information Supply Chain consists of the following items (Figure 4-6)

- Item 1: The analytics team are storing and retrieving their sources of data.
- Item 2: This data is stored in the Deposited Data repositories.
- Item 3: Similarly, the LOB teams can store and retrieve their files and the files that others (such as the analytics team) have shared with them.

Deposited data provides an effective mechanism for the analytics team to share the results of their work with the LOB. Analytics that have ongoing value can then be pushed into production in the data reservoir. This is a process that involved the Data Reservoir Operations team to do the following activities:

- Verify that there is no unexpected impact of the analytics on the existing production environment.
- Provision any additional data that the analytics requires in the data reservoir repositories.
- Arrange for the deployment of the analytic model in the appropriate data reservoir repository or data refinery process.
- Ensure that the resulting insight can be used by the business.

4.3 Implementing information supply chains in the data reservoir

Designing the information supply chains for a specific data reservoir deployment requires specific consideration of the existing sources and the use cases for the data reservoir. A wide range of integration technology can be used to access, move, transform, and store data. Any of this technology can be used in a data reservoir, if it is able to honor the governance rules defined in the catalog and produce appropriate auditing and lineage reporting to ensure that the data reservoir can demonstrate its compliance.

The following section covers some of the architecture considerations for a specific data reservoir deployment.

4.3.1 Erin's perspective

Erin Overview, the enterprise architect, needs to be reassured on her thinking around the data stored in the data reservoir. As the enterprise architect, Erin needs to design the data reservoir so that the data is well-managed and governed along each of the information supply chains. Doing so helps the people using the data reservoir to understand where data came from and how trustworthy it is.

The data reservoir typically starts empty, and data repositories and sources are added as required in an ongoing manner. Erin is looking to understand more around the information supply chains. She wants to know how the data moves from outside the reservoir into and around the reservoir. She also wants to use the concept of information zones in the data reservoir to help her manage the data usage for the various users.

Erin shows simpler versions of the data to Tom Tally, the accounts manager. She would like Tom to see the data in a way that is natural to him. Erin would like to introduce a semantic layer with simple views to ease the data consumption for Tom. At the same time, she would like to give Callie Quartile access to the raw data she needs to understand how the new drugs they are developing work.

4.3.2 Deciding on the subject areas that the data reservoir needs to support

A subject area is a category or topic of information that has business value. Erin decides which subject areas the data reservoir needs to support from the use cases and the users. She identifies patients, suppliers, customers, orders, genomes, clinical trials, manufacturing schedules, deliveries, and invoices as the subject areas that will appear in the Eightbar Pharmaceuticals data reservoir.

Of each subject area, she asks the following questions:

- Where are the likely sources of this information? If Erin does not know, then what investigations does she need to do to identify likely sources?
- Are there any cultural or ethical considerations around accessing each source?
- What is the scope (instances), coverage (attributes), and quality of each of these sources? On what basis does Erin know this?
- Who owns the source? Who does Erin need to ask for permission to access the source? Does Erin need to arrange new agreements with the owner?
- What technology is required to access the data sources? How does the data reservoir connect to it? Does Erin need to arrange for new credentials to be set up for access?

- What is the operational load and availability of the source? When can Erin get the data, and what are the restrictions? On what basis can Erin be sure that this level of service will meet her near-term and future needs?
- What is the volume of this data? What is the volatility of this data? These answers tell Erin how much it would cost to copy and how much ongoing refreshing is required.
- Who are the likely users of this data?
  - What are their requirements for scope, coverage, and quality?
  - What tools are they using? What structure (formats) do they need the data in to support their tools?
  - When do they need this data (the required availability) and how much processing are they going to perform on it?

Typically, the information supply chain is different for each subject area. However, the simplicity of the data reservoir architecture means that many information supply chains will follow the same path through the data reservoir components. This analysis will show the mix of the different subject areas in each data repository and the different subject areas that need to be extracted from each source.

4.3.3 Information sources: The beginning of the information supply chain

This section covers more details about the information sources. Sources are the supplier of much of the data for the data reservoir.

Systems of record, engagement, and insight

Systems of record are operational systems that support the day-to-day operations of the business. There are many systems of record, each specialized for a single purpose. As a result, their data is tightly scoped to the business functions that they support. These systems are used to hold the version of the truth that the part of the business it supports needs to operate. The data associated with the many systems of record is vital to have in the data reservoir.

Systems of engagement are the systems that interact directly with people and support various activities. This ability makes them an important source of information for the data reservoir to understand what an individual is doing and how they are doing it.

Systems of insight (such as the data reservoir) also generate important information that needs to be fed into the system of record and systems of engagement. For example, the data reservoir can generate scores that indicate customer churn rates and who the high value customers are, which would be important for the system of record that is supporting a customer service organization.

What data is needed from a source?

To answer this question, look at the data.

All data

While deciding what data you want to bring into the data reservoir, you might decide that you want all available data in the enterprise. Taking in all data gives the organization the greatest amount of freedom to process that data and you will not be missing any information. But what do you mean by all data? Do you want to bring in every data source that has ever been created?

Pragmatic selection of data

Typically, the data reservoir contains all of the data that the organization would like to share or retain for future analysis.

As you bring in data from new and existing information sources, any data that you can ignore means that you avoid the following situations:

- Moving it with data refinery services
- Being concerned about exposing it inappropriately
- Tracking that data in the reservoir or dealing with lifetime considerations.

It makes sense to filter out data that is unlikely to be used. As you use the data in the data repositories to make insights, you can improve how you filter data, based on usage statistics. For streaming sources (such as Twitter feeds) there will be a large amount of data, much of which you are not interested in. In this case, you need to target the data you require by using rules or analytics models to pick out subsets of the data.

Characteristics of sources

The data reservoir can use many styles and types of data.

Structured data

The structured data can be readily queried, and joined and merged by, for example, using SQL on relational databases. There can be many different types of structured data, including relational and dimensional, each with their own set of data types. Structured data coerces applications to store specific data, which can be useful to force data into a consistent representation. Applications can then readily use this standardized structured data.

The prevalence of mobile devices (such as smartphones, tablets, and wearable devices) and the rise of the *Internet of Things* means that there is a continual creation of data. Many devices typically produce structured data. However, this structure is typically not correlated with other sources.

Semi-structured data

Most of the new data created each year is semi-structured data such as social media, email, documents, video, audio, and pictures. For example, people typically produce text-based data. In order to be useful, this data has associated headers (describing where and when the data was created) along with the text-based data.

Documents can be stored in JavaScript Object Notation (JSON). JSON has structure but is not as rigid as structured data (relational databases have defined schemas). Each JSON document can have a different shape. To query JSON document stores, it can be useful to use a SQL-like paradigm and pick out parts of documents that fit a particular shape. This approach makes the JSON easier to use, by applying a structured lens on the data to pick out what is interesting.

With this data, there is often a small amount of identification information such as a social media handle. Using social and mobile technologies, individuals are empowered to create their own content and interactions. This data can create a better sense of someone's life than ever before. This data is increasingly used by organizations to get the fuller view of a person. You can use the data reservoir to store this data and correlate it with other sources.

Techniques such as text analytics are required to pick out relevant information. This semi-structured data can take the following forms:

- Documents (such as a web page) or a word-processing document.
- Streamed data (such as data from your car). Streamed data often needs to be accumulated over time in a data repository so that a historical view can be taken. If the amount of data is very large, rules are often needed to look for and extract meaningful data. This extracted data is stored in the data reservoir. In addition to existing structured information about an entity, new attributes can be found in the streamed data.

  Time series data is streamed data that is accumulated over time in the data reservoir. Analytics done on time series data allows trends and repeating patterns to be identified. These trends and periodicities allow predictions and forecasts to be made.

  Recent information and roll ups of information are useful for real-time analytics. Information decays, so its value decreases if it is not acted on in a timely manner. Ideally the time window that is used for recent information needs to be large enough to show short-term interesting patterns, but also be performant. Data from different subject areas will become obsolete (decay) will at different rates. This rate of decay will determine how frequently.

4.3.4 Position of data repositories in the information supply chain

There are often questions on whether existing analytical repositories should be adopted into the data reservoir or kept as information sources with selective content copied into the data reservoir as required.

Should your warehouse be in the data reservoir?

If the organization sees their existing warehouse as an analytics source, then this data needs to be represented in the data reservoir.

If the existing information warehouse is well-maintained and managed and there are processes to ensure that the data within it are of good quality, then it makes sense to adopt this information warehouse into the data reservoir and use it as a reliable data repository of data.

If the contents of the information warehouse cannot be confidently relied on, then leave it outside of the data reservoir and treat it a data source. Picking out the pieces of data and refining them in the reservoir might be a good option. If the existing data management practice has failed once, it will fail again unless it is changed. Often, focusing on a smaller scope of data is required to create new well-managed information warehouses in the data reservoir that are fit for purpose.

Enterprise data warehouses are typically large. When adopting an existing enterprise data warehouse into the reservoir, there is a choice on how much of its data to catalog. Some or all of the data in the warehouse can be cataloged. The cataloged data is then available to users and analytics of the data reservoir. If the existing warehouse is poor quality, overloaded, or does not contain relevant data, then it is best left out.

Master data management (MDM) and the data reservoir

In life, when you meet someone new you discover some basic information about them from what you can immediately see and maybe what you have heard about them. You then make assumptions based on that data. If you ask the person's name and then speak to them with someone else's name, they might be offended and feel that you are not interested in them. As you talk with them, you find out more things about them, such as:

- Marital status
- Attitudes about gender
- What job they have

They might feel that some of this information is private and should not be disclosed. As you discover more information, you might have suggestions or insights that might be useful. If all goes well and the person feels respected, you build trust and a relationship. The relationship they have with an organization that knows things about them is similar. You need to be sure that you build relationships based on well managed, appropriate information that you are confident in, so your interactions with them are ethical and wanted.

MDM

Operational data typically is in specialized applications that are each only visible to a small part of the enterprise. Data about people and organization, products, services, and assets that are core to the working of the business is often duplicated in each of these applications and over time becomes inconsistent.

Master data management (MDM) takes a holistic view of the end-to-end information supply chain and considers how to keep this high value data synchronized in the different systems. Often this type of data is being updated in different silos, where a silo has a constrained view of the data. MDM solves this problem by creating a consolidated view of this data in an operational hub that is called the *asset hub*, and provides common ways of accessing that hub that the whole enterprise can use.

MDM has processes that allow these high value entities to be normalized, not duplicated, and synchronized with other copies of the same data.

Information value: High value data about people, organizations, products, services, and assets that drive a business is needed in the data reservoir because it is the data necessary to develop high value analytics. This data is typically present in many of the data reservoir repositories.

Companies move from customer data being spread over many silos to a single MDM operational hub for personal data. This operational hub contains the high value business data. You need the master data so that analytics get access to the up-to-date data, but you do not necessarily need a separate MDM hub.

The inclusion of an asset hub inside the data reservoir is decided by these factors:

- Whether the organization has an asset hub
- Whether they need operational access to it.
- Is the data in the correct format for the users?

For example, you know that you need to use customer data as a basis to be able to treat each person effectively. Do you put the asset hub for customer data into the data reservoir or set it up as a source?

The answer is a political one, not a technical one. It is about who owns customer data. If it is owned by the operational systems of record team, then it is unlikely that the asset hub itself can be included in the data reservoir. However, the data from it can be included.

The asset hub is often a critical operational system and the change management around it is strict so that the customer data can be replicated into the data reservoir. This copy in the data reservoir is read-only. In addition, there can be established governance practices around the asset hub that differ from those in the data reservoir.

For Eightbar Pharmaceuticals, Erin could have chosen to have one of these configurations:

- Read-only patient asset hub with patient data in the reservoir, and to keep the writable data updated by using an asset hub outside the reservoir.
- An updateable asset hub in the reservoir governed by the catalog.

For Erin, the patient details and medical records are owned by the hospital, which supplies this data to Erin in the medical health record. The medical health record is a read-only replication of the patient data from the hospital.

4.3.5 Information supply chain triggers

Every movement of data is initiated with some form of trigger. This trigger can come from the source system, the target system, or an independent information broker or scheduler. The result of the trigger causes a data refinery process to start. Figure 4-7 shows the different types of triggers.

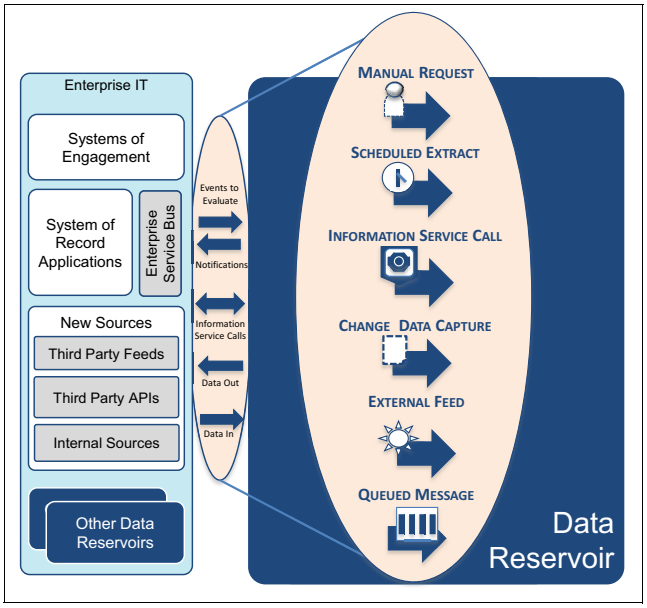


Figure 4-7 Triggering data movement

Triggering data movement can include the following activities (Figure 4-7):

- Manual requests: Allowing authorized users to bring in data to the reservoir.
- Scheduled extract: This could be an overnight extraction of a day's activity into the reservoir.
- Information service call: For an application to push data into the data reservoir.
- Change data capture: As a result of an update to stored data in the information source.

- External feed: A push of data from an external party.
- Queued message: A push of a message or event from an information source.

The trigger starts processing that makes the data available to the enterprise IT interaction subsystem, where it is picked up and processed appropriately.

4.3.6 Creating data refineries

Processes inside the data reservoir called *data refineries*, take data and process it to change its structure, move it between data stores, add new insights, and link related information together. These processes are also used to bring data into the data reservoir and publish data out of the data reservoir.

A core principle of the data reservoir is that users do not have direct access to the data reservoir repositories. The data refinery processes are the only processes that can access and update the data reservoir repositories.

Data refinery processes capture lineage information about the data they are moving and transforming so it is possible to trace where data originated. Data refinery processes also honor the governance classifications defined for the data in the catalog by implementing the associated governance rules. As such, they are key processes in creating robust governance in the data reservoir.

Styles of data refinery

The data reservoir reference architecture supports many types of data refinery process, of which these are some examples:

- ETL bulk load and delta load

  Extract, transform, and load (ETL) is a technology for moving large amounts of data between data stores. ETL jobs are composed of sequences of data transformations (splits and merges are also possible in the flow).

  For example, the initial import of data from a new information source or the population of a new data reservoir repository might be performed by using an ETL job. Then incremental updates to the information sources can also be moved into the reservoir by using ETL jobs. Traditionally, ETL jobs have been run on structured sources. More recently, ETL engines such as IBM InfoSphere DataStage® can also run with Hadoop sources and targets.

- Trickle feed

  Trickle feeds are used when small changes occur in an information source that need to be sent to the data reservoir. Replication and messaging are two approaches in addition to ETL to implement a trickle feed.

- Replication

  Replication monitors changes to the information source and triggers a data flow when new data is stored. In this case, no complex transformations occur. This approach has little effect on the information source system both in terms of performance and change to its logic.

- Messaging

  Messaging is a push mechanism from the information source system. Data is packaged as a message and deposited on a queue that acts as an intermediary. The data reservoir monitors the queue and picks up new messages as they arrive. Using message queues

provides an intermediary between the sender and the receiver, enabling them to exchange data even if they are operating at different times of the day.

A messaging paradigm has these advantages:

- Messages can be sent synchronously or asynchronously. This method allows control over when information flows. Synchronous processing of messages is simple to implement, but the receiving system might not be ready to process the message, so it blocks the sender. Asynchronous message allows the sender to send and not be blocked. The message is processed some time later when the receiver is ready.
- Messaging systems (for example, IBM MQ[2]) ensure delivery and scale.
- It might be more efficient to batch up messages or send them independently.

The data refineries of the data reservoir implement any transformations that are required during the trickle feed.

Validation

Refinery services can be called to validate that data is of the shape and content expected. For example, addresses can be validated to ensure that they represent an actual location. Errors are raised as exceptions that are sent to the team responsible for the original information source so that team can correct them.

Data shaping or transformation

To improve the raw data, you often need to normalize it and make it self-consistent. Data refinery services do this sort of shaping. This action involves filtering, normalizing, and dealing with nulls and trailing blanks.

One way to ensure that data values are normalized is to use reference data. Reference data provides a standard set of values for certain fields. For example, it can be useful to hold a standard definition of country codes and use this reference data to ensure that country code fields comply. Variations of “US”, “USA”, and “United States” could all be resolved to the same reference data value. This means that users get much more consistent data.

It is important to understand that making data consistent is not the same thing as raising the quality of data. Raising the quality of data might take a person who understands the context of the data and can make a meaningful assessment. As a company embraces a semantic definition of concepts, it is possible to make inferences about the meanings of new concepts. Using cognitive analytics technology solutions allows users to interact with the system in natural language in a more meaningful way.

Adding smarts into the information supply chain: Analytics in action

As you create supply chains in your data reservoir, it is useful to think of the information supply chain itself as an analytics project. This means that you should look to capture metrics of the supply chain to assess how it is doing and whether it is meeting the needs of the data reservoir.

So what do you want from the information supply chain? Each information supply chain has an expected service level agreement (SLA). The following items could be considered:

- Operational efficiency: Collecting metrics around the amount of data, the time for transformation, the time taken, and the number of concurrent jobs means you can see how the supply chain is doing. A real-time dashboard or report provide descriptive analytics. It is worth taking the time to understand which roles might need to see which reports and make sure that the reports are matched to the role.

[2] For more details about IBM MQ go to this web address: http://www.ibm.com/software/products/en/ibm-mq

- Seeing how quality and consistency metrics change.
- Using analytics to highlight possible fraudulent activity in the supply chain.
- Using analytics to answer the business questions such as am I getting a return on investment from the supply chains I have?
- Using metrics to spot when you are running out of capacity such as processing power, network bandwidth, or memory, which indicates that more data repositories might be required.
- Metrics and lineage information from the supply chains and their interpretation are part of the story to ensure that audits can be passed, legal standards are complied with, and ethical practices are occurring.

After you have the data and can spot anomalies in the supply chain, you can start making decisions on how to change the supply chain. Deploying analytic scoring models into the supply chain to give up-to-date information to enable real-time decisions to be made should improve the supply chain efficiency.

4.3.7 Information virtualization

Information virtualization is a technique that provides consumable data to people and systems irrespective of how the data is physically stored or structured. Information virtualization is used for two purposes in the data reservoir:

- The View-based Interaction subsystem uses information virtualization to provide consumable versions of data to its users.
- The Data Reservoir Repositories subsystem uses information virtualization to augment the data that is stored in the data reservoir repositories.

The techniques to implement information virtualization are varied and depend on these factors:

- The location of the data
- How compatible the format of the stored data is to the needs of the user
- How frequently the requests for data will be
- How much capacity is available in the systems that store the data to service these queries

There are two basic approaches:

- Accessing the data in place through a simplifying view or API
- Copying the required subset of data to a new location where the users can access it

Many organizations want to avoid copying data, particularly when a required information source is too large and too volatile to make it feasible to copy. The data reservoir can represent this data as cataloged information views over its contents so that data can be located, understood, and accessed on demand. For this to work, the information source must be available when the data reservoir users need it, and able to support the additional workload that the data reservoir brings.

If the information source is not able to support the needs of the data reservoir users directly, then its data should be copied into the data reservoir repositories. For example, Eightbar Pharmaceuticals chose to use an object cache for the information for patients and medical staff for use by the systems of engagement. This choice was made because the availability requirements for systems of engagement are high and it is easy to achieve high availability with a simple object cache. Also, the object cache limits the data that is exposed to the systems of engagement.

When combining data from multiple sources, there are two approaches to consider:

- Copy sources to a single data reservoir repository and join the data sources on demand
- Use federation to query and create federated views, leaving the data in place

Copy sources centrally and join

Creating a copy of each source centrally and then joining them together means that all the data is in the same ecosystem, for example Hadoop or a relational database. After the data is copied, it is located together so joins do not involve excessive network traffic.

Federated Views

Federation returns data from multiple repositories using a single query. It manages logical information schemas that the caller uses. Using federation to query and create federated views, leaving the data in place, means that data does not have to be moved. This approach can be simpler and can handle a variety of data sources both in size and structure. Federation is an up-to-date (within practical limits) “don't move the data” paradigm. Also, federated views can help with authorization by only showing the parts of the data that are allowed to be seen.

Using federation is not magic, so it is worth thinking about sampling the data to reduce the amount of data that needs to flow. Being careful that complete table scans on large tables are not required to support the federated view. Consider caching as a way to keep some data locally to aid performance at the expense of the data being up-to-date. Consider copying the data using ETL, if you do not want the data to change underneath you. This gives you a point in time snap shot of the data that might be more useful to train an analytic model on in combination with the federated view.

Pragmatically for proof-of-concept and discovery activities, federation is often the quickest way to get going with data.

4.3.8 Service interfaces

Service interfaces provide direct access to a part of an information supply chain. They represent points in an information supply chain where information can both be injected and extracted.

In the data reservoir, the service interfaces are one of these types:

- RESTful information services provide simple operations (such as create, update, query, and delete) on named data objects. These data objects are typically stored in the shared operational data repositories.
- Services to initiate processes to manage the data reservoir.
- Services to run particular types of analytics.

Following is an example of using the service interfaces.

Harry Hopeful's data needs

Many people use spreadsheets to hold their data successfully. They are easy to use. The disadvantage is that a spreadsheet is a personal store of data that is not designed for sharing. As individuals exchange spreadsheets, multiple slightly different versions can be produced as each person changes it. It is possible that this spreadsheet data is never shared or stored for future analytics.

This section shows how Harry can move from working purely with spreadsheets to a mobile application that uses cloud services to manage his data in the data reservoir. This means that the data is moved into historical data repositories, so analytics can be performed. The information services around the reservoir allow mobile applications to access the data.

This change of emphasis from data locked up in a silo (such as a spread sheet) to the data being liberated in the data reservoir resonates across many industries and roles (Figure 4-8).

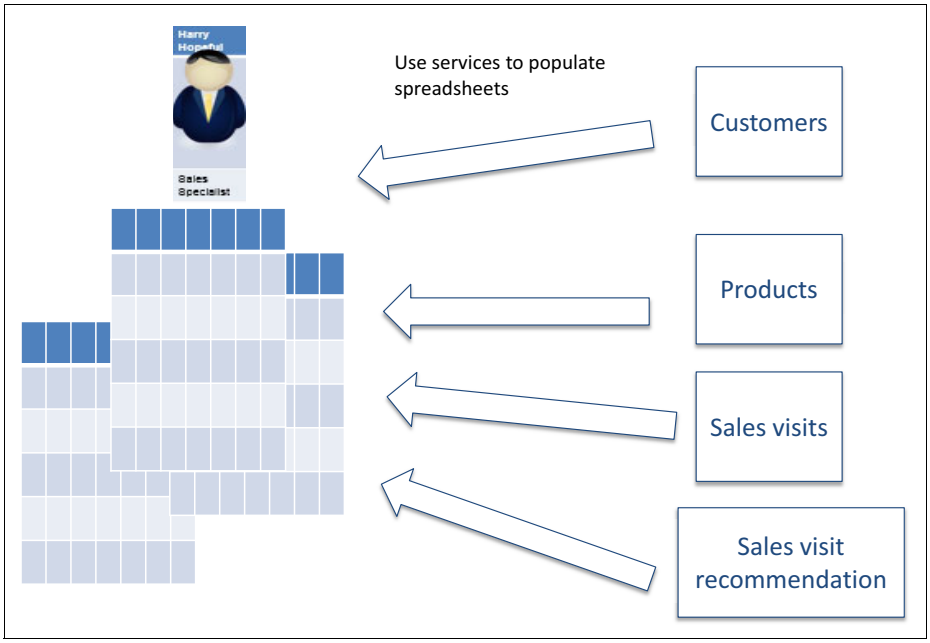


Figure 4-8 Harry Hopeful's spreadsheet usage

Harry Hopeful visits consultants in hospitals and their own practices to sell them products. He is used to working with spreadsheets that he keeps up-to-date manually. Locked in these spreadsheets are years of experience on how and when to approach different types of sales situations. Harry is approaching retirement and is happy to make his spreadsheet data available in the data reservoir so it can be used to develop some analytics that will help guide and train less experienced sales people.

The data reservoir holds the history of Harry's sales visits. Current data, along with recommendations, are fed into the object cache, which in turn feeds information to Harry's tablet. Harry works more on his tablet, and has an app on the tablet that uses location services to determine who Harry is visiting, in which case a draft sales visit is created for Harry to confirm.

The plan is to send a recommendation to Harry and have him confirm whether they are valid or useful. The aim is to improve the recommendations through Harry's feedback so they can be used with other, less experienced sales people.

This is an example of using data and calculated insight published from the data reservoir for new applications. These applications feed the results of using the data reservoir's content back into the data reservoir to improve the analytics.

4.3.9 Using information zones to identify where to store data in the data reservoir repositories

You have considered the subject areas in the data reservoir and the likely information sources. This section considers how the users will use the data reservoir and how data needs to be organized in the data reservoir repositories.

An information zone is a collection of data elements that are prepared for a particular workload, quality of services, and user group. In this way, data is grouped for the same usage: Scoping content in the catalog so that different types of users see the data that is most relevant to their needs.

Information zones are helpful when thinking about information supply chains as they identify the different destinations that the information supply chain must serve.

Information zones are also an implementation time hint to the data reservoir architect as to what SLA should be present in the target data repository, ensuring the data repository is deployed onto infrastructure with the correct characteristics and services. Some information zones overlap on the same data element. However, it might be necessary for the same data element to be copied (and possibly reformatted) and stored in a different repository to make it more consumable in another information zone.

Data reservoir repositories: The data reservoir repositories can each exist in multiple information zones, meaning the information zones overlap. A large number of overlaps is a sign that there are many opportunities to share data. This must be balanced with the need to structure and place data on an infrastructure platform that suits the workload using it.

There are three broad types of information zones:

- Traditional IT information zones: These information zones reflect the publication of enterprise IT data into the data reservoir. This zone includes the deep data, information warehouses, and their reporting data marts.
- Self-Service information zones: These information zones provide data to the line of business (LOB) users that have simple structure and are labeled using business relevant terminology. The self-service zones should not require deep IT skills to understand the data and its structure.
- Analytics information zones: These zones contain the data that underpins all the steps in the analytics lifecycle.

Traditional IT zones

The traditional IT zones describe how data flows from the enterprise IT systems into the data reservoir. They include these zones:

- The landing area zone
- The integrated warehouse and marts zone
- The shared operational information zone
- The audit zone
- The archive zone
- The descriptive data zone

The landing zone

The landing area zone contains raw data just received from the applications and other sources. Figure 4-9 shows the landing zone in the context of the data reservoir. This data has the following characteristics:

- Minimal verification and reformatting performed on it.
- Date and time stamps added.
- Optionally, information is added around where the information came from and who authored it.

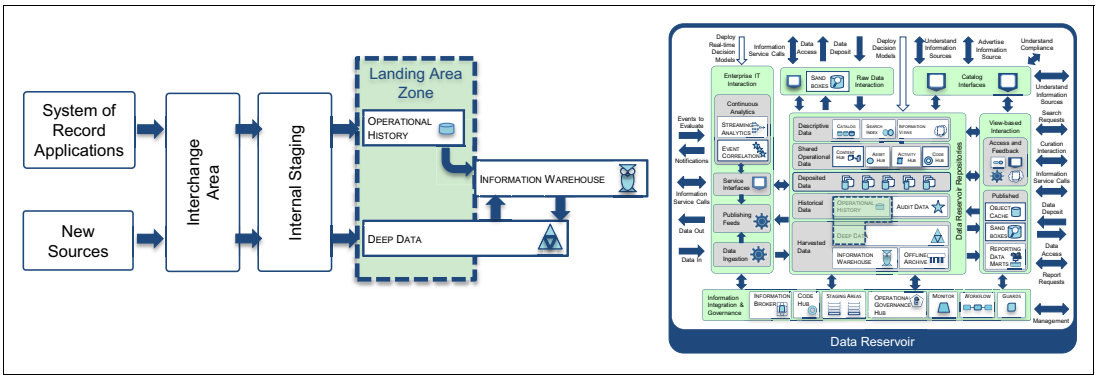


Figure 4-9 The landing area zone

The systems of record do not know about the staging areas, only the interchange area that is on the boundary of the reservoir. Only the data refineries use the staging areas. The interchange area is not in the landing zone because this is not a place where users can get data.

The integrated warehouse and marts zone

The integrated warehouse and marts zone contains consolidated and summarized historical information that is managed for reporting and analytics. It spans the information warehouse and reporting data marts repositories.

This zone is populated and managed using traditional data warehousing practices.

Figure 4-10 shows Operational History and Deep Data repositories feeding into the Information Warehouse. The Information Warehouse can feedback derived information into the Deep Data repository.

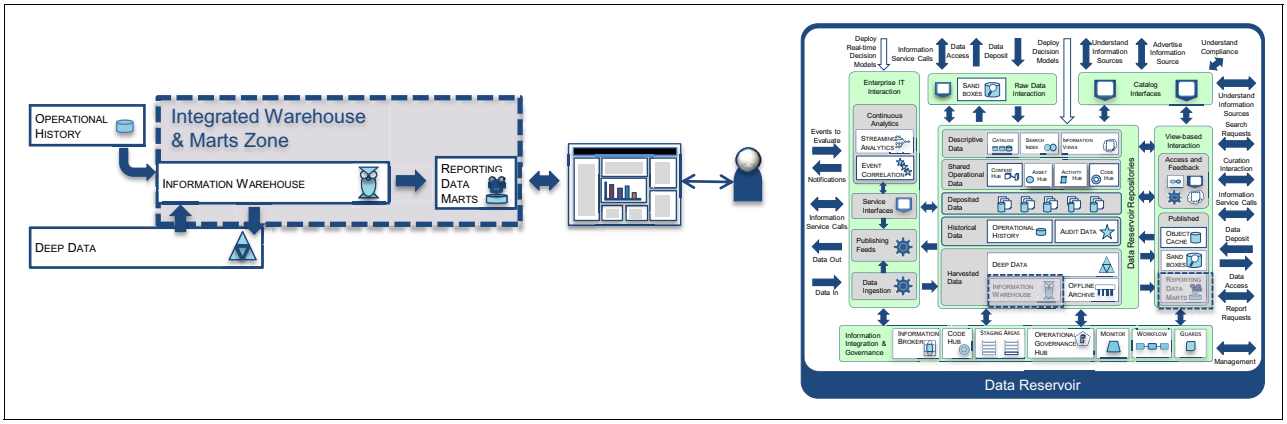


Figure 4-10 Integrated warehouse and marts zone

Reporting data marts are created from the Information Warehouse to create simple subsets of data suitable to drive reports.

Shared operational information zone

The shared operational information zone has the data reservoir repositories that contain consolidated operational information that is shared by multiple systems. This zone includes the asset hubs, content hubs, code hubs, and activity hubs.

Service interfaces are typically available to access these repositories from outside the data reservoir.

Data from the shared operational information zone is also used to help correlate and validate data inside the data reservoir. Data can also be fed to this zone through Data Ingestion and distributed from this zone through the Publishing Feeds. It can also receive new data through the service interfaces.

Figure 4-11 shows three interactions between the shared operational data and deep data.

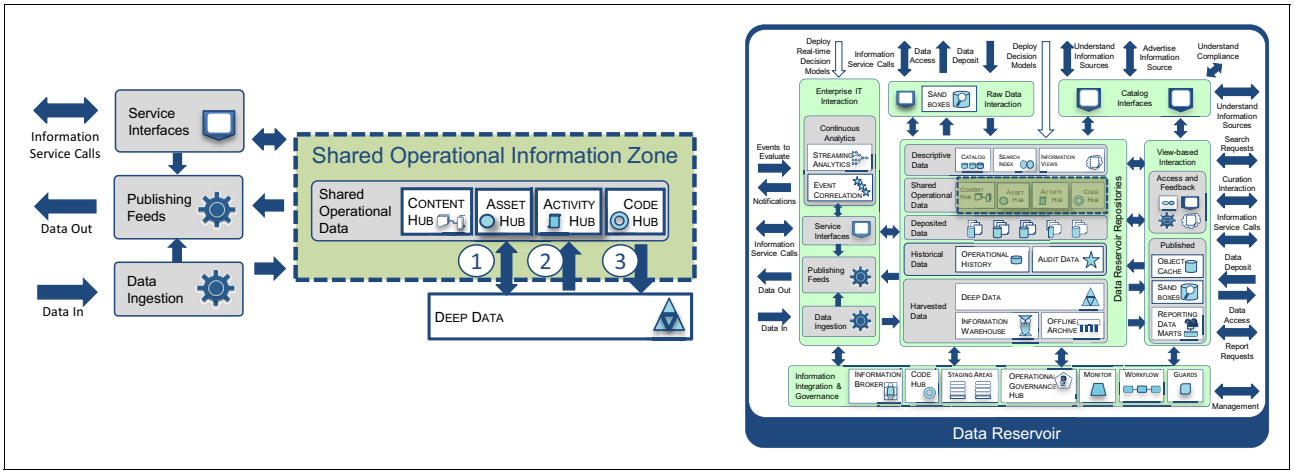


Figure 4-11 Shared operational information zone

The shared operational information zone consists of these items (Figure 4-11):

- Item 1: Service interfaces between the asset hub and the deep data repository, matching master data with new sources.
- Item 2: Pushing recent activity from deep data into the activity hub. The following are other patterns of keeping the activity hub and deep data up-to-date as new activity comes into the reservoir:
  - A refinery process updates the activity hub, and the deep data is then periodically updated depending on how up to date the activity in deep data needs to be. It might be that deep data only needs to be updated overnight.
  - As a new activity comes into the data reservoir, refinery processes update the activity hub and the deep data. This keeps deep data as up to date as the activity hub.
- Item 3: Standardize and map code values. For example, standardizing country codes means that data can be more easily correlated. If possible, standardization of code values should be pushed back to the source systems. Over time, the data coming into the reservoir becomes more standardized, reducing the standardization work that the reservoir needs to do.

The audit data zone

The Audit Data Zone maps to the audit data repository in the Historical Data subsystem. It is fed from the governance processes and logs the activity of the data reservoir. The audit log data comes predominantly from the information brokers, monitors, and guards (Figure 4-12).

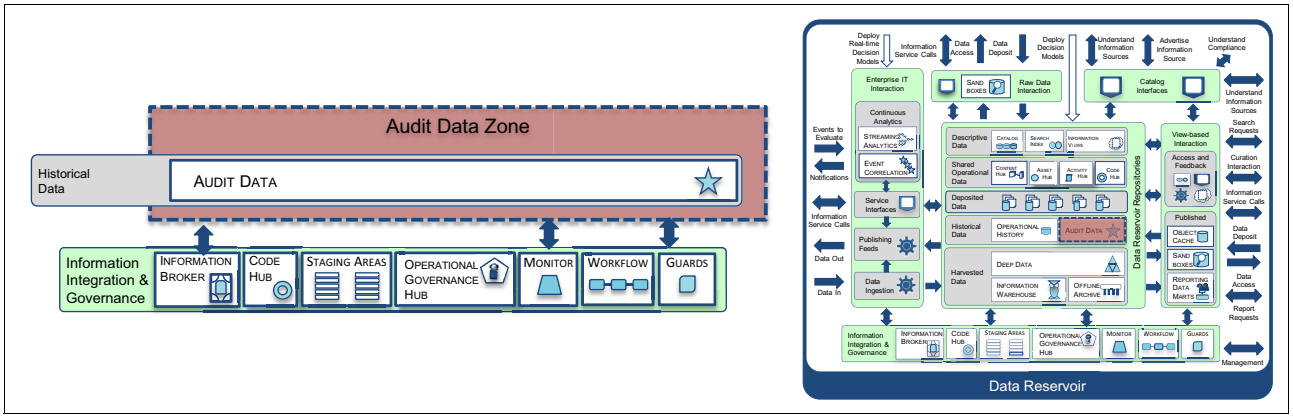


Figure 4-12 Audit data zone

The archive data zone

The operational systems (system of record or system of engagement) run the daily business. They are continually updating data and deleting old data that they do not need any more (for example, completed activities). For analytics, you need to capture this historical data and store it for investigations, audit, and understanding historical trends. This ability is one of the roles of the data reservoir.

From the perspective of the operational systems, the data reservoir is its archive. The data reservoir itself has an archive as well. The archive data zone holds data that is archived from the data reservoir (Figure 4-13).

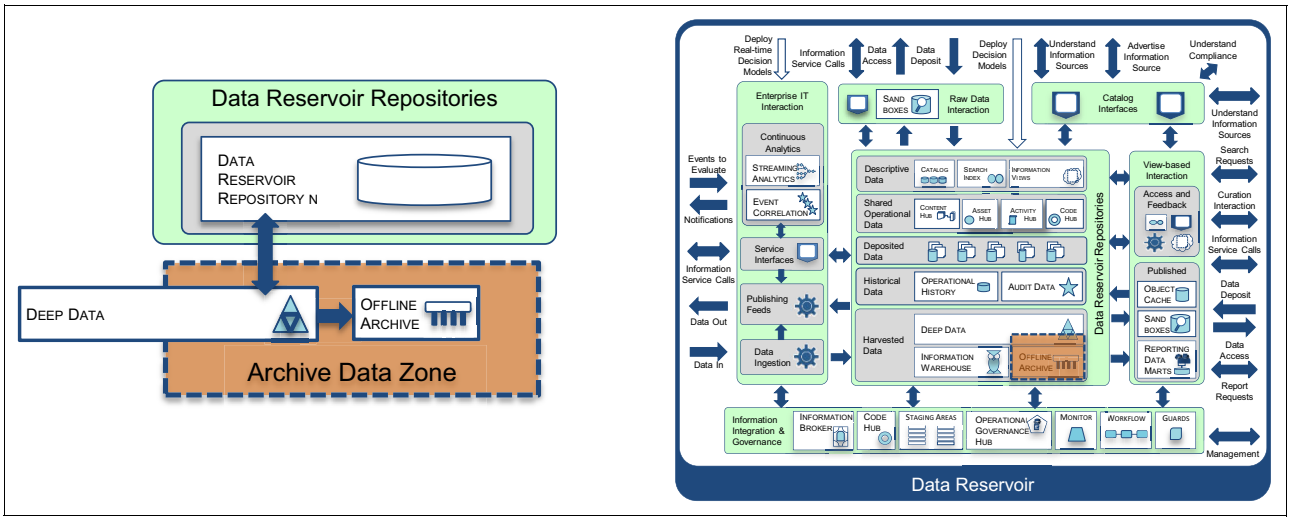


Figure 4-13 Archive data zone

Data is kept online in the data reservoir while its users need it. Then it is shifted to the offline archive (Figure 4-13) in the harvested data subsystem. The Archive Data Zone of the data reservoir is the offline archive plus space in the Deep Data repository for data that is about to go into the offline archive.

From the perspective of the data scientist building analytical models, the data reservoir is much like a museum. Artifacts that are no longer being used in daily life are curated and can be accessed by visitors to gain an understanding of the past. In a similar way, data from the operational systems is curated and stored in the data reservoir. Data scientists and reporting mechanisms accesses this data to gain an understanding of the past. When data scientists are performing data mining, they are looking to learn the lessons from the past to predict how best to react in similar circumstances when they occur in the future.

Not all artifacts are displayed in a museum. The museum needs to store many artifacts in its archive (an area that visitors cannot access without special permissions). Similarly, the data reservoir has an offline archive where data no longer useful to the data scientist or reporting or other analytical processing is stored. This offline archive in the data reservoir is only retained for regulatory reasons and is unlikely to be needed by the data scientists.

Data is moved into and deleted from the archives based on retention classifications set in the catalog. The archiving process is managed by the data refineries and monitored from information integration and governance.

The descriptive data zone

The descriptive data zone contains the data stored in the descriptive data repositories.

The purpose of this zone is to provide descriptions of the reservoir content. This zone includes the catalog content, plus search indexes and information views used by the View-based Interaction subsystem (Figure 4-14).

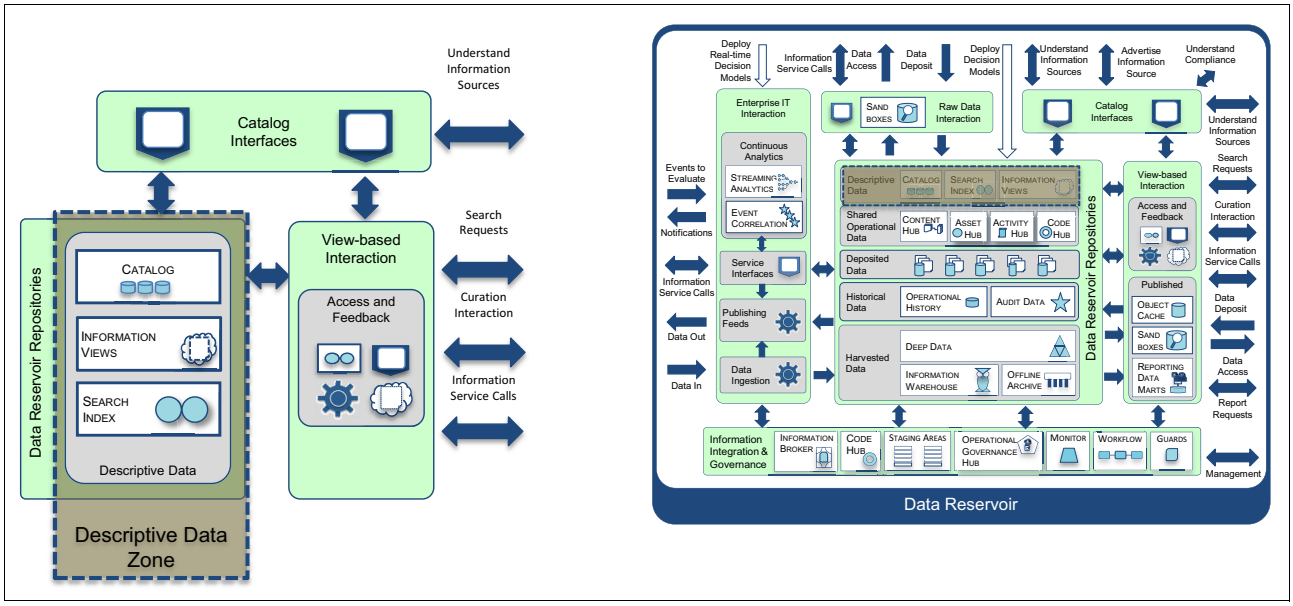


Figure 4-14 Descriptive data zone

Self-service zones

The self-service zones contain the data that is used by LOB teams as they access the data reservoir through the view-based interaction subsystem. These zones include:

- The information delivery zone
- The deposited data zone
- The test data zone

The information delivery zone

The information delivery zone stores information that has been prepared for use by the lines of business. Typically this zone contains a simplified view of information that can be easily understood and used by spreadsheets and visualization tools.

This zone is in the published data stores in the View-based Interaction subsystem, or includes data accessed through the information views (Figure 4-15).

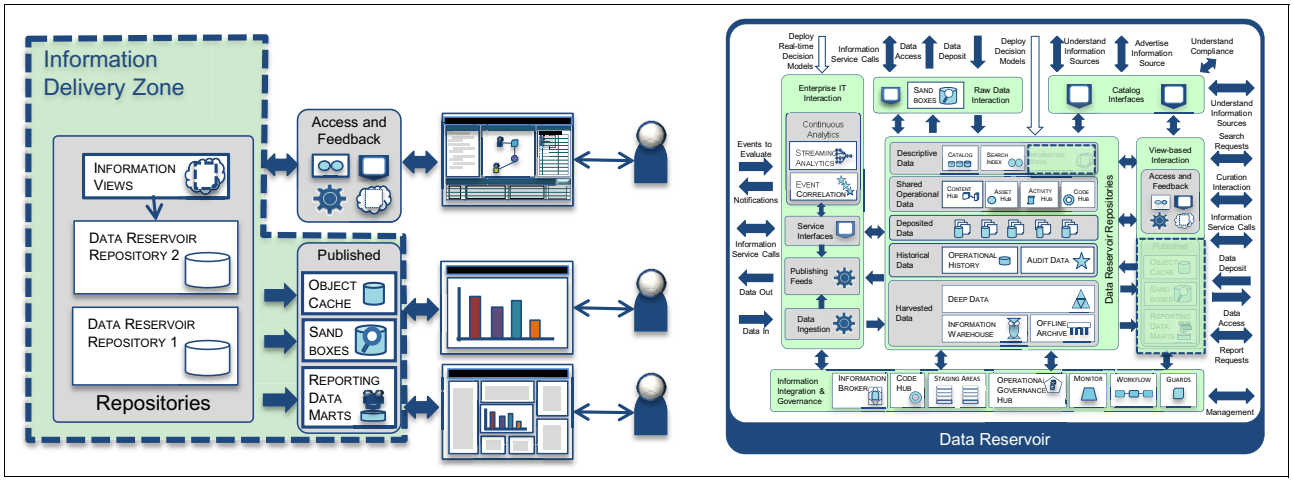


Figure 4-15 Information delivery zone

The deposited data zone

The deposited data zone is an area for the users of the data reservoir to store and share files. These files are stored in the Deposited Data subsystem.

The person depositing the data is its owner (until ownership is transferred to someone else). That person is responsible for the correct classification of the data in the catalog so that access control and other protection mechanisms operate correctly.

Deposited data can be accessed from both the Raw Data Interaction and View-based Interaction subsystems (Figure 4-16).

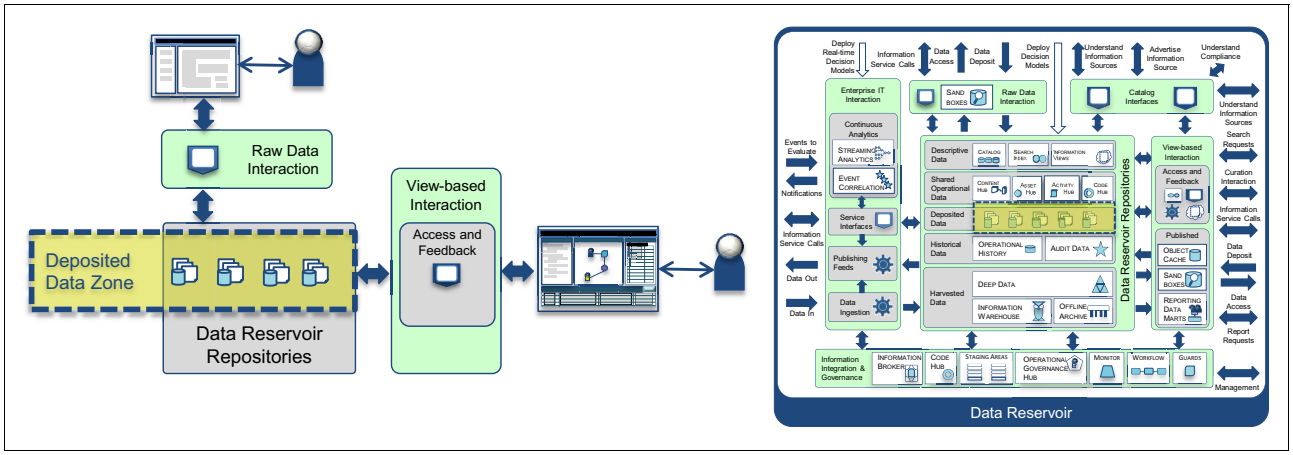


Figure 4-16 Deposited data zone

The test data zone

The test data zone (Figure 4-17) is in the deep data repository. It provides test data prepared for application developers.

- This data can be subsets of other information collections in the data reservoir that have been selected to test a broad range of conditions in new application code.
- It can be selectively masked or jumbled to hide personal information.

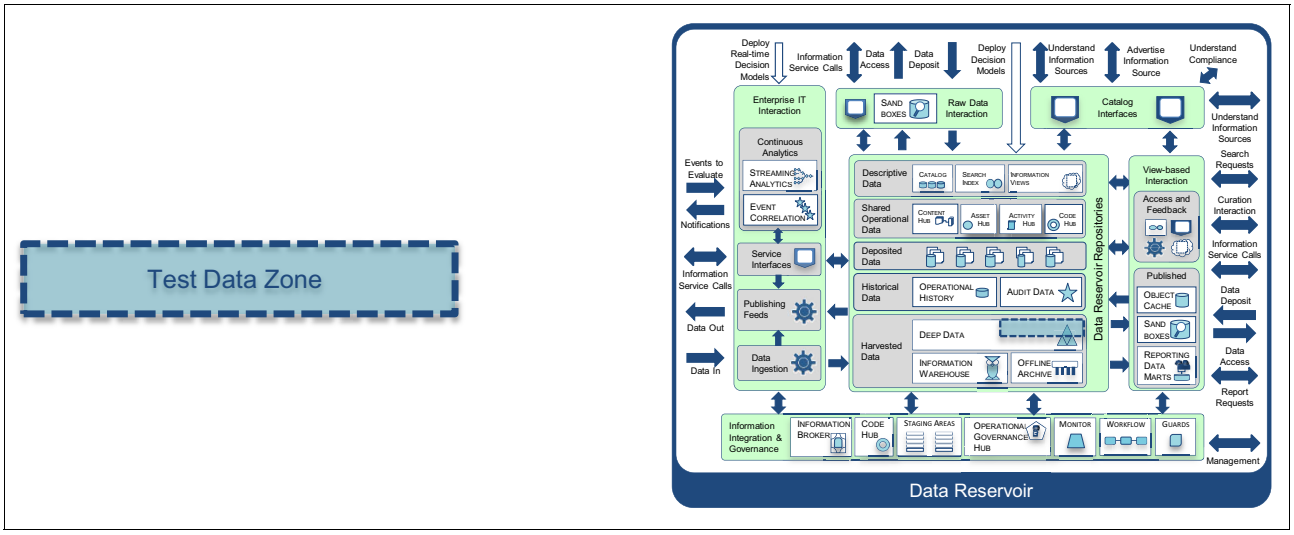


Figure 4-17 Test data zone

Analytics zone

The analytics zones show the data usage during the development and execution of analytics.

The Discovery Zone

The discovery zone (Figure 4-18) contains data that is potentially useful for analytics. Data scientist and experienced analysts from the line-of-business typically perform these steps:

- Browse the catalog to locate the data in the discovery zone that they want to work with.
- Understand the characteristics of the data from the catalog description.
- Populate a sandbox with interesting data.

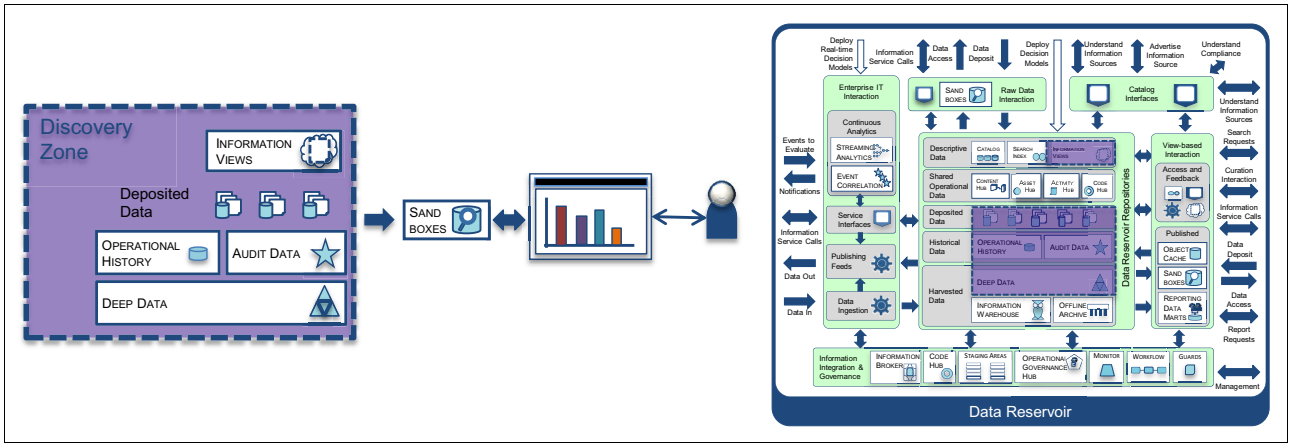


Figure 4-18 Discovery zone

The exploration zone

The exploration zone contains the data that the analysts and data scientists are working with to analyze a situation or create analytics. This data is stored in sandboxes that are managed by the data reservoir. These sandboxes are fed from the discovery zone.

The users of this zone browse, reformat, and summarize data to understand how a process works, locate unusual values (outliers), and identify interesting patterns of data for use with a new analytical algorithm. This is a place where data can be played with, so different combinations of data can be brought together in new ways to drive data mining. This zone can work with subsets of the data (for example, using different sampling strategies) to create the new combinations quickly and facilitate the iterative way of working that the data scientist needs when finding and preparing data, and training analytics models (Figure 4-19).

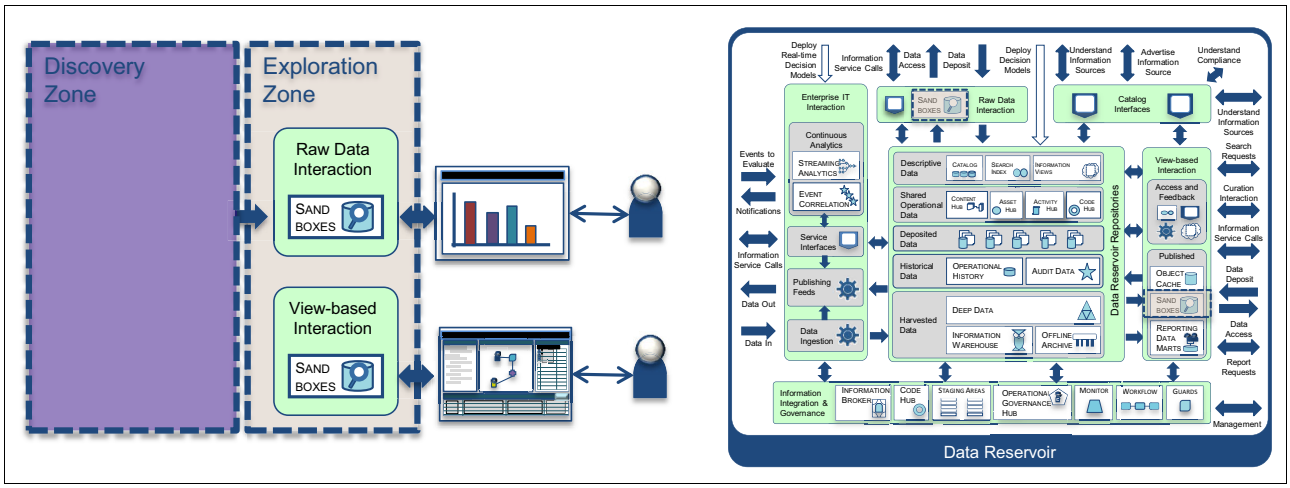


Figure 4-19 Exploration zone

The analytics production zone

The analytics production zone defines where the analytics production workloads are in the data reservoir repositories.

This can vary between each data reservoir. In the initial deployment, it might even be empty. As new analytics are deployed into production, the scope of this zone grows.

The value of this zone is in identifying where production SLAs needs to be maintained. See Figure 4-20.

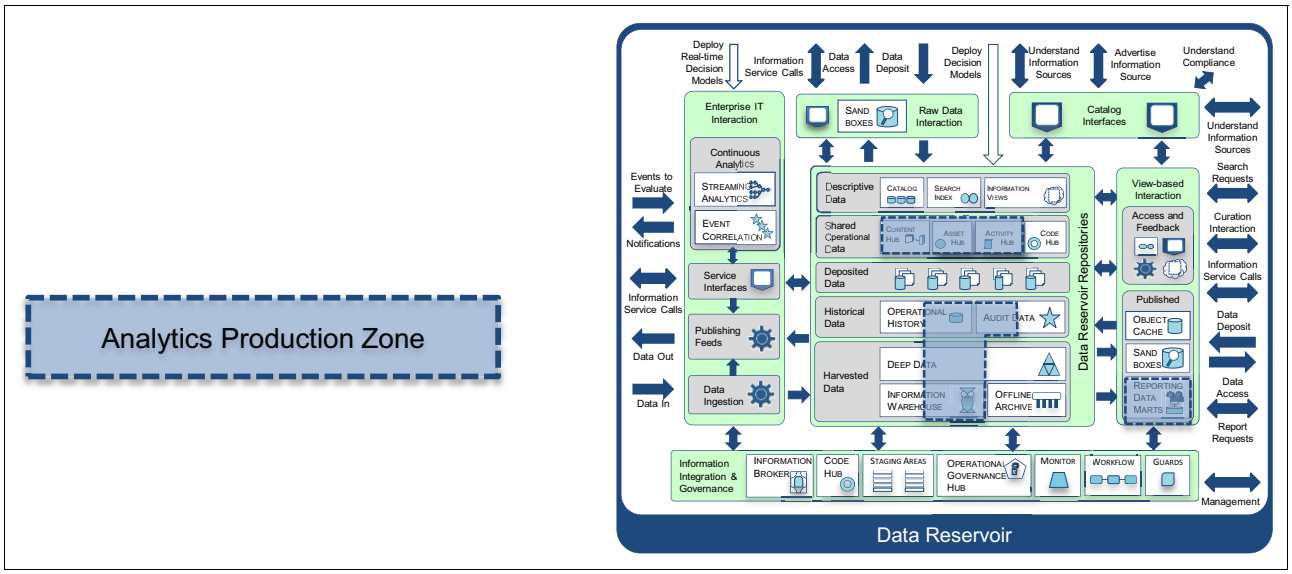


Figure 4-20 Analytics production zone

The derived insight zone

The derived insight zone identifies data that has been created as a result of production analytics. This data is unique to the data reservoir and might need additional procedures for backup and archive.

This data might also need distributing to other systems, both inside and outside of the data reservoir, for the business to act on the insight. Examples of this insight are the propensity to be fraudulent of a supplier or the propensity to leave the trial for a patent. See Figure 4-21.

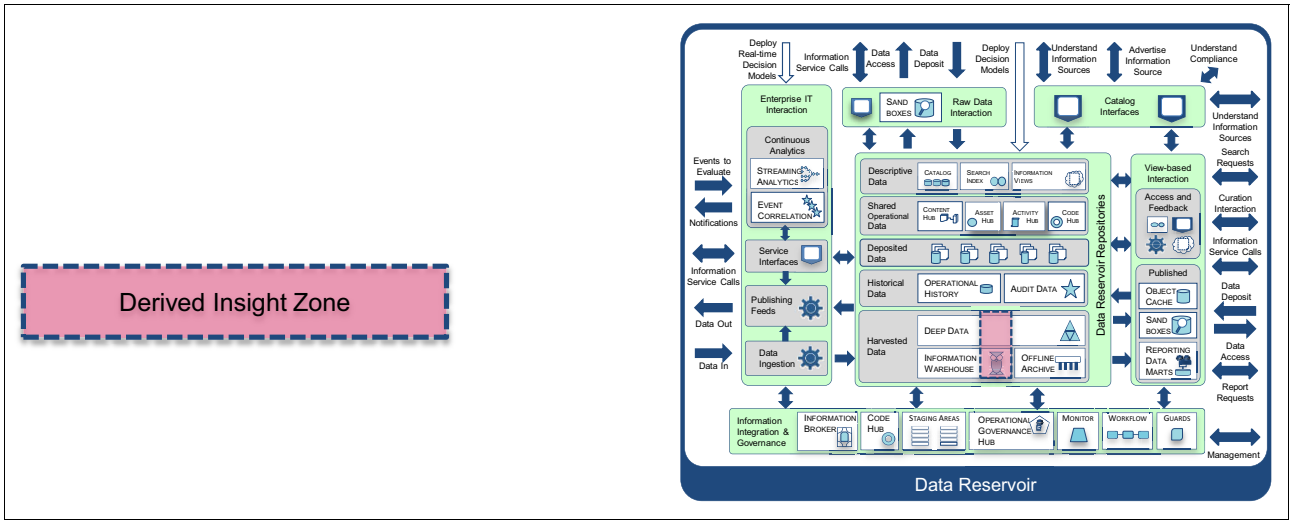


Figure 4-21 Derived insight zone

Information supply chains and zones

So how does the information supply chain fit with the information zones? The information zones show where the data used by each type of user is located. Thus, the information zones show which of the data reservoir repositories (destinations) that the information supply chains must serve.

4.4 Summary

Information supply chains are the flows of data from its original introduction in an IT system to its user. This chapter describes how to identify the subject areas that meet the needs of the use cases and users of the data reservoir. Next, the usage of the reservoir was considered with the information zones, followed by identifying data sources for the subject areas. Information source characteristics were reviewed. Matching up the usage and the sources allows you to identify gaps that the information supply chain fill. The various information supply chains that involve the data reservoir were described for enterprise data, metadata, deposited data, and audit data.

Master data can be an anchor used to hang information, handling identifiers and correlation centrally. The chapter covered information refineries, the information processes in the data reservoir. Finally, it described consumption of data from the reservoir by information services and information virtualization.

5

Operating the data reservoir

This chapter focuses on the day-to-day operational aspects of the reservoir. It explains how the workflows defined as part of the reservoir ecosystem can combine to provide a self-service model. This model helps ensure that data is governed appropriately, that curation is effective, and that security is managed in a proactive manner. Also covered is how a comprehensive collection of workflows allow you to gain further insight from both the operation of the workflow and the data within it by using monitoring. Insight gained from this can contribute to a significant savings in system burden for data stewards. It discusses how governance rules can significantly lighten the system burden of managing the data reservoir, and how monitoring and reporting can be used to ensure proactive management of the environment.

This chapter includes the following sections:

- Reservoir operations
- Operational components
- Operational workflow for the reservoir
- Workflow roles
- Workflow lifecycle
- Types of workflow
- Self service through workflow
- Information governance policies
- Governance rules
- Monitoring and reporting
- Collaboration
- Business user interfaces including mobile access
- Reporting dashboards

5.1 Reservoir operations

When examining a data reservoir from an operational perspective, it is important to also look at it from a governance capability perspective. Chapter 1, “Introduction to big data and analytics” on page 1 shows how the data reservoir provides a trusted governed source of data to an enterprise, avoiding the pitfalls of a data lake and minimizing the risk of creating a data swamp. This chapter covers the components of the reservoir supporting the governance of the data and underpinning the operation of the reservoir.

5.2 Operational components

There are a number of key operational components that relate to the operation of the data reservoir. These components are responsible for supporting the governance program associated with the reservoir and are key to self service.

Figure 5-1 describes the governance capabilities that exist within the data reservoir. The following are the governance capabilities:

- Workflow
- Business policies
- Governance rules
- Mobile and user interfaces
- Collaboration
- Monitoring and reporting

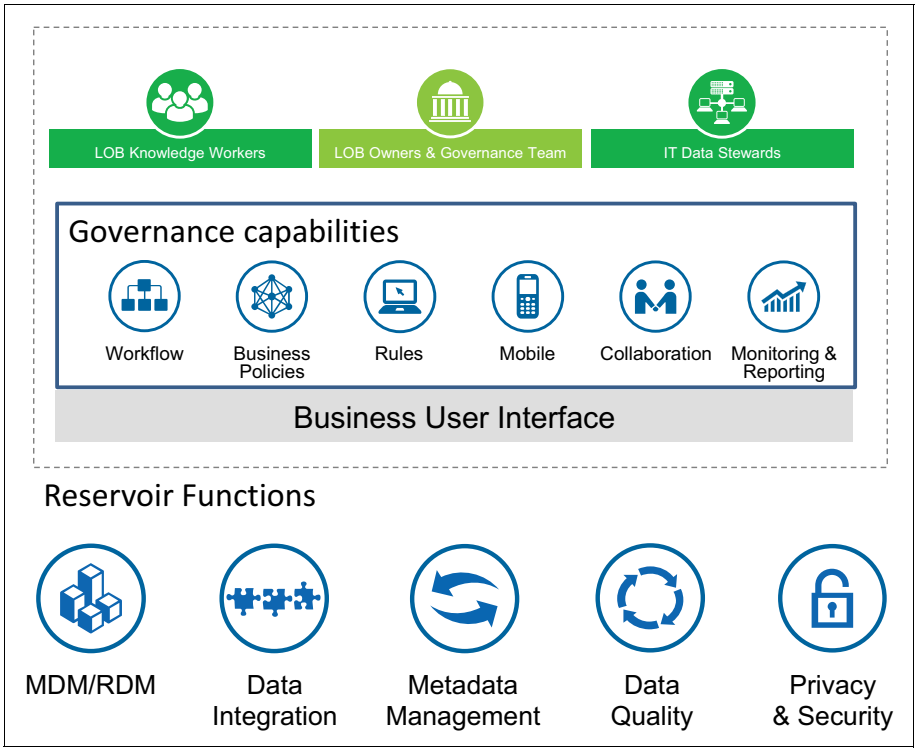


Figure 5-1 Logical architecture of the governance capabilities of the reservoir

These capabilities work together to support the underlying reservoir functions and its users. This chapter scrutinizes each of these operational components and discusses how they are

applicable to supporting the reservoir functions and the underlying operation of the data reservoir.

The governance capabilities provide a uniform set of features that break down silos between the lines of business users, the governance team, and the data stewards. They provide a mechanism to allow these groups of users to collaborate with each other, and support, contribute to, and extract value from the data that is managed by reservoir functions.

5.3 Operational workflow for the reservoir

Operational workflow defines the actionable steps that should be followed to achieve the wanted results. These steps are either system actionable or human actionable. This section takes a deeper look at the role of workflow in underpinning the operations of the data reservoir.

Workflow is essential to coordinate the computer and human activity that are required to operate and maintain the reservoir, and the data and metadata within it. Workflow not only controls the flow of the data within the reservoir, but also encourages individuals within the reservoir to collaborate to ensure that the data is accurate, useful, and secure.

The use of workflow provides these capabilities:

- Guides a user of the reservoir through a predetermined path to complete a particular action within the reservoir
- Automatically notifies users of the reservoir of actions that need to be taken for others to complete
- Provides a comprehensive set of historical data that can be used for building reports on the operational patterns of the reservoir

All of these capabilities make it a fundamental piece of self service and for the sustainability of the reservoir. At the highest level workflow provides a platform for self service that underpins how information is shared, governed, and used within the reservoir.

Each interaction a user has with the reservoir will typically be as a step within a predefined operational workflow. Workflow plays a role at the point where users share and use data from the reservoir and plays a vital role in ensuring the governance of the data while it is present within the reservoir. The point at which these workflows run fits into one of three groups.

Figure 5-2 showing the flow of a document through the reservoir and how workflow manages it throughout each group.

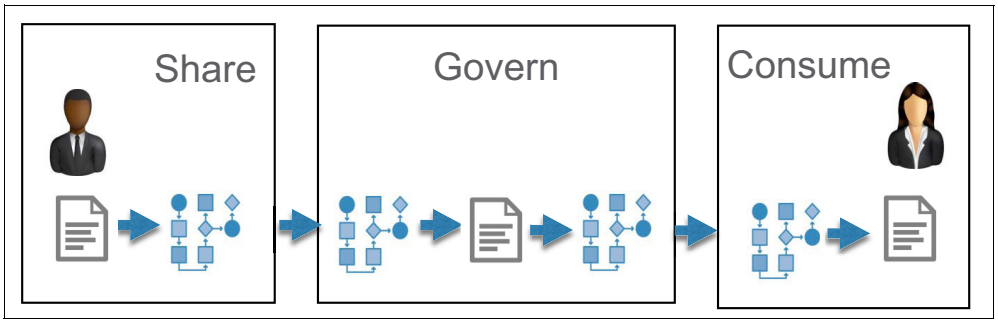


Figure 5-2 Flow of a document through reservoir with workflow

5.3.1 Share

Workflow can be used at the point at which data is shared with the reservoir:

- Coordination of the human and system actions that are required to prepare data for publishing into the reservoir. This activity ensures that the source is cataloged, the data it contains is properly classified, and all relevant approvals are in place before the data is published and *live* within the reservoir.
- Enforcing governance rules on the data whose classification requires action at the point of ingestion into the reservoir. Minimizing exposure to risk of the data not having appropriate rules applied.
- Ensuring that the data is classified correctly in the data reservoir repositories at the point of ingestion into the reservoir. This step makes the data immediately more discoverable and usable by reservoir users and assists with the ongoing curation of the data.

5.3.2 Govern

Workflow can be used to manage the data while it sits within the reservoir. Workflow can be used to apply the governance program to the data within the reservoir in an ongoing manner:

- Enforcing appropriate governance rules upon the data that exists within the reservoir such as retention lifecycles.
- Automatically bringing data back into compliance with governance policies or routing tasks to data stewards notifying them of policy exceptions and guiding them through the steps required to remediate or exempt the policy exception.
- Providing notification and automatic prioritization of work items ensuring the most critical policy violations are addressed first.
- Verifying and renewing exemptions to policies.

5.3.3 Use

Workflow also plays an important role at the point at which the data is used by the users of the reservoir by performing these actions:

- Enforcing rules at the point of consumption, for example ensuring that sensitive data is masked for certain users
- Allowing people to identify issues with the data or catalog, notifying a curator that changes should be made, and then supporting the curator in making those changes

5.4 Workflow roles

Each of the workflows that exist within the reservoir has a number of different personas that interact with it in various forms. Roles define the responsibility, management, and execution of the workflow.

5.4.1 Workflow author

A workflow author is a somewhat technical user responsible for the creation of the workflow implementation. This individual will likely be an experienced user of business process management (BPM) software, such as IBM Business Process Manager. They use the tools provided by the BPM software to model the workflow to provide an implementation of the business processes required to operate the reservoir. The workflow author receives requirements from the workflow owner.

5.4.2 Workflow initiator

A workflow initiator can be a human or a system within or outside of the reservoir that starts a new instance of a workflow, resulting in one or more workflow executors being required to complete the workflow. A workflow initializer can manually trigger a new workflow. For example, a user of the reservoir clicks a link or a system triggers a workflow on a schedule. However, it is also possible for a user to trigger a workflow without realizing it. For example, when taking an action against a piece of data that has a trigger assigned to start a workflow if the data is updated.

5.4.3 Workflow executor

A workflow executor can be a human or a system within the reservoir that has a step within a workflow assigned to them for action. A step that is assigned to a human can be referred to as a task. Often tasks are displayed within a task list or an inbox for action.

5.4.4 Workflow owner

The workflow owner is typically an individual or group of individuals responsible for the operation of the workflow. They have intimate knowledge of the intent of the business process and the problem it is solving. They will typically monitor the effectiveness of the workflow in achieving its required objectives and be responsible for the change management of the workflow. The workflow owner will likely be responsible for a number of processes within the reservoir, which is aligned to their area of responsibility or domain expertise. They are responsible for producing reports for their collective workflows up to their stakeholders.

5.5 Workflow lifecycle

The workflows defined within your data reservoir should continue to evolve with the usage of your data reservoir. What seems like a comprehensive workflow supporting the operations of your reservoir during the initial rollout will require continuous iterative improvements. Doing these improvements ensures that the reservoir will continue to provide the benefits required to support the ongoing operations and growth of your environment. For this reason, a lifecycle can be defined for the workflows. The workflow lifecycle defines the phases that each workflow can go through to support the ongoing operations of your environment. Each workflow iterates many times through the lifecycle, each time evolving to suit the changing requirements of the reservoir.

Figure 5-3 shows the workflow lifecycle's these phases.

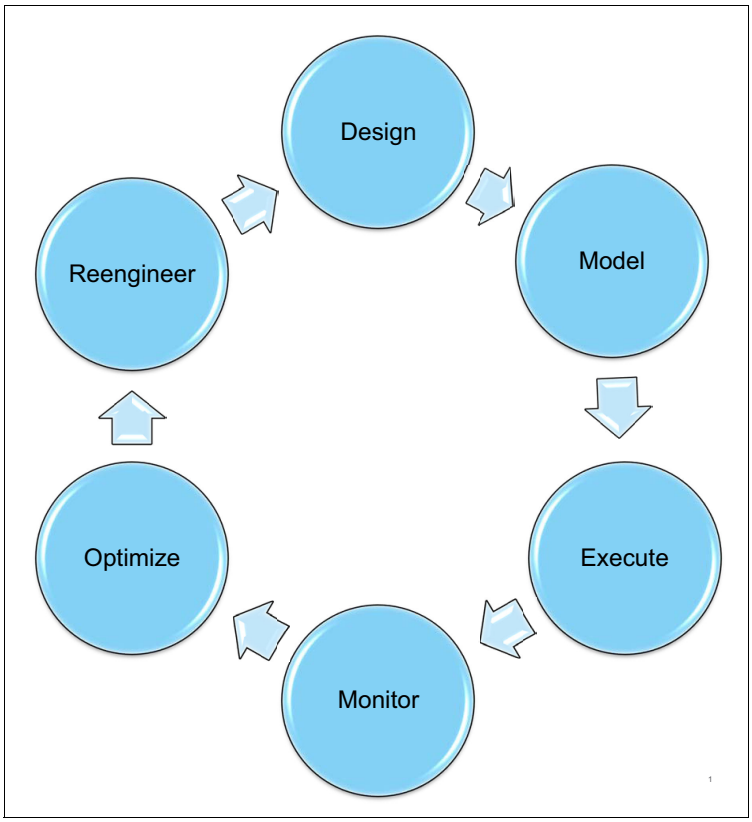


Figure 5-3 Workflow lifecycle phases

The following are the workflow lifecycle phases:

- Design phase

  The design phase allows the workflow owner to evaluate the purpose of the workflow and the requirements that need to be satisfied. They will typically use reports that have been generated from the *Monitor* phase to allow them to determine how the workflow can be fully optimized and determine how the new workflow fits into the existing workflows within the system, while addressing the requirements and how the workflow can best support the operation of the reservoir.

- Improve phase

  The improve phase allows the workflow author to translate the requirements from the workflow owner as laid out on the *design* phase. The workflow author uses a process design tool build the workflow artifacts. After they are built, these artifacts will go through the software development lifecycle steps and receive signoff from the process owner before being deployed.

- Execute phase

  The execute phase consists of the deployment of the new workflow to the reservoir and its execution within it. At the point of deployment, the new workflow is enabled within the system. If the workflow is a newer version of an existing workflow, then the new workflow replaces the old workflow. The users and systems of the reservoir become workflow executors, and use the workflow to complete their operations within the reservoir.

- Monitor phase

  The monitor phase is perhaps the most important of the phases within the lifecycle. It is in this phase that the workflow owner is able to monitor the effectiveness of their workflow. Each instance of the workflow gathers statistics within its data warehouse or can be added to the reservoir, capturing information about the data that the workflow is operating on and the users that are operating on it. Custom reports can be created to allow a workflow owner to monitor the effectiveness of this workflow, identify potential problems with the workflow or the operation of the reservoir, and provide a platform for analysis. These reports can then be used to improve the workflow in the future. For more information, see “Monitoring and reporting” on page 123.

- Optimization

  Optimization is the iterative improvement of a workflow instance while it is live within the system. The workflow author in partnership with the workflow owner identifies improvements that can be made to the process. They do this through changing configuration parameters, optimizing performance characteristics, and using the reports from the monitor phase to identify areas that can be revisited in the reengineering phase.

- Reengineering

  After optimization is complete and no further improvements can be identified with the existing workflow implementation, the workflow author works with the workflow owner to further refine the process. At this stage, a significant amount of information is stored within the workflow data warehouse that can be used to improve the components of the workflow. They typically identify human-centric steps that can be streamlined and where possible fully automated through governance rules. As part of the reengineering phase, superseded workflow implementations are versioned and archived.

5.6 Types of workflow

As previously stated, workflow underpins the operations of the data reservoir. A reservoir has many workflows that support the various functions, operations, and users of the reservoir. Typically, these workflows can be categorized into five distinct areas:

- Data quality management
- Data curation
- Data protection
- Data lifecycle management
- Data movement and orchestration

5.6.1 Data quality management

Responsible for the quality of the data and the usage of the data, these workflows play a significant part in the data governance program of the organization and enforcing that program on the data within the reservoir. Typically responsible for the enforcement of the rules defined as part of the data governance program, these workflows are responsible for handling policy exceptions.

These workflows are responsible for enforcing data quality, by capturing that data is in violation of a particular validation rule. The workflows either automatically correct the data quality issue or notify a subject matter expert of the data quality issue prioritizing the issue. The workflows also allow the individual or business unit to correct the issue. These fixes can be on the data within the reservoir or on the data held in the original source.

Governance rules and collaboration are key to ensure that the workflow is efficient at routing tasks to the correct individuals for action and ensuring that the individuals are able to make the correct decisions in minimal time.

Data quality management workflows provide a mechanism to facilitate adherence to regulatory laws. For example, workflows can be built to ensure that *right to forget* laws are enforced automatically without assigning a team specifically to this task. Reports that are generated from this workflow can record compliance to this legislation on an automatic schedule.

Governance policy violation

An example of a workflow supporting data quality and compliance is one where a rule is tied to a governance policy stating that any customer record under the age of 18 should have a guardian associated to it. Data that violates this rule would trigger a workflow. On validation that the rule had been violated, the workflow tries to correct the data automatically, based on actions taken on similar types of tasks and data. If that does not work, the workflow creates a task for an individual to manually inspect the data and provide them a mechanism to bring the data back into compliance with the policy so that the rule is no longer violated. This activity can either be within the reservoir or at the original source.

Figure 5-4 shows an example of a simple workflow supporting the data quality management of the reservoir. In this example, data is imported into the reservoir and governance rules are used to validate the data of a particular classification. If a rule fails, then a remediation task is created for a data steward, who can then use the remediation user interface to bring the data back into compliance or record an exemption.

After the data has been corrected, it is then revalidated against the governance rules. Finally, after the data is compliant with the rules, the compliant data is saved to the data reservoir.

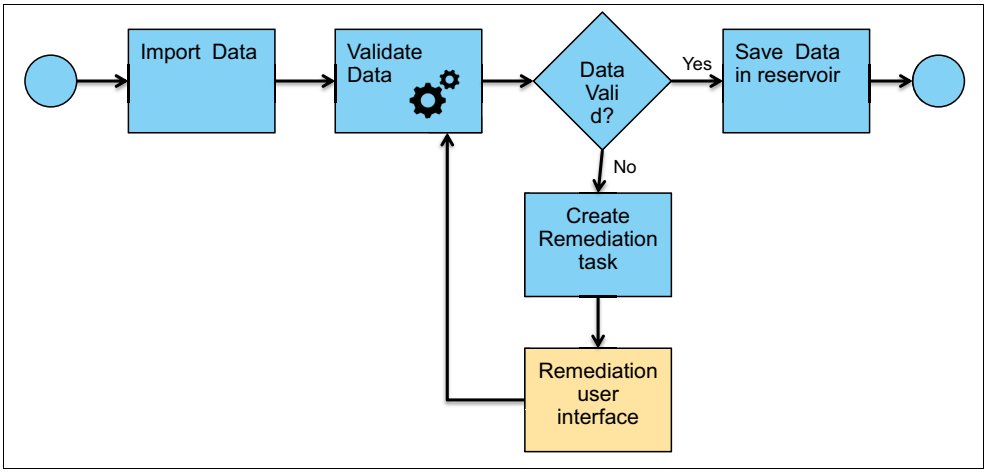


Figure 5-4 A simple workflow

Data quality management workflows typically run in each of the operational workflow locations. For more information, see “Share” on page 108, “Govern” on page 108, and “Use” on page 108. Multiple workflows are used to test data quality either on the way into the reservoir, from within the reservoir, or at the point the data is used from the reservoir. The specifics of where the governance policy is enforced are down to the specific industry, regulatory compliance requirements, level of information maturity, and type of data within the reservoir.

5.6.2 Data curation

Data curation is the *touching* of the data by users of the reservoir. Typically data curation manipulates the catalog data rather than the actual reservoir data itself. Data curation is an important activity to maintain the currency of the metadata and make it more findable, descriptive, and relevant to users of the reservoir. Without data curation, as the volumes of data grow, the challenges to find relevant data, the risk that sensitive information might inadvertently be exposed, and the risk of finding data that has now gone stale are significantly heightened.

Workflow can support data curation and encourage effective data curation. A reservoir deployment includes a number of workflows to support the data curation activities that take place by the users of the reservoir. Done correctly, workflows allow data curation to become business as usual, encouraging users to curate information as they go, which results in up-to-date metadata to support the rapid and effective locating of data within the reservoir.

Invalid tagging of data

Workflows supporting data curation are typically started by an individual rather than being started by a system. A typical example would be when a users are searching for data, they might find that some irrelevant results have been returned. Further investigation of the returned data shows that it has been tagged incorrectly. The user can initiate a curation workflow stating that the data has been tagged incorrectly. If permissions allow, the workflow either allows the user to correct the error themselves, or creates a task for an information curator to investigate and correct the issue.

Figure 5-5 shows a simple workflow that supports one curation activity within the reservoir. In this example, a user of the reservoir clicks a link that starts a workflow due to invalid metadata being discovered. This workflow creates a task for an information curator to investigate and notifies that person of this new task. The information curator then makes any required changes to the metadata before saving the changes back into the reservoir. A real world example might bring in rules to determine which information curator a particular curation activity should be routed to or determine whether the user themselves can make the change to the metadata. However, even in this simple example, you can see how the workflow provides the mechanism to notify the curator of a task and provide that person with the mechanisms to correct the metadata.

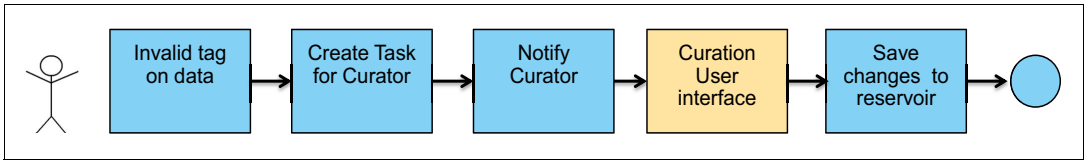


Figure 5-5 Stepping through a simple workflow

Curation workflows can also provide a level of control over the curation activities. Workflows can be created to enforce a safeguard against curation being performed on particular types of data or curation activities being performed by certain groups of users. For example, governance rules could be used as part of a curation workflow to state that any metadata change performed by a set of novice users on data classified as being sensitive must have the change approved by the data owner. The integration of workflow and governance rules facilitates this capability and underpins the ongoing effective curation and the efficient usage of the data reservoir.

5.6.3 Data protection

Protection workflows support the operation of the reservoir by monitoring the operations of the reservoir and automating the initialization of steps to respond to incidents as they occur, ensuring that security and compliance officers are notified. Data protection workflows are typically started automatically by the system based on user behavior rather than being started directly by a user of the reservoir.

Data protection workflows rely on the collection of information about user interaction with the reservoir, analyzing that data and applying governance rules that provide thresholds as to when certain actions should be investigated by the security team.

Data protection workflows cover a broad spectrum of use cases underpinning the operation of the reservoir, of which these are two different examples:

- Being of potentially non-malicious intent
- Being of potentially malicious intent

Both of these scenarios are important to build trust in users of the reservoir. It is important to have a robust set of workflows to supporting the secure operation of the reservoir so that users know that their sensitive data will be maintained and there are safeguards in place continuing to enforce these rules while the data exists.

Classification violation

In this scenario, new data has been added to the reservoir, and the data has been classified as *sensitive*, *classified*, *confidential*, or *private*. However, the data does not have any security associated with it. This document has been published to the reservoir and has full public access to all users of the reservoir by mistake. A workflow can be used at the point of ingestion (shared) to the reservoir to recognize that the tags associated with this document indicate that security should be applied. The workflow can temporarily lock down this document so that only the document owner can see this document and notify them that they must validate that the security permissions for this document are correct.

Figure 5-6 shows a simple implementation of this scenario. The workflow handles the import of data into the reservoir and analyzes its metadata to determine whether the document has been classified as *sensitive*. If it has, the workflow checks to see whether the permissions associated with sensitive data have also been applied. If they have not, the data is locked from view within the reservoir and the information owner is notified about the security exposure that needs to be rectified.

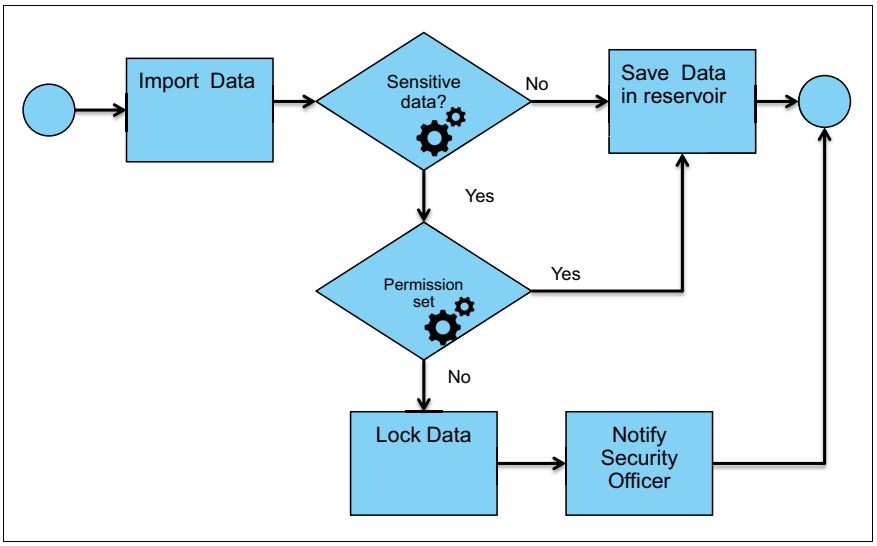


Figure 5-6 Workflow for sensitive data processing

Workflow also allows this security to be enforced on documents that are already within the reservoir. If sensitive information exists within a document, a curator tags the document as *sensitive*. In this scenario, a workflow could be triggered to recognize that this tag now exists but the document is still available with no security enforced on it. A workflow could notify the document owner and request that they take immediate action to rectify.

Malicious data discovery

In this scenario, a disgruntled employee is searching to see what he can find of interest that he could use for malicious reasons to harm the company. He begins by searching for financial information, looking for employee salary information and budgeting information. Next, he searches for information related to special projects that are due to be worked on by the research department and download large volumes of information about these projects. These searches by themselves could be completely innocent and provide no significant security exposure. However, when looked at as a whole the searches could allow a user to determine the budgets being assigned to an important new project, the individuals working on it, and the remuneration package they are receiving. This information could be taken by the employee to a competitor, exposing the project, the budget, the key people, and how much it might cost them to hire the employees away. This situation provides the potential for unmeasurable damage to the research project and the company's future profitability.

Using a combination of audit mechanisms to track what data users are accessing and downloading, governance rules to capture when certain usage patterns or thresholds are met, and workflow to initialize tasks and notify a security officer to provide further investigation, these events can be investigated quickly.

These events can be investigated quickly by using these capabilities:

- Audit mechanisms to track what data users are accessing and downloading
- Governance rules to capture when certain usage patterns or thresholds are met
- Workflow to initialize tasks and notify a security officer to provide further investigation

In this scenario, a pattern of activity, defined within an analytical model would have been met to state that this individual had been downloading large amounts of new project data and was looking at *sensitive-financial* data. A task would be sent to a fraud investigation officer immediately identifying that the threshold defined by the business rule had been met. This action allows the fraud investigation officer to begin an investigation and monitor the activity of the user. The speed at which this can take place would mean that the fraud investigation officer would be able to catch the perpetrator in the act and limit the risk of harm to the company. The combination of workflow and rules provides a level of self service to the effective operation of the reservoir. The workflows and rules can be continuously iterated on to tailor the thresholds to match the type of data and threat level associated with it.

5.6.4 Lifecycle management

Workflows supporting lifecycle management typically occur on data within the reservoir rather than at the point of ingestion into or output from the reservoir. They are responsible for managing the lifecycle of all data within the reservoir. This includes the lifecycle of the data and the lifecycle of the governance definitions such as the policies, rules, workflows, and reference data required to support the operation of the reservoir.

Note: Lifecycle management workflows manage both the lifecycle of the data and the governance definitions associated with the reservoir. This is important to maintain the artifacts that support the governance program.

Lifecycle management workflows are key to maintaining the relevance of the data in the data reservoir repositories. They support the mechanism to request changes, make changes, and approve the results. Data that has expired is removed from the reservoir.

Workflows support the evolution and maturity of the data reservoir definitions. Policies evolve or new ones are required due to various factors:

- Changes in company structure
- Changing regulatory requirements
- Expanding scope of the data reservoir

For these and other factors, workflow provides lifecycle management to ensure that approvals are enforced to manage changes that are made. This scenario applies to all aspects of the governance program, whether they are policies, rules, reference data, or the workflows themselves.

Reference data lifecycle

Imagine a scenario where a large multinational organization has implemented a data reservoir. The data reservoir is accessed by large numbers of users across the entire span of the organization's operating geographies. The data reservoir includes a reference data set specifying country codes. This reference data set is used by other systems and documents that are critical to the financial reporting, organization structure, and budgeting of the organization.

During a period of political unrest, a disputed territory that is claimed by one of the countries that the organization operates in is unofficially renamed by that country. Individuals within that

country now refer to it by its new name. Individuals outside of that country refer to it by its original and legal name. An individual resident of the country claiming the disputed territory might choose to update the reference data set to represent the new name for the country, even though this change is not recognized by the organization nor by the United Nations. Making this change could cause a huge amount of disruption to the data. This disruption could result in incorrect financial reporting and potentially damage the brand of the organization should reports be released that name the disputed territory under an unofficial name.

Figure 5-7 describes a simple workflow that could be used to enforce lifecycle management on a piece of reference data. The workflow is initialized when a user requests a change to a piece of reference data. This action creates a task for an approver and notifies them of this new task. The approver starts the task and is presented with a user interface that allows them to determine what the suggested change is and what the reason is for the change. If the approver approves the change, then the reference data is updated within the reservoir. If the change is not approved then the requester is notified that their change has been rejected and the reason for the rejection.

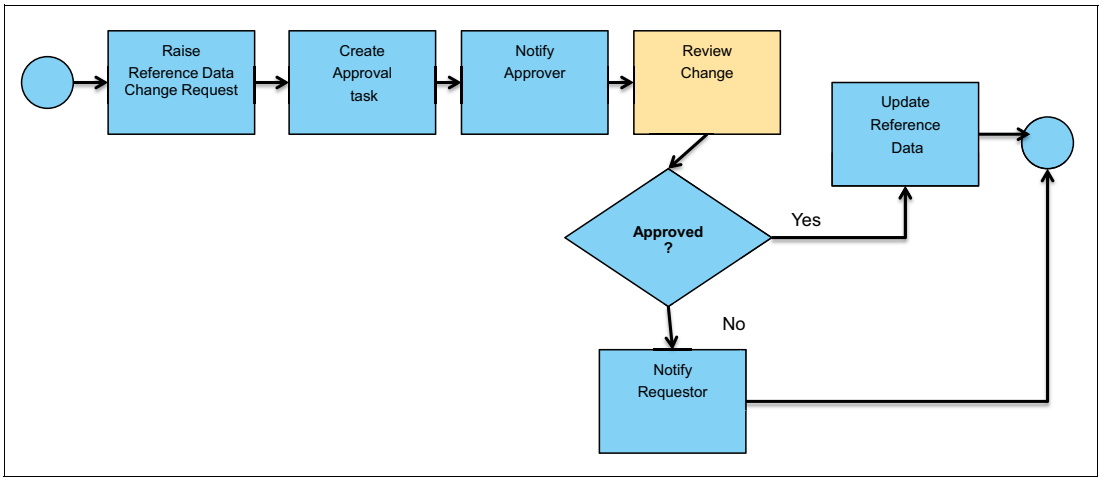


Figure 5-7 Workflow enforcing lifecycle management

Workflow supports the lifecycle management of the reference data. In this scenario, a workflow would have been initiated on the change to the reference data set. This situation would have routed an approval notification to one or more individuals, requiring them to approve the change before the change was implemented, stopping the potential damage to the organization's brand. This level of management can be applied to data and governance program components through workflows. These workflows allow data custodians to be confident in sharing their data in the reservoir and allowing it to be used in a controlled manner, without putting the company or the data at risk.

5.6.5 Data movement and orchestration

Closely linked to the lifecycle management type workflows, data movement and orchestration workflows support the movement of data as it is fed through the reservoir. Workflow can be used to define and control the various paths that need to be followed to ensure that the data is moved between the various groups within the reservoir. This approach ensures that the data is in the correct format and shape for consumption.

Lifecycle management type workflows typically support the *state* of the data, such as whether it is published, stale, awaiting approval, or private. Data movement and orchestration workflows define the steps in which the data must flow through as it interacts with other

services provided within the reservoir. These services can change the format of the data, change the classification of the data, mask the data, or merge the data with other reservoir data.

Workflows that fall under this category are typically started either by a user, a schedule, or pushed from the data source. They can be started at indigestion to the reservoir or on data that is already within the reservoir. Workflows of this type prepare the data for consumption by users of the reservoir. It is unlikely that workflows falling under this category will be started by the consumption of the data from the reservoir, because the processing should have already been completed.

As with the other workflow types, governance rules add an important layer of intelligence to the operation of these workflows. Governance rules determine the transformation that is required on the data depending on its classification. Through the governance rules, these workflows automatically apply the correct rules to the data and ensure that the data is fit for use by the users of the reservoir.

Lifecycle management: Lifecycle management workflows support the state of the data, whereas data movement workflows define the paths that the data should follow to affect the state**.**

Sensitive data ingestion

Imagine a scenario whereby the payroll department of a large organization needs to upload a data set to the reservoir that contains employee payroll information. The information within this data set is considered extremely valuable to select individuals within the organization and therefore would be valuable to be shared. However, to ensure that employee privacy is maintained, elements within the data set such as employee name, employee ID, and bank details are masked when data scientist work with the data. Workflows have been configured within the reservoir to route data through a data masking service at the point of ingestion into the reservoir.

Figure 5-8 shows a simple implementation of such a workflow. As the data is imported into the reservoir, governance rules determine whether the data might be payroll data. If it is, the data is automatically sent on to the masking services, which have been configured to strip out the employee name and ID values. The masked data is then saved into the reservoir.

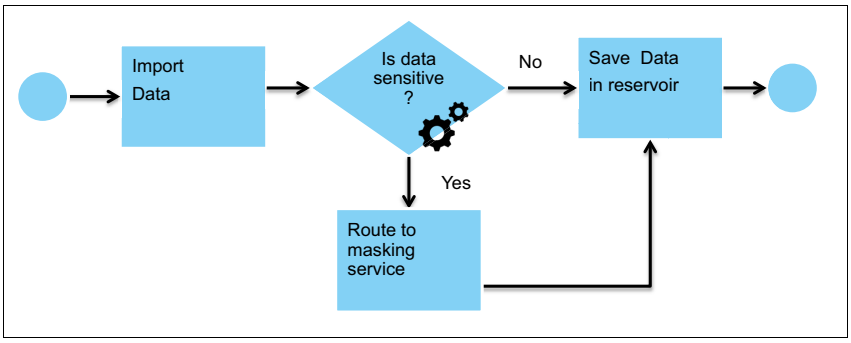


Figure 5-8 Workflow with data masking

In this scenario, the user importing the data does not even need to know the requirements that need to be enforced on this data. The workflow is responsible for ensuring that the sensitive data masking is an automated process and does not require a human to remember to mask the data. The workflow can be initialized in two ways:

- Under a user initialized implementation, at the point of upload to the reservoir the user can select which elements of the data contains sensitive information. The workflow ensures that these data elements are fed through the masking service.
- System-initiated style, the workflow triggered interrogates the catalog to discover the data's classification. Using the governance rules within the workflow, decision points determine whether the data should be routed through the masking service or not.

Periodic checks can be made on the classifications of data in the data reservoir by surveying workflows that seek out particular types of sensitive data.

5.7 Self service through workflow

“Types of workflow” on page 111 showed how workflow provides the operational underpinnings of the data reservoir. Workflow is the component that ensures these items:

- Data is secure
- Governance program is enforced
- Data is findable, valid, and useful

Without these important elements, the reservoir will become untrusted, not useful, and in the worst case a liability and open to abuse and security violations.

Maintaining these operational aspects of the reservoir can be a time consuming role. This challenge increases as the number of users of the reservoir grows, the skill levels, experiences, and expectations of the users broaden in scope and as the volume of the data increases. Automating these workflows provides a mechanism to manage this challenge as the operational dependencies increase. The number, size, and complexities of these workflows can grow as the number of users and volume and type of data within the reservoir grows. It is typical for a few key workflows to exist in early stages of a data reservoir. These workflows typically include a larger number of human-centric steps to be done as part of the workflow. Over time as the reservoir grows, more workflows will be required, and new and more complex rules can be developed to support a much more automated set of workflows. This capability significantly reduces the maintenance burden of the reservoir.

For users of the reservoir, operation workflows provide the self-service mechanisms that they require to extract value from the data. Business users want information now, they want to be able to accomplish these goals:

- Find data quickly
- Trust that the data is accurate and up-to-date
- Ensure that their data is secure
- Be able to overcome issues quickly and without assistance

Ensuring that workflows are provided to cover these categories is key to providing the self service that is expected by today's savvy users.

5.7.1 The evolution of the data steward

The shift towards the growth in the understanding of the importance of the data reservoir is occurring at a time when the role of a traditional data steward is also shifting (Figure 5-9). These two dramatic shifts are complimentary and are important to each other to aid self service.

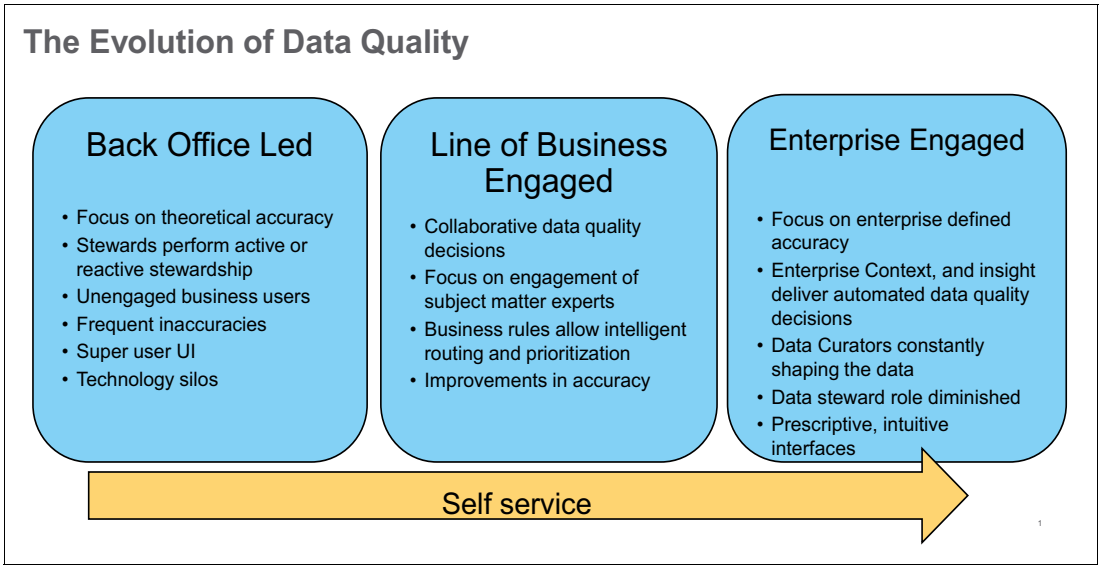


Figure 5-9 Data steward

Traditional data stewardship

In some organizations, data stewardship is solely handled by a team of data stewards who belonging to the back-office IT department. In the past, individual data steward roles were called out by an organization, and they managed the data and ensure its accuracy. These operations were performed from a perspective of theoretical accuracy with no real understanding of the customers, products, and accounts that the data was describing. Worse still was that there was little engagement with the lines of business users who understood the customers, products, and accounts that the data was describing.

Data quality decisions were being made by a group of individuals that understood the function of the data rather than the data itself. This situation led to data stewards contributing to the inaccuracies of the data. It also meant that all data quality issues discovered within the data became a bottleneck within an organization. Data quality tasks would be added to a data steward's backlog quicker than they could correct them. In some organizations, it was not uncommon for 10,000 new data quality tasks to be created each day. A team of data stewards with a capacity of completing 500 tasks a day could never be able to address this ever-increasing backlog. It became *OK* that these tasks were never going to be resolved. This led to significant staff retention issues across the data steward organization, but also created a huge number of data quality issues that were never going to get fixed, leading to low data quality and low trust in the data. Applying this mechanism to stewardship of data within a data reservoir, where the volumes of data and usage of the data is typically very large, causes significant problems when managing the reservoir. This circumstance causes large amounts of maintenance that must be performed and the risk of bad business decisions being made from the inaccurate data.

Imagine a scenario where a data quality issue has arisen on an important piece of master data that is held within the data reservoir. In this scenario, a duplicate customer record has been found for a *gold customer*. Gold customers are the organization's top 5% revenue

generating clients, and are therefore eligible for special offers and dedicated account management. The data steward notices that the duplicate record does not have the gold customer flag set and includes a new contact listed as chief procurement officer. The data steward collapses the records by moving the new contact into the master record and moves on to the next task. The steward has done the job correctly by correcting the data and removing the duplicate records within the system. However, the data steward has no understanding of the impact of the new contact being added to the customer record. Because this new contact is the chief procurement officer, this new individual could have a significant influence over the future revenue stream from the customer. Traditional data stewardship allows the data quality issue to be corrected, but does allow you to address the underlying business opportunity that might exist.

Line of business engaged stewardship

More recently there has been a shift to increasing avenues of collaboration between data stewards and the line-of-business knowledge workers. Data stewards are typically still involved in managing large volumes of data quality decisions. However, lines of business users are now also expected to actively contribute to or own the data quality decisions within their domain of expertise. In some scenarios, the task will still be routed to a traditional data steward. However, the tools that are used by that data steward allow them to easily identify the line-of-business subject matter experts (SMEs) who are associated with that data. This feature allows them to engage and collaborate with the SMEs in real-time to ensure that the correct decision is made upon that data. In an increasing number of scenarios, the data quality task is routed directly to the SME for that data. Workflow enables the task data to be interrogated and a business rule applied to route the task to the correct SME for processing. After the line-of-business SME has corrected the data, the updated data can either be persisted directly to the data source or can again use the workflow to notify a data steward that this change needs to be approved before it can be applied to the data source.

The ability to use workflow and apply governance rules to determine whom a task should be routed to for processing significantly lightens the workload for a data steward. This action cuts down the number of tasks being routed through the data steward's bottleneck. It also puts the data quality decision directly into the hands of the subject matter experts for that data, allowing for faster, more informed decisions to be made. Within a data reservoir, the volumes of data and users can become very large. Having a robust mechanism to route the tasks to the subject matter experts for the data and removing the bottlenecks caused by a back-office data stewardship team is extremely important to maintaining control over the data.

Looking back at the data stewardship example about the our customer, using line of business engaged stewardship, rather than adding yet another task to the data stewards backlog, allows you to apply governance rules to the workflow. The workflow interrogates the data and determines that this suspect record relates to a gold customer. The workflow then routes the task directly to the gold account customer service team. The gold account customer service team (as the SMEs) apply their domain knowledge to determine that this new chief procurement officer should be contacted immediately to build a relationship and inform them of the special service and preferential rates they are entitled to as a result of their gold status. It is noted that this new procurement officer is replacing an individual who moved to another company. You can then update the customer record to remove the individual as a contact and contact that individual at their new company and inform them that you would like to continue your relationship and help them achieve gold customer status. By more effectively managing the burden of operating the data reservoir (by engaging the SMEs in the data quality decisions) the company is able to make much more effective business decisions and provide the potential to drive additional revenue.

Enterprise engaged stewardship

An emerging trend in this space is a further shift to fully enterprise-led stewardship. In this scenario, the lines of business are fully engaged in owning the data quality decision. Data quality tasks are automatically routed to the SMEs in the lines of business. Governance rules are heavily used to ensure that data quality decisions are being routed to the correct individuals. It becomes the responsibility of the lines of business to manage the data quality issues. Comprehensive workflows coordinate efforts across individuals and lines of business, ensuring that approvals, escalations, and critical path management are accounted for. The lines of business are responsible for ensuring that data quality service level agreements (SLAs) are met for the tasks that they are accountable for. Reports are provided to help the line-of-business managers monitor the effectiveness of the team in managing their data quality responsibilities. Simple, prescriptive user interfaces provide the capability for the lines of business to manage their tasks and provide the reports on the task resolution statistics to which they can be held accountable.

Under enterprise engaged stewardship and particularly when implemented as part of the operations of a data reservoir, data curation is an important aspect in supporting the shift to enterprise stewardship. It can feel uncomfortable when shifting the responsibility for the data quality away from the traditional stewardship organization and into the hands of the lines of business. Data curation plays an important role in easing this concern. Within a reservoir, information curators are engaged to constantly touch and shape the data to support the enterprise context. Users of the reservoir are constantly updating, tagging, flagging, and rating the data that they are using. These curation activities contribute to the quality and usefulness of the data as an ongoing concern. The expertise that becomes absorbed within the data as a result of this constant curation, coupled with the data quality decisions being made by the SMEs within the lines of business, can result in the level of data quality being *good enough*. Data stewards will still exist, but they are likely to be focused on a specific set of tasks that require back-office IT involvement rather than subject matter expertise. It is not uncommon that a data steward will still be required to approve certain changes to the data made by a line-of-business user.

5.8 Information governance policies

The information governance policies defined as part of the governance program also describe the manner in which the data reservoir operates. Collectively the policies provide a set of operating principles for the data within the reservoir and the users of the reservoir.

The information governance policies are stored within the catalog and provide a hierarchical view of all of the policies defined for the reservoir. It holds the low level department policies and how they roll up into high level corporation wide policies or industry regulatory policies.

The policies defined within the catalog are owned by the Chief Data Officer or governance leader. They have the responsibility for ensuring that the policies are enforced within the reservoir through the rules and classifications.

The catalog provides the centralized point for housing the various implementation components of the governance program and the operation of the reservoir. From the catalog it is possible to drill down from the policies, into the specific rules that implement those policies for each classification, and then into the operational workflows that execute the rules. The catalog provides the lineage of data throughout the operational components of the reservoir. Operational workflows such as those defined in the section “Lifecycle management” on page 116 provides change request management on artifacts within the catalog.

5.9 Governance rules

Governance rules provide the functional implementation of the policies for each governance classification. It is the information governance policies that provide the logic that needs to be executed when users or systems interact with the data reservoir.

Rules can be implemented in a rules engine such as, IBM Operational Decision Manager. However the rules will be referenced to from the policies that are enforced by these rules from within the catalog. The linkage between the policies and the rules is important as it is typical that policies require multiple rules to be enforced, one for each classification and for each situation where action is appropriate for the classification.

It is the governance rules that provide the *intelligence* to the operation of the reservoir. The rules are embedded within the operational workflows and are responsible for interrogating the data and catalog, making decisions on which systems or which individuals should be notified about specific types of event. They provide the smart infrastructure required for the efficient operation of the reservoir and are vital to maintaining the quality of governance and quality of service expected of the reservoir.

It is typical for initial deployments of a data reservoir to be somewhat light on governance rules. Overtime as the data volume and veracity increases the rules will be monitored for effectiveness. Using the monitoring components of the reservoir the rule implementations will be iteratively improved and enhanced to meet the changing operational requirements of the reservoir and further aid self service.

5.10 Monitoring and reporting

The monitoring and reporting capabilities of the reservoir are important to the owners of the reservoir who are responsible for its continuing operation and growth. It is also important for the executives within the governance organization who need to know the effect the data reservoir has on the compliance of the policies within the governance program.

This section describes the various aspects to the monitoring and reporting capabilities that are required of the reservoir.

5.10.1 Policy monitoring

Perhaps most important to the operation of the reservoir is the ability to monitor the data against the policies that are defined within the governance program. Monitoring the data against these policies is important for a number of reasons including:

- Regulatory and compliance
- Monitoring financial exposure or risk
- Providing an incremental view on the overall effect of the governance program on the data within the reservoir

Monitoring of the policies is key to proving to the business that the data within the reservoir is accurate, trusted, and compliant with the corporate governance policies that are in place across an organization. The need to ensure that the content of the reservoir is governed to encourage users to release their data into the reservoir is important. It is the ability to monitor the effectiveness of these policies on the data and build reports based on these metrics that proves to the business that the reservoir is able to satisfy these requirements.

5.10.2 Workflow monitoring

"Operational workflow for the reservoir" on page 107 described how workflows can provide the operational underpinnings of the data reservoir. The workflows manage the interaction of users and systems with the data and metadata within the reservoir. They also provide a mechanism to automatically respond to events as they occur, heavily using rules to add the intelligence required for self service.

An important capability of any workflow technology is the ability to monitor the effectiveness of each instance of the workflow that gets initialized. These important statistics, among others, can be extracted from the execution of these workflows:

- Amount of time it takes to complete each workflow
- Overall number of instances of each workflow
- Number of tasks that were not completed within agreed SLAs
- Amount of time that a particular step in a workflow takes to complete

These metrics can be analyzed by the reservoir operations team to identify areas of improvement that can be made to the effectiveness of these workflows, or can be used to identify issues as they arise.

Another important aspect with regards to workflow monitoring concerns critical path management. It is possible for alerts to be associated with particular thresholds. When a metric within the workflow monitoring meets a particular threshold, a reservoir administrator can be notified. Notifications can also be associated when a particular exception path within a workflow instance is initialized. These capabilities significantly lighten the processor burden required to operate the data reservoir.

5.10.3 People monitoring

Closely aligned to workflow monitoring is the ability to monitor the interaction of people with the operational workflows. While a workflow is running, users of the workflow will be interacting with the various steps of the workflow, completing their tasks, and moving the workflow onto the next step.

People monitoring allows you to monitor the interactions that each user is making with the operational workflows. You can monitor how long it takes a user or group of users to complete a step within a workflow, how many tasks are assigned into the inbox of each user or group, and how many tasks a particular user or group has completed over a prescribed time.

Similar to workflow monitoring, alerts can be sent to individuals when specific thresholds are met, such as when a task with a status of *high priority* must be actioned within one hour. If the one hour time limit is exceeded, an alert is sent to the person responsible for that task. And if there is no action after a further 30 minutes, the task is automatically assigned to their manager or one of their peers.

The ability to monitor and report on the metrics produced by human interaction with the workflows is important when judging the overall efficiency and health of the data reservoir. Large volumes of tasks assigned to an individual might mean that the person is overloaded or out on vacation. *What is the impact of that person not being able to act on those tasks within the wanted time frame?* Being able to identify these issues through monitoring allows steps to be taken to avoid this situation, perhaps adding to the governance rules to reassign a task if it has not been begun within 24 hours.

5.10.4 Reporting

Reports can be generated from the three types of metrics. Different business units might require reports on one or more aspects of the reservoir. For example, the chief governance officer will be more interested in a monthly report that focuses on the policy metrics. A business line manager might be more interested in monthly reports that highlight inefficiencies in the teams or people that they are responsible for, and could use the information to identify training requirements for the team.

Extracts can be taken from these metrics and custom reports built to provide the information that is wanted by the various lines of business that require this information. Operational workflows can be used to compile and schedule these reports, ensuring that reports are sent to interested individuals as required.

5.10.5 Audit

The data that is required to provide the monitoring and reporting capabilities of the reservoir is automatically captured through a combination of the workflow engine (backed by a data warehouse capturing the workflow information) and the users interactions with the workflows, as well as the changing shape of the data and metadata and its compliance with the governance policies defined within the catalog. This data is built up throughout the lifetime of the reservoir and will become more comprehensive over time as the intelligence, users, and volume of data grows.

This information can be combined to provide a fully audit-able history of the operation of the reservoir, capturing the events, the changes to the data or metadata, and the users' interaction with the reservoir. This information can be used to satisfy the audit requirements for high regulatory industries and environments.

5.10.6 Iterative improvement

This chapter has repeatedly mentioned the need to allow the reservoir operations to grow and adapt to the changing volumes and veracity of the data within it. As the volume of users increases or the policies within the governance program evolve, the reservoir must be able to shift to adapt to these changing requirements. Ensuring that a robust monitoring capability is implemented as part of the reservoir is key to enabling this.

Bringing together the metrics that are produced by the three types of reporting:

- Policy
- Workflow
- People

It is possible to identify inefficiencies within the operation of the reservoir and take corrective action on the operational components to correct the inefficiency. To assist with the identification of these issues, dashboards (Figure 5-10) can be used to provide visual reports on the operation and health of the reservoir operation. Reservoir administrators can use this information to determine what course of action should be taken to improve the efficiency.

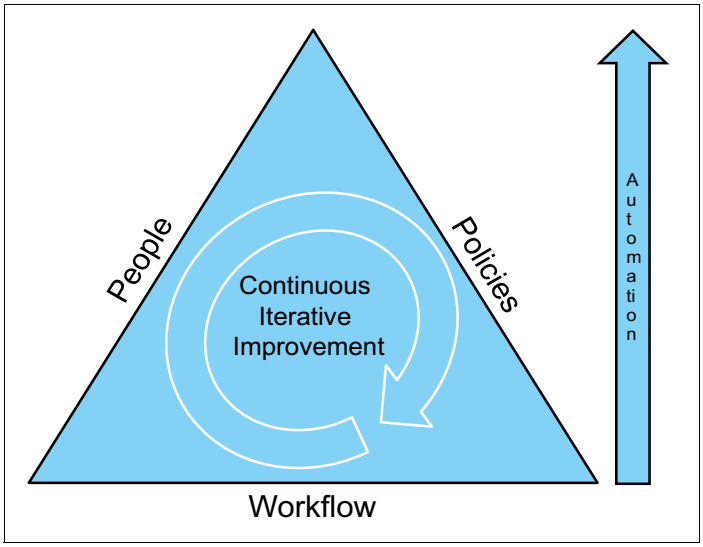


Figure 5-10 Continuous iterative improvement

The metrics that are provided by the dashboards provide the *Monitor* capability as explained in “Workflow lifecycle” on page 109. It is important that the workflow lifecycle is used to continuously improve the rules and workflows that run the reservoir. Doing so aids automation and self service for the users of the reservoir and drastically improves the operational efficiency of the reservoir.

Taking an iterative approach to improvement of the workflow will quickly identify small areas of improvement that can be made to the workflows and drive an increase in the effectiveness of the governance rules. Over time, this will lead to an increase in the number of automated steps within a workflow and a decrease in the number of human-centric steps, leading to further operational efficiencies.

5.11 Collaboration

This chapter describes the role of the evolving data steward and how a drive to engage the lines of business in the data quality decision is important to ensure that the correct data quality decision are being taken, and to minimize the bottlenecks that can occur from traditional data stewardship. It also describes how information curators hold the key to being able to shift the enforcement of the governance program fully towards the enterprise. Key to making a success of the engagement with lines of business and providing a mechanism to encourage constant curation is collaboration. This section covers various aspects of collaboration that are important to the successful operation of a data reservoir.

5.11.1 Instant collaboration

For collaboration to be fully used, users must be able to collaborate at the click of a button. The means to collaborate need to be easily accessible to the users of the reservoir and

encourage the breaking down of organizational silos. Instant messaging provides a means to offer this collaboration. An enterprise's instant messaging system should be integrated into the data reservoir operations to help break down the organization's silos and encourage easy, instant collaboration of people around the data and the processes that operate within the data reservoir.

The degree of integration of the corporate instant messaging service varies based on organization policies, lines of ownership, and technical constraints of the instance messaging application adopted by the organization.

At its simplest level, collaboration can be enabled by ensuring that the user interfaces of the data reservoir describe users in a manner that is reflected within the corporate instant messaging system. Authors of artifacts published to the data reservoir should be displayed in the same format as users of the instant messaging system. A search within the reservoir for documents authored by Joe Doe should return the documents that have been authored by the same individual that is returned when searching the corporate instant messaging directory for Joe Doe. Inaccuracies or inconsistencies between the format that users are represented within the reservoir and against other corporate systems can result in miscommunication and become a hurdle to collaboration across an organization.

Moving on from loose integration is a mechanism where the collaboration tools are slightly more integrated with the reservoir. In this scenario, the data reservoir user interfaces are used to embed the corporate instant messaging client within it. In this scenario, the instant messaging tools are always accessible and displayed when a user is accessing reservoir data within the same window.Ensuring that the tools to enable collaboration are on the window and accessible provides a constant cue to encourage users to collaborate. It is still important in this configuration that the identities of users in the reservoir and corporate instance messaging be identical.

Taking the integration a step further, the instant messaging capabilities are enabled directly on the data reservoir components that are displayed in the user interface. In this scenario, the users or department names displayed on the reservoir data become hot links to start an instant messaging conversation with that user. Real-time awareness support is added so that if a user has an instant messaging status of *available*, then the link is clickable. If the instant messaging status of a user is *busy*, then a warning can be displayed before the instant messaging conversation begins. Further levels of integration can be included that allow a user to point to the name of a reservoir document author to display a pop-up that shows profile information for the user from the corporate directory. This advanced level of integration with the corporate instant messaging system fully breaks down the organizational barriers that affect the successful engagement of users with the reservoir. Reservoir users can see the availability of other users within the reservoir, and can easily collaborate to curate data for maximum effectiveness and contribute to the quality of the data contributing to the success of an enterprise engaged data stewardship model.

5.11.2 Expertise location

Expertise location is another aspect that provides significant value to users of the data reservoir. The various forms of collaboration above assume that the information about who to collaborate with is known by the users of the reservoir. This can be easy if an artifact has an author associated with it. However, there are frequently scenarios where the author of an artifact is not the best person to comment on its contents or where no author is listed. Ensure that these scenarios are equally as accessible for collaboration on within the reservoir.

Expertise location allows the organization or the system to identify individuals who should be collaborated with for particular types of artifacts. Using expertise location, rather than only

displaying the author of an artifact as being available to collaborate with, the *experts* also appear to the users of the reservoir as being available to contact.

5.11.3 Notifications

Throughout the usage of the reservoir, users should be notified of significant events that are relevant to them. Notifications can occur as part of the workflows that are underpinning the operation of the reservoir. Perhaps a new artifact is published to the reservoir that has a tag of *finance* and *confidential*. This publishing can trigger a workflow that notifies the security team of this artifact's existence within the reservoir so that it can be checked to ensure that it has the correct security permissions applied to it.

Notifications can be surfaced to a user of the reservoir in a number of different styles. More than one style will typically be used within a reservoir, depending on the type of notification and the audience to which it is sent. Users might also be able to configure which notification styles they want to receive for which types of events. The following are some of the most common types of notifications:

- Instant message

  A visual notification appears on the window of the recipient through the integration of the corporate instance messaging system. This mechanism is good for informational messages or broadcast messages from the reservoir meant for a wide audience.

- Email

  Emails sent to individuals or groups containing relevant information about the reservoir or its contents. Daily digests containing a complete list of all of the day's events can also be sent.

- Task inbox

  Many users of the reservoir have a task inbox available to them. This inbox contains tasks that are awaiting their action. The task inbox provides the user interaction point with the workflows that under underpin the operation of the reservoir. As an operational workflow runs, steps within that workflow can get assigned to a user to take some action. The user is notified of this required action by selecting the task in their inbox.

- Short Message Service (SMS)

  Text messages can be sent by the reservoir, typically to a set of key users who need to be notified of important events such as a message to a security officer informing them of a high priority security violation that needs urgent attention or to an account manager who needs to be notified of a change to some critical data.

- Social media

  The boom in social media platforms and levels of engagement that they offers their subscribers, means that increasingly users are turning to their social media portals as their platform of entry into their environment. User will turn to their chosen platform and expect to be notified of all the events occurring that are related to their interests. This can be in the form of receiving news bulletins in their social media portal, but also can be events sent from the data reservoir. This also has the added convenience of typically being accessed on a mobile device.

- Log

  Notifications can be written to log files within the reservoir. Typically data written to log files is informational and aids the reservoir operators in diagnosing issues as they arise.

It is commonplace for more than one notification type to be used for a single event, such as sending an instant message and an email at the same time.

5.11.4 Gamification in curation

Curation is important when it comes to supporting the ongoing operation of the reservoir and the quality of the data within it. It is vitally important that users of the reservoir are empowered to *curate as they go* when using the data within the reservoir:

- It should be easy for users to curate, so minimizing the number of clicks and processor burden associated with performing a curation activity is key.
- Give users a reason to want to curate, beyond it just being the right thing to do.

Gamification is a reference mechanism to drive engagement and increase user uptake in particular activities, especially those activities that are deemed to be more mundane.

When applied to data curation, gamification can significantly increase the level of engagement of the reservoir users. Driven by statistics that capture the number, type, and effectiveness of curation activities that a user performs. Simple leaderboards and star ratings can be applied to user profiles and shared across the user community. Providing levels such as *power-curator* for the top 10% of active curators can radically drive curations activities as users strive to improve their rating and maintain their status after it is achieved.

This mechanism of gamification can further enhance the automatic selection of experts. For more information, see “Expertise location” on page 127.

5.12 Business user interfaces including mobile access

There can be a number of different user interfaces used to interact with the different operational and user aspects of the data reservoir. This section describes some of the key ones.

5.13 Reporting dashboards

The need to monitor the policies, workflows, and people operating against the reservoir is important to satisfy the ongoing regulatory and operational requirements as the data increases in size and veracity and as the users of the reservoir grow. To satisfy this monitoring and reporting requirement, a set of interactive dashboards are used to provide a snapshot of the status of reservoir.

These dashboards report on the three dimensions of reservoir operations (policy compliance, the workflow status, and the person interactions) with the reservoir. Separate dashboards can be provided for each dimension or dashboards can include information from multiple dimensions as required by the reservoir administrators.

The dashboard capability should be configurable allowing custom reports and metrics to be *mashed-up* allowing new insight to be gained from the metrics that are available.

The dashboards provide a mix of different types of data:

- Live data: Updated in real time as events change
- Static data: Data that is updated on a schedule, such as nightly
- Historic data: Showing trends in the reservoir operations over a period of time

The dashboards include access control that ensures that only authorized users are able to gain access and include tools that allow a user to scroll back in time to display the reservoir

status at a particular time in history. Also included is the ability to export reports from the dashboard over prescribed periods of time.

5.13.1 Catalog interface

The catalog interface is used by the governance team to define the policies that should be enforced on the reservoir operations and its content. It contains a number of policy trees that define how policies are related to each other and points to the rules that implement the policies. The governance team uses this interface to manage the governance program and be able to link to reports provided by the reporting dashboards to determine how enforced the policy is against the data. Change request management is supported through workflow to support changes to the data in the catalog.

For more information about using the catalog, see “Information governance policies” on page 122.

5.13.2 Mobile access

For various reasons, employees require access to corporate systems from devices external to the traditional corporate network so that they are able to access systems from anywhere at any time, including these reasons:

- Growth in smartphone ownership over the past decade along with technological leaps forward in the areas of mobile web development technologies
- Increased 3/4G network coverage and acceptance

It has now become a necessity that users can access these systems from their smartphones and tablets, 24 hours a day.

For this reason, each user interface provided by the data reservoir is mobile enabled through a combination of mobile web user interfaces and native and hybrid apps available for download either through commercial app stores or enterprise managed app stores. Using access through a corporate virtual private network (VPN), users can interact with the reservoir, manage catalog content, and monitor and export reports from all of their mobile devices. Recent advancements in enterprise-ready mobile device management technologies take care of the security considerations that occur when opening up your data to mobile users.

5.13.3 Summary

This chapter described the various operational aspects of the reservoir and at the core components that underpin its operation. The chapter discussed the importance of workflow, governance rules, and information governance policies to support various operational aspects and the importance of monitoring the effectiveness of these components overtime.

The key takeaway from this chapter are that combining the capabilities identified in this chapter and supporting their adoption across the users of the reservoir is vital to the success of a reservoir implementation. Policies must be enforced to allow users to trust that they can safely share their data within the reservoir and avoid regulatory noncompliance and risk. Workflows and their rules must be optimized iteratively to support the growth and evolution of the reservoir. Collaboration should be encouraged at each opportunity to engage users with curation, minimizing the requirements for data stewards. The overall system should be monitored through a series of rich reports and dashboards proving compliance to the policies of the governance program and iterative improvements against data quality targets.

These capabilities combine to provide the long-term benefits of the data reservoir to an enterprise and empower business users through efficient self service while reducing operational cost.

6

Roadmaps for the data reservoir

The data reservoir reference architecture describes a complex, comprehensive environment for managing all types of data for analytics and decision making. This chapter describes some options for sequencing the building of a data reservoir. The chapter begins with the process of establishing the fabric of the data reservoir that provides the governance and management capabilities. It covers the rollout of number of use cases that each use different subsets of the capabilities of the data reservoir:

- Data warehouse augmentation
- Master data management
- Operational data for systems of engagement
- 360º view of customer
- Self-service data for business users
- Data distribution

Each organization makes its own choice in how much of the data reservoir it needs to achieve the business value it needs and thus can choose to partially follow a roadmap.

Implementation Note: The technology that underpins a data reservoir is constantly evolving and can consist of a combination of previously purchased and deployed technology, open source software, and new technology purchased from various vendors.

The IBM portfolio of hardware and software products and offerings is also evolving. As a result, implementation details and links to further information are shown as implementation notes highlighted in gray boxes. These details should be checked to make sure that they are current as time passes.

This chapter includes these sections:

- Establishing the data reservoir foundation
- Data warehouse augmentation use case
- Operational data for systems of engagement use case
- 360 degree view of customer use case
- Self-service data use case
- Data distribution use case
- Summary

6.1 Establishing the data reservoir foundation

The data reservoir needs governance and change management to ensure that information is protected and managed efficiently.

6.1.1 Deploy the integration and governance fabric

The first step in creating the reservoir is to perform the following activities (Figure 6-1):

- Establish the information integration and governance components.
- Establish the staging areas for integration.
- Create the catalog, the common data standards, and audit areas of the data reservoir.

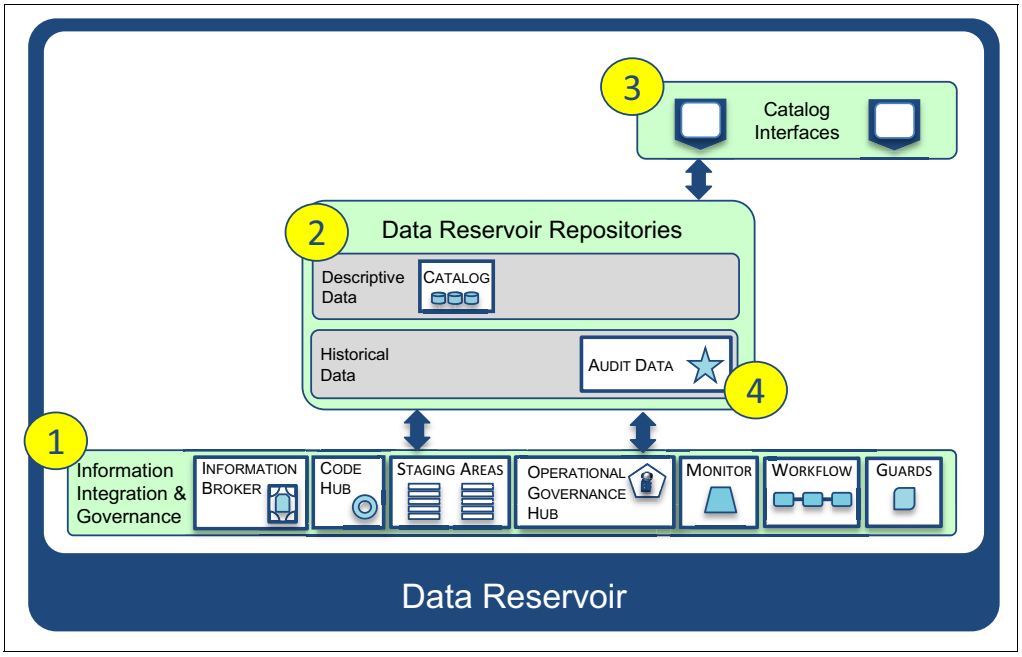


Figure 6-1 Data reservoir integration and governance fabric

The integration and governance fabric is a combination of middleware libraries and servers, plus the repositories that are used to manage and govern the data in the data reservoir (Figure 6-1):

- Item 1: Shows the middleware that provides integration and governance libraries and servers. This is described in detail in Chapter 3, “Logical Architecture” on page 53.
- Item 2: Shows the catalog repository where governance definitions are linked to descriptions of the data and its structure, location, and contents.
- Item 3: Shows the user interfaces and APIs for access to the catalog.
- Item 4: Shows the audit data repositories for storing details of the activities that people and processes are performing with the data in the data reservoir. The monitors and guards of the data reservoir populate these repositories. They are analyzed and monitored by the security and audit teams.

Implementation Note: At the time of writing, the IBM InfoSphere, IBM InfoSphere Optim, IBM Guardium® and IBM WebSphere Connectivity portfolios provide this capability:

- IBM InfoSphere Information Governance Catalog provides the catalog function that is part of IBM InfoSphere Information Server.
- IBM InfoSphere Information Server also provides an information broker (IBM InfoSphere DataStage), various operational governance hubs for operational monitoring, stewardship, and compliance monitoring.
- IBM InfoSphere Master Data Management provides the code hub through IBM InfoSphere Master Data Management (MDM) Reference Data Management.
- Other types of information brokers in use in the data reservoir can be IBM InfoSphere Data Replication, IBM InfoSphere Federation Server, and IBM Integration Bus.
- The IBM InfoSphere Optim and IBM InfoSphere Guardium portfolios provide various guards and monitoring capabilities for protecting data in the data reservoir.

6.1.2 Setting up the governance program

After the integration and governance fabric is in place, the information governance definitions are established as shown in Figure 6-2.

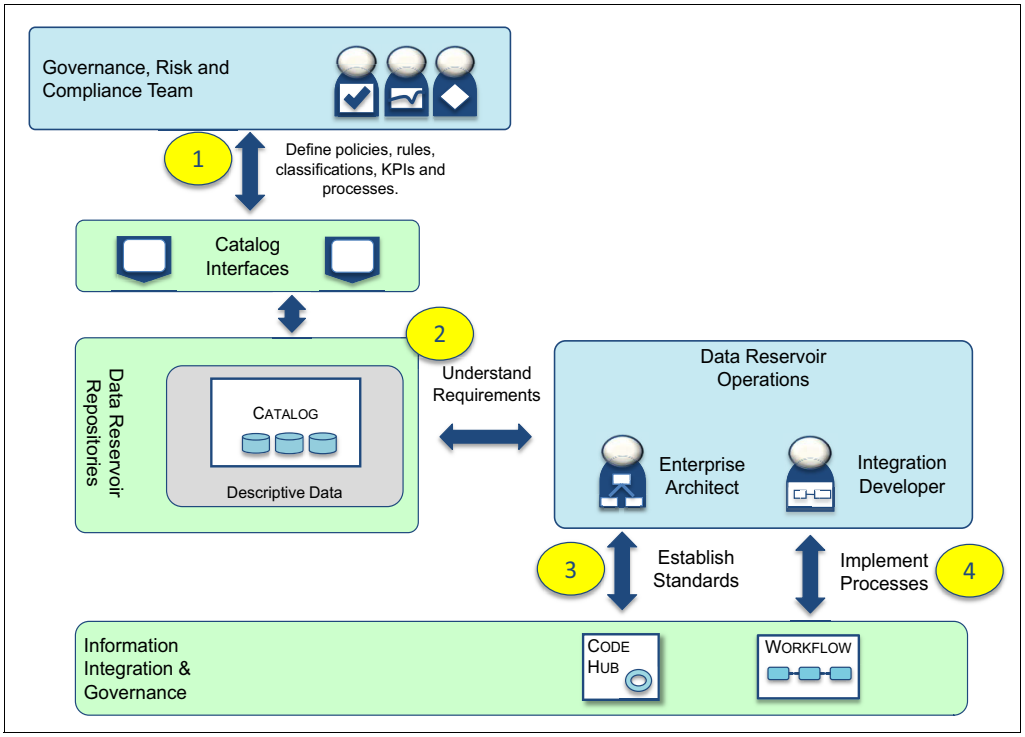


Figure 6-2 Setting up the governance program

The governance program is defined in the catalog and then implemented in the integration and governance fabric (in Figure 6-2):

- Step 1: Shows the definition of the policies, governance rule, classifications, KPIs, and the kinds of processes needed to manage and govern the data reservoir. These definitions are described in detail in Chapter 2, “Defining the data reservoir ecosystem” on page 29.

- Step 2: Shows the definitions that provide requirements to the technical teams setting up the data reservoir. They will also guide the users of the data reservoir after it is up and running.
- Step 3: Shows the enterprise architect establishing standards around technology choices, how data is structured, and how the services around it should be implemented.
- Step 4: Shows the integration architect implementing basic workflows for the data reservoir operations such as these:
  - Access control management
  - Feedback and requests for help on the catalog content
  - Security and quality reporting

These processes are described in detail in Chapter 5, “Operating the data reservoir” on page 105.

Implementation Note: An initial tutorial about governance definitions can be found in developerWorks at:

http://www.ibm.com/developerworks/data/library/techarticle/dm-1412infosphere-governance/index.html

Typically, each subject area (topic) of data that is present in the data reservoir needs a chief data officer (owner) who is responsible for defining the governance policies and rules for that kind of data. Some of these tasks can be delegated to team members, the compliance team, or the enterprise architect team. However, they should include these activities:

- Definition of concepts in the subject area
- Definition of classification scheme for data elements in the subject area
- Definition of policies and rules for data elements with each classification

Data elements from the subject area should not be stored in the reservoir if they are not covered by these definitions.

6.1.3 Adding a data repository

The data reservoir needs at least one repository in which to store its data. Adding a repository to the data reservoir is illustrated in Figure 6-3.

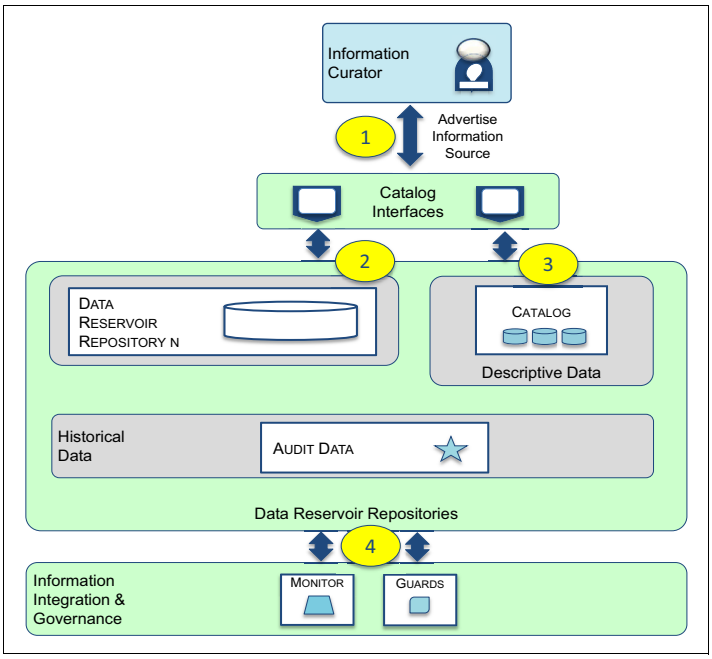


Figure 6-3 Adding a repository to the data reservoir

Adding a repository involves these steps shown in Figure 6-3:

- Step 1: An information curator accessing the catalog.
- Step 2: The import of the physical metadata description of the repository.
- Step 3: The classification of the repository by zone and capability.
- Step 4: The operations team enable monitoring and security guards on the policy that will feed the audit data.

The same process is used to define the sandboxes and published data stores.

Implementation Note: Figure 6-3 step 2 uses the IBM InfoSphere Metadata Asset Manager (IMAM) application and Figure 6-3 step 3 uses the IBM InfoSphere Information Governance Catalog from IBM InfoSphere Information Server.

6.1.4 Adding an information source

The data reservoir is now ready for data. This data initially comes from external sources such as from internal systems, external sources, or files loaded by the people using the data reservoir.

An information source is assumed to have an owner who agrees to share the information. The owner is responsible for ensuring that their data is properly classified before it enters the data reservoir.

Each source of information needs to be documented in the data reservoir's catalog to ensure that the data reservoir can provide lineage information for the data it manages.

Figure 6-4 shows the process for describing an information source. This process is called curating the information source.

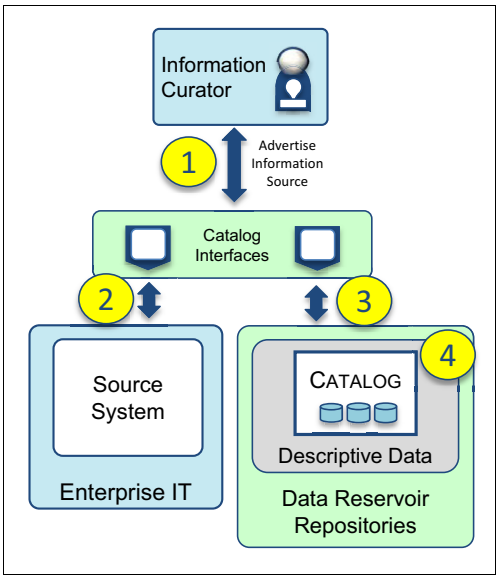


Figure 6-4 Curating an information source

Curating a new information source is a similar process to describing a repository (Figure 6-4):

- Step 1: The information curator accesses the catalog.
- Step 2: The information curator imports the physical metadata description of the source system to understand the data that it contains.
- Step 3: The information curator categorizes the data that the information source contains and adds descriptions about the information source.
- Step 4: The resulting metadata is stored in the catalog.

Implementation Note: Figure 6-4 step 2 uses the IBM IMAM application and Figure 6-4 step 3 uses the IBM InfoSphere Information Governance Catalog from InfoSphere Information Server.

6.1.5 Provisioning data from an information source

After the information source is curated in the catalog, data can be provisioned from the information source into the data reservoir.

Typically, provisioning involves copying data from the information source into the data reservoir repositories. There is an initial copy of the existing data in the information source into the data reservoir repositories (called the initial load), and then an ongoing copy of new data that is added to the information source. This copy can occur immediately as new data arrives (trickle-feed) or updates can be sent in batches (incremental load).

Alternatively, provisioning can be implemented by using service interfaces. The data reservoir extracts data directly from the information source whenever it is requested. This second method works well when the information source's data is large in volume, changing rapidly, and the information source has the capacity and availability to support the requests from the data reservoir. Otherwise, copying the data gives predictable availability of the data in the data reservoir.

Figure 6-5 shows provisioning an information source using the copying approach.

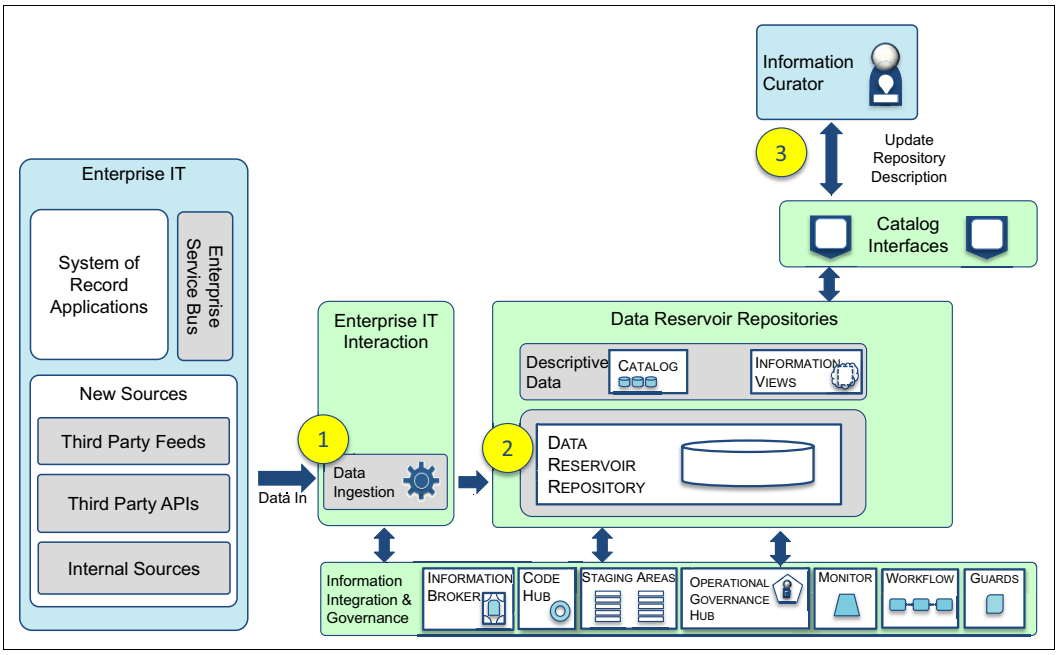


Figure 6-5 Provisioning data from an information source

The following are the steps needed to provision an information source (Figure 6-5):

- Step 1: A provisioning process is added to the data ingestion subsystem to bring the data into the data reservoir. Typically, there is an initial load of existing data and an ongoing trickle feed of new data as it is added to the original source.
- Step 2: This data is distributed to each of the reservoir repositories where it is to be stored.
- Step 3: Additional information about the new data is updated in the catalog description for these repositories:
  - Classifications that are related to the kind of data
  - Descriptions about the scope of the repository
  - Updates to catalog queries and collections

Implementation Note: Provisioning data using the copying approach is implemented by using IBM InfoSphere DataStage.

6.1.6 Enabling an information view

Information views are virtualized interfaces to data in the data reservoir. They provide consumable subsets of information that are helpful to line-of-business users who are unfamiliar with complex data models.

Information views are supported directly by relational databases or by federation capability that can create views over many types of data sources.

Each information view must be recorded in the catalog so they can be located and selected by the users of the data reservoir.

Information views are set up to support both queries and the provisioning of a sandbox through the information view. A sandbox provisioned through an information view should

conform to the structure of the information view regardless of the structure of the underlying repository.

Figure 6-6 shows the process of creating an information view.

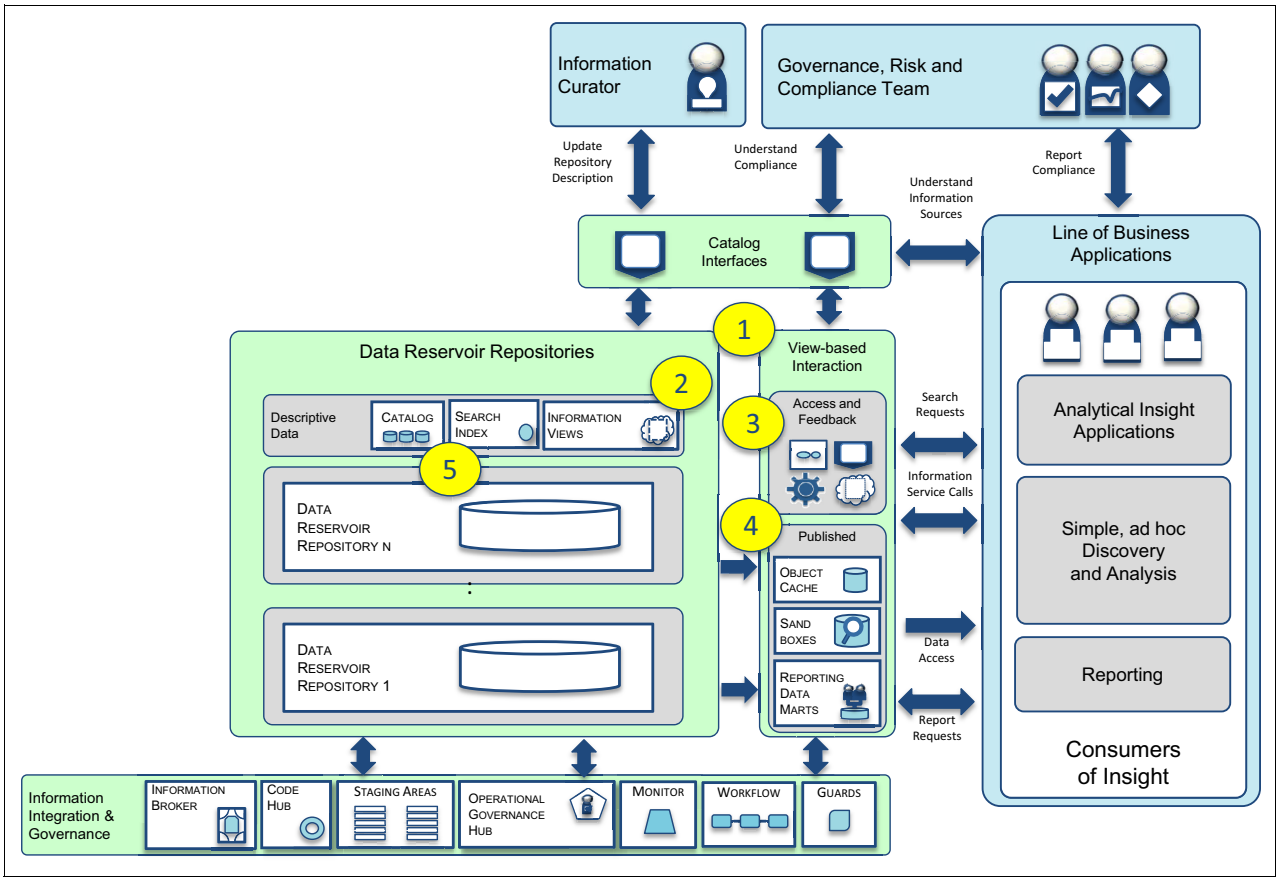


Figure 6-6 Building an information view

Complete the following steps to build an information view (Figure 6-6):

- Step 1: View-based interaction provides access to data in simplified, consumable formats. There are two approaches. The first is described in Steps 2 and 3. The second approach is described in step 3.
- Step 2: Provide an information view over the data using SQL mapping technology (such as Apache Hive or IBM Big SQL), database federation, or API technology.
- Step 3: The information views are accessed through the access and feedback subsystem.
- Step 4: Specialist stores can be provisioned from the data reservoir repositories to provide read-only materialized subsets of data that is formatted for particular types of queries. These stores sit in the Published subsystem.
- Step 5: A search index can be built over text-based data to provide simple access to data values.

Implementation Note: IBM has a technology called Big SQL that is included with IBM InfoSphere BigInsights™ for creating SQL views over data that is stored in IBM InfoSphere BigInsights. IBM InfoSphere Federation Server is able to expand this capability to build views across many types of data.

The first part of this chapter has covered the basic capabilities of a data reservoir. The sections that follow consider different use cases related to the data reservoir.

6.2 Data warehouse augmentation use case

The data warehouse augmentation use case describes how an existing data warehouse could be extended using the data reservoir to support a wide range of data and provide self-service access to data with governance.

At the start of the use case, the organization is assumed to have a successfully operating data warehouse that contains key data about the operation of the organization such as details of customers, products, and the business transactions that surround them (Figure 6-7).

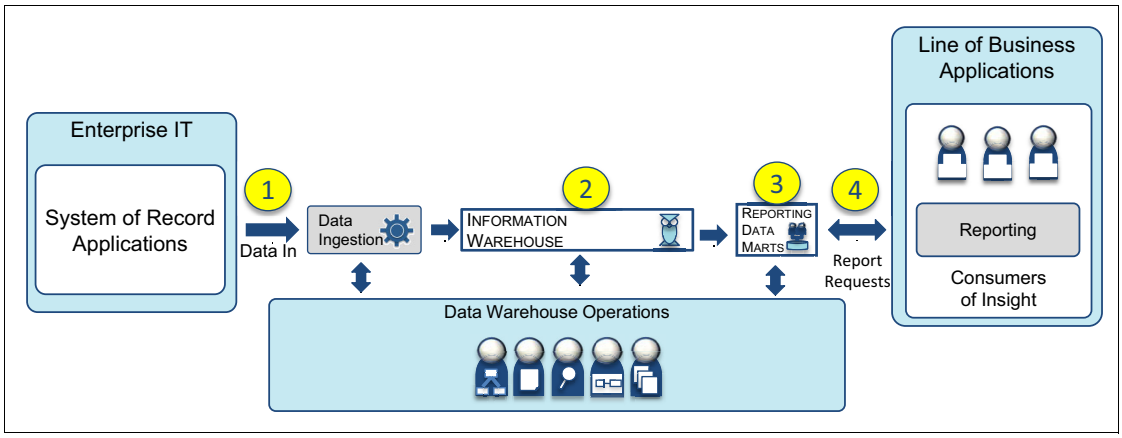


Figure 6-7 Before the data warehouse augmentation

There are four main steps to providing data to the business using the data warehouse (Figure 6-7):

- Step 1: Data from the system of records is continuously moved into the information warehouse.
- Step 2: Within the information warehouse, data is linked, consolidated, and aggregated to create useful summaries of the workings of the organization.
- Step 3: Subsets of the data are copied into reporting data marts to satisfy requests for data from the line-of-business.
- Step 4: Reports are built to query the reporting data marts to display data to the business.

6.2.1 Adding the data reservoir around the data warehouse

The first stage in augmenting a data warehouse is to wrap the data warehouse in a data reservoir as shown in Figure 6-8.

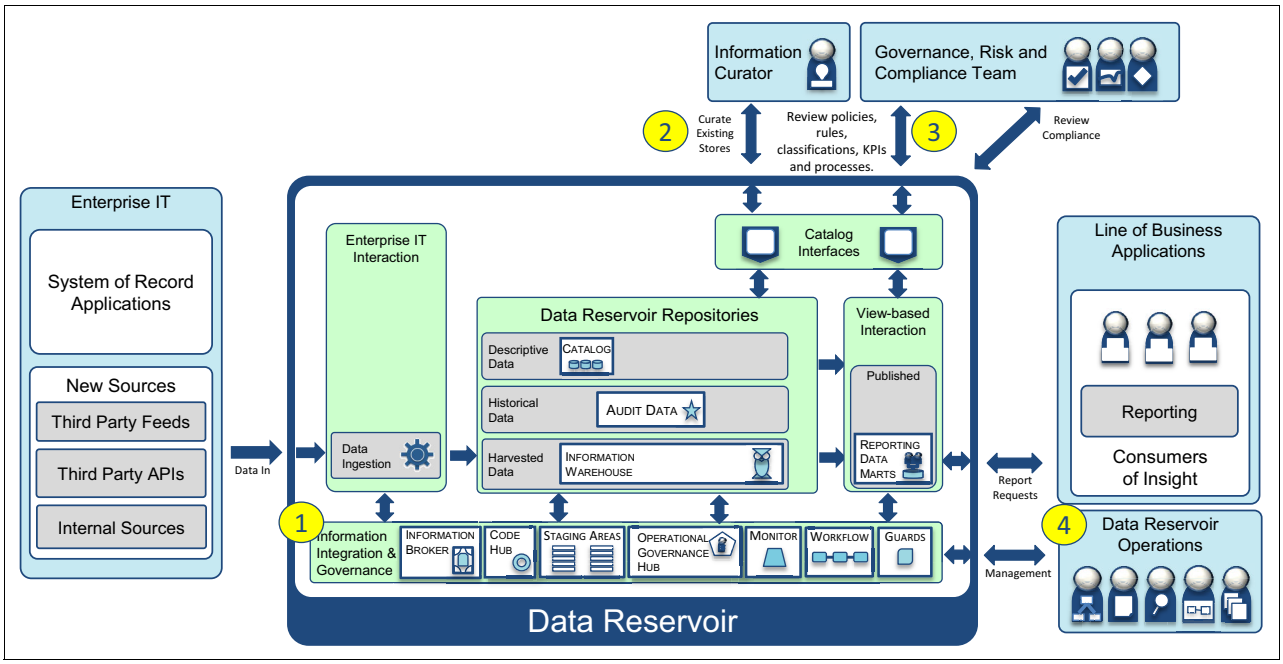


Figure 6-8 Adding the data reservoir around the data warehouse

The steps involved are as follows (in Figure 6-8):

- Step 1: Deploy the integration and governance fabric.
- Step 2: Describe existing sources and the data in the information warehouse and reporting marts in the catalog.
- Step 3: Add the governance program.
- Step 4: Set up monitors and guards.

These steps are described in more detail in “Establishing the data reservoir foundation” on page 134.

6.2.2 Working with new data

Although the data reservoir is a production environment, it provides the development environment for new analytics models or the enhancement of existing analytical models.

In this stage of the data warehouse augmentation, additional information sources are used to provision the data reservoir with new data. The analytics and data scientists are given access to this data so they can produce new insights.

When the analytics model is complete, it can either be run against the sandbox of ad hoc analysis or deployed into the data reservoir or into another operational system. Deploying models into either the data reservoir or an operational system requires a software development lifecycle (SDLC) process to accomplish these goals:

- Verify that the model works as expected
- Create the data flows that capture the model execution and outcome and send them to the data reservoir for future analysis

See Figure 6-9.

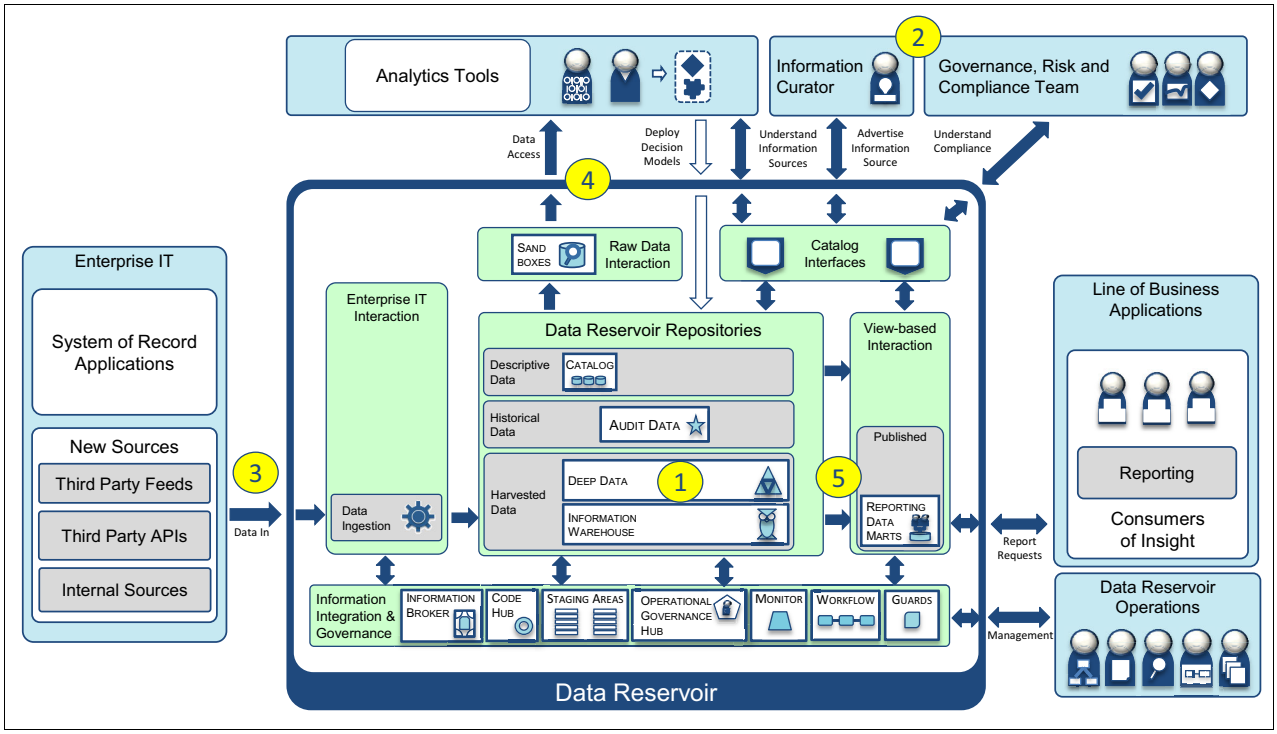


Figure 6-9 Enabling the data scientists to work with the new data

The following are the steps needed to enable this stage (in Figure 6-9):

- Step 1: Add a deep data repository to store the new data.
- Step 2: Add descriptions of the new repository and the new data contained within it in the catalog. Also, describe the new information sources that will feed deep data to the catalog.
- Step 3: Provision the data from new information sources into the deep data repository.
- Step 4: Enable the data scientists to locate the data they need and provision it into a sandbox. They can then work with the data in the sandbox to develop new analytical models.
- Step 5: The insight from the new analytical models that have been deployed into the data reservoir can be added to reports.

6.2.3 Enabling business access to new insight

Access to the new data can be extended to the business community as shown in Figure 6-10.

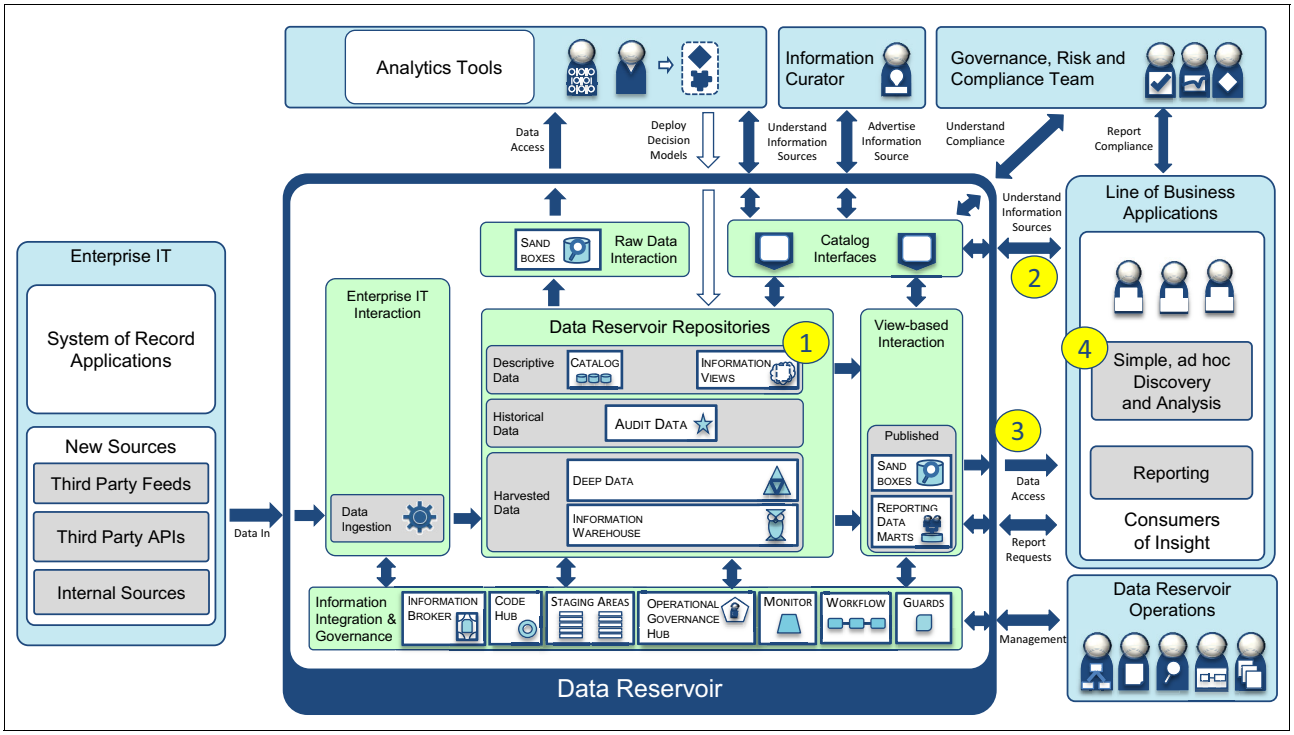


Figure 6-10 Enabling the data scientists to work with the new data

Complete these steps (in Figure 6-10):

- Step 1: Simplified views of the data that is held in either of the data reservoir repositories can be created and added to the catalog in the information delivery zone.
- Step 2: The information delivery zone provides details of the data that is suitable for the business users to access.
- Step 3: A business user can select some data from the catalog and have it provisioned into their own sandbox for their own use.
- Step 4: The data in the sandbox can be loaded into visualization tools for analysis.

6.3 Operational data for systems of engagement use case

The next two use cases cover the data reservoir's role in supplying data to operational systems. The data reservoir can include shared operational data repositories such as:

- Asset Hubs: Storing data about people, organizations, and products
- Code Hubs: Storing code tables, transcoding mappings and hierarchies
- Content Hubs: Storing controlled documents
- Activity Hub: Storing details of recent activity

These repositories support services interfaces and are designed for real-time online transaction processing (OLTP) access.

In this first use case, the data from an asset hub is required for a system of engagement such as a mobile application. Systems of engagement need to be available wherever and whenever its users want it. They are developed and evolved rapidly, but have a short lifetime. As such, it is important that they do not introduce their own data silos and use shared operational data as much as possible.

In this first use case, an object cache is used to push relevant data from the data reservoir to the systems of engagement. The object cache provides simple, flexible interfaces to data for the systems of engagement based on the data maintained by the data reservoir.

Figure 6-11 shows the initial state for this use case.

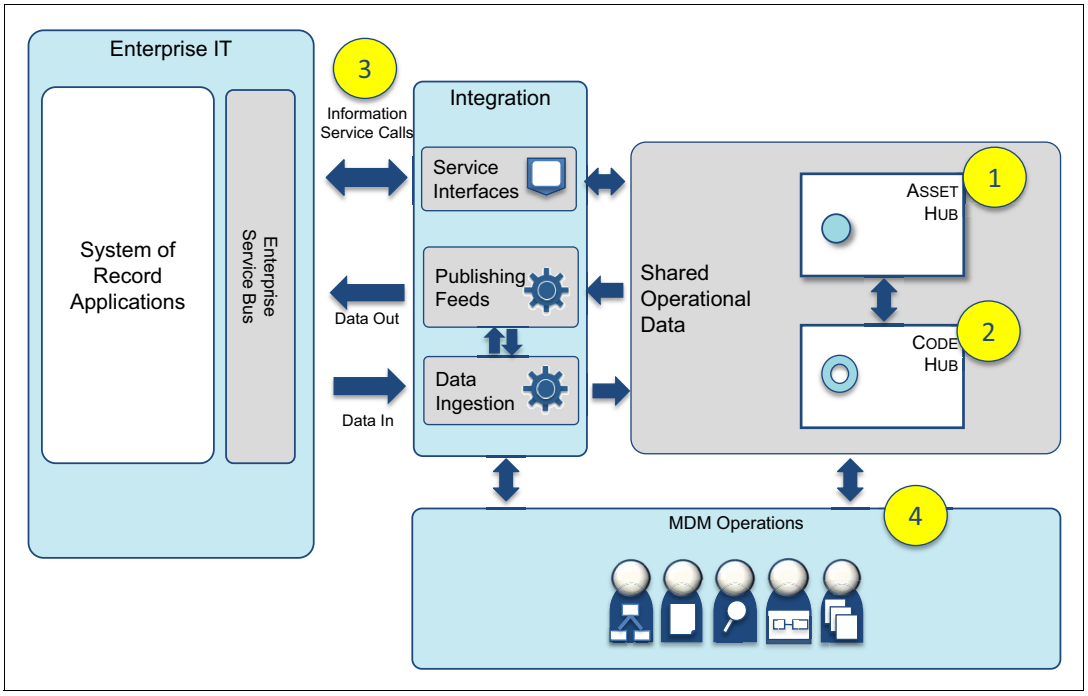


Figure 6-11 Shared operational data before the data reservoir

The following are the key items in Figure 6-11:

- Item 1: The asset hub provides a specialist server for consolidating data about people, organizations, products, and accounts. It is able to match records from different sources and link, or merge them when they relate to the same “entity”.
- Item 2: The code hub provides a specialist server for managing code tables and hierarchies and the mappings between them
- Item 3: Both the asset hub and code hub are accessed through service interfaces.
- Item 4: The MDM operations team include information stewards to manage quality issues in the data these hubs are receiving.

Implementation Note: IBM InfoSphere MDM provides implementations of both the asset hub and code hub along with the service interfaces.

InfoSphere Information Server provides the integration subsystem.

6.3.1 Adding the data reservoir around the shared operational data

Figure 6-12 shows the standard process for adding the data reservoir around the existing systems.

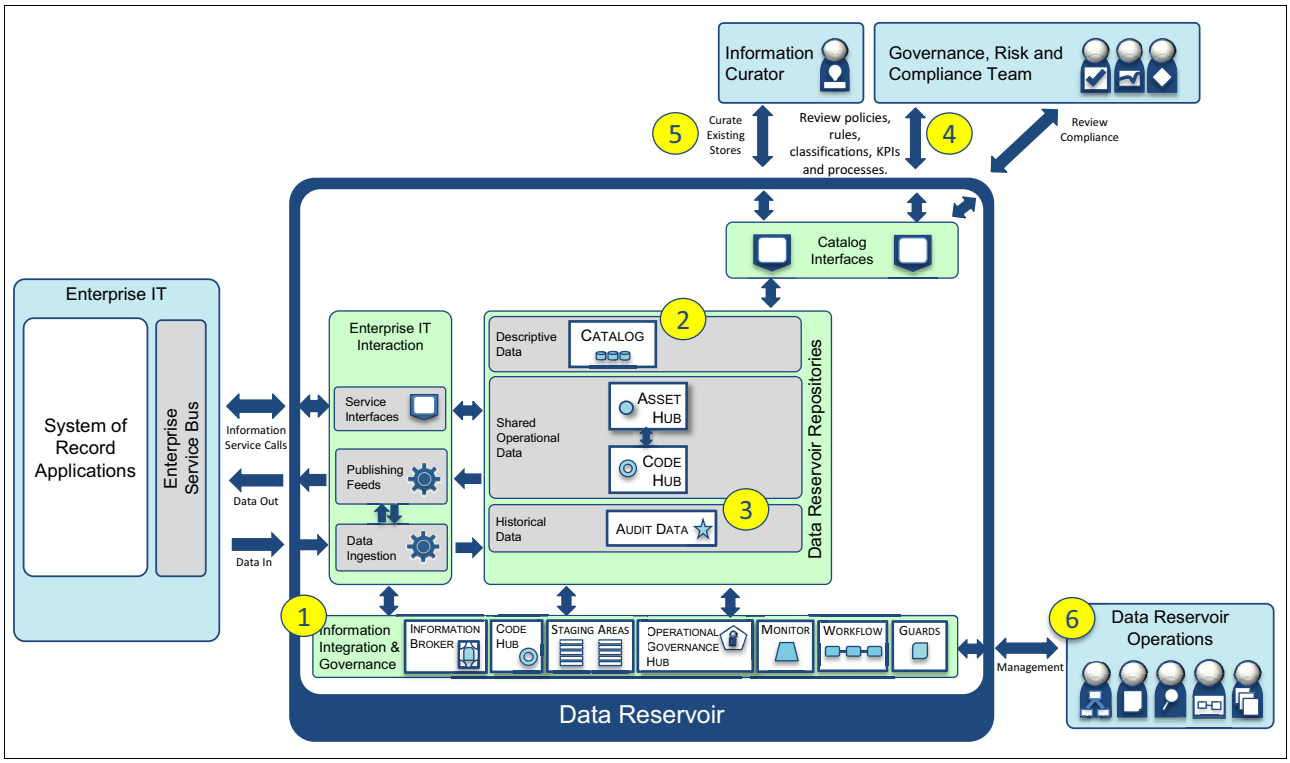


Figure 6-12 Adding the data reservoir around the shared operational data

Complete the following steps (in Figure 6-12):

- Step 1: Deploy the integration and governance fabric.
- Step 2: Add the catalog repository.
- Step 3: Add the audit data repository.
- Step 4: Describe the asset hub, code hub, and audit data repository in the catalog.
- Step 5: Add the governance program.
- Step 6: Set up monitors and guards.

These steps are described in more detail in “Establishing the data reservoir foundation” on page 134.

6.3.2 Adding the object cache

Setting up the object cache is shown in Figure 6-13.

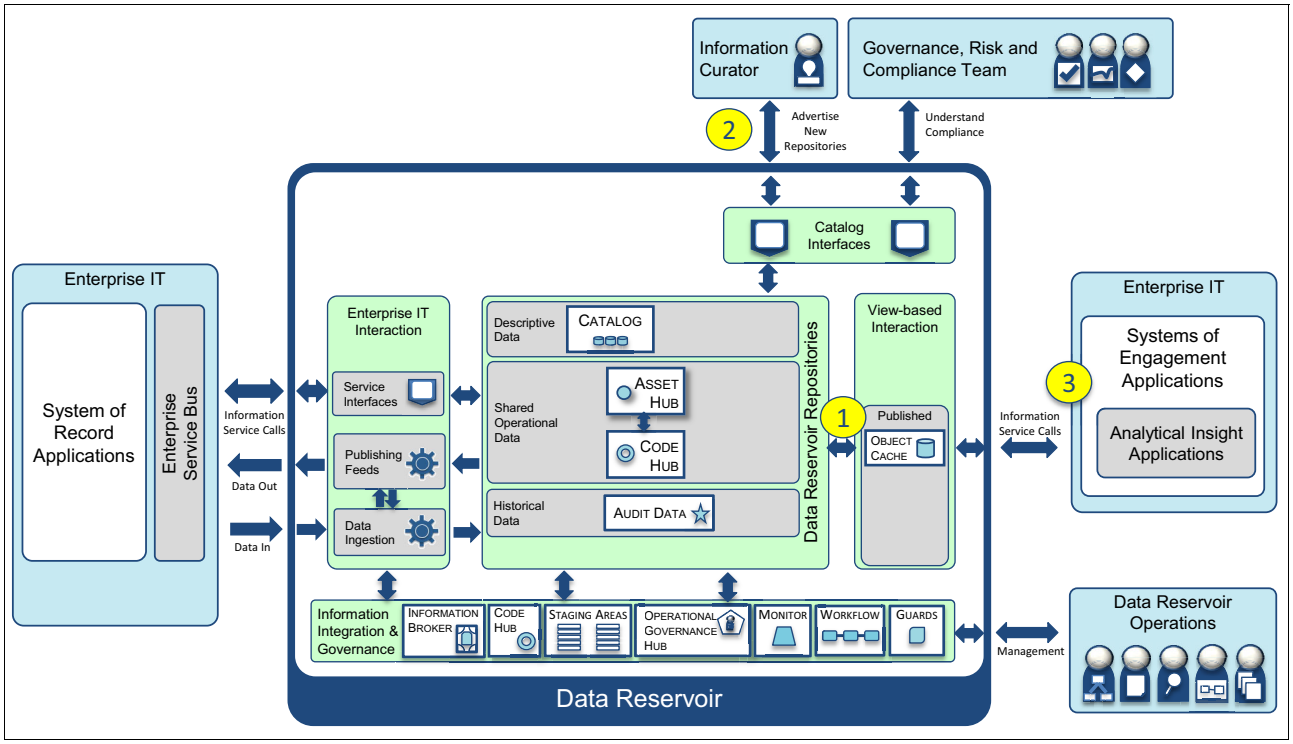


Figure 6-13 Adding the object cache

There are three steps when adding the object cache (in Figure 6-13):

- Step 1: Add an object cache to the reservoir and populate it with useful data from the asset hub and code hub.
- Step 2: Advertise the object cache in the data reservoir.
- Step 3: The object cache is available for use by the systems of engagement applications through service interfaces (typically JSON).

At this point, the data flow is outbound from the data reservoir to the systems of engagement. The systems of engagement produce new data, and this needs to be fed back into the data reservoir for analysis, particularly if the organization wants to build a 360 degree view of their customers.

6.4 360 degree view of customer use case

With the systems of engagement in place, the organization is able to capture more interaction data about their customers. This use case covers how to augment shared operational data about a customer to create a 360 degree view.

Figure 6-14 shows the types of data that can be used to build a 360 degree view of a customer.

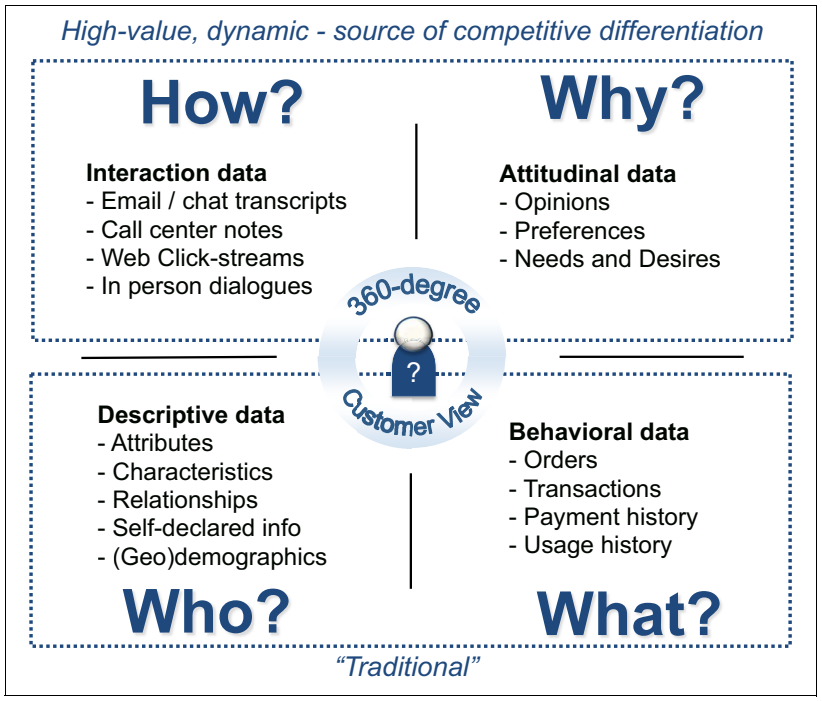


Figure 6-14 Types of data required for a 360 degree view of a customer

Figure 6-15 shows that this data comes from different systems. Although some is structured, the rest is semi-structured or unstructured.

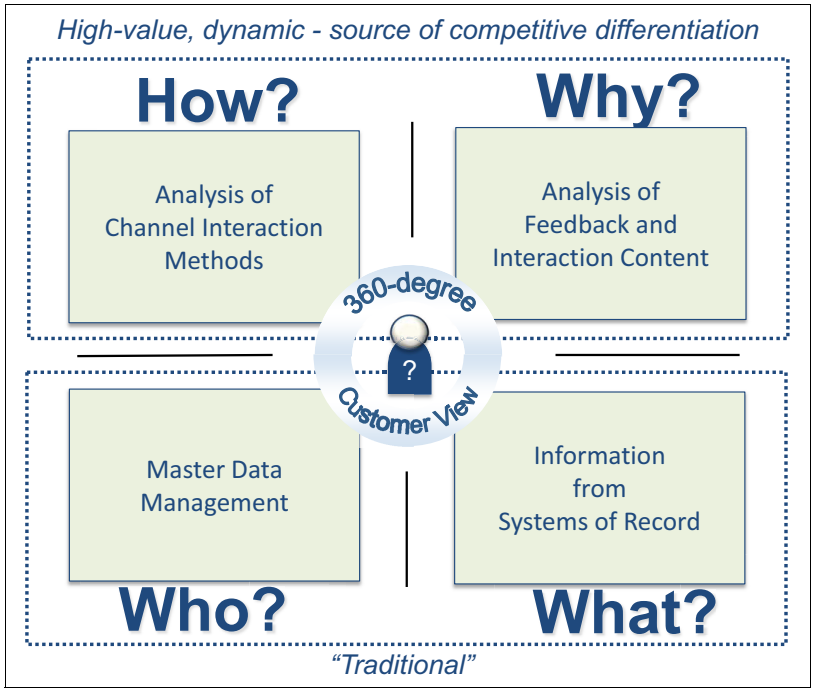


Figure 6-15 Sources of data required for a 360 degree view of customer

The role of the data reservoir in creating the 360 degree view is to accommodate the different types of data, and to enable analytics to parse, correlate, and aggregate this data together.

6.4.1 Adding new data reservoir repositories

The data reservoir needs more repositories for assembling the 360-degree view. See Figure 6-16.

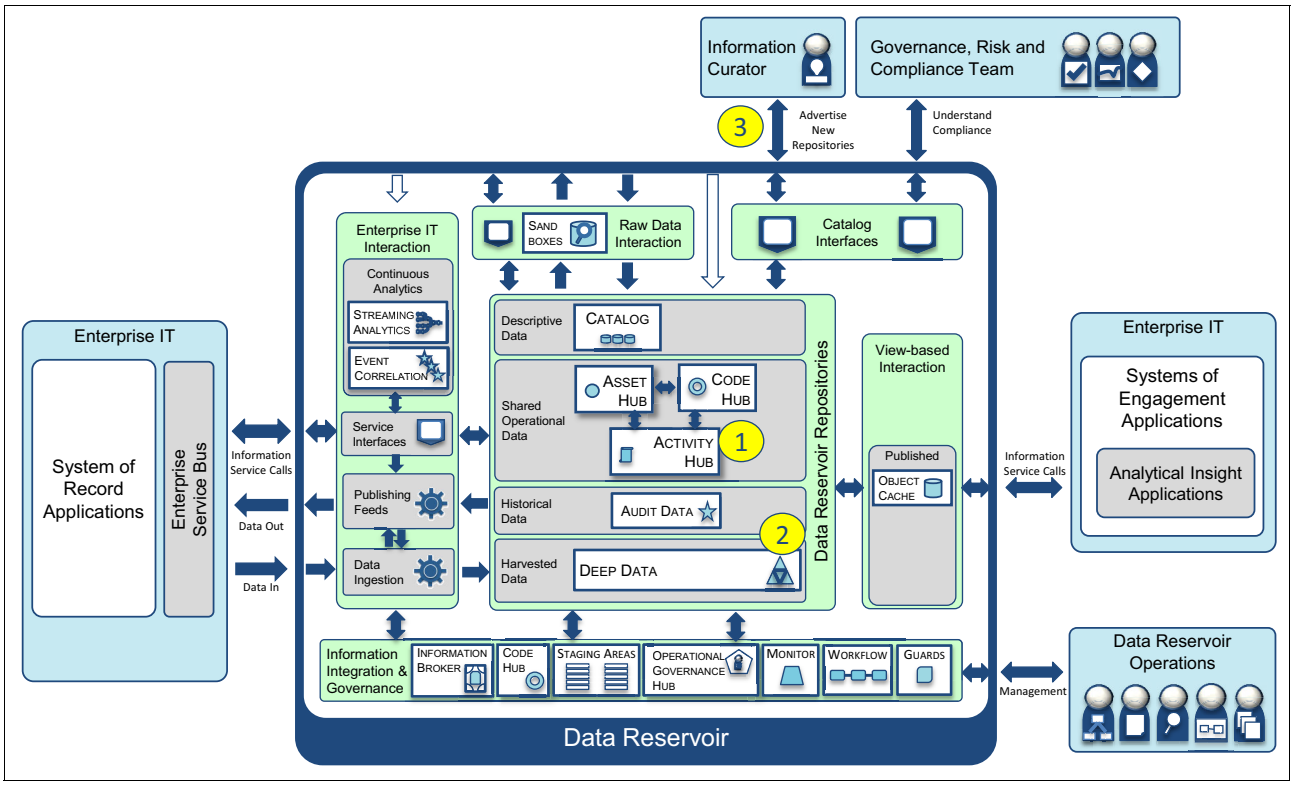


Figure 6-16 Adding new repositories for a 360-degree view of the customer

Complete the following steps to add the additional repositories (in Figure 6-16):

- Step 1: Add Activity Hub to capture customer activity detected by analytics.
- Step 2: Add Deep Data repository for storing raw activity data.
- Step 3: Advertise new repositories in the catalog.

Implementation Note: The activity hub can be implemented as an operational data store using a database. For more sophisticated use cases, the InfoSphere Custom Domain Hub product provides tooling and server libraries for building a service-oriented server that is linked to the InfoSphere MDM asset hub.

Deep data is typically implemented using an Apache Hadoop platform such as InfoSphere BigInsights.

6.4.2 Adding new data from additional information sources

The new repositories are populated from new types of information sources such as (Figure 6-17):

- Social media services
- Website click-logs
- Transcripts from call center conversations
- Activity tracked by systems of engagement

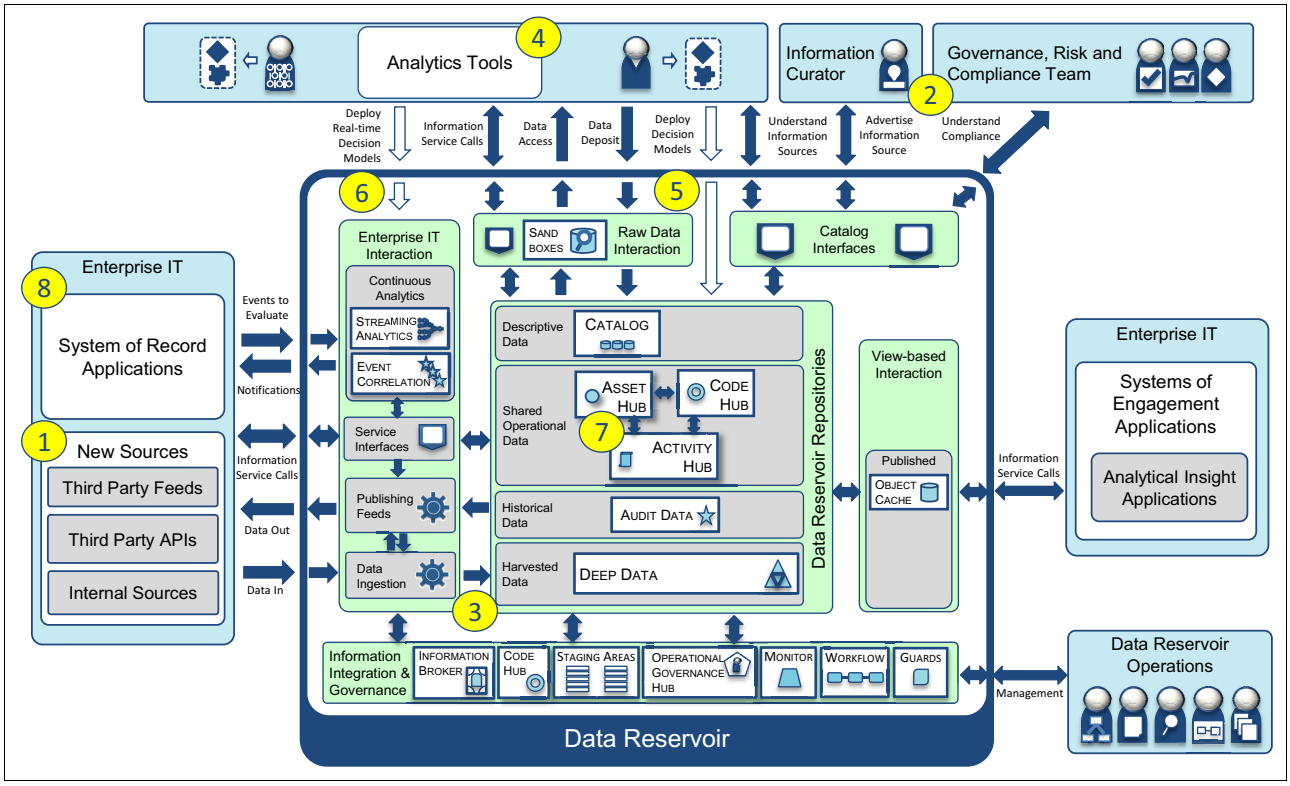


Figure 6-17 Add new sources to the data reservoir into 360-degree view

The following are the steps required to add sources to the data reservoir and incorporate them into the 360-degree view (in Figure 6-17):

- Step 1: The data reservoir is ready to receive new sources of information.
- Step 2: Details of these sources are added to the catalog.
- Step 3: The data is provisioned into the data reservoir repositories (typically Deep Data)
- Step 4: Analytical discovery and exploration are performed on this new data to discover analytical models that can extract information about customers' characteristics, activities, and potential needs.
- Step 5: These analytical models can be deployed into the data reservoir repositories to automatically generate this insight about the stored data.
- Step 6: Similarly, analytical models can be deployed into the refineries working on incoming data to the data reservoir
- Step 7: The insights related to recent customer activity are stored in the activity hub. Insight that is factual about the customer, such as a social media account, is stored in the asset hub. This makes the insight available in real time through the services interface.

- Step 8: The system of record applications can use the insights stored in the asset hub and activity hub to drive decisions made during business transactions.

The data reservoir now offers a comprehensive view of the characteristics and activity of the organization's customers. The critical next step is to ensure the business changes to act on this new insight. Typically, this action is driven by a “Next Best Action (NBA)” solution. *Smarter Analytics: Driving Customer Interactions with the IBM Next Best Action Solution*, REDP-4888, provides a description of this solution.

6.5 Self-service data use case

This use case concerns a data reservoir that is primarily focused on providing ad hoc access to data for analytics. Initially much of this data is sourced from the analytics team. However, as the use case develops the data reservoir provides access to managed data sets from the enterprise systems and the business value grows.

6.5.1 Self-managed data

In the first stage, the aim is to provide governance around the data that the analytics teams are sourcing and managing themselves as shown in Figure 6-18.

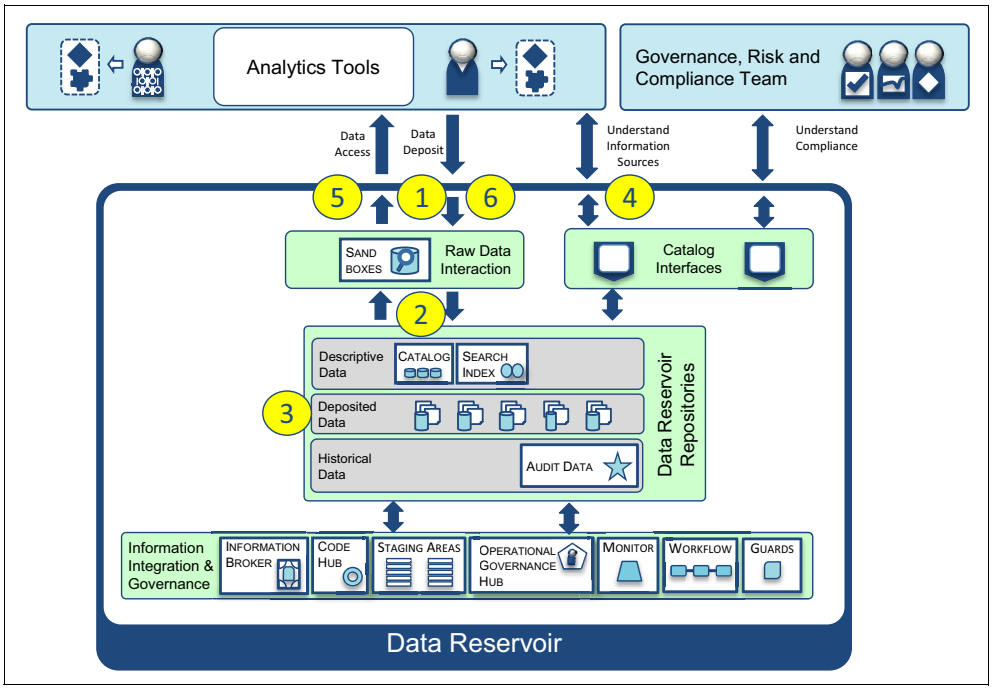


Figure 6-18 Supporting self-managed data

Complete these steps to support self-managed data (in Figure 6-18):

- Step 1: The teams using the data reservoir deposit the data through the Raw Data Interaction services.
- Step 2: Each deposited file is cataloged as part of the data deposit process.
- Step 3: The deposited file is stored in deposited data.
- Step 4: Using the catalog, a person can find the data that they need.

- Step 5: This data can be extracted into a sandbox for exploration and creation of analytics.
- Step 6: Any newly generated data can be deposited as a new file in the data reservoir.

6.5.2 Adding enterprise data to the data reservoir

At this stage, the analytics team, who are familiar with how to work with data in the data reservoir, classify and govern their own data. However, they are still requesting extracts from the enterprise systems to populate the deposited data repositories.

The next stage is to automate the provisioning of the enterprise data (Figure 6-19).

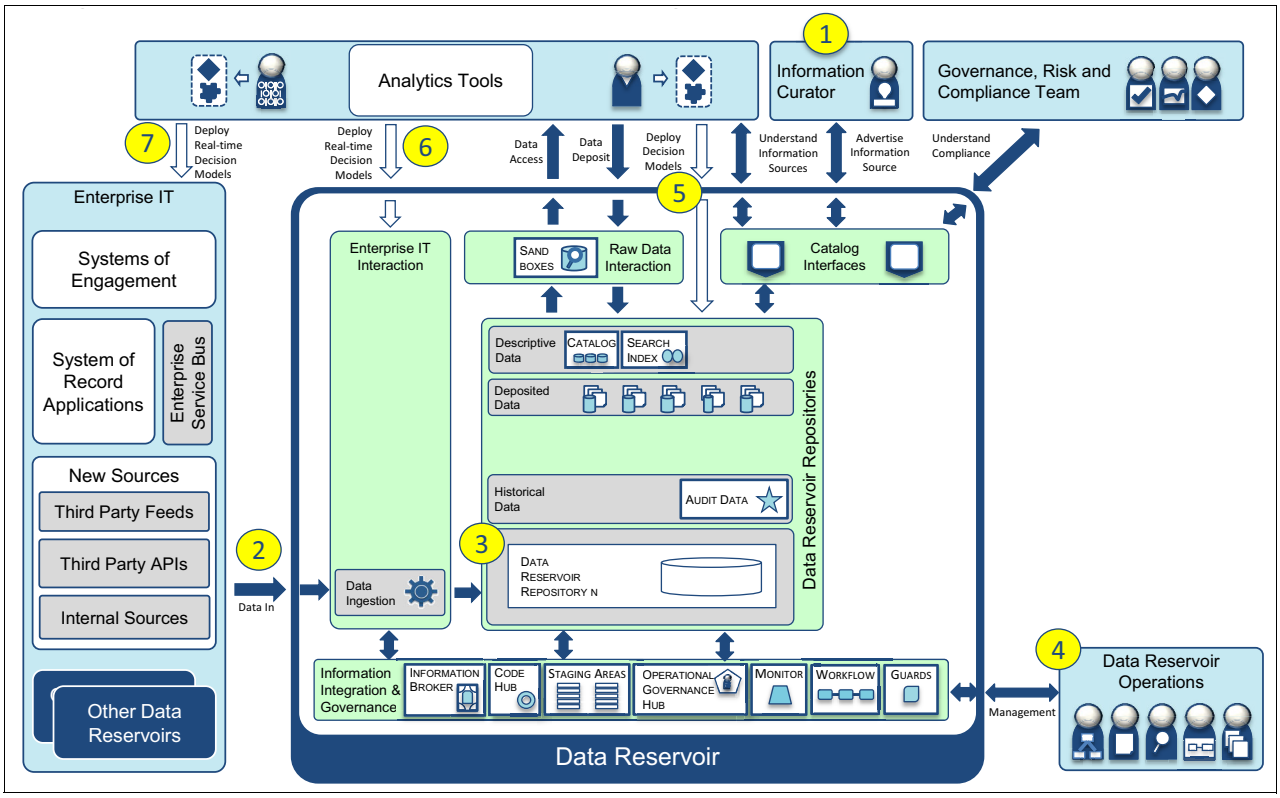


Figure 6-19 Adding enterprise data to the data reservoir

Complete these steps to add enterprise data (in Figure 6-19):

- Step 1: Enterprise data adds to the variety of data that is available for the data scientists. The enterprise sources must be described in the catalog.
- Step 2: Data from the enterprise sources are fed into the data reservoir repositories.
- Step 3: The data must be stored in the discovery zone for the data scientists to find the data.
- Step 4: An operations team manages the data and analytics for the data reservoir so more can be done with the data.
- Step 5: Analytical models can be deployed into the data reservoir repositories to create new insights (data) from the data stored.
- Step 6: Analytical models can also be deployed into the data ingestion processes to generate new insights from incoming data.

- Step 7: Finally, because analytics is being performed on enterprise data, the analytical models can also be deployed in the original sources so the results can be used in real time within the business transactions.

After this stage is completed, the analytics team has an opportunity to delete the enterprise data snapshots they have added to deposited data in the past.

6.5.3 Giving access to business users

At this stage, the data reservoir is acting as a dynamic real-time environment for analytics (Figure 6-20).

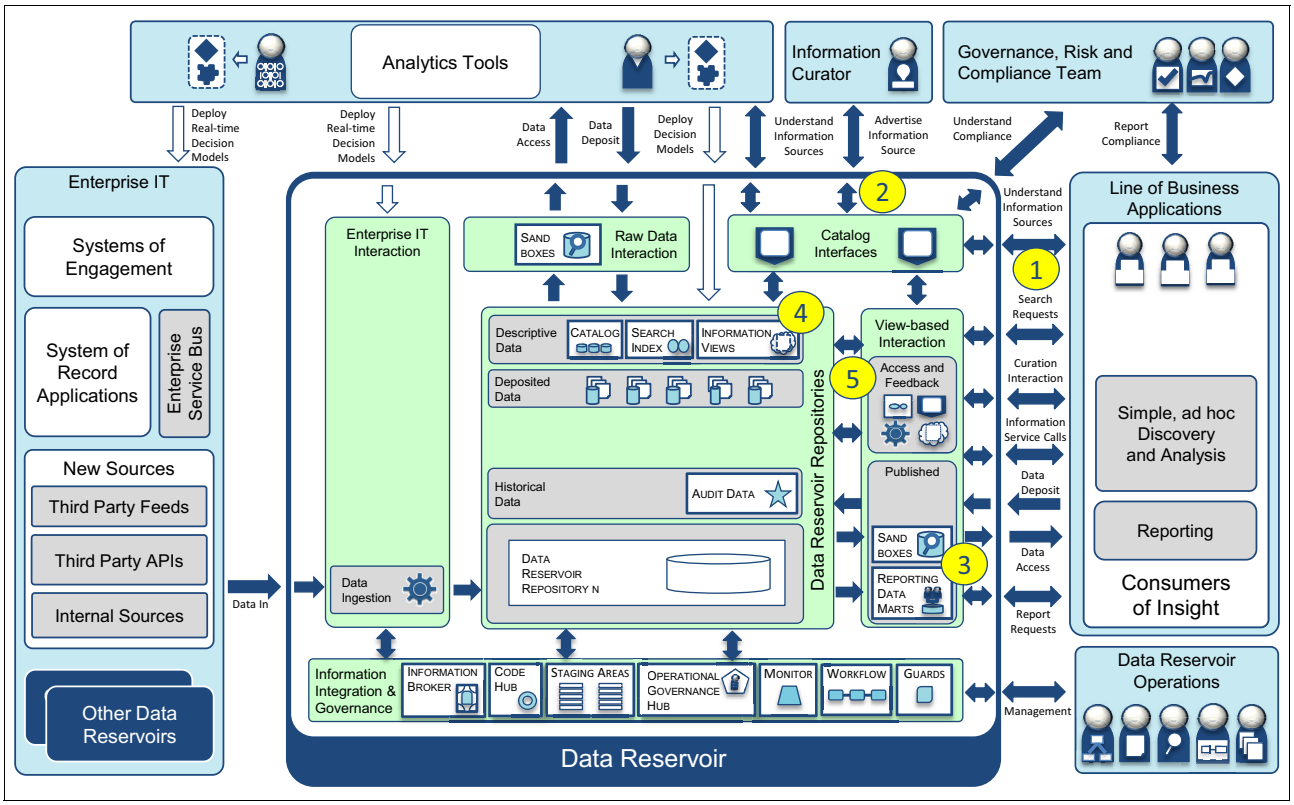


Figure 6-20 Adding information views to the data reservoir

6.6 Data distribution use case

Data distribution is a common requirement for many organizations that is used to synchronize the data in different systems. The data reservoir is a collection point for much of an organization's data, and as such is a convenient point to manage data distribution from. It can also act as the first use case of the data reservoir because it populates the data reservoir with interesting data as a by-product of collecting and sending the data around.

Data distribution copies data between systems along well-defined information supply chains. The data that is distributed can originally be from the sources that feed the data reservoir, or from new data derived from the analytics running in the data reservoir. Chapter 4, "Developing information supply chains for the data reservoir" on page 75 covers the design of information supply chains through the data reservoir and onto other systems.

Figure 6-21 shows the data reservoir acting as a data distribution hub.

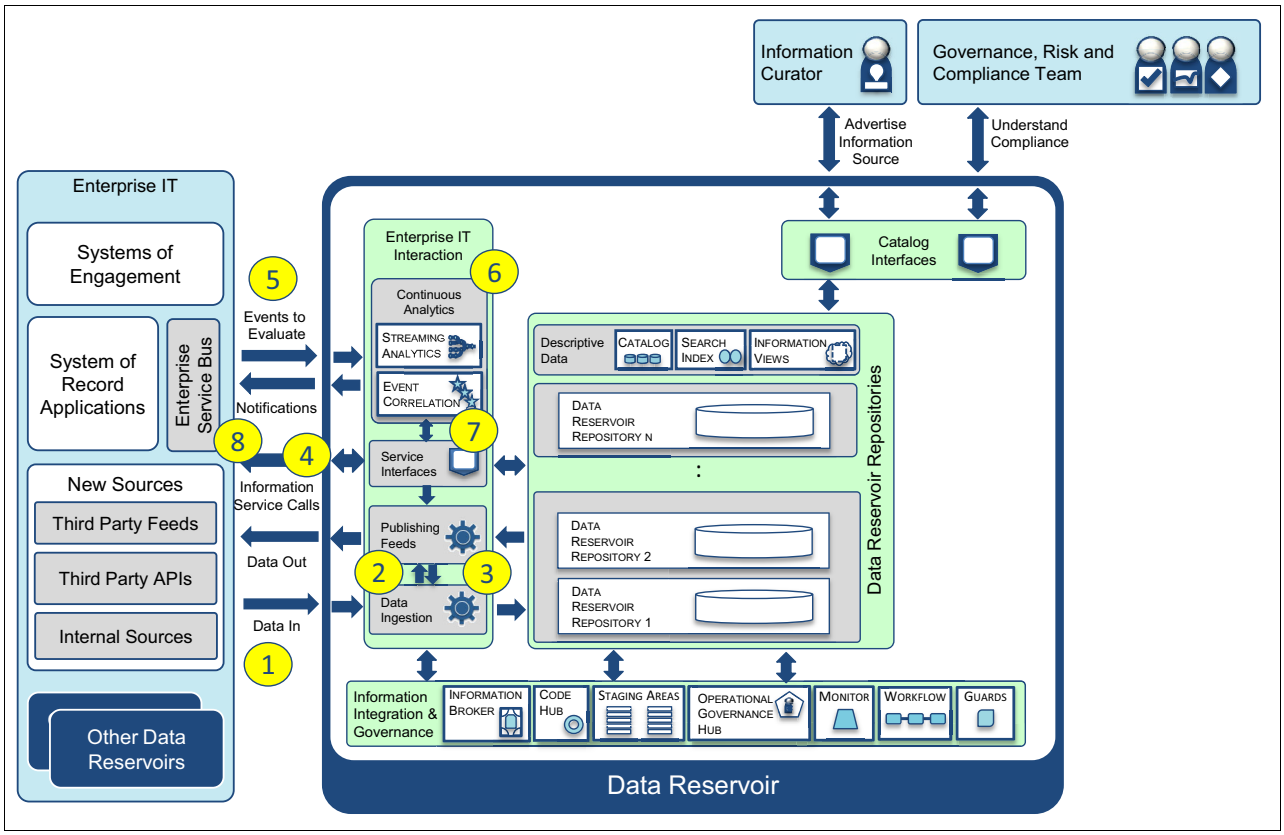


Figure 6-21 Data distribution with the data reservoir

Complete these steps to set up data distribution (in Figure 6-21):

- Step 1: Data is received by the data reservoir from various sources through the data ingestion subsystem. The data ingestion subsystem copies the data to one or more of the data reservoir repositories.
- Step 2: Data ingestion can also move data to the publishing feeds subsystem for immediate distribution.
- Step 3: New or aggregated data derived from the data reservoir's analytics is pushed from the data reservoir repositories to the publishing feeds for distribution. The publishing feeds component is responsible for distributing the data that it receives from downstream systems.
- Step 4: Data can be added, updated, or retrieved through information services that access the data reservoir repositories. This is typically data from the shared operational information zone.
- Step 5: The data reservoir is able to process real-time feeds, either event-based or streaming data.
- Step 6: These are processed by the Continuous Analytics subsystem. The continuous analytics engines use analytics to process the incoming data.
- Step 7: They use the service interfaces to extract reference data and store results.
- Step 8: These results can also be published to systems outside the data reservoir.

6.7 Summary

The roadmaps discussed in this chapter illustrate that the data reservoir is a componentized solution that provides flexibility in the order that its capabilities are rolled out. It is also not necessary to implement all components to get business value from the data reservoir.

The next chapter builds on the implementation notes included with the roadmaps to describe the current technology available to implement the data reservoir.

7

Technology Choices

Due to the wide variety of technology available from many vendors, the data reservoir reference architecture is primarily a logical architecture.

This chapter covers some of the technologies available from IBM to implement the data reservoir. It supplements the implementation notes in Chapter 6, “Roadmaps for the data reservoir” on page 133.

This chapter includes the following sections:

- Technology for the data repositories
- Technology for the integration and governance fabric
- Technology for the raw data interaction
- Technology for the catalog
- Technology for the view-based interaction subsystem
- Technology for the continuous analytics subsystem
- Summary

7.1 Technology for the data repositories

One of the hardest design decisions for the data reservoir is to determine how and where data will be stored in the data reservoir.

The data reservoir repositories in the data reservoir reference architecture are logical repositories. The intent is to characterize the different dispositions that data in a data reservoir is likely to have. The following are some examples of dispositions:

- Shared operational data repositories are designed to hold consolidated data for use in real time. Therefore, these repositories have data structures that are optimized for online transaction processing (OLTP) access.
- Deposited data is stored in whatever the owner of the data chooses.
- Operational history repositories are formatted in the same way as the original source system that produced the data. The only change is the addition of data and time stamps showing when the values were copied to the operational history. People familiar with the data in the source systems can then use these repositories easily because they understand the context of the data values.
- Audit data is in the format produced by the information protection tools that generate it. The data must be organized for the convenience of the analysts who are looking for suspicious activity.
- Deep data typically has raw data from unstructured and semi-structured sources plus reference copies of enterprise data for correlation and validation during analytics processing.
- Information Warehouses are structured stores that focus on creating a consolidated historical view of the organization.

It is possible that core data, such as data about customers, products, and key activities in the organization, is present in multiple data reservoir repositories that are formatted for different workloads.

Multiple data reservoir repositories of different types can be on the same infrastructure. Data repository infrastructure is undergoing a boom at the moment with new types of technology appearing every few months. Figure 7-1 shows a typical mapping of the data repositories to the IBM technology available in this publication.

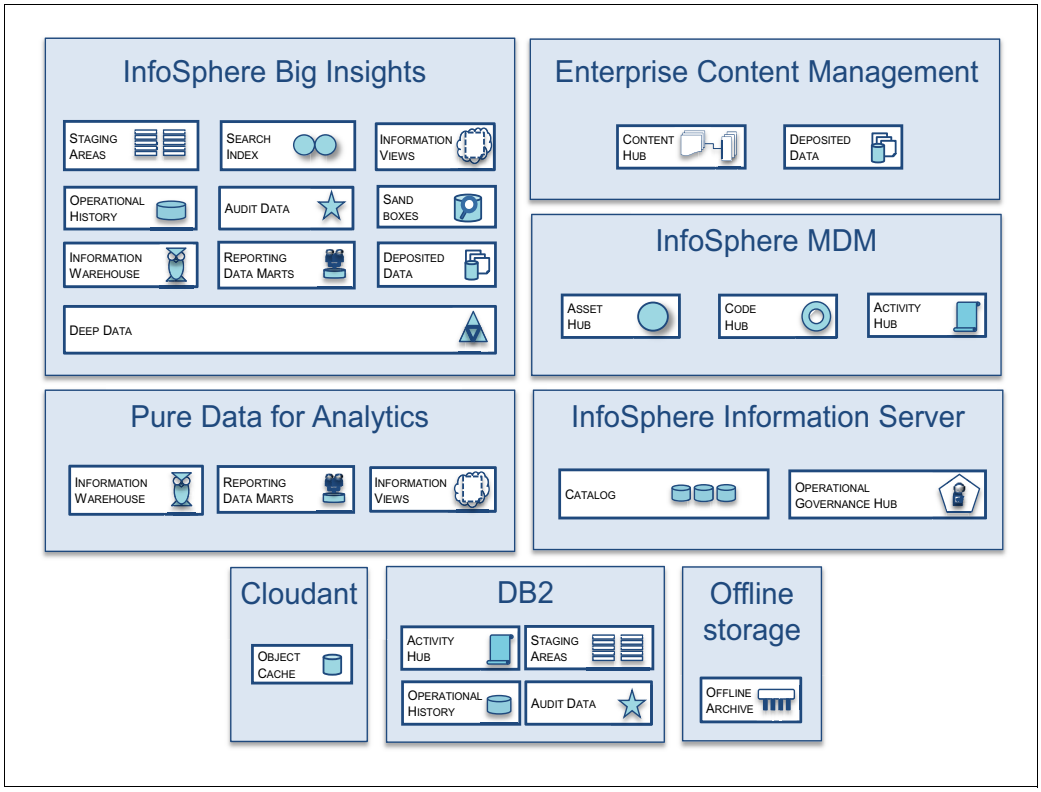


Figure 7-1 Data reservoir repository mapping to technology

The Apache Hadoop-based IBM InfoSphere BigInsights is a versatile data management platform. It can store many types of data reservoir repositories. However, it is a fairly slow execution environment that is designed for batch workloads.

Where repositories, such as the information warehouse, appear on multiple technologies, it means that there is an implementation choice. The choice depends on the tools required and the non-functional requirements. So for example, the Information Warehouse can be implemented on IBM PureData™ Systems for Analytics if there are analytics deployed to it that need the specialist hardware acceleration that PureData Systems for Analytics offers.

Not all of the repositories in the data reservoir need to be collocated. For example, IBM Cloudant® is recommended for the object cache. Cloudant typically runs as a public cloud offering, making it a good solution for providing data to systems of engagement. Cloudant databases can be part of a data reservoir where the rest of the data repositories are on-premises.

Similarly, operational history stores for IBM CICS and IBM IMS™ systems can be in a z System DB2® database that uses the IBM DB2 Analytics Accelerator. The Accelerator provides fast access to this data for analytical queries. It enables the operational history repositories to be collocated with the original sources while still having them cataloged and available as part of the data reservoir.

The roadmaps had examples of existing data repositories being incorporated into the data reservoir. This is a natural approach. The key requirement is that these repositories are

cataloged and conform to the governance program associated with the data reservoir. The repositories with a data reservoir can encompass multiple technologies from multiple vendors.

7.2 Technology for the integration and governance fabric

The data reservoir reference architecture assumes that the technology implementing the integration and governance fabric is IBM InfoSphere Information Server. The reason is because it encompasses both the governance philosophy for the data reservoir and the recognition that a big data environment is going to involve heterogeneous data and data stores.

Within information server, you can have these characteristics:

- IBM InfoSphere Information Governance Catalog provides the catalog function.
- IBM InfoSphere DataStage provides an information broker.
- IBM Business Process Manager (BPM) provides the workflow engine.

IBM InfoSphere Information Server also includes various operational governance hubs for operational monitoring, stewardship, and compliance monitoring.

Information Server is complemented with the IBM InfoSphere Reference Data Manager product that implements the code hub.

Other types of information brokers in use in the data reservoir could be IBM InfoSphere Data Replication, IBM InfoSphere Federation Server, and IBM Integration Bus.

The IBM InfoSphere Optim and IBM InfoSphere Guardium portfolios provide various guards and monitoring capabilities for protecting data in the data reservoir. For example, IBM InfoSphere Optim Data Privacy provides masking libraries and IBM InfoSphere Guardium Data Encryption provides encryption of data both at rest and in motion.

InfoSphere Guardium also provides monitoring services for the data reservoir repositories to alert the data reservoir operations team if data is being accessed under suspicious circumstances.

7.3 Technology for the raw data interaction

IBM InfoSphere Information Server can also implement the raw data interaction subsystem:

- IBM InfoSphere Information Governance Catalog provides the ability to locate the raw data that the data scientist or analyst requires.
- IBM InfoSphere Data Click populates a sandbox with the data (or a sample of that data) and catalogs the sandbox, through a simple wizard.

7.4 Technology for the catalog

IBM InfoSphere Information Governance Catalog also provides the catalog repository and catalog interfaces (Figure 7-2).

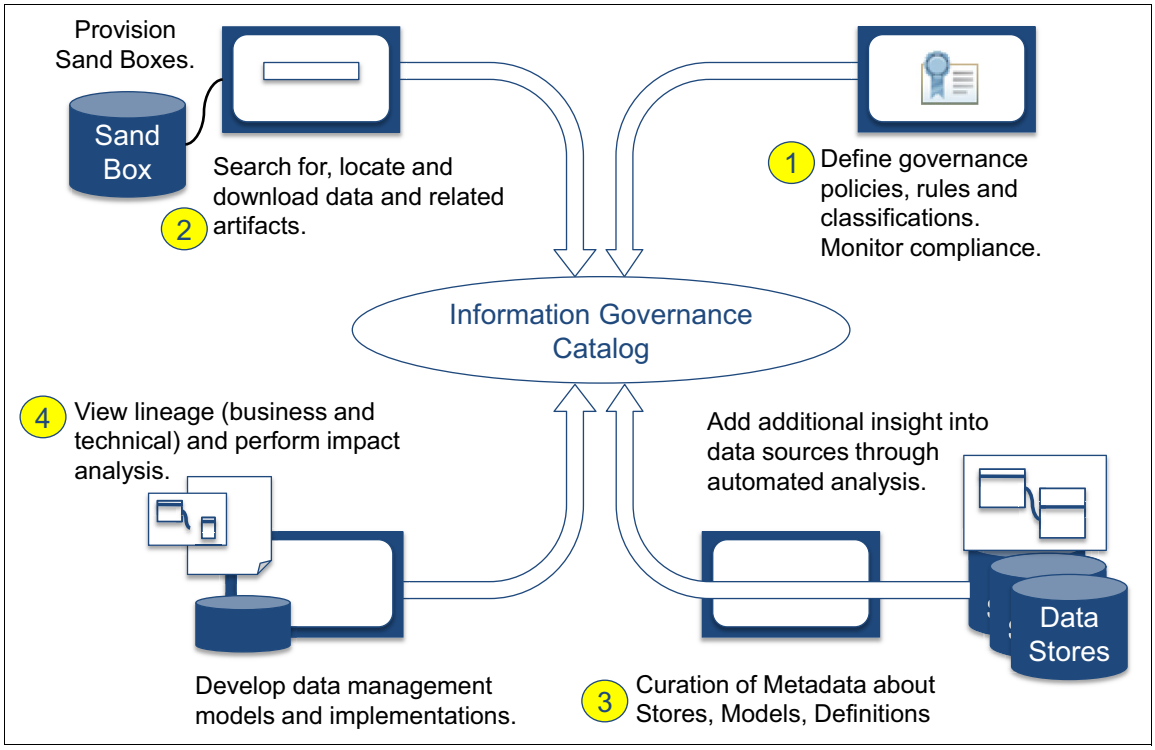


Figure 7-2 The roles of the Information Governance Catalog in the data reservoir

The information governance catalog has four roles:

- Setting up the governance program
- Locating data and provisioning sandboxes
- Curating repositories and sources of information
- Viewing lineage to understand the origin of data

7.5 Technology for the view-based interaction subsystem

The view-based interaction subsystem has perhaps the largest number of options in terms of how it is implemented. This subsystem reaches to the business communities. The goal in its implementation is to enable the tools that the business users want to use. Many of these tools work with simple files such as CSV files or relational databases. These formats are typically used in the Published data stores found in view-based interaction.

For the assess and feedback part of view-based interaction, often a search engine, such as Apache Solr (supported by IBM InfoSphere BigInsights), is used to enable business users to search text-based data. This capability can be augmented with information virtualization technology such as IBM InfoSphere Federation Server to provide simplified views to the data.

7.6 Technology for the continuous analytics subsystem

Within the continuous analytics subsystem are two types of engine. The streaming analytics engine is designed for processing a constant stream of information. It is looking to detect the occurrence of patterns within that data. IBM InfoSphere Streams is an ideal product for implementing this engine.

The event correlation engine processes discrete messages or events. For simple cases where the events are discrete and can be processed by a stateless engine, IBM Integration Bus is a good choice. Where events need to be correlated together, IBM Operational Decision Manager (ODM) is a better choice.

7.7 Summary

This chapter provided a high-level mapping of the data reservoir components to IBM technology. During a data reservoir deployment, this information can be used as a starting point. However, the data reservoir requires that a proper operational model is developed to ensure that its technology meets the non-functional requirements of the organization.

8

Conclusions and summary

This book is a comprehensive description of the data reservoir reference architecture. The data reservoir is a big data solution that provides self-service data to an organization for decision making and analytics. In addition, it supports the synchronization and distribution of data between an organization's IT systems.

The data reservoir supports the deployment and execution of analytics, making it a system of insight. It can act as a data and insight exchange mechanism between the systems of record and systems of engagement.

With the data reservoir, an organization has the fundamental capability to become data driven, improving their customer service and operational efficiency through better use of data and insight.

This chapter includes the following section:

- Summary of the data reservoir reference architecture
- Further reading

8.1 Summary of the data reservoir reference architecture

The data reservoir reference architecture provides guidance on how to build a data reservoir. It focuses primarily on the logical architecture and business processes that support it. In particular, there is a strong focus on information governance. Many organizations need to be sure that as their data is consolidated together and more people are given a broader access to data, that the data is used both effectively and appropriately.

Chapter 1, “Introduction to big data and analytics” on page 1 introduced the data reservoir reference architecture, covering the key subsystems. It introduced a case study about a pharmaceutical company that is keen to improve their use of data and analytics as part of their strategy to move to personalized medicine.

Chapter 2, “Defining the data reservoir ecosystem” on page 29 covered the processes that surround a data reservoir.

Chapter 3, “Logical Architecture” on page 53 provided detailed descriptions of the components within the data reservoir.

Chapter 4, “Developing information supply chains for the data reservoir” on page 75 overlaid a couple of critical perspectives on the data reservoir components: The information supply chains and the information zones. Figure 8-1 shows the relationships between the concepts in chapter 3 and chapter 4.

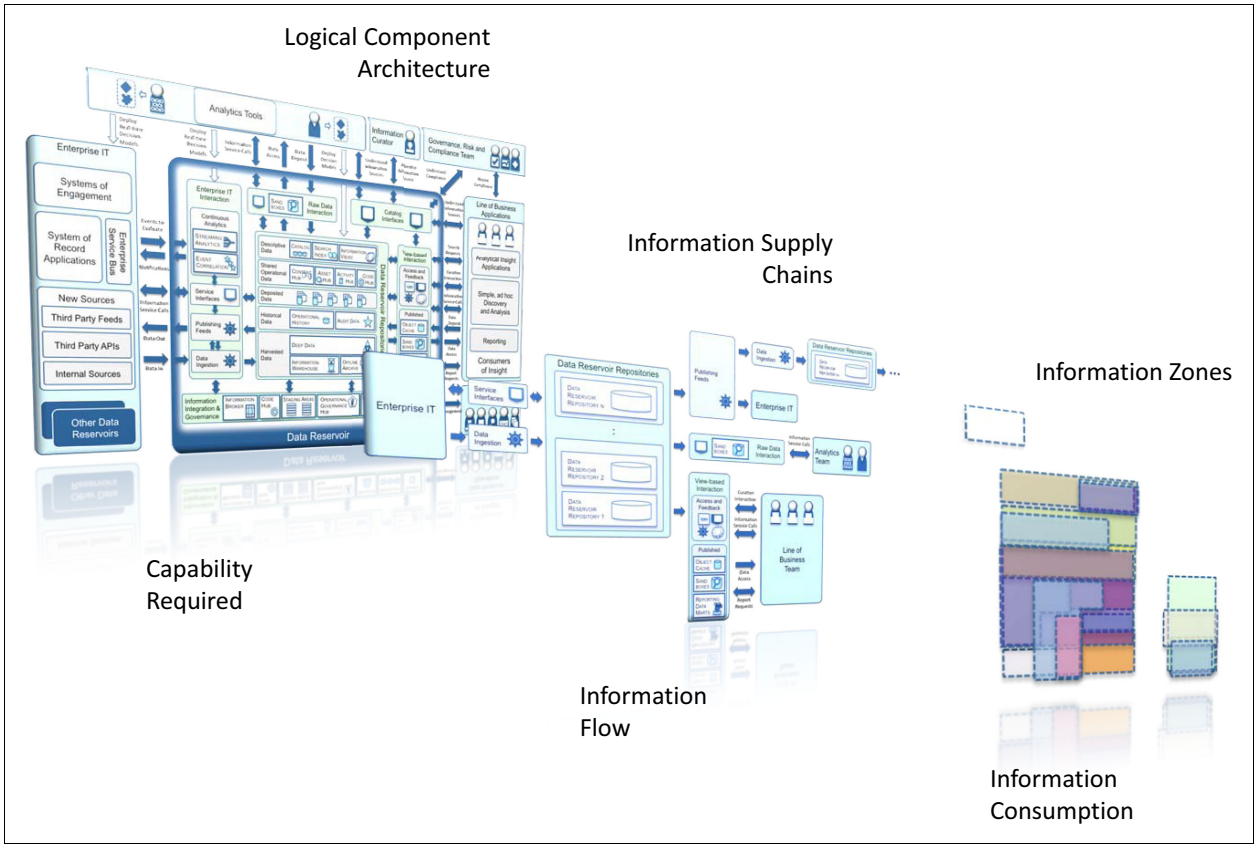


Figure 8-1 Comparison of data reservoir logical component architecture, information supply chains, and information zones

The logical component architecture presented in Chapter 3, “Logical Architecture” on page 53 shows the functions required to support a complete data reservoir. The information supply chains explain how data flows through the different components, making it available in different formats and on platforms that are designed to support the variety of workloads that are running within the data reservoir. The information zones show which subsets of data the different groups of users and workloads use. The overlap of these zones shows where stored data is being used for multiple purposes and as a result the underlying repository must support the non-functional requirements from these multiple zones.

The content of Chapter 3, “Logical Architecture” on page 53 and Chapter 4, “Developing information supply chains for the data reservoir” on page 75 together provide the detailed behavior of the data reservoir.

Chapter 5, “Operating the data reservoir” on page 105 adds more detail about the different types of processes that support the data reservoir.

Chapter 6, “Roadmaps for the data reservoir” on page 133 covers the rollout of the data reservoir, describing the order that the components of the data reservoir might be deployed, and includes implementation notes on the current IBM technologies used to implement the data reservoir.

Chapter 7, “Technology Choices” on page 157 expands on the implementation notes from the roadmaps to describe some of the IBM technologies available to implement the data reservoir.

8.2 Further reading

The data reservoir reference architecture is built from the design patterns presented in the IBM Press book called *Patterns of Information* by Mandy Chessell and Harald Smith (ISBN-13: 978-0-13-315550-1). This book provides details about the behavior and interaction patterns for the components within the data reservoir.

For a high-level description of the data reservoir for executives, see *Governing and Managing Big Data for Analytics and Decision Makers*, REDP-5120.

Related publications

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

IBM Redbooks

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

- *Governing and Managing Big Data for Analytics and Decision Makers*, REDP-5120

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

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- *Patterns of Information* by Mandy Chessell and Harald Smith, IBM Press (ISBN-13: 978-0-13-315550-1)

Online resources

These websites are also relevant as further information sources:

- Intelligent business process management:

  http://www-03.ibm.com/software/products/en/category/BPM-SOFTWARE
- Data Protection Handbook:

  http://www.dlapiperdataprotection.com/#handbook/world-map-section/c1_CN/c2_DK

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Redbooks

Designing and Operating a Data Reservoir

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

SG24-8274-00

ISBN 0837440663

Printed in U.S.A.