IBM Analytics

Prepare and maintain your data

Data quality and master data management in a hybrid environment

ENTER »

Table of contents

Introduction: Governance matters in the new normal

Cloud-based data presents a wealth of potential information for organizations seeking to build and maintain a competitive advantage in their industry. However, as discussed in "The truth about information governance and the cloud," most organizations will be confronted with the challenging task of reconciling their legacy on-premises data with new, third-party cloud-based data. It is within these "hybrid" environments that people will look for insights to make critical decisions.

A hybrid environment blends data and computing from both public cloud sources and on-premises systems. The fact that hybrid environments generally grow without much advance planning makes the task of managing ever-growing data stores even more difficult. Scalable data platforms such as Hadoop offer unparalleled cost benefits and analytical opportunities. However, while Hadoop and Hadoop-based solutions have their advantages when it comes to addressing big data volumes, Hadoop itself is not designed to ensure good data quality. Data quality issues can undermine even the most carefully planned analysis. Further, data quality problems grow over time, increasing the complexity of the data management problem quality issues present.

Yet, there is a way to make sense of the chaos. As always, the first step is to understand the nature of the problem. The focus needs to be on the data itself, and much less on the source of the data and on systems used to manage the data. If you make data and ownership of the information derived from the data the highest priority, everything else falls into place quickly.

The four pillars

How can your organization realize the financial benefits of the cloud while ensuring information culled from cloud sources is secure and trustworthy? The answer is governance.

Good hybrid information governance rests on four key priorities for IT and the business:

1. **Broad agreement on what information means**, including metadata on common policies and plain-language rules for the information the business needs and how it will be handled.
2. **Clear agreement on how owned-information assets will be prepared, maintained and monitored**—for example, operational data quality rules to master data management in on-premises systems.

Table of contents > Introduction: Governance matters in the new normal > The four pillars

3. **Enterprise- and departmental-standard practices for securing and protecting strategic information assets**, such as articulating role-based access to information, creating rules governing how information is shared and protecting sensitive information from third parties.
4. **An enterprise data integration strategy** that includes lifecycle management, architecting how data will flow and be assembled into strategic information, and also understanding how that information will be maintained over time.

These components form the foundations of information governance in a hybrid environment. In each case, you need a blend of process, organizational and technical enablers for success. With these pillars in place, your organization will have the flexibility to move with speed and confidence.

This e-book will focus on the second pillar: Preparing and maintaining your data.

Roll over the icons below for more on the top priorities for good hybrid information governance.

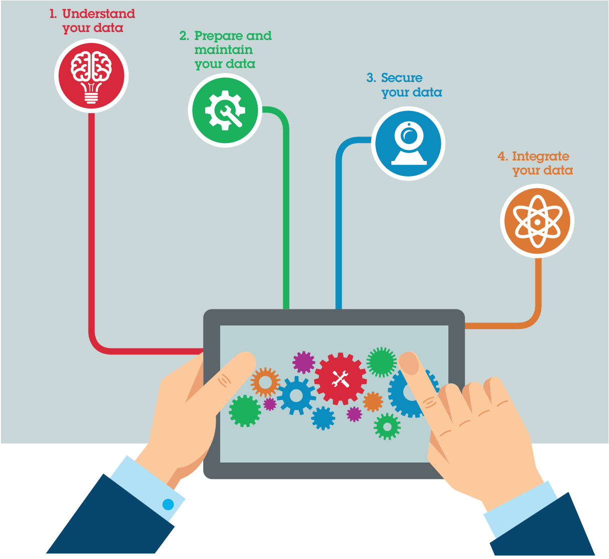

Table of contents > Taking ownership of strategic information > Quality stewardship: Managing the information supply chain

Taking ownership of strategic information

Adopting a hybrid environment does not imply you must have your IT strategy completely worked out. In fact, cloud-based aspects of the environment will evolve rapidly in response to business priorities. However, even if only a small percentage of data is flowing in from cloud-based sources, IT needs a plan for data integration and security. IT must help the organization ensure it “owns” the information created from all data and processing, no matter where it is located.

The hybrid infrastructure and decentralized computing are means to the ultimate end of creating strategic information assets. Embracing this fundamental notion lends clarity to what IT should be concerned with, and importantly, how IT can more effectively partner with the business users.

Quality stewardship: Managing the information supply chain

To create strategic assets, an organization must be able to manage its supply chain of information, and then integrate and analyze it to make business decisions (see Figure 1). Unlike a traditional supply chain, an information supply chain has a many-to-many relationship. For example, data about the same person can come from many places—that person may be a customer, an employee and a partner. The information can end up in many reports and applications, and various systems may define the same information differently.

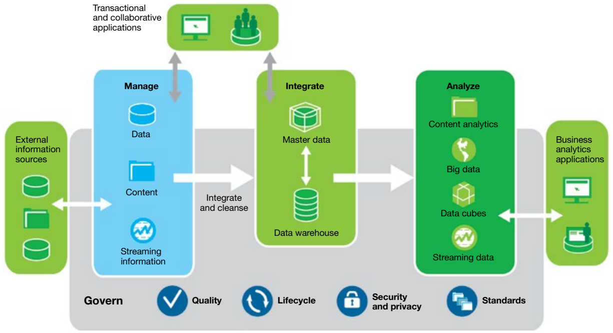

Figure 1. Governance enhances the quality, availability and integrity of the information supply chain.

Given this complexity, integrating information, ensuring its quality and maintaining a master record are crucial tasks. Information needs to be transformed into a trusted asset and governed to maintain quality across its lifecycle. The underlying systems must be cost-effective, easy to maintain and perform their workloads well, even as information grows at exponential rates.

Effective information governance can enhance the quality, availability and integrity of your data by fostering cross-organizational collaboration and structured policy making. Governance balances functional silos with enterprise-level oversight, directly affecting several factors critical to any organization: increasing revenue, lowering costs, reducing risk and boosting confidence. Excellent data quality is achieved through these essential attributes:

- **Completeness:** All related data must be linked from all possible sources.
- **Accuracy:** Data must be correct and consistent, with common problems remediated, such as misspellings or abbreviations.
- **Availability/timeliness:** Data must be available upon demand, regardless of proprietary or open source platforms being used to manage and store the data

The ability to prepare, maintain and monitor these data attributes is essential in a hybrid data environment.

What's holding back your information-gathering and analysis efforts?

- **Several technical and operational factors can disrupt attempts to deploy information-driven solutions or adjust systems to enhance information access.**
- **Multiple versions of the truth prevent organizations from effectively complying with information-centric regulations or achieving a single view of customers, products, accounts or locations.**
- **Multiple databases and applications with little governance make it difficult to maintain consistency or accuracy.**
- **Information overload prevents workers and systems from prioritizing and differentiating data.**
- **A lack of metadata, which provides information about the data itself, prevents IT and business from collaborating on the specific meaning and usage of information and terminology.**
- **An inability to efficiently standardize, merge and correct information from multiple sources creates a lack of trust, breeds disdain for inaccurate information and delays adoption of new business applications.**
- **Flexibility and agility problems, due to information that is tightly coupled to specific applications and processes, prevents the natural evolution of IT architecture as systems are added, updated and retired.**

Table of contents > Preparing your data

Preparing your data

Organizations want to tap into the rich and varied data sources available through the cloud. But if an organization's IT department has not touched that cloud-based data, its trustworthiness is automatically suspect.

The key to creating and maintaining confidence in your information is a sustainable, agile and governable enterprise information architecture. You need the ability to deliver trusted information wherever, whenever and however it is needed—across a full range of business requirements. This requires a comprehensive information integration solution that can:

- **Connect to relevant applications**, data and content, and recognize and respond to data changes in those sources whether they are structured or unstructured, mainframe or distributed, internal or external
- **Discover, model and govern** information structure and content
- **Standardize, merge and correct information** to provide authoritative, consistent and complete views of business information and its relationships across the extended enterprise
- **Effectively and efficiently collect, combine and restructure** high volumes of data for new uses
- **Synchronize, virtualize and move information** for in-line delivery
- **Flexibly publish and manage reusable information services** in a service-oriented architecture (SOA) model

Table of contents > Preparing your data > Ensuring trustworthiness

It is important to apply data cleansing to any data used in a hybrid environment so you can establish confidence in your data and the resulting analytics. There are two steps to this process: investigation, which involves discovering and governing information structure and content, and post-investigation, which involves standardizing, matching and surviving appropriate information. Together these keep data stores free of duplicate and erroneous data.

Ensuring trustworthiness

To be confident in your data, you must be able to trace its path from various internal and external sources, through systems and to the final reports and data repositories. This allows you to see where the data came from and how it was manipulated.

It's important to have a governance solution that can support this level of transparency. **But to ensure high-quality data, it is also critical to have information analysis capabilities that enable data stewards to test data quality.**

For example, stewards might perform a simple null check to ensure all the fields and tables they are analyzing actually contain data. In another scenario, they might run their data against sophisticated algorithms to determine its validity. This information is especially useful in a dashboard view, so business analysts can quickly determine whether there are any quality issues and easily access details.

Maintaining your data

Trusted data must be maintained after preparation. One approach to consider is a master data management (MDM) system. Historically, these systems focused on internal, structured data. While this is highly useful, the hybrid data environment demands that the old view of MDM be broadened to include external, cloud-based data sources.

The goal of an MDM solution is to improve operational and analytical decisions by delivering the most complete information about a customer, product or other domain—creating a "single source of truth." MDM expands this goal for a hybrid data environment by including data and applications from both on-premises and cloud sources.

Three categories of data sources

Looking at MDM in this hybrid context requires determining the source of the data. It will typically fall into three categories:

1. **Internal data:** This is traditional "behind the firewall" data. It is typically transformed and cleansed, and then maintained in a virtual or physical MDM hub.
2. **External, trusted data:** This is information that you would like to include in your single view, but you must take it as is and cannot make changes. Dun&Bradstreet data is an example of this data type. However, by correctly preparing the external data, you can ensure it is formatted properly, has been analyzed for completeness and accuracy, and is suitable for ingestion by analytical software along with your internal data.
3. **External, untrusted data:** In this case, the data is typically lower quality and raises concerns about using it in a widespread manner. It may still be valuable, but should not be treated like data from the previous two categories. Social media data is one example of external, untrusted data. Another example is prospect lists that may contain hundreds of names and addresses, and could pollute a company's customer relationship management (CRM) system. However, this data could be used to determine if someone in a household already has an account with an organization.

If you assume a majority of data analysis initiatives focus on improving the customer experience, it makes sense to consider tapping into less-trustworthy cloud-based data sources in order to:

- Improve the 360-degree view of the customer by adding insight from social media
- Discover other relationship links based on insights from documents and unstructured text
- Augment traditional product information with dynamically derived product traits based on web and social media feedback

MDM: Use cases to consider

Hotel rewards program

MDM can help improve holistic views of the customer by enhancing master data with unstructured content. For example, Facebook postings may tell a hotel chain that a high proportion of its business guests have too many children for standard reward rooms. The hotel can respond by providing reward privileges on larger suites for these high-value customers, potentially increasing their loyalty.

SaaS applications

Like most CRM systems, Salesforce.com users often have a difficult time limiting duplicate accounts. In hybrid data environments, MDM can help alleviate the problem when it is used as part of the account creation process. When a user wants to create a new account, the MDM system can conduct a real-time search of already-established accounts to uncover potential matches. This dramatically reduces the number of duplicate accounts in a software-as-a-service (SaaS) CRM application.

Table of contents > Monitoring your data

Monitoring your data

After investing the time and resources to prepare and maintain data in a hybrid environment, ongoing monitoring is essential to help ensure the data remains trustworthy over time. Data stewards are generally tasked with this effort, and as these individuals become increasingly responsible for improving the value of an organization's data assets, they need capabilities to help them manage these new requirements.

To tackle these governance challenges, data stewards need clearly documented and, if possible, automated workflows that integrate into both existing and newly supplied data flows. These capabilities help data stewards come to agreements on the meaning and value of data, what "good" means in relation to various sources and types of data, and how to handle exceptions.

A good monitoring system will provide pre-built yet customizable data rules that work in both batch and real-time data streaming scenarios, so exceptions can be captured and alerted on the fly. To accomplish this, a system should provide visualization tools so stewards can monitor data health at a glance, drill into any areas of concern and remediate exceptions.

Remediation for on-premises data can be as simple as making direct changes, if changes are captured and logged properly. For external cloud-based data, remediation may take the form of internal conversations to study the impact of data problems, and external discussions with data providers to address data concerns at the source.

Empowering data stewards

Today's data stewards are being asked to handle a diverse set of scenarios, including:

- **Collaborating across multiple lines of business to build information policies supporting regulatory requirements**
- **Assessing the cost of poor data quality and managing data quality issues to closure**
- **Engaging subject-matter experts through business processes to review and approve corporate glossary changes**

IBM solutions for data preparation, maintenance and monitoring

IBM® InfoSphere® Information Server for Data Quality provides rich capabilities for cleansing data and monitoring quality on an ongoing basis, helping turn data into trusted information that can be used to inform business decisions and streamline the execution of business processes. The software delivers comprehensive and customizable data cleansing capabilities in batch and real time to automate source data investigation, information standardization and record matching—all based on business rules you define.

Information Server for Data Quality helps you scope your data quality projects, develop metrics to form a complete picture of data quality and continuously monitor data health using an easy-to-understand dashboard. The artifacts delivered by Information Server for Data Quality enable data owners to focus on detecting and responding to critical data quality issues, and to deliver trusted data to the enterprise. Creating and reusing rules across multiple data sources helps improve time to value and deliver more consistent and correct data.

With Information Server for Data Quality, your organization can create and maintain an accurate view of master data entities. The development environment includes a flexible set of capabilities and an intuitive, “design-as-you-think” user interface. The software matches data using probabilistic algorithms designed to help ensure the information needed to run the enterprise is accurate and trustworthy. It processes global data on a massively scalable parallel platform for optimal performance in demanding situations.

You can manage and maintain data quality through several core functions:

- **Investigation:** Understand the nature and extent of data anomalies and enable more effective cleansing and matching.
- **Standardization:** Create a standardized view of customer, partner or product data. This capability also enables global address cleansing, validation and certification (for significant postal discounts in select localities), and data enrichment through geocoding.
- **Probabilistic matching:** Provides an industry-leading matching engine to help ensure the best match results possible; it is built on a platform enabled for high connectivity and scalability.
- **Survivorship:** Helps ensure the optimum consolidation, householding or linked view of record information; enables a consolidated and accurate view of customers, partners, products and more.

IBM InfoSphere Master Data Management (InfoSphere MDM) is a complete, flexible and proven MDM solution that creates trusted views to improve operational business processes, hybrid data and analytics. Part of the InfoSphere platform, it supports all domains, architectural styles and use cases across industries, and offers quick time to value through pre-built and customizable data models and business services. InfoSphere MDM complements data analytics tools and helps your organization deliver trusted information to win sales, satisfy and retain customers, improve operations and increase compliance.

IBM Stewardship Center works with both InfoSphere Information Server and InfoSphere MDM to provide a single, collaborative environment for business users to define and monitor compliance with information governance policies and manage data quality issues to resolution.

IBM Stewardship Center includes the following capabilities that address the needs of users across the governance organization:

- Workflows and rules to support data governance and quality activities
- Dashboards to monitor activity and progress
- Integrated social collaboration tools
- A browser-based, customizable interface
- The Stewardship Center Application Toolkit, which you can use to design custom business process management workflows for managing and resolving data quality issues in your organization

The Data Quality Exception Console provides a unified view of data quality issues that are collected from data integration and data quality activities in IBM InfoSphere Information Server. You can monitor the application of data quality policies and work toward resolving these issues.

Using the Data Quality Exception Console, you can complete the following tasks:

- View exceptions from InfoSphere Information Analyzer, InfoSphere DataStage® and InfoSphere QualityStage®
- Filter exception sets
- Drill down to see details about specific exception sets and the exception records they contain
- Set the priority of exception sets
- Send exception sets to be managed in IBM Stewardship Center

Next steps: Continuing the cloud governance discussion

Cloud-based data and processing services present too much opportunity for business users to ignore, and IT is charged with maintaining the integrity of internal, on-premises transactional and reporting systems. Charting a governance strategy for a hybrid environment is not something to consider at a future date.

It needs to happen now.

This e-book discusses the role of data preparation, maintenance and monitoring in a hybrid data environment. For a look at other pillars of information governance in hybrid environments, download one or all of the e-books in this series:

- The truth about information governance and the cloud
- Make sense of your data
- Securing data in the cloud and on the ground
- Developing a data integration and lifecycle management strategy for a hybrid environment

For more information on IBM governance thought leadership and supporting technologies, visit: http://www.ibm.com/analytics/us/en/technology/agile/

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

© Copyright IBM Corporation 2016

Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
September 2016

IBM, the IBM logo, ibm.com, DataStage, InfoSphere, and QualityStage are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at **ibm.com**/legal/copytrade.shtml

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

Please Recycle

« HOME