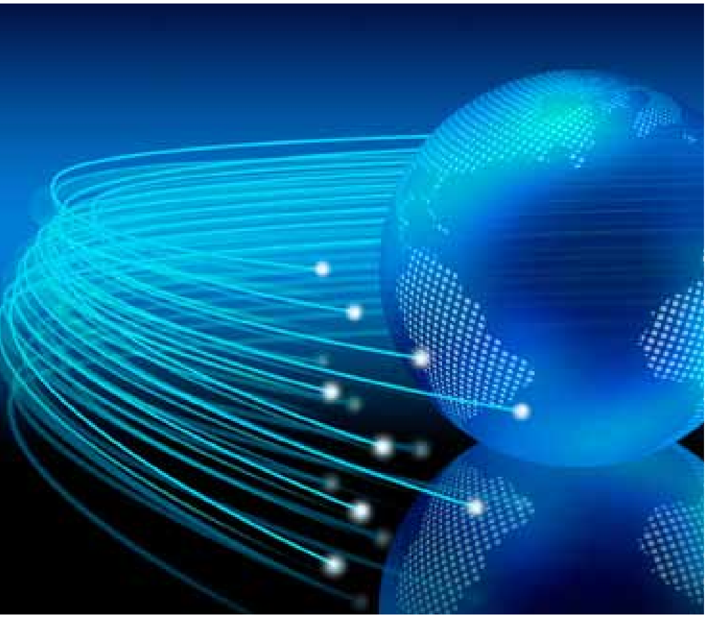
Understanding big data so you can act with confidence

More data, more problems? Not if you have an agile, automated information integration and governance program in place

Introduction

Business leaders are eager to harness the power of big data. However, as the opportunity increases, it becomes exponentially more difficult to ensure that source information is trustworthy and protected. If this trustworthiness issue is not addressed directly, end users may lose confidence in the insights generated from their data—which can result in a failure to act on opportunities or against threats.

Take, for example, the case of a large multinational corporation whose CFO asked a very simple question every month: How did sales this month compare to the same month last year? The answer turned out to be quite complicated. Sales were calculated and reported in various systems: material planning, financial, rent and royalty, real estate and so forth. Each system reported a different number because each system defined a “sale” slightly differently, and therefore extracted dissimilar data from the millions of daily transactions generated around the world. So what was the truth? No one actually knew.



The sheer volume and complexity of big data means that the traditional method of discovering, governing and correcting information using manual stewardship may not apply. Information integration and governance must be implemented within big data applications, providing appropriate governance and rapid integration from the start. By automating information integration and governance and employing it at the point of data creation, organizations can boost confidence in their big data.

A solid information integration and governance program should include automated discovery, profiling and understanding of diverse datasets to provide context and enable employees to make informed decisions. It must be agile to accommodate a wide variety of data and seamlessly integrate with various technologies, from data marts to Apache Hadoop systems. And it must automatically discover, protect and monitor sensitive information as part of big data applications.

IBM® InfoSphere® is designed to do all of these things by evolving information integration and governance to meet the challenges presented by big data.

The four Vs: Challenges inherent to big data

Every second of every day, businesses generate more data. Researchers at IDC estimate that by the end of 2013, the amount of stored data will exceed 4 zettabytes, or 4 billion terabytes. That’s 50 percent more data than the digital universe held at the end of 2012, and four times as much as in 2010.[1] All of that big data represents a big opportunity for organizations.

Companies recognize that their big data contains valuable information. They are eager to analyze it to obtain actionable insights that could help them deepen customer relationships, prevent threats and fraud, optimize operations and identify new revenue opportunities.

Unfortunately, extracting valuable information from big data isn’t as easy as it sounds. Big data amplifies any existing problems in your infrastructure, processes or even the data itself. To make the most of the information in their systems, companies must successfully deal with the **“four Vs” that distinguish big data: variety, volume, velocity and veracity**. The first three—variety, volume and velocity—define big data; when you have a large volume of data coming in from a wide variety of applications and formats and it’s moving and changing at a rapid velocity, that’s when you know you have big data.

The fourth V, veracity, is a measure of the accuracy and trustworthiness of your data. Veracity is a goal—one that the variety, volume and velocity of big data make harder to achieve.

Your company can take advantage of the opportunities available in big data only when you have processes and solutions that can handle all four Vs.

According to IBM, “Every day, we create 2.5 quintillion bytes of data—so much that 90 percent of the data in the world today has been created in the last two years alone.”[2]

Variety

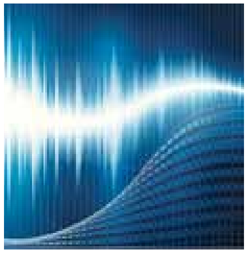
Enterprise data comes from many sources. Customers and salespeople generate data every time they make a sale. Accountants produce data every time they create a spreadsheet. Marketers create data every time they write a report.

In addition, machine data—from sources such as meter readings and IT system logs—can increase data volumes enormously. Organizations also get constant streams of data from external sources such as suppliers, consultants, research firms and news feeds.

All of this data is stored in a dizzying array of formats. Some gets stored in structured formats in databases or enterprise resource planning (ERP) applications, but much of it is in documents, email messages or even photos and videos—unstructured formats that are challenging to manage.

And data doesn't rest once it is in storage. It must be moved from application to application and from system to system so managers and executives can interpret the data and come to meaningful conclusions.

What data should an organization care about? How will it find that data?



Volume

Increasing data volume is at the heart of the big data challenge. Large data volumes can cause many obvious technical problems, such as excessive batch processing times, bottlenecks and so on.

However, the problem goes deeper than that. What data should an organization care about? How will it find that data? How will it know how to leverage that data for business use? All of these questions point to a fundamental need to understand the data—and manual techniques of examining data as it comes in are simply unequal to the task of working with big data.

Velocity

In almost every industry, data is coming in faster than ever. In activities such as stock trades, fraud detection or patient health monitoring, seconds—sometimes microseconds—can be critically important.

Organizations must be able to understand data as it is streaming in. In most cases, getting that all-important information requires moving data quickly from one application or repository to another, where it can be processed and analyzed. Unfortunately, many data integration solutions lack the high performance that big data projects require. There is not enough time to collect the data and process it in batches—particularly if a real-time or near real-time response is warranted.

In most cases, getting that all-important information requires moving data quickly from one application or repository to another, where it can be processed and analyzed.

Veracity

The first three Vs—variety, volume and velocity—define big data. In contrast, the fourth V—veracity—is a desirable aspect of data that organizations want, but don't always get.

Veracity is directly related to the trustworthiness of an organization's data. The first step toward building trust into data is to understand which data is needed and for what business purpose. Can knowledge workers be sure that the data is accurate, current and complete—and therefore be confident in any decisions that are based on that data?

Big data, big opportunities

Big data presents significant opportunities for increased growth and profitability—and forward-thinking organizations are already realizing some of these benefits:

- **A retailer actively uses social media data to analyze customer sentiment and improve loyalty, helping to drive revenue growth**
- **A transportation company uses machine data to optimize logistics, thereby reducing shipping costs**
- **A telecommunications company slashed developer costs by over 50 percent by articulating clear terms and policies for the data it needed**
- **A manufacturing company enhanced the trustworthiness of its reports after it discovered and eliminated 37 unique definitions of "customer" across its enterprise, and agreed to a single, standard definition**

Capturing these sorts of benefits from big data requires knowing what the business needs and being able to find key items within the larger mass of big data.

Begin by articulating the goals of the business. Determine the analytics and reports necessary to support those business objectives. Now you can make decisions about the data needed to inform the reports.

Standard terminology is a critical piece of this process. For example, there should be no misunderstanding about what constitutes a "sale" or a "customer." With standard terms, organizations can create standard policies, such as defining the data that comprises a complete customer record.

Armed with objectives and standard terms and policies, teams can create data processing flows that automate the extraction of relevant data from the mass of big data. Importantly, this understanding also includes knowing at all times where data is coming from, how it is being manipulated and where it is going. This chain of understanding builds trust into information—and promotes confident decision making.

You can’t govern what you don’t understand

It’s an unfortunately common refrain across business and IT users when it comes to discussing data: “That’s not what I meant!” Organizations, departments and sectors often use different terms or labels for the same information, leading to confusion over details such as what constitutes a customer (is it a household or an individual?).

In one case, a large manufacturer had 37 different definitions of the term “Employee ID” across multiple divisions. Using spreadsheets to record and compare the representations, the company discovered at least 15 different data types and formats, with different characteristics and usages—a result of multiple legacy systems, acquired companies with their own systems and home-grown applications. This meant that IT spent extra time and used undocumented knowledge to determine which version to use in which reports. Developers and analysts then spent more time determining which data to use in which tests, examples and tasks. And business users were hampered by data quality issues coming from duplicated or missed data.

A large European telecom company faced similar issues when it discovered not just terminology problems but also an inability to trace data back to its source. Using InfoSphere Business Information Exchange, the company standardized business terms and policies, which helped increase confidence in reports. Next, it redesigned integration processes and built a clear blueprint, reducing time required to write actual integration jobs by 50 percent. Finally, it used the InfoSphere tool to establish a clear metadata lineage—from source to target, including all transformations—that uncovered redundant data. Consolidating and archiving that data led to a 75 percent reduction in storage space.

Understanding: The key to successfully leveraging big data

Successful businesses depend on trusted information. IBM InfoSphere Business Information Exchange helps organizations create an understanding of big data that enables it to be converted into trusted information. It allows business users to play an active role in information-centric projects and to collaborate with technical teams—all without the need for technical training. The end result: an organization with a consistent understanding of information, what it means, how it is used and why it can be trusted. Decisions are more accurate and business opportunities can be readily captured.

Defining "truth"

In a philosophy class, students might consider the idea that truth is ultimately unknowable or that truth is constantly changing. Most organizations do not have that luxury. Enterprises must have agreed-upon definitions for important terms so they can monitor and act on key metrics.

Consider a global financial institution that has grown rapidly through mergers and acquisitions. Because each of the company's various lines of business grew independently, each has its own unique processes, IT systems and definitions for important terms. It isn't uncommon for the CIO, CFO and CMO to use different definitions, which can lead to confusion when it comes to analytics and reports.

InfoSphere Business Information Exchange includes a business glossary that enables business and IT to create and agree on definitions, rules and policies. InfoSphere Business Information Exchange also includes data modeling capabilities that allow data architects to determine where each piece of data will come from and where it will go. This capability helps ensure that everyone involved in the big data project knows exactly what key metrics mean and where the data should originate—establishing the "truth" as it relates to their business data.

Defining "trustworthy"

Unfortunately, it isn't enough just to establish policies and definitions and hope that people will follow them. To be truly confident that their data is trustworthy, organizations must be able to trace its path through their systems, see where it came from and understand how it was manipulated.

InfoSphere Business Information Exchange provides this level of transparency. It includes data lineage capabilities that enable users to track data back to its original source and to see every calculation performed on it along the way, increasing the data's accuracy and trustworthiness.

Defining "good"

The "goodness" of data depends upon its usefulness in analysis, reporting and decision making. Therefore, developing an understanding of big data requires companies to separate useful "good" data from unhelpful "bad" data. In other words, organizations must be able to extract only those bits of data necessary to support a particular business objective, and set aside the rest. By filtering data in this way, unnecessary data is kept out of data warehouses and Hadoop file systems, creating a more efficient processing environment and lessening hardware and software costs.

The data quality and data governance capabilities of InfoSphere Business Information Exchange help companies know for certain that their data is good, trustworthy and true. That certainty allows them to be more confident in the results of their analytics activities and to take action quickly and assertively. When they have a solid foundation of good data, business leaders can build a more flexible and agile company that will be better able to identify and capitalize on business opportunities.

Developing an understanding of big data requires companies to separate useful "good" data from unhelpful "bad" data.

InfoSphere Business Information Exchange: The foundation for understanding big data

The first step toward proper information governance involves establishing correct data definitions that the entire organization can use to better understand information. While a full understanding of business context and meaning resolves ambiguity and leads to more accurate decisions, users often require more detail behind their data.

InfoSphere Business Information Exchange offers a solution to this problem. Integrated with InfoSphere Information Server metadata, it enables organizations to create and then link business terms to technical artifacts.

Here are some key features:

- **Web-based management of business terms, definitions and categories** enables the creation of an authoritative and common business vocabulary for technical and business users
- **Integration with InfoSphere Information Server metadata** helps ensure that technical and business information is always connected and consistent
- **Security permissions** help protect sensitive business terms and definitions from unauthorized users
- **Customizable features and attributes** enable business users to define unique parameters for their specific organizational and business environment

- **Collaborative environment and feedback mechanisms** encourage organic growth and allow different glossary users to jointly develop or improve the glossary content
- **Powerful glossary import and export capabilities** enable administrators to combine existing fragmented and home-grown glossaries into a single enterprise glossary for use by a wider business audience
- **Data stewardship** empowers ownership of business-term integrity and its governance
- **Terms** can be linked in an easy-to-use web interface to create policies that govern information objects
- **Metadata lineage** is captured and maintained so that information contained in reports and applications can be easily traced back to original sources for validation (a critical step in meeting compliance requirements for regulations such as Basel II and the Sarbanes-Oxley Act)
- **Globalization and translation support** for simplified Chinese, traditional Chinese, Japanese, Korean, French, German, Italian, Spanish and Brazilian Portuguese allows customers around the world to use InfoSphere Business Information Exchange in their native language

To make matters even simpler for businesses embarking on a data governance and business glossary solution, IBM provides packaged business terms and definitions for six major industries: banking, financial services, retail, telecommunications, healthcare and insurance. These content offerings help accelerate the implementation and deployment of your business glossary by immediately providing rich, industry-specific business terms and definitions.

Why InfoSphere?

The IBM InfoSphere platform for information integration and governance combines the capabilities necessary for creating trusted information from traditional and new sources of big data. It provides an enterprise-class foundation for information-intensive projects, offering the performance, scalability, reliability and acceleration needed to deliver trusted information faster. These capabilities include:

- **Metadata, business glossary and policy management:** Defining both metadata and governance policies with a common component used by all integration and governance engines is a critical task. InfoSphere Business Information Exchange contains capabilities for data discovery, metadata management, governance policy definition and management, and governance project blueprint design, as well as a business glossary of terms and definitions.
- **Data integration:** The InfoSphere platform offers multiple integration capabilities for batch data transformation and movement (InfoSphere Information Server), real-time replication (InfoSphere Data Replication) and data federation (InfoSphere Federation Server).
- **Data quality:** InfoSphere Information Server for Data Quality has the ability to parse, standardize, validate and match enterprise data.
- **Master data management (MDM):** InfoSphere MDM manages multiple data domains, including customer, product, account, location, reference data and more. It handles any domain or style and offers the flexibility to define custom domains as required.
- **Data lifecycle management:** IBM InfoSphere Optim™ manages the data lifecycle from test data creation through the retirement and archiving of data from enterprise systems.
- **Data security and privacy:** IBM InfoSphere Guardium® activity monitoring solutions continuously monitor data access and protect repositories to prevent data breaches and support compliance. The InfoSphere Optim Data Masking solution masks data in applications to help protect sensitive data.

Additional resources

To learn more about the IBM approach to information integration and governance and the IBM InfoSphere platform, please contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/software/data/information-integration-governance

Please also check out the other e-books in this series:

- **Mastering big data with MDM**
- **Integrating trusted and protected information for big data**
- **Managing the lifecycle of big data**
- **Protecting and monitoring sensitive big data**

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
July 2013

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 performance data discussed herein is presented as derived under specific operating conditions. Actual results may vary.

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.

Actual available storage capacity may be reported for both uncompressed and compressed data and will vary and may be less than stated.

[1] Gantz, John and Reinsel, David. IDC. “THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East.” December 2012.
[2] **ibm.com**/software/data/bigdata

♲ Please Recycle