IBM Software
White Paper

Integrating and governing big data

David Corrigan
Director, Product Marketing
IBM InfoSphere

The new era of computing requires trusted information

With the dawn of a new era of computing and the growth of big data, information integration and governance is more important than ever before. Trusted information is the foundation for analytics and business intelligence—without it, decisions will not be made and insight will not be acted upon.

Three unique characteristics of the new era of computing make information integration and governance especially relevant in today's business world:

Information from everywhere: Information is exploding, both in volume and variety. The amount of digital data is set to grow to 44 times its current volume by 2020.[1] And by 2015, nearly 80 percent of data—including transaction data, social data, content and machine data—will be uncertain,[2] meaning many aspects about it may be uncertain, such as its origin, quality, source and accuracy. Exponential growth in the number of data sources will continue to make gathering trusted information a challenge. The role of information integration and governance is to mask that complexity and make it as easy to govern a complex architecture as it is to govern a single database.

Radical flexibility: The real question facing organizations is whether their governance is agile enough to meet rising business expectations. Business users will proceed with new applications, new data marts and reports as they see fit—working around enterprise rules, creating more complexity and increasing ungoverned information. An agile information integration and governance platform can quickly adapt to users' changing business needs over time to help ensure that information remains trusted.

Extreme scalability: The growth of data demands extreme scalability for all applications—especially those that address information integration and governance. As data volumes grow, more data passes through the integration hub at the heart of your infrastructure. With extremely scalable information integration and governance technologies capable of high performance, businesses can continue to meet the rising demand for trusted data.

By controlling how information is created, shared, cleansed, consolidated, protected, maintained, retired and integrated within your enterprise, information integration and governance strategies turn uncertain data into trusted information. The result? Analytic and operational applications can overcome the dual challenge of rising uncertainty in data combined with overwhelming growth in the volume and variety of information. As the foundation for certain, actionable insight, sound information integration and governance enables organizations to trust before they act.

Information has an inherent value

Information—especially good information—has an inherent value. Analytics help unlock this value, but the quality of the information determines the worth of the resulting insight: bad data results in poor recommendations; trusted, high-quality data results in a more complete and accurate analysis.

Unfortunately, bad data has a massive impact on businesses' bottom line. The Data Warehousing Institute estimates the cost of untrustworthy "bad" data at USD600 billion annually—and that is only for businesses in the United States.[3] Researchers predict the volume of data will grow 50 times in the next eight years,[4] and it is reasonable to assume the cost of bad, untrustworthy, uncertain data will grow by a similar factor.

How does such an important issue get so large and so painful without being addressed? The "cost" of untrustworthy data is experienced through other applications—ones that create and use information, such as data warehouses, big data analytic applications, business intelligence (BI) and enterprise operational applications. Their return on investment is often hobbled by untrustworthy information. Worse still, the adoption of those applications may be stifled if users don't trust the information and insights coming from those systems.

In the new era of computing and with the rise of big data, organizations must get serious about governing their information or they will face staggering costs and missed opportunities.

The inherent value of data is unlocked only by governing it, which helps reduce the costs of big data—including the time and effort required to track down multiple, duplicate records; expensive data breaches; reputation-killing security issues; and more.

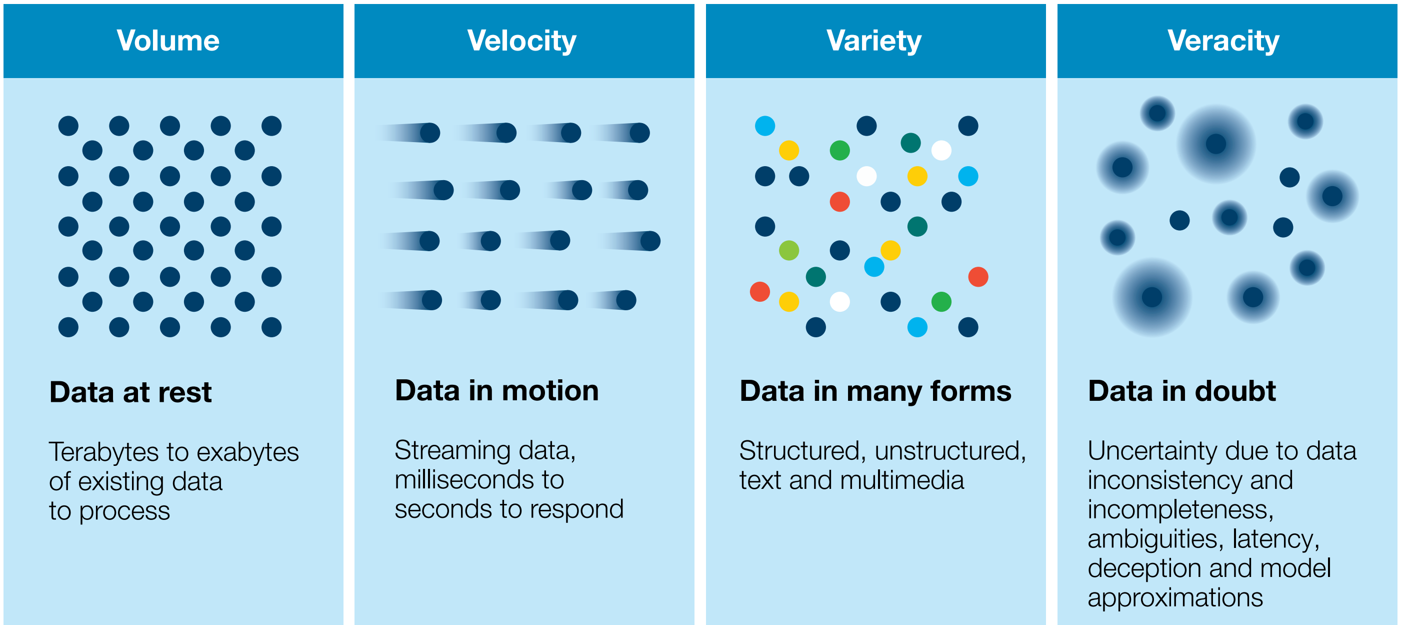
The outcomes that businesses want to achieve with big data analytics are not possible without trusted data. Information integration and governance technologies help organizations truly manage information as an asset, and they are the foundation of successful analytics, BI and big data initiatives.

With the dawn of a new era of computing and the growth of big data, information integration and governance is more important than ever before.

Big data is a phenomenon, not a technology

Cloud, mobile and social technologies are often mentioned alongside big data. However, it is crucial to think of big data as a phenomenon rather than a singular technology.

In every system, the volume of data is increasing, data is being produced at an increasing velocity, data types and formats have more variety, and data veracity is becoming more uncertain. That means big data affects every application in your enterprise in four areas: volume, velocity, variety and veracity (see Figure 1).



Volume

Data at rest

Terabytes to exabytes of existing data to process

Velocity

Data in motion

Streaming data, milliseconds to seconds to respond

Variety

Data in many forms

Structured, unstructured, text and multimedia

Veracity

Data in doubt

Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations

Figure 1: The definition of big data involves the volume, velocity, variety and veracity of information.

Big data affects every aspect of your IT organization

The big data phenomenon (the increasing volume, variety and velocity and the decreasing veracity of data) affects every IT system and project. Although sometimes big data is discussed solely in relation to Apache Hadoop systems, most IT professionals now realize that the phenomenon of big data is affecting all of their systems, and therefore brings a new set of requirements. The interest in information integration and governance for new big data sources has mirrored the maturity curve of new big data technologies—as they moved from research to production, suddenly the question of trust and security became much more critical. Today, the answer is clear: new sources of big data must be governed.

There are five categories of problems addressed by information integration and governance, all of which are affected by the phenomenon of big data:

- Trusted information for data warehousing and big data analytics
- Improve application efficiency
- Protect and secure enterprise data to ensure compliance
- Consolidate and retire applications
- Act on a trusted view

Trusted information for data warehousing and big data analytics

Analytics applications rely on a big data platform to process and analyze information. In turn, the analytic engines of the big data platform rely on a common foundation of information integration and governance for trusted, certain information in order to return accurate, actionable results, and to integrate those insights into other enterprise systems.

Improve application efficiency

Data is growing in every application—ERP, CRM, transactional systems, data warehouses and more. Expanding data volumes impact performance and efficiency in the form of longer queries, slower processes and unproductive business users. What's more, higher data volumes mean greater complexity for application upgrades, leading to more downtime and lost productivity. Information integration and governance improves application efficiency by archiving according to compliance regulations, while also streamlining test data management to ensure faster implementation times for applications.

Protect and secure enterprise data to ensure compliance

Because sensitive data is increasingly shared among many systems for a growing number of uses, it must be masked or redacted. New technologies such as Hadoop must be monitored to protect against data breaches. In addition, compliance reporting must be automated to support cost-effective compliance. Information integration and governance provides an agile approach to data security that can protect data no matter when and where it is used—today and in the future.

Consolidate and retire applications

Maintaining efficiency means consolidating and retiring systems to keep growing complexity in check. Many Fortune 1000 companies have thousands or even tens of thousands of applications and repositories of information. Organizations need to develop a core competency in application retirement, which includes archiving and retiring old applications as well as integrating data into new systems. With information integration and governance, application consolidation and retirement can be fast, simple and compliant.

Act on a trusted view

More systems mean more fragmentation. Organizations need a consolidated view of data that is spread across thousands of repositories to derive true insight from new analytic applications. Because complexity will only increase, now is the time to start building a single view of customers, products, suppliers, accounts and other key business entities.

Big data technologies (see Figure 2) include:

- Hadoop-based systems for processing and analysis
- Stream computing to analyze data in motion
- Massively parallel processing data warehouses and appliances
- Federated discovery and navigation to visualize and search big data repositories

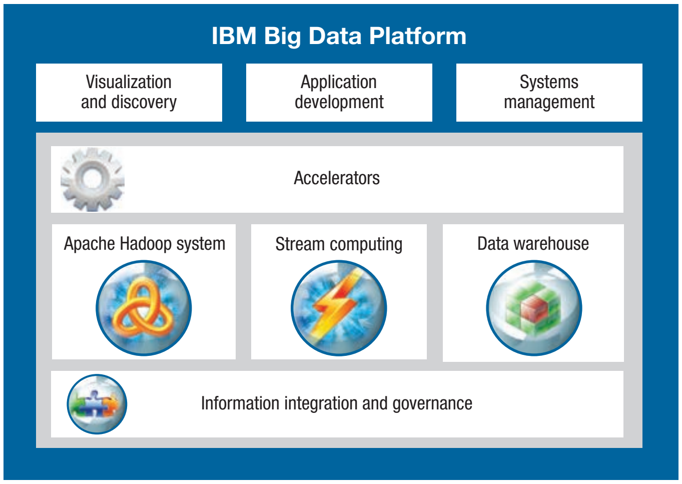


Figure 2: The IBM Big Data Platform features capabilities to manage, analyze and navigate high volumes of data.

All of these capabilities act with certainty by relying on the foundation of trusted information provided by information integration and governance.

Getting started on governing big data

Big data represents an evolution for information integration and governance technology. The notion that existing information integration and governance capabilities can be leveraged for big data governance must be evangelized within your organization. You must identify a business problem that will drive a big data project in order to identify your first tactical big data governance project. It is critical to outline the tactical first phase and second phase (develop a road map), quantify the business value, describe the overall strategy for information governance and establish a governance organization to drive the project toward completion (see Table 1).

When correctly planned, information integration and governance initiatives will improve the return on investment for analytic and operational applications by helping to resolve the underlying data trust and organizational complications that can cripple successful deployment.

Manage change to build trust

Information governance always involves change; by definition it involves determining and implementing policies that ensure the trustworthiness of data for the enterprise. But governance does not have to be a burden. Finding the right balance between processes and people—and the technology to support them—for an initial phase while being mindful of long-term goals results in agile information governance.

Table 1: Six best practices for successful big data governance initiatives

Identify a business problem	Locate a business problem that requires big data analytics and a supply of trusted, governed information. You may find in-progress big data projects that have not yet factored in information governance; those are an excellent potential first big data governance project.
Identify an executive sponsor and gain their support	Get the buy-in of a sponsor with the organizational clout to mobilize change.
Develop a road map and a plan	Show the organization the long-term objective and provide details on the first steps to get there.
Develop a business case	Quantify metrics and the overall business value, and tell qualitative stories of the impact in Phase 1.
Define metrics	Define metrics and targets for the key drivers of your business case and get buy-in to measure them.
Establish a governance organization	Include both business and IT constituents to discuss and resolve governance policies, and establish an executive steering committee to oversee status and help remove roadblocks.

A governance board or council is key to success. By setting up a governance structure, you can proactively avoid complications and ensure the buy-in of business and IT constituents. Agile information governance allows for multiple "sprints" to define policies, implementing them via information integration and governance technology and then moving on to the next sprint of policies. A governing body keeps agile information governance moving at full speed.

Measure and communicate

Post-implementation measurement and communication are critical to ensure ongoing success for governance and big data projects. But measuring everything is not always reasonable. Monitor the top five metrics that support the business case (for example, the number of addresses standardized and records merged for a data quality initiative), and then communicate those findings early after the first phase to key stakeholders. Customer stories and anecdotes from users can also help get the story out to the broader organization, building the buy-in necessary for information governance initiatives to be a true success.

Big data and governance: When, how and how much?

The type of big data project you plan to implement is the major factor in determining how you govern your data.

Value and usage of data are two very useful ways to plot big data projects. Some data has short-term value and expires quickly. Other data has long-term value and must be retained for years. Similarly, data usage provides a general guideline for governance. Some data is used for aggregate or anonymous analysis, whereas other data is analyzed or used at the individual record level. Information integration and governance needs can vary depending on retention, perception, recognition and preservation requirements, as shown in the quadrants on Figure 3.

Perceive

The projects in this quadrant assemble data to help identify trends—to pinpoint consumer sentiment using social media analysis, for example. Data accumulates quickly and has a short half-life; therefore, it must be integrated quickly.

The role of information integration and governance for this category is to deliver data, ensure requisite consistency, protect sensitive data and ensure it is deleted or archived in a timely manner. Data lifecycle policies are often applied at the aggregate level (for example, delete last month's data but archive and retain the items that were used to determine a particular insight). In these cases, retention and archiving policies are important for controlling data growth. Sensitive data should be masked to ensure that it remains realistic while protecting privacy and security. Data quality may be applied to ensure some level of consistency to facilitate analysis, but not all aspects of quality must be rigorously applied.

Retain

These types of projects are similar to the Perceive quadrant, except that the data is retained longer for historical analysis. In general, the longer data is retained, the more governance is required. Examples include demographic analysis and inventory forecasting.

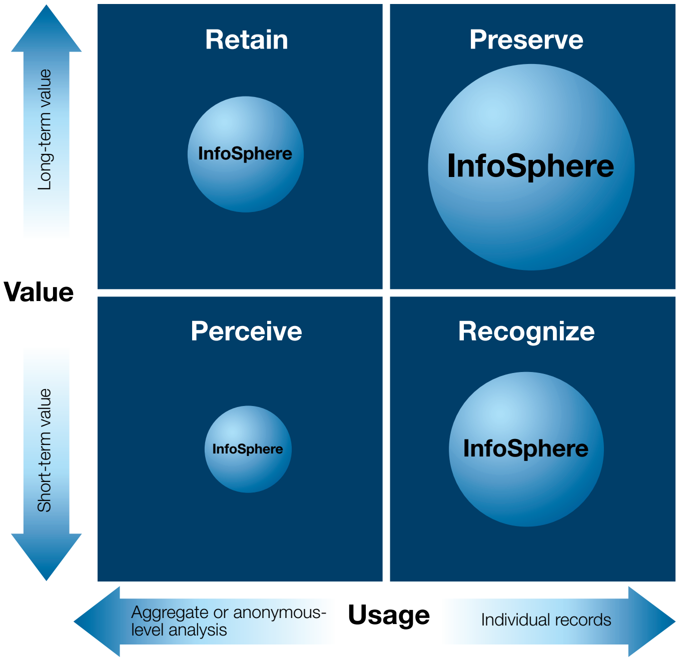


Figure 3: As data's value and usage requirements change, it requires a greater level of governance.

Projects in this quadrant are focused on greater consistency—data quality capabilities are applied to help ensure data is stored in a more consistent format. Test data management helps to ensure that the system and upgrades are streamlined with both right-sized efficiency and protected test data. An expanded data set means that integration requirements may grow beyond traditional batch extract, transfer and load operations to include a mixture of replication and/or federation. Data lifecycle management also remains an important capability to keep data growth in check.

In this quadrant, information integration and governance helps to ensure greater consistency and the inclusion of data from many more sources via multiple integration methods.

Recognize

The projects in this quadrant—machine data or campaign analysis, for example—are fairly similar to the Perceive quadrant in that the half-life of the data is very short. However, these projects are distinctly different in that they may focus on recognizing individual records rather than broad trends.

The scope of data quality is broadened for these projects, so the role of governance goes beyond consistency to ensure correctness. Data validation and matching are employed, and master data management (MDM) provides unique master entities from a fragmented set of data sources. Archiving to control data growth, efficient test data management and integration of various types (batch, replication, federation) remains important. In addition, agility is a key concern because of the short timespan for data analysis; the faster data can be integrated and governed, the sooner an organization can capitalize on its value.

In this quadrant, information integration and governance includes agile integration, data retention and archiving; expansion of security and privacy to mask sensitive individual data record attributes; and a growth in the scope of quality and master data to recognize individual records.

Preserve

These projects, ranging from mission-critical enterprise applications to big data analytics and financial reporting systems, have the most exacting requirements for governance—individual records that must be preserved for the long term.

In this quadrant, individual records must be preserved and they must be absolutely correct. MDM plays a prominent role in ensuring that trusted data is accurately maintained, and data quality initiatives help ensure that information is standardized and validated. The focus for lifecycle management shifts from aggregate policies (deleting and selectively archiving a block of data) to the individual record (archiving particular customer records). Business objects—not just data tables—are archived and retrieved.

Privacy and security are prominent concerns in this quadrant. Data repositories must be monitored to ensure that there have been no breaches, and sensitive data must be masked or redacted as it is shared among multiple systems. Data integration must cover batch processing, replication and federation to help ensure that the individual records are accurate and up to date.

As with any generalization, this model is only meant to approximate different types of information integration and governance for different types of big data projects. But without exception, every big data project requires some level of information integration and governance capabilities.

A single platform for information integration and governance

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 (see Figure 4).

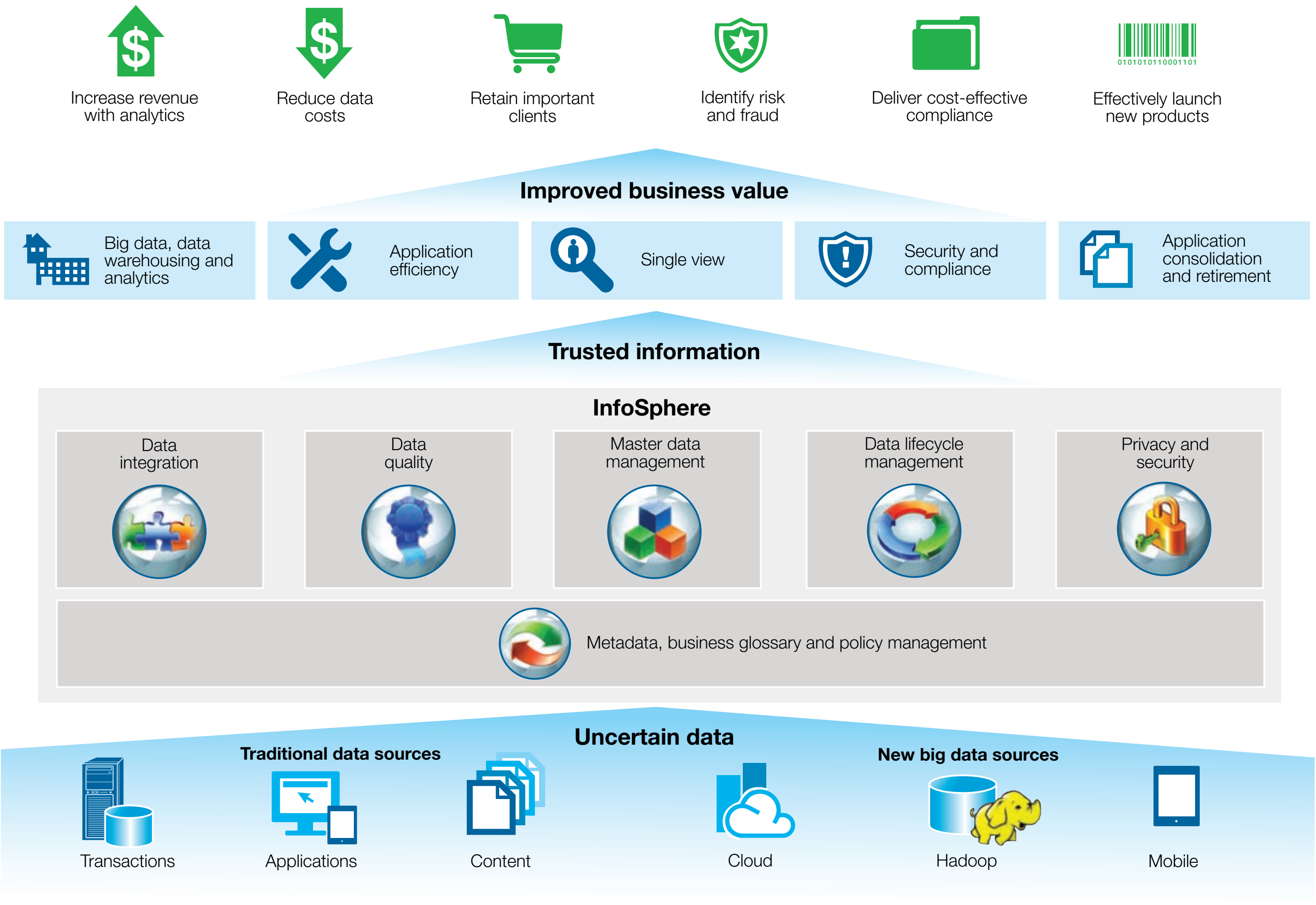


Figure 4: The IBM InfoSphere platform creates trusted information from uncertain data sources.

These capabilities include:

- **Metadata, business glossary and policy management**—Define both metadata and governance policies with a common component used by all integration and governance engines. IBM InfoSphere Business Information Exchange contains capabilities for data discovery, metadata management, business glossary of terms and definitions, governance policy definition and management and governance project blueprint design.
- **Data integration**—The InfoSphere platform offers multiple integration capabilities for batch data transformation and movement (IBM InfoSphere Information Server), real-time replication (IBM InfoSphere Data Replication) and data federation (IBM InfoSphere Federation Server).
- **Data quality**—IBM InfoSphere Information Server for Data Quality has the ability to parse, standardize, validate and match enterprise data.
- **Master data management**—IBM InfoSphere MDM manages multiple data domains, including customer, product, account, location, reference data and more. InfoSphere MDM handles any domain or style and provides 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.
- **Privacy and security**—The IBM InfoSphere Optim Data Masking Solution masks data in applications to ensure sensitive data is protected. IBM InfoSphere Guardium® monitors repositories to prevent data breaches and support compliance.

Build a reliable foundation of trusted information for big data analytics

In the new era of computing, the growing complexity of data sources and variety of data types creates a unique set of challenges. Left unchecked, these expanding volumes of data are becoming more uncertain and more difficult to navigate. Information integration and governance counters this trend by creating trusted information from new sources of big data and traditional sources of data, and delivering it to the applications that run your business. By creating trusted information that becomes the basis for true insight and action, InfoSphere serves as the foundation for key enterprise projects for big data analytics, BI and enterprise applications.

For more information

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/infosphere

© Copyright IBM Corporation 2013

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
January 2013

IBM, the IBM logo, ibm.com, Guardium, InfoSphere and Optim 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 client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. 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. Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

[1] IDC. *The Digital Universe Decade - Are You Ready?* May 2010. http://www.ameinfo.com/231603.html

[2] IBM Research. www.research.ibm.com/new-era-of-computing.shtml

[3] Eckerson, Wayne W., *Data Quality and the Bottom Line*, Report of the Data Warehousing Institute, January 2002.

[4] IDC. *2011 Digital Universe Study: Extracting Value from Chaos*. June 2011. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Please Recycle