# Integrating and governing big data

*Does big data spell big trouble for integration?*
*Not if you follow these best practices*

IBM

# Introduction

Business leaders are eager to harness the power of big data. However, as the opportunity increases, ensuring that source information is trustworthy and protected becomes exponentially more difficult. 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.

To make the most of big data, you have to start with data you trust. But the sheer volume and complexity of big data means that the traditional, manual methods of discovering, governing and correcting information are no longer feasible. 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 big data confidence.

A solid integration and governance program must include automated discovery, profiling and understanding of diverse data sets 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 diverse technologies, from data marts to Apache Hadoop systems. And it must automatically discover, protect and monitor sensitive information as part of big data applications.

**Big data is a phenomenon, not a technology**

**With all the hype about big data, it's easy to think that big data can solve all your problems. But big data isn't a technology—it's a phenomenon. To leverage it effectively, you must be able to integrate and govern key data throughout your enterprise.**

# Integration and governance requirements for big data

Whenever the topic of big data arises, discussions often turn to analytics and Hadoop. Interestingly, big data analytics have been shifting recently toward structured data and away from its origins in unstructured data. But while analytics and Hadoop are important for both structured and unstructured data, they represent just one piece of the big data puzzle.

Forward-thinking IT professionals now realize that the phenomenon of big data is affecting all of their systems, creating a new set of requirements that impact the results of data warehousing and big data and analytics initiatives. To ensure the best results, data from big data sources must be integrated, governed and trusted.

Many of the most common challenges associated with big data aren't really analytics problems. In many cases, these problems are fundamental—even "traditional"—information integration problems, and they can be avoided or addressed with an agile, enterprise-class data integration and governance solution.

Additionally, new big data sources are not useful if they exist in a silo—they must be integrated into your enterprise architecture.

The best solutions form a solid, integrated foundation that facilitates the analytics work that yields valuable, actionable business information.

Appropriate solutions for integrating and governing big data should:

1. **Be agile**
2. **Be built on a high-performance, scalable architecture**
3. **Support greater efficiency**
4. **Help create confidence and trust in the veracity of data**
5. **Meet your requirements for flexible, agile delivery of data**

# Best practices: Integrating and governing big data effectively

Several best practices for integration and governance can help you make the most of big data in your organization.

## Embrace IT agility for performance and scalability

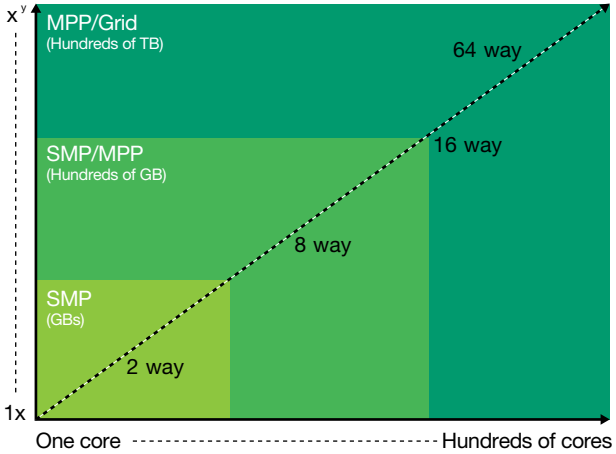
Big data streams in at high velocity—so performance is key. Data changes rapidly, and it must be fed to various applications in the system quickly so that business leaders can react to changing market conditions as soon as possible.

To successfully handle big data, organizations need an enterprise-class data integration solution that is:

- **Dynamic** to meet your current and future performance requirements.
- **Extendable and partitioned** for fast and easy scalability.
- **Integrated with Hadoop.** Hadoop itself is not an integration platform, but it can be leveraged as part of an integration architecture—to land and determine the value of data, as well as for balanced optimization.

Scalability is one of the most challenging big data integration requirements, since business requirements can evolve very quickly. Consequently, when tackling big data integration, it's important that you have a product that can achieve data scalability across all architectures with the same function and with linear speedup, scaling "N way" without issue.

**Data scalability across hardware architectures**




- Same functionality
- All architectures
- Linear speedup
- Decision (at runtime) to scale "N way"

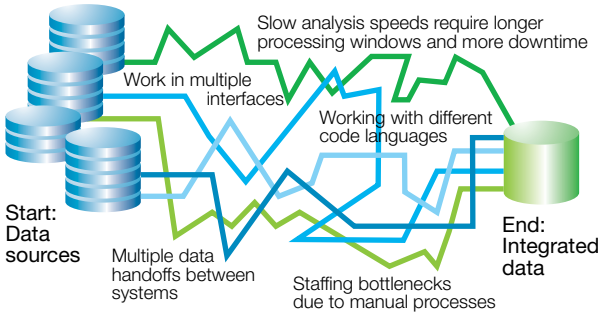
## Work smarter, not harder—and control costs

Employee time is a valuable and costly resource. An integration solution for big data that supports employee productivity and efficiency helps to improve the enterprise's bottom line, eliminate bottlenecks and enhance agility.

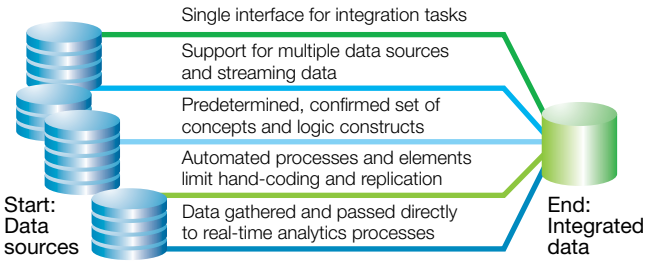
For IT departments, service-level agreements (SLAs) are often impacted by inefficiencies. As data volume, variety, velocity and veracity grows, the time required to process data integration jobs frequently exceeds the window allowed by SLAs, meaning that IT is no longer meeting the needs of internal customers.

To improve productivity, it's important to create design logic for Hadoop-oriented data integration efforts using the same interface, concepts and logic constructs as for any other deployment method. This eliminates the need to learn new coding languages as they evolve and perform hand-coding and replicating work.

**Working harder**




**Working smarter**



For big data projects focused on real-time analytical processing, it is also critical to quickly and easily integrate with systems that support streaming data (also known as "data in motion"). Big data integration solutions should be "smart" enough to allow standard data integration conventions to gather and pass data directly to real-time analytics processes.

## Create confidence with accurate, timely data

Companies usually tackle big data to augment and improve their existing analytics capabilities, either through analyzing new data sources or by tackling greater volumes of data—neither of which is possible with traditional technologies.

**However, analytics insight is only as good as the underlying data. If companies aren't feeding their analytics systems with quality data, the insight they gain is invalid.**

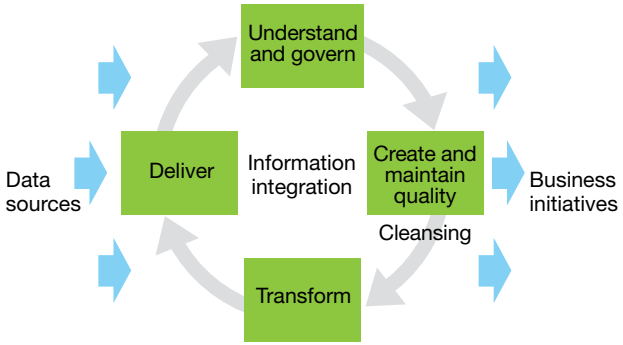
Without the ability to agree on and leverage common definitions for key business terms, businesses simply cannot be responsive and adaptable. When departments have inconsistent definitions for key terms, decisions cannot be made with the necessary speed and accuracy. For example, what happens when someone on the marketing side asks for "customer" data to analyze—but receives just a subset of the data they actually need to make a decision because the IT team defined "customer" as a household instead of an individual?

Unfortunately, it isn't enough to simply establish definitions and policies for information, and then hope that people will follow the rules. To be confident that their data is trustworthy, organizations must be able to trace its path through their systems so they can see where it came from and how it was manipulated. It's important to have a big data integration solution that can support this level of transparency.

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 most useful in a dashboard view, so business analysts can quickly determine whether there are any issues and easily get into the details.

It is important to apply data cleansing to any big data you want to retain so you can establish confidence in your data. Confidence in data quality enables confidence in analytics results.

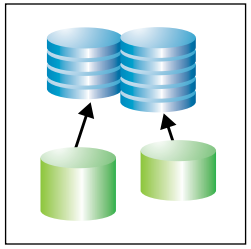
**Applying data cleansing in the integration and governance workflow**




**Cleansing data as part of the information integration cycle helps ensure data quality for the rest of the process.**

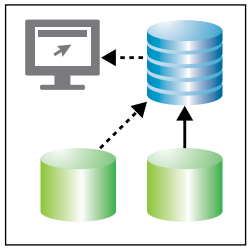
## Deliver data appropriately

When approaching big data integration projects, you want to achieve high performance and scalability for real-time data processing, as well as for bulk or batch movement. In many cases, organizations also need to leverage data replication or virtualization as part of their larger big data integration solution. This is true for traditional data integration as well as for big data integration. **Here are several good styles of data delivery that can be used along with big data platforms:**



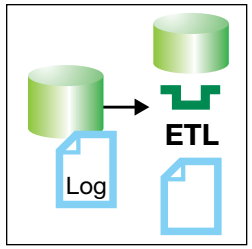
IBM InfoSphere Information Server for Data Integration

**High-speed bulk data delivery,** including ETL, ELT and dynamic integration that leverage Hadoop to support information exchange with big data sources.



IBM InfoSphere Federation Server

**Virtualized access** to and delivery from diverse and distributed information allows virtual consolidation of big (and regular) data.




IBM InfoSphere Data Replication

**Real-time integration** provides flexibility for transactional integrity plus high-volume, low-latency replication for continuous business availability.

IBM InfoSphere Data Click

**Self-service data integration** enables line-of-business and other nontechnical users to get information whenever needed to fuel their analytics.

## Leverage data replication

As the amount and variety of data in your environment builds, it will become less practical to maintain physical pools of data. To remain flexible and agile in the big data world, enterprises must leverage different technologies—including incremental data delivery—to ensure they have the data they need. Data transformation and delivery requirements have broadened from batch and bulk data movement to also include real-time data transfer based on data replication capabilities—specifically around change data capture (CDC). Whereas batch and bulk data movement happens relatively infrequently, real-time data delivery occurs whenever data at the source changes. The changed data is captured, transferred and transformed, and then loaded into the target.

Three factors impact the performance and scalability of real-time data transformations:

1. **The approach used to capture a change at the source or sources.** The most flexible and efficient option for capturing changes at the source is for a CDC mechanism to "push" changes as data streams. As soon as source data is modified, the mechanism becomes aware of the alteration and forwards the data.
2. **The mechanism used.** Many mechanisms can be used for CDC. When properly implemented, a log-based capture approach often has a lower impact on the source database, resulting in higher overall performance.
3. **Temporary data persistence.** Whether data is temporarily persisted also impacts CDC performance. Ideally, an organization would be able to stream changes without persisting them to increase performance (since data does not need to be written to disk and then accessed by a transformation engine).

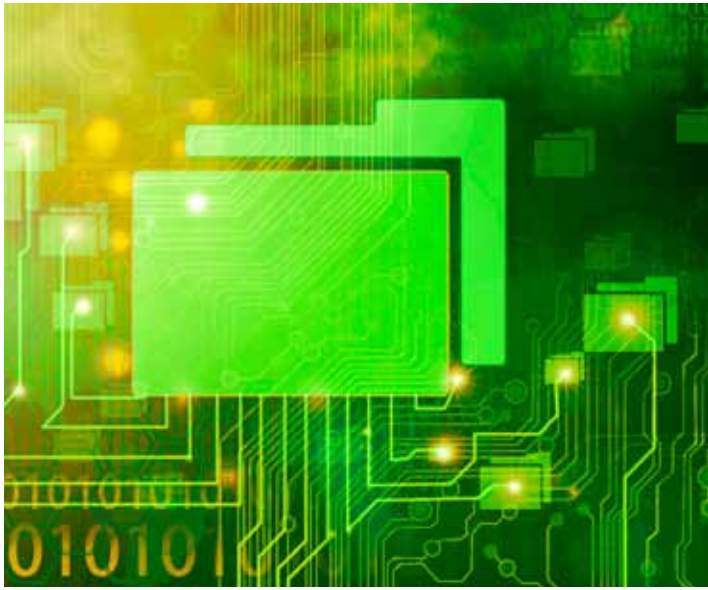
## Virtualize data

Given the massive upswings in the volume, variety, velocity and veracity of data, the question of data access is more relevant than ever before. Data virtualization technologies can help create the pool of data you need to support your business.

Data virtualization focuses on simplifying access to data by isolating the details of storage and retrieval and making the process transparent to data consumers.

By doing so, data virtualization reduces the time required to take advantage of disparate data, which makes it easier for users and processes to get the information they need in a timely manner.

Two primary strategies exist for data virtualization: data federation and data services. In both cases, data is exposed to make it more consumable, accessible and reusable by users, customers or business processes throughout the enterprise.



# IBM InfoSphere delivers the confidence to act on big data

While the term "big data" has only recently come into vogue, IBM has designed solutions capable of handling very large quantities of data for decades. The company has long led the way with data integration, management, security and analytics solutions that are known for their reliability, flexibility and scalability.

The end-to-end information integration capabilities of IBM® InfoSphere® Information Server are designed to help you understand, cleanse, monitor, transform and deliver data—as well as collaborate to bridge the gap between business and IT. InfoSphere Information Server enables you to be confident that the information that drives your business and your strategic initiatives, from big data and point-of-impact analytics to master data management and data warehousing, is trusted, consistent and governed in real time. In fact, InfoSphere Information Server is 10-15 times faster than Hadoop for data integration.[1]

## Be fast and agile

Organizations working with big data need unlimited data scalability from their integration software. InfoSphere software is designed from the ground up to optimize the usage of hardware resources, allowing the maximum amount of data to be processed per node. It has powerful data transformation and delivery capabilities, enabling clients to process data on massively parallel systems, eliminating bottlenecks and dramatically improving time-to-value.

**InfoSphere Information Server in action: Watch the demo**

**Want to see more about how InfoSphere Information Server V9.1 capabilities help you support agile integration, business-driven governance and sustainable data quality? Check out this video demonstration: ibm.com/software/data/integration/info_server/demo.html**

# The University of Arizona speeds data access with InfoSphere Information Server

With over 38,000 students and faculty, the University of Arizona's infrastructure carries a heavy load. To stay competitive, it needed to replace aging administrative computer systems that couldn't handle the demand for BI information. According to Manav Mehra, senior manager of information integration for enterprise information and analytics at the University of Arizona, the organization wanted a single source of data that users could easily query at their convenience and get results in a timely manner.

The University's BI team used InfoSphere Information Server to build that single source of trusted data; the team employed the software to understand, cleanse, transform and deliver data from source systems into its enterprise data warehouse.

The solution includes tools to help BI staff:

- Discover, model, visualize, relate and standardize diverse and distributed data sets
- Capture and define business requirements in a common familiar format to support the development of extract, transform and load (ETL) jobs
- Gain insight into data source analysis, ETL processes, data quality rules, business terminology, data models and BI reports

"On average, InfoSphere Information Server software saves us around six hours per developer in terms of data modeling and ETL job creation," says Mehra. "We had two graduate students from our MIS department help us create ETL jobs and they were able to build close to 9,000 ETL jobs from a template within three months. But more important for me is the amount of time it saves in finding and fixing data problems."

Mehra says the team can run more than 22,000 nightly ETL jobs in 2.5 hours compared to 9 hours before they introduced InfoSphere Information Server. And in the six months since deployment, use of the organization's enterprise data warehouse has significantly increased—a sign that users are finding the information they need.

**Read more about the University's experience here.**

**InfoSphere Information Server for Data Integration**

**Integration is the name of the game when it comes to boosting accuracy and efficiency. Watch this video and find out how InfoSphere Information Server helps you bring data sources together. Download the video at ibm.co/13jL5mr**

## Be efficient

InfoSphere Information Server includes capabilities that help make better use of employee time. For example, Version 9.1 includes InfoSphere Data Click, which greatly simplifies self-service data integration and provisioning. As a result, line-of-business staff can do these tasks themselves, while skilled IT engineers focus on higher-value efforts.

InfoSphere Information Server also saves developer time by providing a single design palette in a shared application environment. Developers don't have to flip through different interfaces, since everything they need is easily accessible. In addition, every InfoSphere Information Server component uses the same metadata layer. This makes it simple to track job progress and diagnose problems quickly. There is also a dashboard to provide a unified view of the environment.

As big data stores continue to grow, these high-performance and time-saving features become even more important. For IT departments, they can mean the difference between meeting SLAs or not, between having time to work on innovative new projects or low-value efforts just managing existing systems. For the business, it can mean quicker, more informed decision making—leading to stronger profits, better service for customers and competitive advantage.

## Be confident

Many enterprises have improved data quality by implementing data governance. Ideally, a data governance initiative will encompass three functions: definition of terms, cleansing of existing data and data quality monitoring.

To help individuals across the organization reach a shared understand of key terms, InfoSphere Information Server provides a data glossary, which allows business and IT to create and agree on definitions, rules and policies. Data modeling capabilities also are included, enabling data architects to determine where each piece of data will come from and where it will go. These tools allow organizations to establish "truth"—at least as it relates to business data.

To be truly confident that data is trustworthy, however, organizations must also be able to trace the path of data through their systems. To support this level of transparency, InfoSphere Information Server provides metadata and lineage capabilities that allow users to track data back to its original source and see every calculation performed on it along the way.

Furthermore, InfoSphere Information Server provides data quality capabilities to help cleanse data and monitor quality on an ongoing basis. Cleansing capabilities include sophisticated tools for investigation, standardization, matching and survivorship, enabling data stewards to fix any problems they find during their analysis. For example, names should be matched automatically, so that "William Smith" and "Bill Smith" are listed as a single customer.

**InfoSphere Information Server for Data Quality in action**

**Learn more about the four phases of cleansing and standardizing data for maximum quality—and how InfoSphere Information Server for Data Quality brings them all together. Download the video and watch now: ibm.co/17yl8nC**

## Be flexible

In many cases, “fast” isn’t enough. Delivering real-time integration is about flexibility, not just speed. One way to obtain that data is to run queries against the databases or applications where the data resides. However, that approach can slow down transactional systems so much that business leaders find it unacceptable.

A better approach: use a solution, such as InfoSphere Data Replication, that quickly captures ever-changing data and sends it wherever it needs to go, giving business managers an up-to-the-second view of vital information without slowing business-critical processes. InfoSphere Data Replication uses a “push” CDC mechanism as data streams to provide flexibility and efficiency. It also employs log-based capture to reduce source database impact, and it streams changed data without data persistence to increase performance.

Depending on your big data integration requirements, data federation can also help you adapt to your big data requirements. IBM InfoSphere Federation Server quickly creates a consolidated view of your data to support key business processes and decisions. You can access and integrate diverse data and content sources as if they were a single resource—regardless of where the information actually resides.

**The four Vs of big data**

**How do you deal with the volume, velocity, variety and veracity of big data? InfoSphere Data Replication gives you the near-real-time capabilities you need to adjust product offers, deliver reliable data and more. Download the video and learn more about the power of flexibility: ibm.co/11cy27N**

# Why InfoSphere?

As the foundation of the IBM big data platform, InfoSphere provides market-leading functionality across all the capabilities of information integration and governance. InfoSphere creates confidence in big data by making it trusted and protected. It is designed to handle big data: optimal scale and performance for massive volumes, agile and rightsized integration and governance for velocity, and support for many data types and big data systems to address the variety of data sources. InfoSphere helps make big data and analytics projects successful by delivering the confidence to act on insight.

InfoSphere capabilities include:

- **Metadata, business glossary and policy management:** Define metadata, business terminology and governance policies with IBM InfoSphere Business Information Exchange.
- **Data integration:** Handle all integration requirements, including batch data transformation and movement (InfoSphere Information Server), real-time replication (InfoSphere Data Replication) and data federation (InfoSphere Federation Server).
- **Data quality:** Parse, standardize, validate and match enterprise data with IBM InfoSphere Information Server for Data Quality.
- **Master data management (MDM):** Act on a trusted view of your customers, products, suppliers, locations and accounts with InfoSphere MDM.
- **Data lifecycle management:** Manage the data lifecycle from test data creation through retirement and archiving with IBM InfoSphere Optim™.
- **Data security and privacy:** Continuously monitor data access and protect repositories from data breaches, and support compliance with IBM InfoSphere Guardium®. Ensure that sensitive data is masked and protected with InfoSphere Optim.

## Additional resources

To learn more about the IBM approach to information integration and governance for big data, please contact your IBM representative or IBM Business Partner, or check out these resources:

- **ibm.com**/software/data/information-integration-governance
- **ibm.com**/software/data/infosphere/information-integration-big-data
- **ibm.com**/software/data/integration/info_server
- InfoSphere Information Server: A Forrester Total Economic Impact Study
- Delivering Trusted Information for Big Data and Data Warehousing: A Ventana Research Report
- Gartner: Hadoop Is Not a Data Integration Solution
- ITG: Business Case for Enterprise Data Integration Strategy: Comparing IBM InfoSphere Information Server and Open Source Tools

© Copyright IBM Corporation 2013

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
July 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 **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.

It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs. 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.

[1] IBM internal testing.

Please Recycle

IMM14125-USEN-00