IBM Software

Thought Leadership White Paper

October 2013

Who's afraid of the big (data) bad wolf?

Survive the big data storm by getting ahead of integration and governance functional requirements

Once upon a time…

Once upon a time, and not so long ago, data grew slowly over time, and it grew in a linear manner. Today, data volumes are exploding in every facet of our lives. Business leaders are eager to harness the power of big data, but before setting out into the big data world, it is important to understand that as opportunities increase, 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 traditional, manual methods of discovering, governing and correcting information are no longer feasible. Information integration and governance must be implemented to support big data applications, data warehouses and data warehouse augmentation initiatives, providing appropriate governance and rapid integration from the very start. (Once you get behind, it's like trying to catch a wolf by the tail.)

By automating information integration and governance and deploying 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 also be agile to accommodate a wide variety of data and seamlessly integrate with diverse technologies, from data marts to Apache Hadoop systems.

Build your house of straw or sticks, and it blows away

Today, performing big data analytics is strategic, and being able to supplement (or augment) data warehouses with this key information is essential. Apache Hadoop technology is an integral part of this process. Apache Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer. Hadoop clearly changes the economics and the dynamics of large-scale computing.

While Hadoop and Hadoop-based solutions have their advantages when it comes to addressing big data volumes, Hadoop is not designed for data integration. Data integration carries its own unique requirements (such as supporting governance, metadata management, data quality and flexible data delivery styles) for success. In a recent paper, Gartner states "As use of the Hadoop stack continues to grow, organizations are asking if it is a suitable solution for data integration. Today, the answer is no. Not only are many key data integration capabilities immature or missing from the stack, but many have not been addressed in current projects."[1]

Using Hadoop to build out a data integration solution is like building a house of straw. You might think you have the basic shape right, but once you've factored in the costs to make the house functional and sound by developing all the necessary data integration functionality, you might as well let this house blow away in the wind.

Be a smart architect: Insure against unpredictable winds

A house built of bricks (or, in this case, a solid and robust data integration platform) will stand up to all the elements—including the storm of increasing data volumes, velocity and variety created by the big data phenomenon.

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 business leaders can react to changing market conditions as soon as possible. To successfully handle big data, you 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)
- Part of a rich set of transformation and data quality components available out of the box to help accelerate your team's development cycle

A foundation for high-performing technology

The most critical requirement for processing large, enterprise-class data volumes for big data integration is massive data scalability (MDS). MDS provides the ability to process vast quantities of data in parallel, dramatically reducing the amount of time it takes to handle various workloads. Unlike other processing models, MDS systems optimize the usage of hardware resources, allowing the maximum amount of data to be processed per node (see Figure 1).

The limits of Hadoop data integration

Hadoop has quickly become the tool of choice for enterprise big data projects—but succeeding with Hadoop requires an understanding of what it does well and what it doesn't.

Hadoop is essentially a file system. It is a way to store big data so that it can be analyzed. The performance issues associated with using Hadoop for data processing are well known:

- Hadoop is written in Java, which is slower than frameworks in C or C++.
- MapReduce/Hadoop File System lands data between reduce steps, which is a huge performance constraint.
- Hadoop File System centrally manages an index necessary to map tasks to the data distributed throughout the nodes—a documented bottleneck.
- Operations requiring data collocation to compute the result (joins, aggregations, sorts, deduplications and so on) will run inefficiently when the data distribution is different from the index.
- Hadoop is not a good choice when real-time, low-latency processing is required because there is no "real-time" version of Hadoop available.
- Job start-up is slow, which can be a big performance penalty, particularly for small, bursty jobs.

Data integration solutions such as IBM InfoSphere Information Server offer better performance for integration workloads. And InfoSphere Information Server provides full connectivity to Hadoop, so customers can maintain their data using Hadoop while taking full advantage of the unlimited data scalability offered by InfoSphere Information Server. This offers customers the best of both worlds: inexpensive data storage with Hadoop along with unlimited data scalability, and therefore, less expensive data processing.

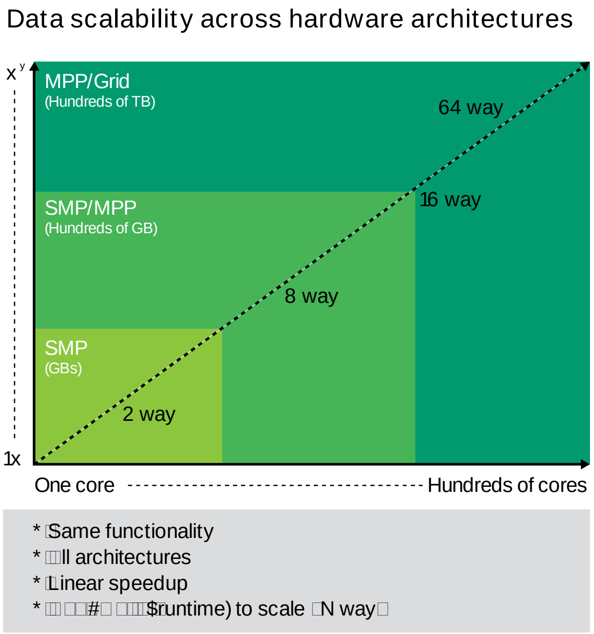


Figure 1. The essential characteristics needed to support MDS requirements enable processing of unlimited data volumes.

MDS is important because processing unlimited data volumes makes it possible to solve many high-value business problems for the first time while ensuring that a hardware platform will yield predictable benefits.

How do you know if your house is strong enough to hold back the winds of big data? It's all in the architecture. MDS systems have four key things in common:

1. Feature a *shared nothing* architecture
2. Are implemented using *software dataflow*
3. Leverage *data partitioning* for linear data scalability
4. Use a *design isolation* environment

□□□□□□□□□□□□□ architected software is designed from the ground up to exploit a shared nothing, massively parallel processing (MPP) architecture by partitioning data sets across computing nodes and executing a single application with the same application logic executing against each data partition (see Figure 2). This means there is no single point of contention, or processing bottleneck, anywhere in the system. Therefore, there is no upper limitation on data volume, processing throughput, or number of processors and nodes.

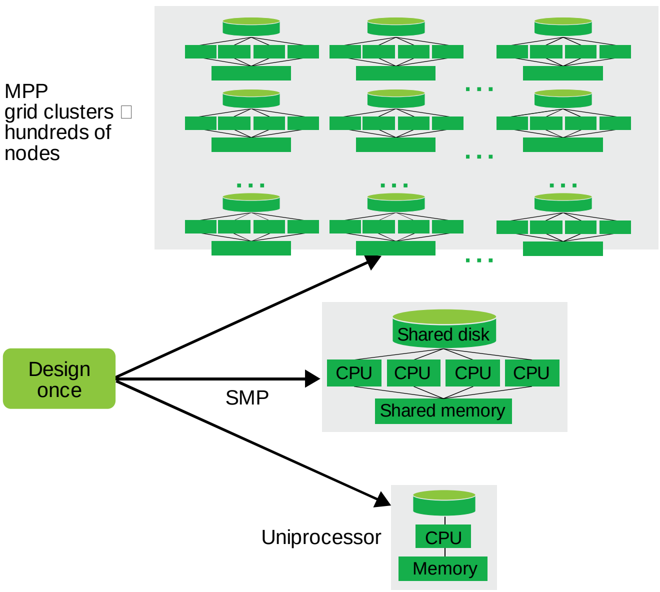


Figure 2. An example of a shared nothing architecture.

□□□□□ □□□□□□□□□□□□□ allows full exploitation of shared nothing by making it easy to implement and execute data pipelining and data partitioning within a node and across nodes (see Figure 3). Software dataflow also hides the complexities of building and tuning parallel applications from users.

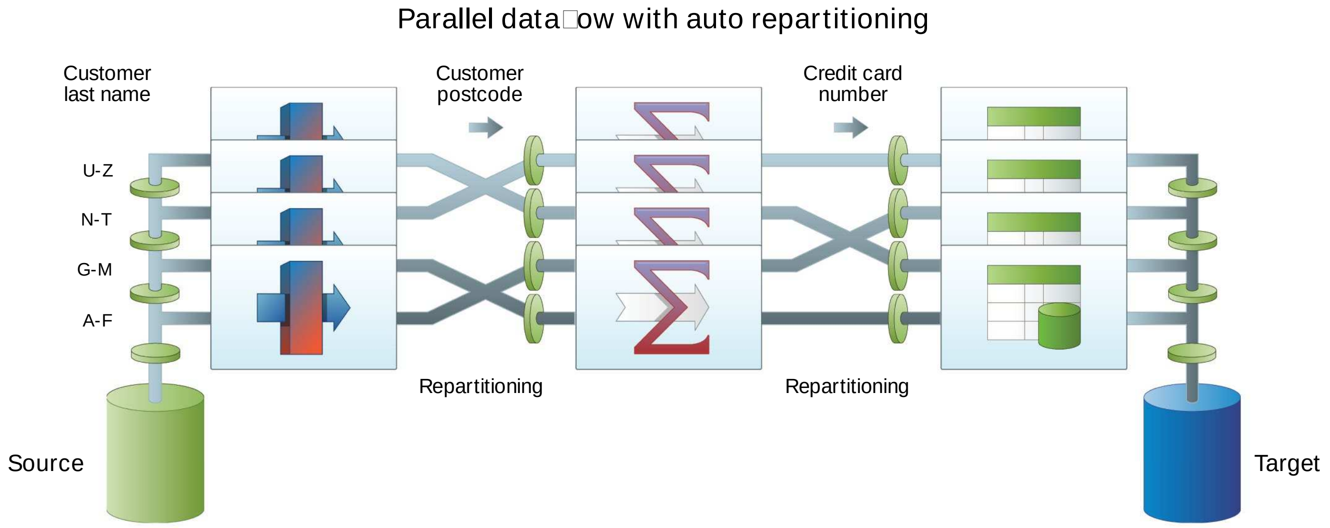


Figure 3. Software dataflow architecture.

Software dataflow is the best architecture for exploiting multi-core processors within an SMP server (application scale-up) and for scaling out to multiple machines (application scale-out). The architecture:

- Supports pipelined and partitioned parallelism within and across SMP nodes
- Provides a single mechanism for parallelization across all hardware architectures, helping to eliminate complexity
- Reduces the complexities of building, tuning and executing parallel applications
- Has no upper limit on data volumes, processing throughput and numbers of processing nodes

[illegible] means that large data sets are partitioned across separate nodes and a single job executes the same application logic against all partitioned data (see Figure 4). Other approaches, such as task partitioning, cannot deliver linear data scalability as data volumes grow because the amount of data that can be sorted, merged and aggregated is limited to what can be processed on one node.

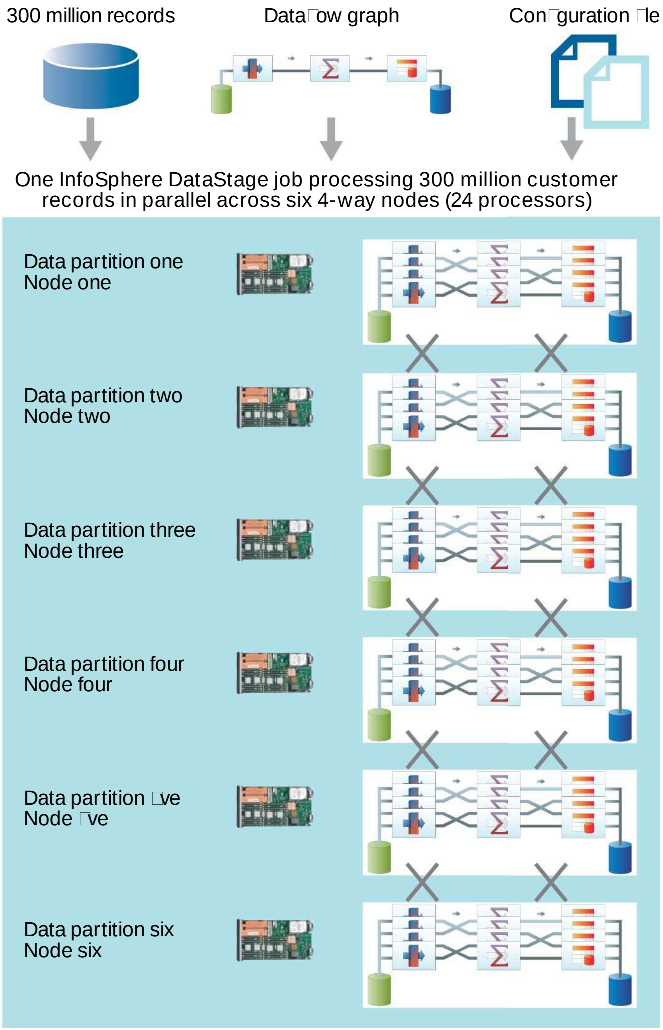


Figure 4. Data partitioning architecture.

Characteristics of systems with data partitioning include:

- Distributes data partitions across nodes
- Executes one job in parallel across nodes
- Enables pipelining and repartitioning between stages and between nodes without landing to disk
- Exploits low-cost grid hardware for big data

Finally, □□□□□□□□□□□□□□□□ means a developer can design a data processing job once, and use it in any hardware configuration without needing to redesign and retune the job (see Figure 5).

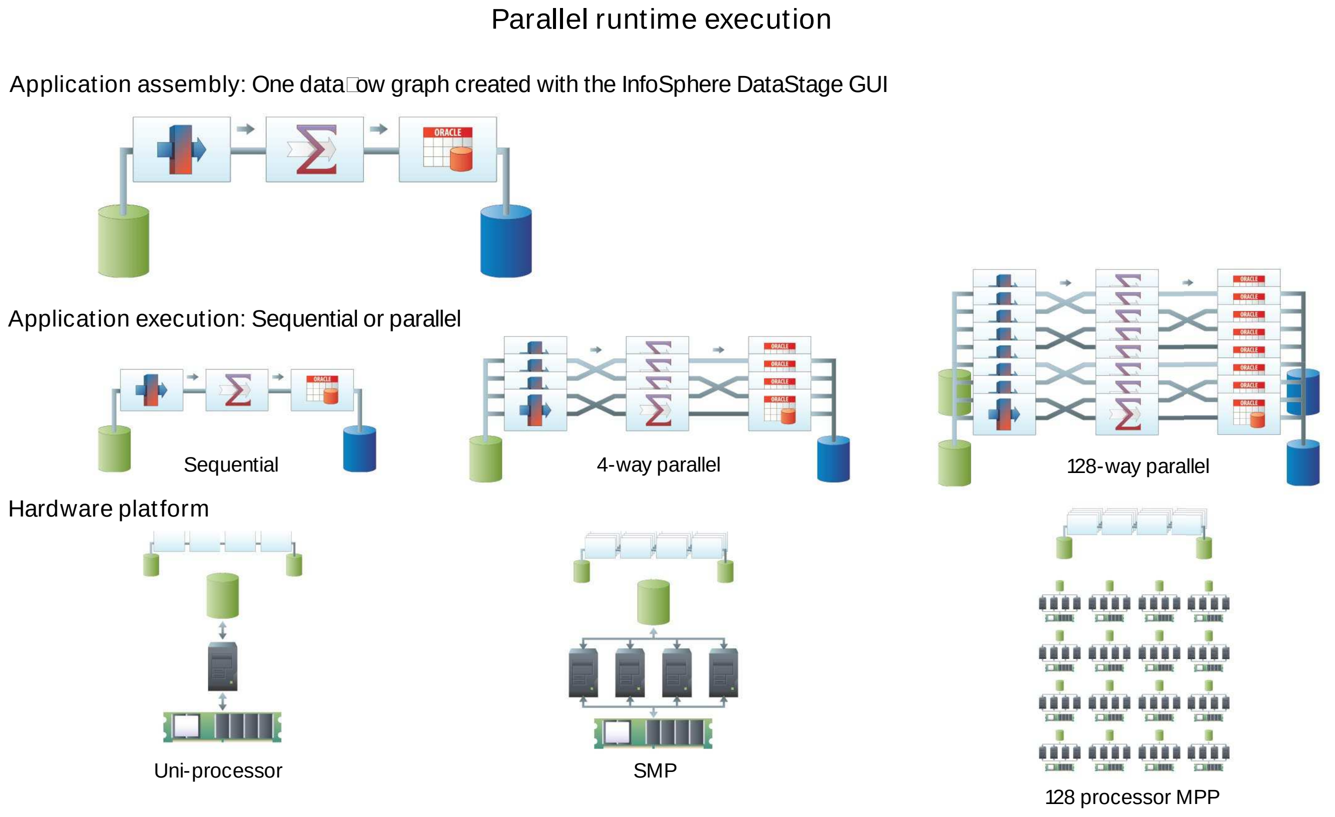

Parallel runtime execution
Application assembly: One dataflow graph created with the InfoSphere DataStage GUI
ORACLE
Application execution: Sequential or parallel
Sequential
4-way parallel
128-way parallel
Hardware platform
Uni-processor
SMP
128 processor MPP

Figure 5. Design isolation architecture.

Characteristics and benefits include:

- Build once and run without modification anywhere
- Reduce complexity with one unified mechanism for parallelizing
- Achieve a clean separation between the development of a job and the expression of parallelization at runtime
- Eliminate the need for performance tuning every time you change hardware architecture
- Add hardware with no data scalability upper limit

All of the most scalable platforms (IBM® Netezza®, IBM PureData™ System, IBM DB2® Database Partitioning Feature, Teradata, Hadoop and IBM InfoSphere® Information Server) have been built from the ground up to support these four characteristics and can seamlessly exploit MPP and commodity grid architectures.

Enhancing staff efficiency

Exploiting more efficient big data integration technology helps employees make better use of their time. The development environment should be designed to improve efficiency by providing a single design palette in the shared application environment. Developers should never have to flip through a lot of different interfaces; everything they need should be easily accessible.

An integrated, shared metadata environment is another significant time-saver. It needs to be quick and easy for developers to track job progress and to diagnose any problems. Additionally, visualizing data through a dashboard will save countless hours by giving developers and IT staff a unified picture of what's going on with their data.

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 service-level agreements (SLAs) or not, between having time to work on innovative new projects or being caught in the drudgery of managing existing systems. And for the business, it can mean quicker, more-informed decision making, which can lead to stronger profits, better service for customers and competitive advantage.

Build your house of bricks to ensure confidence

For the purpose of big data integration, what does it mean to "build your house of bricks?" It means ensuring you are being as efficient and cost-effective as possible, while also ensuring that you can get the data you need to where you need it, as you need it (that is, supporting flexible data delivery styles).

Efficiency and reduced costs

Easy self-service data integration

Wherever information resides, you need to integrate it efficiently, quickly and flexibly, easily managing information provisioning to or from data warehouses, for augmenting data warehouse projects, or for integrating big data. Self-service information provisioning, which may also be thought of as self-service data integration, is often a recurring drain on the enterprise's IT staff, requiring technically skilled extract, transform and load (ETL) engineers.

In today's fast-moving business environment, organizations must empower nontechnical, line-of-business users so that they may easily retrieve data and populate new systems (such as business intelligence and analytics systems), freeing up expert technical resources to focus on more strategic activities that deliver higher returns for the business. Additionally, nontechnical users need to be able to do this simply and on demand across large quantities of data (for example, moving thousands of tables, loading a data mart and so on). Design and operational metadata also need to be supported as part of this process so that the information may be governed and remains consistent.

Organizations need a solution that specifically supports self-service data integration needs. Such a solution needs to be able to intelligently select the right data integration mechanism for moving information—such as batch workloads that leverage the underlying information integration infrastructure or real-time workloads based on data replication capabilities. Ultimately, this type of solution should support processing optimization, help reduce the time and processing required from source to target, reduce network traffic and make data available when a full extract is not possible. A high degree of automation, an easy-to-use, intuitive interface, and broad connectivity for data sources and targets will help you realize greatest time-to-value when leveraging self-service data integration.

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, SLAs are often impacted by inefficiencies. As data volume, variety and velocity grow, the time required to process data integration jobs frequently exceeds the window allowed by SLAs. That means 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 investment in extra time and resources from learning new coding languages as they evolve, or falling back on older methods of performing hand-coding for data integration work (see Figure 6).

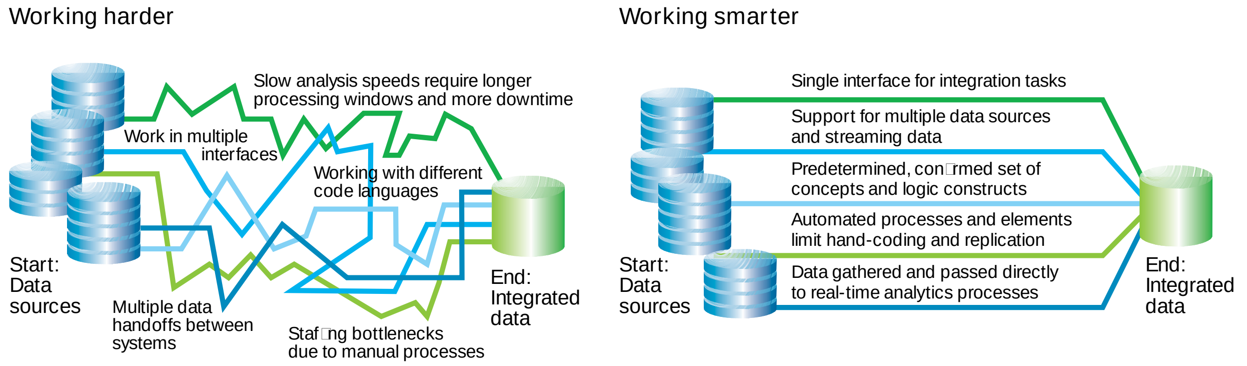

Working harder
Start: Data sources
Work in multiple interfaces
Slow analysis speeds require longer processing windows and more downtime
Working with different code languages
Multiple data handoffs between systems
Staffing bottlenecks due to manual processes
End: Integrated data
Working smarter
Start: Data sources
Single interface for integration tasks
Support for multiple data sources and streaming data
Predetermined, confirmed set of concepts and logic constructs
Automated processes and elements limit hand-coding and replication
Data gathered and passed directly to real-time analytics processes
End: Integrated data

Figure 6. By streamlining processes with automation, pre-set templates and other strategies, organizations can significantly enhance their efficiency.

Support a variety of big data sources and types

Organizations exploring big data analytics and using technologies such as Hadoop for data at rest, or streaming technology for data in motion, face many of the same challenges as in other analytical environments, including:

- Determining the location of the information sources needed for analysis
- Assessing how that information can be moved into the analytical environment
- Deciding how information must be reformatted to make it easier and more efficient to explore
- Determining which data should be persisted to quickly get to the next level of analysis

You don't want to get stuck trying to do any of this manually.

To achieve the greatest efficiency, your information integration platform needs to effectively handle the wide—and growing—complexity of heterogeneous enterprise information sources and types with a common, seamless architecture. Supporting new and emerging big data source types is essential and must include everything from Hadoop Distributed File System (HDFS) for massively scalable and resilient storage, to "Not only SQL" (NoSQL) for record storage optimized for read or write, to InfoSphere Streams for supporting massive-scale, real-time analytics. As new applications begin leveraging these technologies, you also need to ensure your information integration platform supports those systems and data types as well (such as Apache HBase, Java Message Service [JMS], Mongo DB, Hive, Cassandra, JSON and so on).

Optimize big data processing and integration

While it is important to be able to get to the big data you need without driving up costs within your organization, it's also necessary to insulate your teams and IT systems against the significantly larger volumes of data they are now facing. You can do this by creating data integration jobs that exchange data with big data sources and augment data with Hadoop-based analytics. Figure 7 shows how you can access data on the HDFS and augment data with Hadoop-based analytics.

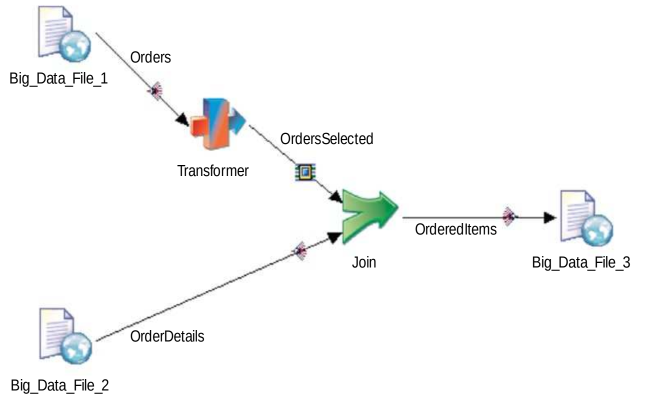


Figure 7. Accessing data on HDFS.

You also need to be able to augment data in a data warehouse with Hadoop-based analytical results. Figure 8 shows an example of how to move analytical data from a Hive data warehouse system to a Netezza data warehouse. Here, a Hive stage runs on top of a Java integration stage and provides a Hive connector to the information integration platform.

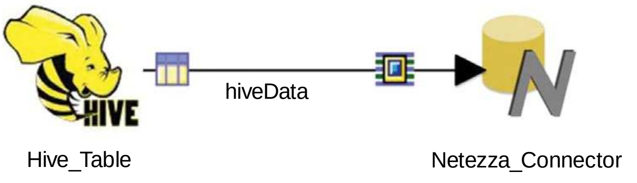


Figure 8. Augmenting data with Hadoop-based analytics.

To get the most efficient use of your big data, you should be able to push the processing to the data wherever it resides (in Hadoop, for example). Ideally, to support this effort, you would be able to use a common set of integration stages and links to build the data integration logic, and then enable your developers to choose how to run the logic (that is, to run the entire logic, or only portions of that logic) as a MapReduce job that would execute directly on the Hadoop platform. When the sources and targets of the integration task are Hadoop data stores, this approach may yield significant performance gains when removing network resource bottlenecks.

Another key capability is the ability to control the sequencing of your big data jobs, along with other processing that needs to occur in support of it. For example, consider the data warehouse augmentation use case where data is pulled from a warehouse, landed in Hadoop, analyzed alongside other unstructured information, and then the result set is put back into the warehouse. These steps all must be coordinated in order from a single point of control. The data integration platform should provide a straightforward method to bridge across such workflows—especially as your IT landscape becomes increasingly more heterogeneous.

Support a variety of data delivery styles

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, you 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. The following are several effective styles of data delivery that can be used along with big data platforms.

Leverage data replication

To maximize the amount of insight derived from big data, enterprises must employ different data delivery styles depending on the use case. Certain use cases require an up-to-the-minute (or up-to-the-second) view of data in order to make trusted decisions (see Figure 9). In certain cases, like fraud detection, inventory analysis across channels and real-time operational analytics, basing decisions on data that is a month, week or even a day old is tantamount to building one's big data house out of straw.

Figure 9. Replication helps deliver up-to-the minute insight for big data.

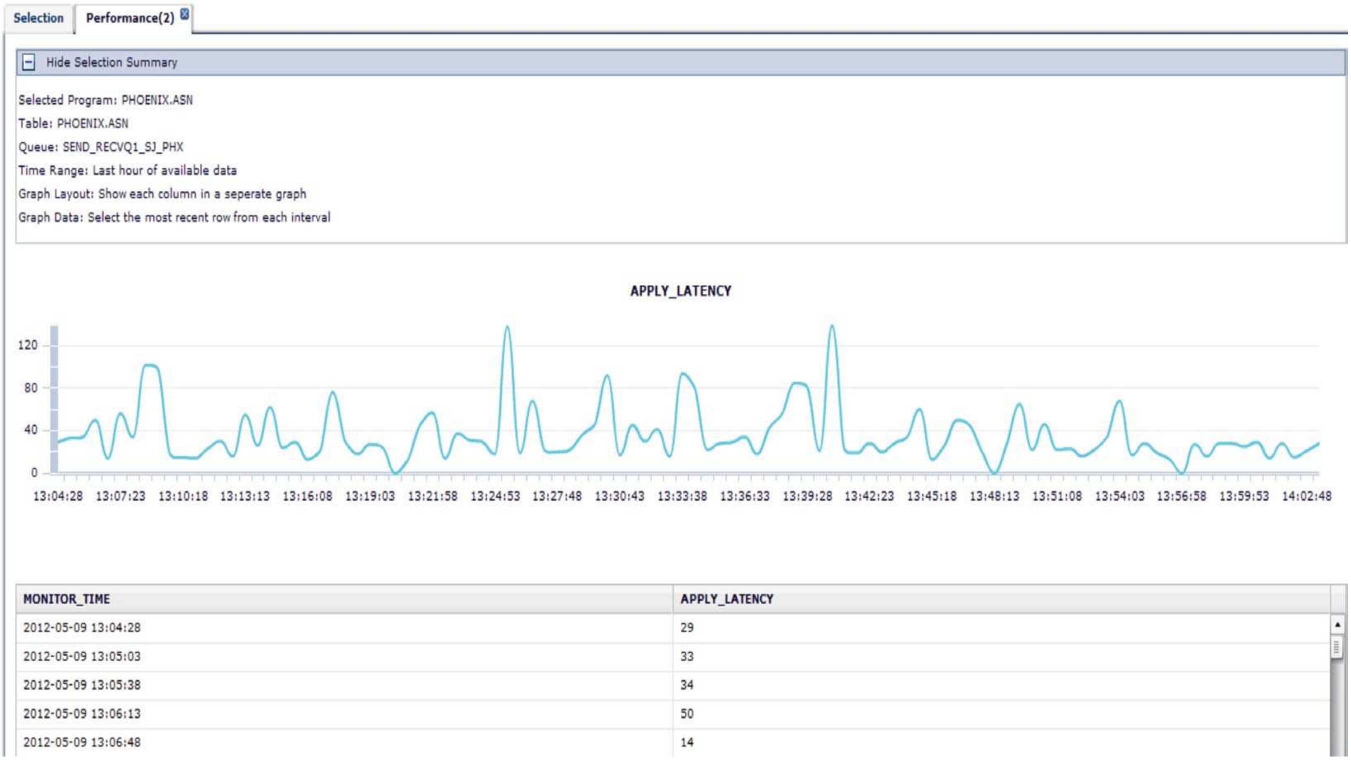
Successful data-centric organizations deliver superior experiences for their customers by leveraging insights from the information they collect faster than their competition and with greater continuity of services. In recognition of this, data transformation and delivery requirements have broadened from batch and bulk data movement to include real-time data transfer based on data replication capabilities—specifically around change data capture. Batch and bulk data movement happens relatively infrequently, but 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.

When it comes to maximizing the performance and scalability of real-time data integration for big data, there are three factors to consider:

1. [illegible] The most flexible and efficient option for capturing changes at the source is for a replication process to capture changes as they're written to a source log. As soon as source data is modified, the mechanism becomes aware of the alteration and forwards the changed data with little to no impact on the source database and application, thereby minimizing the need for large batch windows.

2. [illegible] Many mechanisms can be used for data replication. When properly implemented, a log-based capture approach often has a lower impact on the source database, resulting in higher overall performance.

3. [illegible] Whether data is temporarily persisted also impacts data replication performance. Ideally, an organization would be able to stream changes without persisting them to increase performance (because data does not need to be written to disk and then accessed by a transformation engine), as shown in Figure 10.



Selected Program: PHOENIX.ASN
Table: PHOENIX.ASN
Queue: SEND_RECVQ1_SJ_PHX
Time Range: Last hour of available data
Graph Layout: Show each column in a seperate graph
Graph Data: Select the most recent row from each interval

MONITOR_TIME	APPLY_LATENCY
2012-05-09 13:04:28	29
2012-05-09 13:05:03	33
2012-05-09 13:05:38	34
2012-05-09 13:06:13	50
2012-05-09 13:06:48	14

Figure 10. Streaming changes without persisting them can help increase performance for big data solutions.

Performance isn't the only important factor in big data integration solutions, though. Additional considerations include:

- [illegible] Flexible mechanisms support a wide variety of platforms, sources and targets, including HDFS files for big data. They will also support an equally wide variety of topologies.
- [illegible] The preferred data replication solution will be easily integrated into your existing change management processes, by allowing easy automation through scripting or common programming languages like Java.
- [illegible] Huge learning curves have an impact on self-service data integration. It is no less important to have powerful graphical user interfaces that allow easy configuration and monitoring, which will minimize the time to deriving real insights from big data.

Virtualize data

Given the massive upswings in the volume, variety and velocity of data, the ability to get simple data access is more relevant than ever. Data virtualization technologies can help create a single access point to 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. For example, an organization can create a virtual view across data in the warehouse and data that exists within Hadoop to create a new logical data warehouse. Because end users interact with only one endpoint and use the same access methods they are accustomed to, they can easily take advantage of these new insights from the big data platform without learning new languages and without administrators having to maintain user privileges in multiple locations.

What's the best data virtualization approach for you?

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. To learn more about the two approaches and which one is right for your organization, download the IBM data virtualization white paper at http://ibm.co/16QaBUW

Establish governance and quality rules to keep the big data wolf at bay

It's an unfortunately common refrain across business and IT users when it comes to discussing data: "That's not what I meant!" Organizations, internal departments and sectors often use different terms or labels for the same information. This can lead 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 from acquired companies with their own systems and homegrown applications. This meant that IT took extra time and used undocumented knowledge to determine which version to use in separate 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.

Without the ability to control the quality of the data that's guiding and driving your business, and without the ability to govern that information, you cannot ensure that the data is trustworthy. Companies need information integration capabilities to enable them to standardize business terms and policies and help improve the consistency and quality of data, which in turn can help increase confidence in reports.

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.

To ensure that veracity becomes a pervasive part of the organization, your information integration capabilities should include a business glossary that enables business leaders and IT to create and agree on definitions, rules and policies. You also need 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. To create this transparency, you need data lineage capabilities that let users track data back to its original source and see every calculation performed on it along the way, which increases the data’s accuracy and trustworthiness.

Defining ‘good’

Determining whether data is “good” involves assessing its usefulness in analysis, reporting and decision making. Therefore, developing an understanding of big data requires companies to separate “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 reducing hardware and software costs.

The path to proper information governance

So how can you do this with enterprise-class data volumes and 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. Solutions that help organizations create and then link business terms to technical artifacts provide a way to rapidly develop and deploy an information governance program. Important characteristics of such a solution include:

- Web-based management of business terms, definitions and categories to enable the creation of an authoritative and common business vocabulary for technical and business users
- Integration with metadata to help ensure technical and business information is always connected and consistent
- Security permissions to help protect sensitive business terms and definitions from unauthorized users
- Customizable features and attributes that let business users define unique parameters for their specific organization and business environment
- Collaborative environment and feedback mechanisms to encourage organic growth and allow different glossary users to jointly develop or improve the glossary content
- Easy-to-use glossary import and export capabilities that allow administrators to combine existing fragmented and homegrown glossaries into a single enterprise glossary for use by a wider business audience
- Data stewardship capabilities such as catalogs and dashboards to empower ownership of business term integrity and its governance
- Ability to link terms in an easy-to-use web interface to create policies that govern information objects
- Metadata lineage that 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 for meeting regulatory compliance requirements such as Basel II and the Sarbanes-Oxley Act)
- Globalization and translation support for multiple languages, so global enterprises are not arbitrarily limited

The IBM big data integration platform

While the term “big data” has only recently come into vogue, IBM has been providing solutions that can handle enterprise-class 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.

IBM InfoSphere Information Server is a market-leading information integration platform that helps companies understand and govern data, create and maintain data quality, and transform and deliver data. InfoSphere Information Integration capabilities are delivered by InfoSphere Information Server, InfoSphere Data Replication and InfoSphere Federation Server, and include:

- **Business information exchange:** Helps organizations understand and govern their information, encouraging a standardized approach to discovering IT assets and defining a common business language so they are better able to align business and IT goals.
- **Data quality:** Enables organizations to analyze, cleanse, monitor and manage data, adding significant value by helping them make better business decisions and improve business process execution.
- **Data integration:** Transforms data in any style and delivers it to any system, ensuring faster time-to-value and reduced risk for IT. This package also includes the InfoSphere Data Click feature, which supports self-service data integration.
- **Data federation:** Flexibly delivers data and supports virtual data hub requirements.
- **Data replication:** Supports real-time data replication requirements, and enriches mobile applications and business analytics and big data projects by integrating replicated data.

InfoSphere Information Integration capabilities are essential to the success of any big data project. They also support the IBM big data platform, which includes tools for visualization and discovery, Hadoop-based analytics, stream computing, data warehousing and text analytics.

Tame the big data wolf

This paper shows the value of implementing information integration and governance from the very start of all types of big data projects—whether those projects focus on big data applications, data warehouses or data warehouse augmentation initiatives. The important thing is to deploy the right solution for the right set of requirements, lest you risk a bite from the big data bad wolf.

There are two key things to remember as you set off on your big data journey:

1. [illegible] Information integration requirements go far beyond the purpose Hadoop was meant to serve.

2. [illegible] Your solution needs to truly meet demands for scalability, efficiency, flexibility and breadth of integration. And don’t overlook data replication. When introduced into heterogeneous environments, you can integrate replicated data to enrich mobile applications, business analytics and big data projects to quickly and easily manage growth.

Why InfoSphere for information integration?

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

InfoSphere supports the success of big data and analytics projects by delivering the confidence to act on insight. InfoSphere capabilities include:

- [illegible] Define metadata, business terminology and governance policies with IBM InfoSphere Business Information Exchange.
- [illegible] Handle all integration requirements, including batch data transformation and movement (InfoSphere Information Server for Data Integration), real-time replication (InfoSphere Data Replication) and data federation (InfoSphere Federation Server).
- [illegible] Parse, standardize, validate and match enterprise data with IBM InfoSphere Information Server for Data Quality.

Notes

For more information

To learn more about data information and governance strategies and IBM big data capabilities, please contact your IBM representative or IBM Business Partner, or visit the following website: **ibm.com**/software/products/us/en/category/SWB50

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
October 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.

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.

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.

[1] Adrian, Merv and Friedman, Ted. "Hadoop Is Not a Data Integration Solution." Gartner. 29 January 2013.

Please Recycle