What IT Needs to Know About Data Ingestion and Egression for Hadoop

Informatica PowerExchange for Hadoop
This document contains Confidential, Proprietary and Trade Secret Information ("Confidential Information") of Informatica Corporation and may not be copied, distributed, duplicated, or otherwise reproduced in any manner without the prior written consent of Informatica.

While every attempt has been made to ensure that the information in this document is accurate and complete, some typographical errors or technical inaccuracies may exist. Informatica does not accept responsibility for any kind of loss resulting from the use of information contained in this document. The information contained in this document is subject to change without notice.

The incorporation of the product attributes discussed in these materials into any release or upgrade of any Informatica software product—as well as the timing of any such release or upgrade—is at the sole discretion of Informatica.

Protected by one or more of the following U.S. Patents: 6,032,158; 5,794,246; 6,014,670; 6,339,775; 6,044,374; 6,208,990; 6,208,990; 6,850,947; 6,895,471; or by the following pending U.S. Patents: 09/644,280; 10/966,046; 10/727,700.

This edition published October 2011
Table of Contents

Executive Summary ......................................................... 2

Big Data Processing with Hadoop ........................................... 3
  Processing Big Transaction Data and Big Interaction Data .................. 3
  The Hadoop Data Processing Pipeline .................................... 3

Integrating Hadoop with Enterprise Applications and
Information Management Systems ........................................ 4
  Deliver Timely, Trusted, and Relevant Data to the Business .............. 4
  Leverage Resource Skills and Maximize Business Value .................. 5
  Simplify Data Access and Increase Developer Productivity ............... 6
  Preview Big Data on Hadoop for Processing and Analysis ............... 7

Unleash the Benefits of Hadoop with
Informatica PowerExchange ............................................. 8
Executive Summary

More IT organizations are deploying Hadoop to complement their existing information management systems and using Hadoop to do more with Big Data that they could not do before. However, deploying Hadoop so that it seamlessly integrates with existing information management systems is proving a major challenge for IT organizations because they often lack the skills, software, and tools needed. To unleash the power of Hadoop within an IT organization requires the efficient flow of data among all the enterprise applications and information management systems. This efficiency in turn depends on scalable, reliable, and secure data integration technology to ensure that you can access, integrate, and deliver any and all of your data throughout the enterprise.

For this reason, a lot of companies are turning to Informatica, the data integration leader, to solve their Big Data problems and address the challenges of adopting Hadoop within their organizations. The stakes are high as companies seek to mine golden nuggets of information from large amounts of data in Hadoop. If you cannot quickly access data no matter where it resides, then your Hadoop implementation is at risk of becoming yet another information silo only available to a handful of experts. Informatica® technology provides a comprehensive, open, and unified data integration platform to break down these silos and enable you to access and integrate your data to deliver timely, trusted, and relevant information.

This white paper focuses on one of the first challenges IT organizations face when deploying Hadoop into their information management infrastructure: How do you get all types of data into and out of Hadoop? Informatica PowerCenter® with PowerExchange® enables IT to leverage skills and resources available today to easily, consistently, and reliably ingest and egress any and all data between Hadoop and enterprise applications and information management systems.
Big Data Processing with Hadoop

Processing Big Transaction Data and Big Interaction Data

Hadoop is used to cost-effectively scale to process petabytes of data from a variety of applications, data stores, and platforms. The purpose of processing all of this data is to improve business imperatives such as customer acquisition and retention, fraud detection, predictive analytics, and risk and portfolio analysis. These business imperatives often depend on a variety of datatypes and data processing algorithms running on Hadoop—from the fairly basic (e.g., keyword frequency) to the more sophisticated (e.g., pattern recognition, machine learning, multivariate and Bayesian statistics).

To produce the best results, these algorithms require massive amounts of data from various datatypes, structures, and formats from many different systems. These quantities include Big Transaction Data and Big Interaction Data. Big Transaction Data (e.g., OLTP) has been growing fairly steady over the last few decades while the emergence of Big Interaction Data (e.g., social media, Web logs, machine data, etc.) is growing exponentially. For example, an algorithm used to calculate customer churn indicators for improving customer acquisition and retention requires both transaction and interaction data. Transaction data in this case includes customer account information from a CRM system and customer order and product configuration information from an ERP system. Interaction data includes clickstreams from Web logs, customer support call records, social media data from Facebook and Twitter, and comments on blogs.

The Hadoop Data Processing Pipeline

The data processing pipeline depends on data moving efficiently between systems and high-performance parallelized data transformations and analysis. There are essentially four steps in a typical data processing pipeline using Hadoop (see Figure 1). The first step requires the ability to access a variety of data sources from any platform and any latency (e.g., batch, near real time, real time) and then to ingest the data into Hadoop distributing the data evenly across the nodes in the cluster. Most data today is stored in the Hadoop File System (HDFS) and the structure is declared at run time through data parsing. This step prepares the data for the downstream analysis on Hadoop using various simple and more complex algorithmic transformations. Finally, the resultant output data is only useful if the business can consume it through the business applications and tools it uses daily. So the last step in the data processing pipeline is to deliver output data from Hadoop to these enterprise applications and information management systems.
At each step in the data processing pipeline there is the potential for bottlenecks to occur. The first step—data ingestion—is often the most common bottleneck as data moves from the source systems, through Hadoop for processing, and then into other IT systems. The difficulty with data access is that it requires subject matter experts who understand the source application APIs and data structures along with the system communication and security protocols. At large and mid-sized companies with hybrid IT infrastructures, it is common to access data from multiple systems and applications for an IT project. For some legacy applications and platforms, these resource skills are becoming scarce. Furthermore, they tend to use different tools for each type of data, making it difficult to share data access skills across the project team.

**Integrating Hadoop with Enterprise Applications and Information Management Systems**

**Deliver Timely, Trusted, and Relevant Data to the Business**

The business finds itself waiting much too long for the information it needs. One of the primary reasons for this wait is that the most useful information depends on data that is distributed across application silos. If organizations are not careful in integrating Hadoop with their current information management infrastructure, then their Hadoop platform is at risk of becoming yet another information silo. Or worse yet, Hadoop becomes a “black hole” where petabytes of data go in and information is only available to a few trained experts proficient at getting data out of Hadoop.
Informatica technology ensures that the business has access to timely, trusted, and relevant information. PowerExchange for Hadoop delivers data from Hadoop to virtually any enterprise application, data warehouse appliance, or other information management system and platform whether on-premise or in the cloud. Business users and analysts in this way have immediate access to the information they need through the tools and applications they use every day, preventing Hadoop from becoming another information silo.

With Informatica tools, data is not only timely but also trusted and relevant because PowerCenter with PowerExchange for Hadoop supports postprocessing of data from Hadoop to cleanse, match, and transform the data to the structures and semantics of the target enterprise systems. For example, one customer uses Informatica PowerExchange to deliver processed data from Hadoop, such as customer churn indicators and marketing campaign best offer recommendations, to their data warehouse and MDM systems. Customer support and marketing now have an enhanced 360-degree view of customers through their BI applications and customer service portals. The benefit is increased customer retention and acquisition.

**Leverage Resource Skills and Maximize Business Value**

A common question among IT managers and enterprise architects is “How does Hadoop best complement my enterprise architecture?” IT organizations have invested millions in their information management systems and want to get the most out of what they have today before investing in other Big Data technologies such as Hadoop. Therefore, Hadoop needs to complement, leverage, and enhance an existing enterprise architecture to deliver more business value from the whole and not just the individual components.

To minimize risk and reduce overall costs, IT managers prefer to leverage current resource skills and avoid hiring an army of specialized developers to implement Hadoop projects. One of the first challenges IT managers face is getting all of the data they need into Hadoop. Not only are data volumes growing but so are the number and complexity of datatypes (e.g., social media, smart devices, NoSQL data stores and file systems, etc). IT managers should future-proof their ability to access all data residing in current systems and emerging technologies without having to always depend on subject matter experts and invest in new tools, adapters, and scripts.

Data needs to flow efficiently, securely, and reliably across all enterprise systems to unleash the power of Hadoop within your information management infrastructure. Informatica PowerExchange provides universal access to virtually all types of transaction and interaction data, enabling IT organizations to easily get data into and out of Hadoop at any latency. The product ensures access to all your data today and in the future as new software versions and datatypes emerge, making it easier to maintain data flows.

Informatica PowerExchange is integrated with PowerCenter’s metadata-driven development environment to increase productivity, provide a system of record for data loaded into Hadoop, and data lineage for auditing where the data comes from. Informatica PowerCenter supplies unified administration to schedule, monitor, and manage all jobs that load data into and extract data from Hadoop.
Simplify Data Access and Increase Developer Productivity

One of the first challenges Hadoop developers face is accessing all the data needed for processing and getting it into Hadoop. All too often developers resort to reinventing the wheel by building custom adapters and scripts that require expert knowledge of the source systems, applications, data structures, and formats. Once they overcome this hurdle, they need to make sure their custom code will perform and scale as data volumes grow. Along with the need for speed—security and reliability are often overlooked, increasing the risk of noncompliance and system downtime. In addition, building a robust custom adapter takes time and can be costly to maintain as software versions change. Sometimes the end result is adapters that lack direct connectivity between the source systems and Hadoop, which means you need to temporarily stage the data before it can move into Hadoop, increasing storage costs.

Informatica PowerExchange can access data from virtually any data source at any latency (e.g., batch, real time, or near real time) and deliver all your data directly into Hadoop (see Figure 2). Similarly, PowerExchange can deliver data from Hadoop to your enterprise applications and information management systems. You can schedule batch loads to move data from multiple source systems directly into Hadoop without any staging. Alternatively, you can move only changed data from relational and mainframe systems directly into Hadoop. For real-time data feeds, you can move data off of message queues and deliver it into Hadoop.

Informatica PowerExchange accesses data through native APIs to ensure optimal performance and is designed to minimize the impact to source systems through caching and process offloading. To further increase the performance of data flows between the source systems and Hadoop, PowerCenter supports data partitioning to distribute the processing across CPUs.

Informatica PowerExchange for Hadoop is integrated with PowerCenter so that you can preprocess data from multiple data sources before it lands in Hadoop. This ability enables you to leverage the source system metadata since this information is not retained in HDFS. For example, you can perform lookups, filters, or relational joins based on primary and foreign key relationships before data is delivered to HDFS. Existing Informatica PowerCenter customers can reuse existing mappings by simply configuring the target to be an HDFS Flat File Writer. You can also push down the preprocessing to the source system to limit data movement and unnecessary data duplication to Hadoop. Common design patterns for data flows into or out of Hadoop can be generated in PowerCenter using parameterized templates built in Microsoft Visio to dramatically increase productivity. To securely and reliably manage the file transfer and collection of very large data files from both inside and outside the firewall, you can use Informatica Managed File Transfer.
Data scientists are in demand these days and as such are highly paid professionals trained in advanced mathematics, statistics, and computer science. Their primary responsibility is to design algorithms that maximize the business value of their company’s data assets. Before processing data on Hadoop, the data scientist will visualize the data to get a feel for what the data looks like and its distribution and quality. To test and refine an algorithm often requires time spent getting data into Hadoop, parsing and preparing the data for analysis, and then once the algorithm has been run, viewing the resultant output. This process is repeated as the algorithms are refined and validated. Given their technical acumen, data scientists spend a good deal of time writing code and scripts just to access, prepare, and view the data both prior to and after processing on Hadoop. Their time is much better spent designing and building algorithms to run on Hadoop.

Informatica technology enables organizations to quickly and easily get all the data they need into Hadoop and prepare the data so that data scientists can spend more time designing and building algorithms to deliver the maximum business value. Data scientists, analysts, and developers can leverage Informatica PowerCenter’s self-service and collaboration capabilities to work efficiently together as they design data flows for the preprocessing and postprocessing steps required to perform data analysis on Hadoop. Informatica PowerExchange for Hadoop populates the Hive metadata tables so that data scientists and analysts can simply view the data by executing a SQL-like Hive query. This enables them to better understand the data and test assumptions as data processing algorithms are refined and validated.
Unleash the Benefits of Hadoop with Informatica PowerExchange

More companies are adopting Hadoop to complement their other information management systems as part of their Big Data strategy. A Big Data strategy needs to consider what types of data to store in Hadoop and which data processing workloads to deploy on Hadoop. To avoid Hadoop becoming yet another data silo, Informatica PowerCenter with PowerExchange enables IT organizations to access all their data, remove bottlenecks from the Hadoop data processing pipeline, and deliver the resultant data to enterprise applications and information management systems.

All business and IT stakeholders benefit with Informatica PowerExchange for Hadoop. The business has access to more timely, trusted, and relevant information with data processed and enriched on Hadoop. IT managers and architects reduce costs and maximize business value by standardizing on the Informatica data integration platform to access, integrate, and deliver data across the enterprise including Hadoop and other Big Data platforms, on premise, and in the cloud. You can leverage current resource skills to move data into and out of Hadoop without application and system subject matter experts to access data sources. Developers increase their productivity and simplify maintenance using Informatica PowerCenter’s metadata-driven development environment. Informatica PowerExchange delivers data from any system directly to HDFS and populates the Hive metadata tables so that data scientists and analysts can also benefit by querying data prior to analysis. The Informatica Platform ensures that you have access to all your data today and in the future so you can take advantage of new and emerging technologies such as Hadoop and stay competitive.

To learn more about Informatica’s solutions for Hadoop, including product demos, please visit http://www.informatica.com/hadoop.
About Informatica

Informatica Corporation (NASDAQ: INFA) is the world's number one independent provider of data integration software. Organizations around the world rely on Informatica to gain a competitive advantage with timely, relevant and trustworthy data for their top business imperatives. Worldwide, over 4,440 enterprises depend on Informatica for data integration, data quality and big data solutions to access, integrate and trust their information assets residing on-premise and in the Cloud. For more information, call +1 650-385-5000 (1-800-653-3871 in the U.S.), or visit www.informatica.com. Connect with Informatica at http://www.facebook.com/InformaticaCorporation, http://www.linkedin.com/company/informatica and http://twitter.com/InformaticaCorp.