Data warehouse augmentation, Part 1: Big data and data warehouse augmentation

Sandip Chowdhury May 27, 2014

Current data warehouse tools and technologies cannot handle the load that new data sources and analytic workloads bring to an enterprise data warehouse. To augment existing data warehouse solutions, organizations must implement big data technology within the context of the traditional data warehouse. This series, intended for data warehouse architects and information architects who work with traditional data warehouses, describes how to combine traditional and big data technologies to maximize and augment the effectiveness of existing data warehouses.

View more content in this series

This article describes the big data technologies, which are based on Hadoop, that can be implemented to augment existing data warehouses. Traditional data warehouses are built primarily on relational databases that analyze data from the perspective of business processes.

Part 1 of this series describes the current state of the data warehouse, its landscape, technology, and architecture. It identifies the technical and business drivers for moving to big data technologies and identifies use cases for augmenting existing data warehouses by incorporating big data technologies.

As organizations look for the business value that is hidden within non-structured data, they encounter the challenge of how to analyze complex data. Because business decisions are influenced by many factors, analysis models become increasingly complex to take into account many facets.

A traditional IT infrastructure is not able to capture, manage, and process big data within a reasonable time. It cannot accommodate data sets with volumes that range from a few dozen terabytes to many petabytes.

Traditional data warehouses

Traditionally, data warehouses analyze structured, transactional data that is contained within relational databases. These warehouses apply key performance indicators and model-driven architecture.

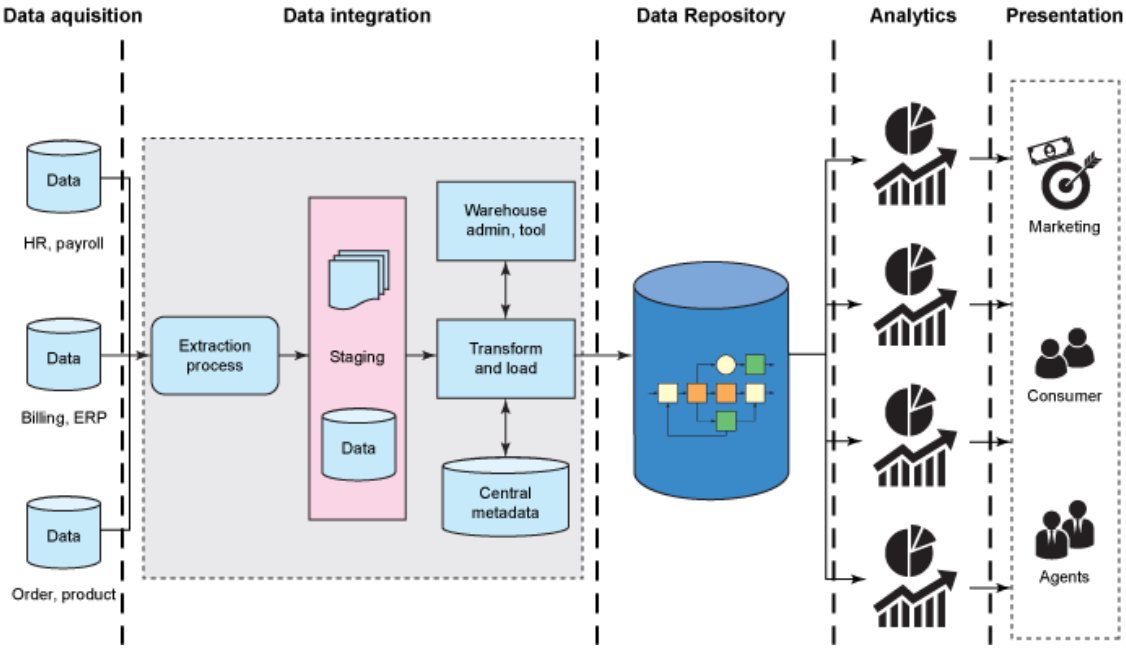
Data management landscape

Until recently, the data management landscape that is shown in Figure 1 was simple.

- Online transaction processing (OLTP) systems supported the enterprise's business processes.
- Operational data stores (ODSs) accumulated the business transactions to support operational reporting.
- Enterprise data warehouses (EDWs) accumulated and transformed business transactions to support both operational and strategic decision making.

Usually, enterprises analyze the structural data sources that are generated within the organization.

Figure 1. Traditional data warehouse reference architecture



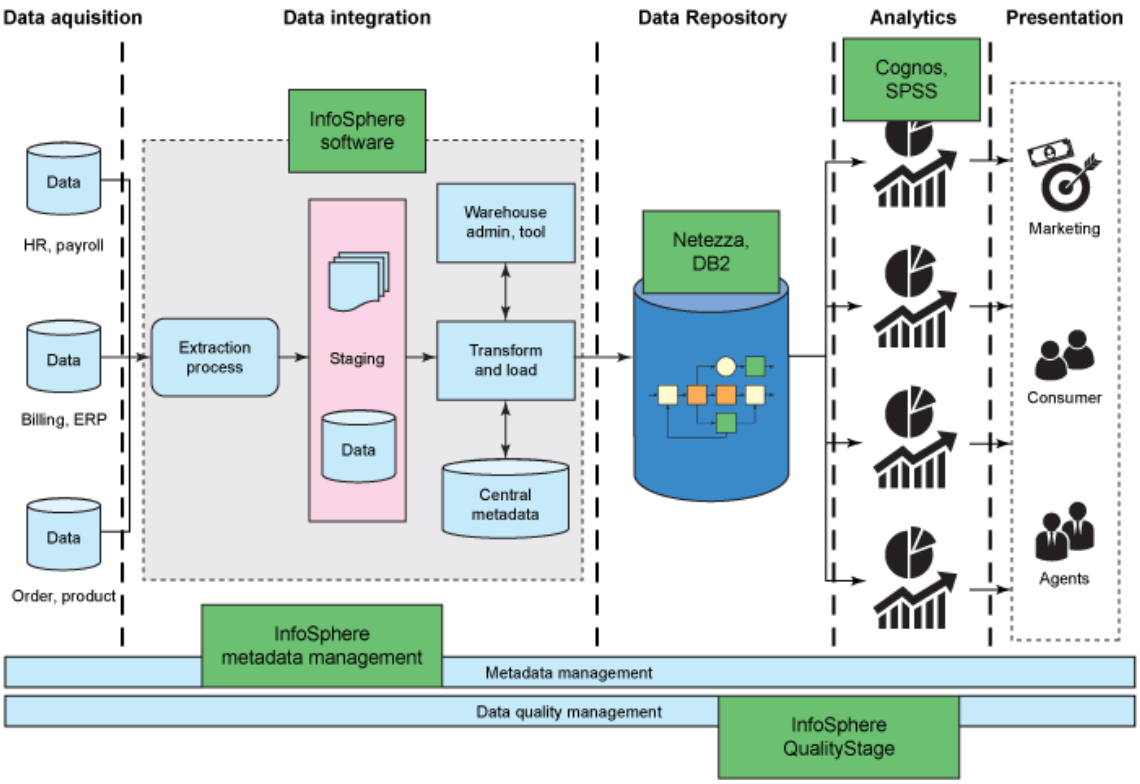
Each layer performs a particular function:

- **Data acquisition layer:** Consists of components to get data from all the source systems, such as human resources, finance, and billing.
- **Data integration layer:** Consists of integration components for the data flow from the sources to the data repository layer in the architecture.
- **Data repository layer:** Stores data in a relational model to improve query performance and extensibility.
- **Analytics layer:** Stores data in cube format to make it easier for users to perform what-if analysis.
- **Presentation layer:** Applications or portals that give access to different set of users. Applications and portals consume the data through web pages and portlets that are defined in the reporting tool or through web services.

The current BI reference architecture that is shown in Figure 2 is supported by many products:

- **IBM® InfoSphere® software:** A set of tools for information integration and management.
- **IBM InfoSphere Metadata Workbench:** The tools, processes, and environment that are provided so that organizations can reliably and easily share, locate, and retrieve information from these systems.
- **IBM InfoSphere QualityStage®:** Helps to create and maintain consistent views of key entities that include customers, vendors, locations, and products. Use it to investigate, cleanse, and manage your data.
- **IBM® PureData™ System for Analytics:** Simplifies and optimizes performance of complex analytics for analytic applications, enabling complex algorithms to run in minutes not hours.
- **IBM® DB2®:** Industry leading performance, scale, and reliability on your choice of platform from Linux, UNIX, and Windows to z/OS. Learn how customers are transforming their data center with DB2.
- **IBM SPSS software:** Predict with confidence what happens next so that you can make smarter decisions, solve problems, and improve outcomes.
- **IBM® Cognos® Business Intelligence:** Provides reports, analysis, dashboards, and scoreboards to help support the way people think and work when they are trying to understand business performance.

Figure 2. Products that are mapped to the reference architecture



Changes in data processing

Changes in demand for data analysis are driving the need to implement technology to handle new requirements. Examples of new demands include:

- An organization's reliance on data analysis to glean insight into customers, customer buying patterns, and supply chains

- An increasingly sensor-enabled and instrumented business environment, which generates huge volumes of unstructured data
- Data flowing through the system in high volumes
- Technical issues that are related to handling data complexity
- Resource-intensive computing demands

The move to big data technologies

Organizations built data warehouses to analyze business activity and to produce insights that enable decision makers to act on and improve business performance and operational effectiveness. Despite the maturity of the market, business intelligence (BI) technology remains at the forefront of IT investment. As more data is created, advances in analytical relational database technology improve BI software.

Businesses are driven to adopt big data technology for many reasons:

- Business demand to analyze new data sources
- Growth in data complexity:
  - Variety of data types
  - Volume of data
  - Velocity of data generation
  - Veracity of data from multiple sources
- Growth in analytical complexity
- Increasing availability of cost-effective computing and data storage

Business requirements drive demand for a big data platform

Decision makers in business organizations can ask themselves the following questions to gauge the need for big data technology:

- Are the current data sets large? Are you limited by your current platform or environment because you can't process the amount of data that you want to process?
- Is the existing warehouse environment a repository of *all* data that is generated or acquired?
- Do you have much cold or low-touch data that is not being used to analyze and derive business insight?
- Do you want to be able to analyze non-operational data?
- Do you want to use your data for traditional and new types of analytics?
- Are you unable to analyze new sources of data because the data does not fit neatly into schema-defined rows and columns without sacrificing fidelity or the rich aspects of the data?
- Do you need to ingest data as quickly as possible? Does your environment require generation of the schema during run time?
- Are you looking for ways to lower your overall cost for analytics?

The situations that are described by these questions can be improved by augmenting the existing data warehouse environment with big data technologies.

IBM big data platform and InfoSphere BigInsights™

For many organizations, Apache Hadoop offers a first step to begin implementing big data analysis. This open source software enables distributed processing of large data sets across clusters of commodity servers.

InfoSphere BigInsights Quick Start Edition

InfoSphere BigInsights Quick Start Edition is a complimentary, downloadable version of InfoSphere BigInsights, IBM's Hadoop-based offering. Using Quick Start Edition, you can try out the features that IBM built to extend the value of open source Hadoop, like Big SQL, text analytics, and BigSheets. Guided learning is available to make your experience as smooth as possible including step-by-step, self-paced tutorials and videos to help you start putting Hadoop to work for you. With no time or data limit, you can experiment on your own time with large amounts of data. Watch the videos, and download BigInsights Quick Start Edition now.

IBM InfoSphere BigInsights combines Apache Hadoop (including the MapReduce framework and the Hadoop Distributed File Systems) with unique, enterprise-ready technologies and capabilities from across IBM, including Big SQL, built-in analytics, visualization, BigSheets, and security. InfoSphere BigInsights is a single platform to manage all of the data. InfoSphere BigInsights offers many benefits:

- Provides flexible, enterprise-class support for processing large volumes of data by using streams and MapReduce
- Enables applications to work with thousands of nodes and petabytes of data in a highly parallel, cost effective manner
- Applies advanced analytics to information in its native form to enable ad hoc analysis
- Integrates with enterprise software

IBM PureData Appliance for Hadoop

To implement Hadoop, you need guidance on how to build, configure, administer, and manage production-quality Hadoop clusters at scale (more than 1,000 nodes, potentially). With IBM® PureData™ for Hadoop, an integrated platform for Hadoop implementation, get access to information and resources to help overcome implementation challenges. PureData for Hadoop offers:

- Built-in expertise
  - Deploys eight times faster than custom-built solutions
  - Built-in visualization, which helps accelerate insight
  - Default analytic application accelerators for social data, machine data, and text analytics
- Simplified experience
  - Single console for full system administration
  - Rapid system upgrades with automation
  - SQL framework, which offers simplified access to unstructured data
- Integrated by design
  - Bidirectional archiving and restore capability
  - Robust security tools

- High availability architecture
- Integration with the InfoSphere BigInsights platform
- Ability to ingest data up to 14 TB/hour

Big data use cases

To explore and implement a big data project, you can augment existing data warehouse environments by introducing one or more use cases at a time, as the business requires. This approach enables organizations to act with agility, reduce cost of ownership, and provide faster time to market with an increased business value and competitiveness.

Consider applying big data technologies in the following ways:

- Use case 1: As a landing zone for source data
- Use case 2: For historical data in the warehouse
- Use case 3: For exploratory analysis

Summary

In the past, inadequate tools and technologies to handle big data forced organizations to build analytics solutions that are based on structured data. Therefore, existing data processing engines and data storage solutions accommodate a low throughput for data, rather than the volume and variety of data that constitutes big data.

Faced with an expanding analytical ecosystem, BI architects need to make many technology choices. Perhaps the most difficult involves selecting a data processing system to power various analytical applications.

With new technologies such as Hadoop, organizations can cost-effectively consume and analyze large volumes of semi-structured data. Big data technology complements traditional, top-down, data-delivery methods with more flexible, bottom-up approaches that promote ad hoc exploration and rapid application development.

Part 2 of this series describes Use case 1: Using big data technologies to build an enterprise landing zone. It also explains how the enterprise can reuse raw data (structured and unstructured) to support ad hoc and real-time analytics.

Related topics

- "Big data architecture and patterns" series (developerWorks, September 2013): Classify big data problems, choose an architecture for a big data solution, and implement the solution with the IBM big data platform.
- *Big data technology and its impact on data warehousing*: Business intelligence expert Wayne Eckerson describes the growing adoption rates of big data technology and why companies are finally recognizing its many benefits.
- *Big data glossary: A guide to the new generation of data tools:* Pete Warden's guide describes 60 of the most recent innovations, from NoSQL databases and MapReduce approaches to machine learning and visualization tools.
- Download InfoSphere BigInsights Quick Start Edition: Available as a native software installation or as a VMware image.