The Modern Data Warehouse—How Big Data Impacts Analytics Architecture

Karen Lopez and Joseph D'Antoni

Karen Lopez is senior project manager and architect at InfoAdvisors. She specializes in practical application of data architecture and data evangelism.
karenlopez@infoadvisors.com

Joseph D'Antoni is a senior architect with over 10 years of experience. He is a solutions architect for SQL Server and big data at Anexinet.
joey.dantoni@joeydantoni.com

Abstract

The advent of big data technologies—and associated hype—can leave data warehouse professionals and business users doubtful but hopeful about leveraging new sources and types of data. This confusion can impact a project's ability to meet expectations. It can also polarize teams into "which one will we use" thinking.

Good architectures address the cost, benefits, and risks of every design decision. Good architectures draw upon existing skills and tools where they make sense and add new ones where needed. We architects always use the right tool for the job.

In this article, we describe the parts of the Hadoop framework that are most relevant to the data warehouse architect and developer. We sort through the reasons an organization should consider big data solutions such as Hadoop and why it's not a battle of which (classic data warehouse or big data) is best. Both can—and should—exist together in the modern data architecture.

Introduction

The concept of data warehousing has been with us for at least 30 years and has reached maturity within IT organizations and among data analysts. In the 1990s, online analytical processing (OLAP) systems allowed analysts to perform operations that might not have been possible in other solutions during that period. However, newer, disruptive technologies have been introduced that change overall system architecture and approaches to large-scale data analysis.

There has been a good deal of us-versus-them controversy in the relational and non-relational database world, mostly due to the mistaken belief that an organization must choose one over the other. As we have seen with many technologies over the decades, finding the right tool for the job is paramount to support business needs. Platform wars rarely benefit our organizations.

Big data technologies have moved beyond the "only for Web start-ups" or "only for scientific use" phase and are now ready to answer real-world business questions.

A Data Story

Many stories used to explain big data and Hadoop use social media and scientific sensor data—all wonderful examples of the divergence from traditional data. However, these examples sometimes leave traditional enterprise users feeling as if there are no applications in their world for these technologies.

Big data isn't just about using new tools; it's about solving problems that could be too expensive to solve in traditional architectures. Let's look at how a retailer with a mature data warehouse might make use of big data solutions.

A typical retailer might support the following types of data analytics in the data warehouse:

- Product sales
- Promotion effectiveness
- Store sales
- Shopping basket mixes and trends
- Customer preferences and purchasing histories
- External customer demographic data
- External daily weather data

In addition, a retailer might want to include analysis of the following:

- Customer traffic and shopping patterns within a store via mobile tracking, shopping cart tracking, or customer interactions with kiosks
- Customer shopping behavior via in-store video analytics and other sensor tracking
- Customer shopping patterns on a website, complete with browsing behavior, ad tracking, and other Web-based logging
- Municipal traffic data and road closure data to identify anomalies in sales patterns
- Consumer tax credits by income and postal code
- Hourly weather data by store
- Sentiment analysis from social media
- Influencer analysis from social media

The latter examples could be technically implemented in traditional data warehouse architectures, but the volume and performance load of all this data would likely require significant hardware upgrades and put performance pressure on existing loads, some to the point of being economically infeasible. This retailer would want to offload all that data into big data clusters that are optimized for processing large data volumes, then load the resulting smaller, post-processed, smarter data into their enterprise data warehouse and marts.

Big data opportunities for more insight abound for all kinds of organizations, not just technology or start-ups.

Hadoop and Its Ecosystem

Hadoop is the technology with the most disruptive potential in the big data space—it started simply as a project at Yahoo! to build a better search engine and process all that data, but has evolved into the centerpiece of a modern data analytics architecture, with a large group of open source components surrounding it.

When Hadoop was introduced, implementation and interaction were a challenge, especially to enterprise IT organizations. Management tools were extremely limited and an installation required managing versions of Java libraries, compiling software, and writing custom code to interact with data—which required a new paradigm for developers to learn and understand. Despite these early limitations, Hadoop's power quickly brought it to the fore for large-scale data processing. At its core, Hadoop is two things—a framework for data processing called MapReduce and a distributed file system known as the Hadoop Distributed File System (HDFS). These technologies combine to allow massive parallelism and fault tolerance while running on commodity hardware.

A common refrain in modern computing is that storage is cheap—this is far from the case with large enterprises utilizing storage area network (SAN) storage. According to the Gartner Group, the average cost for enterprise SAN storage was $4,876 per terabyte in 2011 (Gartner, 2011). Even allowing for some reduction in cost over time, storage is a major part of IT's ongoing operating expense. We can use an analytic architecture that is optimized to process larger data volumes to leverage costs and benefits of storage and processor budgets appropriately.

The Hadoop ecosystem performance approach is different from traditional systems tuning in the following ways:

- **Scale out instead of up.** In the relational data warehouse environment, performance is often improved by using larger and faster hardware (which tends to be exponentially more expensive as it grows in scale) or by purchasing an appliance from a software vendor. In the Hadoop world, we add more nodes (servers) and do the work in parallel.

- **Commodity hardware.** Hadoop is designed around dense, local storage and large sequential reads. It leverages horizontal scale to provide a great deal of aggregate memory (RAM) and I/O operations per second by combining all available resources in a given cluster of nodes.

- **Parallel processing.** Hadoop is architected to manage and support massively parallel processing (MPP), which is optimized for processing very large data sets.

Although MapReduce is a powerful and robust framework, writing Java code in mass scale would have required retraining data analysts and other IT personnel, who are used to working with structured query language (SQL) and scripting. This skills and tools mismatch meant that enterprises were unlikely to adopt Hadoop solutions. The open source community realized these limits and brought together several projects—Hive, Pig, and later Impala—to provide a more user-familiar interface to HDFS.

Hive

Apache Hive refers to itself as a "data warehouse which facilitates querying and manages large data sets residing in distributed storage." Hive functions as a SQL metastore on top of HDFS—users can impose schemas (which look like tables to the user) onto files and then query them using a language called Hive Query Language (HiveQL). This language is based on SQL, so developers and analysts can more easily query HDFS data. When a user runs a query in HiveQL, a MapReduce job is generated and launched to return the data. No Java coding is required.

Pig

Apache Pig also builds a high-level procedural language that acts as an interface to HDFS. Pig is more frequently utilized in extract, transform, and load (ETL) scenarios than for just returning data results. Pig uses a text-based language called Pig Latin, which focuses on ease of use and extensibility.

Impala

Apache Impala is part of a number of second-generation Hadoop solutions (along with Spark and Shark) that leverage memory-based processing to perform analytics. Impala has access to the same data in the HDFS cluster (and typically relies on the Hive metastore for table structures) but it doesn't translate the SQL queries it's processing into MapReduce. Instead, Impala uses a specialized distributed query engine similar to those found in commercial parallel relational database management systems (RDBMS).

YARN

Hadoop has evolved. In the past, the entire operations of the cluster were run using MapReduce. Now, YARN (Yet Another Resource Negotiator) allows for a more distributed, faster architecture. One of the implications of these changes is the need to build HDFS clusters with more memory than was common in the past. It used to be commonplace to use 64–96 GB of RAM in a given cluster data node; today, 256–512 GB nodes are becoming common.

These components make up only a small subset of the entire Hadoop framework, but they are the most relevant pieces for a data warehouse architect to understand. The Hadoop ecosystem is sometimes referred to as the "zoo" in keeping with its elephant-based name. Figure 1 shows these relevant components and how they fit together.

Analytics and Data Warehousing

Traditional data warehousing is focused on operational metrics such as inventory, supply chain, and operational goals. These metrics tend to look at historical and current data, and although they may allow for some forward-looking forecasting, they usually look at internal data only, with limited use of outside data sources. With years of evolution and ever more powerful hardware, data warehouses have become repositories allowing for large-scale reporting and analysis. Pundits have speculated that big data platforms could be the death of the traditional data warehouse; however, there are many regulatory, operational, and financial reporting requirements that will ensure that the data warehouse remains a component of the IT landscape in the future.

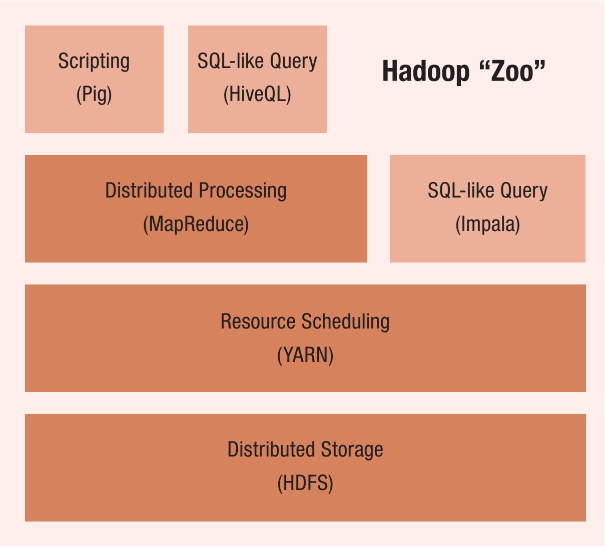

Figure 1: Hadoop data warehouse components.

Although data warehousing asks questions about past business events and does attempt to perform predictive analysis, the RDBMSs at the center of the warehouse were not specifically designed for analytical queries. Online analytical processing and multidimensional capacity added more power to the analysis. However, at larger scales the needs of these systems could only be met by expensive, converged solutions. This was driven by several trends; for example, data volumes increased dramatically and now terabytes are normal and petabytes are becoming more common.

Predictive Analytics

Predictive analytics is an area of data mining that specializes in extracting patterns from past data and applying statistical models to forecast future behavior. These types of analysis have become more widely available to organizations as computing power has become cheaper and their data volumes have increased. In the past, such analysis was limited to credit agencies, financial services, and insurance firms. Now, these types of analyses have become widely available and are used in a variety of industries as diverse as professional sports and medical decision-support software.

Other trends have changed an organization's data landscape. The proliferation of mobile devices, sensor data, and Web logs have led to new forms of data. Frequently called "unstructured," data, this data is most commonly presented in the form of Java Script Object Notation (JSON) or Extensible Markup Language (XML). These data types are not easily ingested by traditional platforms due to their variable structures within the same data set, but they are easily loaded into HDFS and several parsers are available to transform that data into a format that can be easily analyzed. Truly unstructured data is also being analyzed with pattern matching in video and audio files.

Analytics Marts

At the same time these trends have converged, analysts have begun taking advantage of larger volumes of data in order to perform "advanced" analytics. A frequent use case is to build a model for predictive analytics and run it against real-time or near-real-time data. These models will be built over and over again and run many times in an effort to perfect the models, so service times for these solutions must be very good. Hadoop HDFS has not been the platform for these real-time analytics; a more common scenario is to extract data from HDFS and load it into a memory-optimized columnar platform that allows for a high degree of data compression. Many columnar databases still support SQL and offer scale-out MPP on a similar hardware platform to HDFS.

External Data

Another trend in this area is the widespread use of external data sources. The most publicized use cases for this data involve social media data for sentiment analysis and even outage reporting, but external data use cases go far beyond that. Many firms have begun to incorporate weather data, purchased data about their competitors, and income tax and census data. Market research data tends to be very expensive, so the firms using it need to respect it like any key business asset. Many cities have begun open government initiatives in an effort to maintain transparency—this data can also be used for competitive advantage.

Bringing It All Together

The modern enterprise data warehouse (EDW) needs to bring together the technologies and data required to support traditional business needs and stronger predictive analytics, leveraging large data sets. The classic data warehouse architecture features transactional databases, some external data, ETL systems, and business intelligence systems, as shown in Figure 2.

The enterprise data warehouse would typically be implemented in a relational database management system, as would the OLTP and data mart data stores.

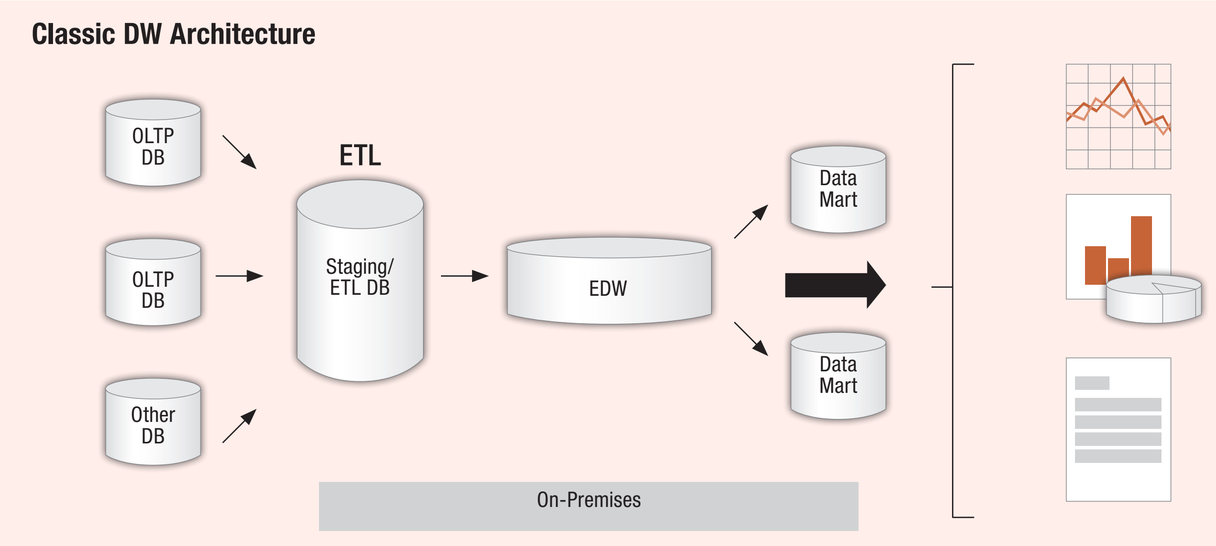

Figure 2: Classic data warehouse architecture.

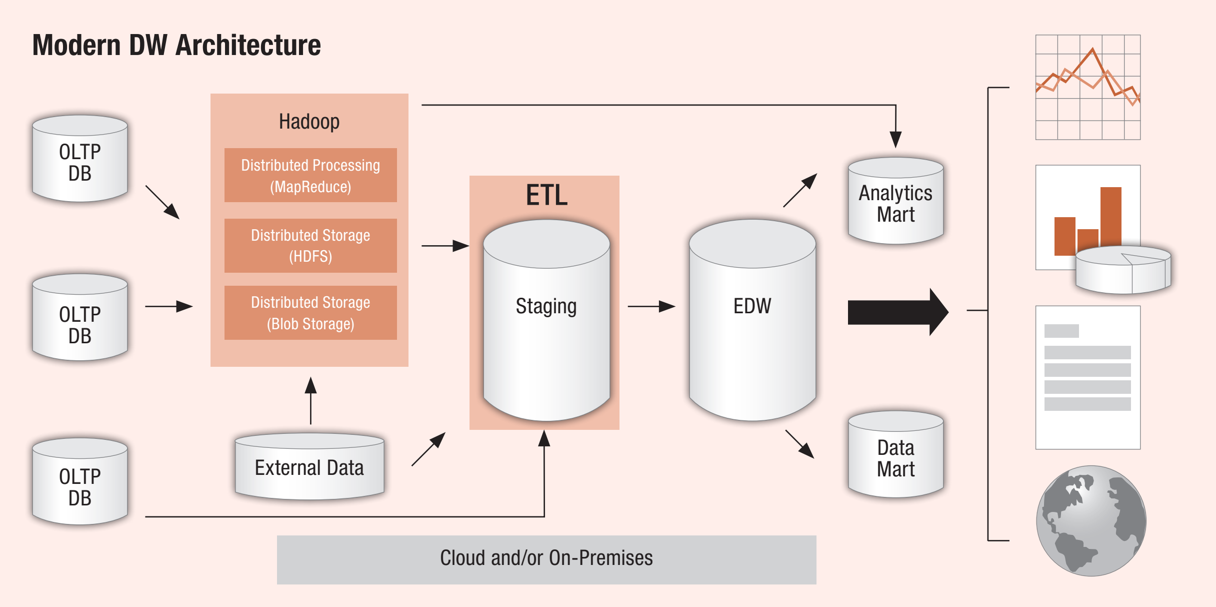

Figure 3: Modern data warehouse architecture.

The modern big-data-enabled warehouse adds to those component systems to support parallel processing, scale out, and analytic marts as shown in Figure 3.

Data Modeling in a Classic Data Warehouse

In a traditional data warehouse development project, dimensional models are prepared for the EDW and data marts, usually derived or inspired by OLTP and external data models and specifications. Data architects optimize these models for data loading and consumption. Data cleansing, denormalization, datatype transformations, and indexing strategies tend to be the focus of data modeling efforts.

Both of these architecture diagrams are highly symbolic. A tailored architecture might contain other components or leave some components out of the solution. Components might also be derived from other components, depending on the business needs and models being addressed. In fact, it is common to see Hadoop components used to process data throughout the architecture: using data from the EDW, analytics mart, or OLTP systems.

Data Modeling in the Modern Data Warehouse

In addition to the efforts described in the classic data warehouse project, data architects can provide value to the Hadoop tasks. Data models for OLTP systems will still be required where that data is used in Hadoop. Data models should be prepared for external data sources. Data architects can assist in the design of HiveQL "tables." Data models of the physical file store in Hadoop (HDFS) aren't required, but logical data models of the data that is managed there for any length of time would be.

Many modern data modeling tools have begun to support Hive schemas, at least for import. These tables can then be documented along with all the other enterprise data assets.

Challenges in Big Data Implementation

Changing hardware and software paradigms has never been easy or inexpensive for IT organizations as evidenced by the large number of firms still using mainframe platforms. In some aspects, new platforms make some IT problems easier—as noted, the hardware is distributed, which eliminates single points of failure. High availability is inherent in the system design; however, when talking about massive amounts of data, backups are always a challenge.

Given the highly available nature of HDFS and the challenges of backing up massive data volumes, many firms choose to forego performing backups of these data volumes, which could leave them vulnerable in a disaster. There are options from some Hadoop vendors for disaster recovery if your organization needs it for its analytic platform.

From a skills perspective, your organization needs the following key abilities:

- Linux system administrators
- Automation engineers
- Java
- Data analysis

Compared to a traditional model, where the database administrator (DBA) manages the data warehouse database, the DBA role does not apply in HDFS. Linux system administration skills are very important, and although there are distributions of Hadoop running on the Windows operating system that are popular with enterprise organizations, the overwhelming majority of implementations are running on Linux platforms, where community support is more available. When dealing with tens or thousands of cluster nodes, automation becomes very important. Software and firmware updates are also candidates for automated processing.

Leveraging Cloud Computing for Big Data

Big data makes for an interesting cloud computing solution—particularly if workloads are highly variable. Like most other cloud computing offerings, there are two types of solutions: platform-as-a-service (PaaS)—basically Hadoop-as-a-service—and infrastructure-as-a-service (IaaS).

Most major cloud vendors have Hadoop-as-a-service offerings—these can be a fantastic way to get up and running with Hadoop and the toolkit within an afternoon. This means that the vendor manages all the underlying infrastructure and you manage the configuration of Hadoop.

The IaaS offerings simply involve spinning up a number of virtual machines (VMs) and building a Hadoop cluster on them. This places more of the onus of configuration onto your staff but provides more flexibility with the tools installed alongside Hadoop.

One major challenge to both of these solutions is getting large, existing data volumes into the cloud. As a result, many vendors provide services allowing you to ship data tapes or hard drives to get them loaded onto their storage. The good news here is that most cloud providers do not charge a fee to upload data.

Like most other cloud computing solutions, the benefits involve flexibility and low initial capital investment. For example, if a firm wants to run a large-scale fraud detection solution that monitors personal behavior and browsing history across thousands of nodes, the cloud is a viable option if the workload is over a short period of time. Even for much smaller configurations, getting up to speed quickly without the hassle of installing software can be incentive enough to use a cloud solution.

Cloud Trade-Offs

The trade-offs with cloud solutions are the ongoing expense, slower performance, and security concerns.

The cloud limits an enterprise's initial capital investment, but for long-term, larger implementations the costs may creep up. Most cloud vendors also charge for outbound data flows, so if your reporting solution is on-premises, that is another expense to be considered.

Performance in a cloud will always be limited by the multitenant nature of the environment. Shared infrastructure is required to offer the cost savings and scale of cloud computing. To meet its financial goals, the provider needs to maximize its hardware usage while meeting its performance service-level agreements (SLAs). This is not to say cloud performance is bad—it simply will not match levels achieved with dedicated hardware in an on-premises installation.

Firms are concerned about security when moving to a cloud computing model; however, cloud providers are going out of their way to address these concerns. Consult your cloud provider for specifics about any privacy or regulatory concerns that apply to your industry; providers update these certifications regularly.

Economics of Big Data Solutions

One of the key drivers of big data in enterprise IT organizations has been the high cost of RDBMS licensing and the infrastructure to support it. Data warehouses tend to require features that only the more expensive "enterprise" editions of RDBMS offer and in some cases require the purchase of additional options. Most major RDBMS packages are licensed by the CPU core, which means as workload increases, so does the license expense. The nature of the RDBMS also limits horizontal scaling, so to address performance concerns larger, more expensive server hardware or faster storage is required. Another expense (though much smaller) is the cost of operating system licensing required to support the RDBMS.

Big data platforms are not totally free, but there are some clear cost advantages. Because performance is achieved through horizontal scaling and aggregate resources, individual nodes do not need to be as powerful as a monolithic server. As addressed earlier, Hadoop (and most other big data and NoSQL platforms) leverage dense, local storage that comes at a much lower cost than enterprise SAN storage. All of these software platforms run nearly exclusively on Linux and most implementations take place on completely free distributions of the operating system.

Hadoop itself is available as a free open source project, but most organizations will choose to go with a commercial distribution for ease of management. The annual cost of support and licensing for the commercial solutions are about $4,000/node/year (Bantleman, 2012), which is not insignificant but is far lower than the cost of a commercial RDBMS. Although RDBMS pricing varies per vendor and individual agreement, costs can be as high as $50,000 per CPU core.

In most scenarios, it makes the most sense to use big data technologies to process and aggregate big data into classic data solutions using the right tool for the job.

A Final Thought

The most important thing a data warehouse professional needs to understand is that Hadoop and other big data technologies are not an either/or decision. Every design decision comes down to cost, benefit, and risk. Those factors change over time, as we have seen since the first release of Hadoop. Right now, we have the opportunity to leverage these special-use technologies within an existing data warehouse architecture to leverage a greater variety of data sources than ever before. ■

References

Bantleman, John [2012]. "The Big Cost of Big Data," *Forbes*, http://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/ (accessed on May 14, 2014).

Gartner [2011]. "IT Key Metrics Data 2012: Key Infrastructure Measures: Storage Analysis: Current Year," Jamie K. Guevara, Linda Hall, Eric Steggman.