Global Knowledge®

Expert Reference Series of White Papers

Big Data and Apache Hadoop Adoption:

Key Challenges and Rewards

Big Data and Apache Hadoop Adoption: Key Challenges and Rewards

Kerry Doyle, MA, MSr, CPL

Introduction

Every data center has unique hardware and software requirements that can pose significant management challenges to IT. Cloud utilization, virtualization, mobile device management, app development, security, and storage networking represent just a few of the management challenges facing IT today. Add to that the responsibility of providing the infrastructure for processing an exponential increase in data and making it available to an increasingly mobile workforce on a daily basis.

This effectively means that big data and today's data centers are on a collision course. The processing power offered by the open-source Apache Hadoop platform offers a solution for analyzing and handling this increase in data loads. However, the learning curve for gaining the skill and knowledge to successfully deploy and implement Apache Hadoop is steep. Fortunately, a number of Apache Hadoop distributions ease deployment with integrated management features within Hadoop. In this white paper, we explore the demands of big data, provide an overview of recent vendor offerings, and explain how they streamline the process for using Apache Hadoop in handling today's data loads.

Challenges for Tackling Big Data

The era of big data has arrived and an increasing number of companies are looking for the right tools to analyze structured and unstructured data. Organizations need capabilities that offer the following features: scalability, cost effectiveness, flexibility, and fault tolerance. According to recent research by the global market research firm IDC, the volume of digital content will grow to 2.7 zettabytes (ZB), up 48 percent from 2011.

In the past, traditional business intelligence (BI) relied on historical analysis of primarily transactional data to glean information about customers. Today's big data analytics require more skill in iterative analysis, including the ability to recognize subtle patterns and shifts in customer behaviors, not to mention the massive computing power required to undertake such analysis.

In addition to the ever-growing demand for storage and compute capacity, companies are faced with making sense of the enormous amount of unstructured data they accrue to reveal these customer trends. While structured data (transactions, inventory, etc.) fits well into relational tables and arrays (i.e., the classic database) newer unstructured data does not. A partial list of this type of data includes:

- Computer logs
- Spatial data/GPS outputs

- Industrial sensor data
- Images, videos, MP3s
- Social Media updates
- Weblogs, forum comments, etc.

It's easy to see how this data can quickly grow. In the meantime, the big data industry is still building the infrastructure needed to integrate structured and unstructured data. Another capability required to handle these massive data loads is scale-out compute and storage capacity.

Apache Hadoop meets those large-scale requirements, because new computing nodes can be added as necessary. IT can use commodity servers to function as these nodes and to conduct massive parallel computing to easily handle any increase.

As the scale-out requirements increase by a factor of ten, companies must have the infrastructure and analysts in place to meet this need. According to recent Gartner research, mobile payments are predicted to reach $171B in 2012. This increase in customer-side transactions represents another area related to data analytics that organizations would like to exploit.

Once companies acquire this data, they have a couple of goals. First, they look to analyze that data and find key trends. A second use case is for online profiling and the continuous monitoring of "sensors," that is data that provides insights into customer behavior and business performance.

Although a well-optimized Hadoop cluster offers fast throughput for this type of large dataset analysis, it lacks fast real-time transaction processing. While this represents a perceived limitation to Hadoop, effective workarounds are possible.

A need also exists to make data available to separate divisions across an enterprise. These silos of information can be used in different ways by various departments, but often the lack of inter-communication limits the ability to achieve more enterprise-wide visibility for that data. In a more perfect data-sharing world, that content would be easily distributed between users and departments.

Moreover, large volumes of big data are simply eliminated and not stored because companies simply lack the tools to analyze and share that data. In some instances, this data is archived to tape, Blu-ray, or disk and later extracted and analyzed at considerable cost in time, effort, and resources.

However, it's no longer adequate to simply archive accrued data, regardless of its format. In addition to complications related to extracting that data later, organizations miss immediate information, insight, and prospects that otherwise could be available.

What Hadoop Has To Offer Big Data

Today, the standard relational database management systems (RDBMS) that have proved so effective for analyzing structured data are ineffective for dealing with the volume, variety, and velocity of big data.

Open-source models such as Apache Hadoop offer capabilities perfectly aligned with the types of file systems that store vast amounts of unstructured data, including event, social, web, spatial, and sensor data. As a consequence, Hadoop can apply that in-depth analytic capability to unstructured data for which RDBMS is unsuitable.

When it comes to Hadoop, Hadoop Distributed File System (HDFS) is an integral feature. Based on Java, this file system enables large amounts of structured and unstructured data to be stored and quickly accessed across large server clusters. In contrast to RDBMS, Hadoop doesn't require the complicated transformation and schema changes that traditional databases require. In fact, Hadoop has the capability to store data in its raw form and has minimal, if any, data model restrictions.

Since Hadoop runs on commodity hardware and storage, it's less expensive to employ than a conventional RDBMS. In a Hadoop cluster, each computer offers its own local computation and storage. Moreover, easy scalability makes Hadoop ideal for analytical workloads, unlike the real-time transaction processing of a relational database.

The Hadoop framework utilizes a process where data is written once then read many times in large volumes as opposed to in single records. Hadoop's ability to rapidly process vast amounts of data in parallel and its capacity to scale to an enormous number of nodes offer a built-in redundancy that offsets individual nodal failure.

It's also useful to point out that the framework is designed to detect and handle failures at the application layer, thus providing a built-in high availability (HA) process. However, for enterprises, which are the key focus of Hadoop adoption, this doesn't quite offer the level of granularity for failover control, fault tolerance, and security to which enterprises have become accustomed.

These limitations are unlikely to pose obstacles because it's readily apparent that enterprises are taking big data analytics seriously. These constraints represent areas that are addressed with the management controls and features offered by commercial providers.

Difficulties of Hadoop Adoption

The scale-out potential of Apache Hadoop is impressive. However, while Hadoop offers the advantage of using low-cost commodity servers, extending this scale-out potential to thousands of nodes can translate to a true expense.

As the demand for compute and analytic capacity grows, so can the machine costs. This has an equal effect on storage since Hadoop spreads out data, and companies must have equal space for increased data storage repositories, including all the indices, and for all the acquired raw data.

Integrating and processing all of this diverse data can be costly in terms of both infrastructure and personnel. While traditional BI relies on evaluating transactional and historical data, today's analytics require more skill in iterative analysis and the ability to recognize patterns.

When dealing with big data, an advanced skillset that goes beyond RDBMS capabilities—both in terms of analysis and programming—is essential. Not only is there need for advanced systems administration and analyst capabilities when working with Hadoop, but learning the MapReduce programming unique to this framework represents a significant hurdle.

In terms of relational databases, moving and modifying large volumes of unstructured data into the necessary form for Extraction, Transformation, Loading (ETL) can be both costly and time-consuming. That's a key reason why Hadoop seems so attractive.

One could argue that the ongoing growth in data volume, velocity, and variety has made the traditional data warehousing architecture less and less viable. However, it is still easier to find experienced RDBMS programmers and developers than those with MapReduce programming capabilities. Part of the difficulty lies in just learning the language beyond having the skills to install and maintain the Hadoop platform.

MapReduce uses a computational approach that employs a Map pre-processing function and a Reduce data aggregation/distillation step. However, when it comes to real-time transactional data analysis, the low latency reads and writes characteristic of RDBMS structured data processing are simply not possible with HDFS and MapReduce.

Of course, as the platform matures, more features will continue to be added to it. While add-on products make Hadoop easier to use, they also present a learning challenge that requires constantly expanding one's expertise. For example:

- **Hive** is the data warehousing component of Hadoop, and it functions well with structured data, enabling ad-hoc queries against large transactional datasets. On the other hand, though workarounds do exist, the absence of any ETL-style tool makes HiveQL, the SQL-like programming dialect, problematic when working with unprocessed, unstructured data.
- **HBase**, the column-based storage system, enables users to employ Hadoop datasets as though they're indices in any conventional RDBMS. It typically allows easy column creation and lets the user store virtually any structure within a data element.
- **PIG** represents the high-level dataflow language, Pig Latin, and requires quite advanced training. It provides easier access to data held in Hadoop clusters and offers a means for analyzing large datasets. In part, PIG enables the implementation of simple or complex workflows and the designation of multiple data inputs where data can then be processed by multiple operators.

As IT organizations consider wholesale adoption of the Hadoop platform for analytics, they must carefully strategize their approach. The platform's specialized methodology, scale-out storage, and powerful processing capacity make it optimal for analytical data loads.

However, the dedication in training competent personnel and machine costs, as well as the framework's inability to function as an RDBMS replacement, should prompt careful consideration.

The Commercial Platform Approach to Apache Hadoop

As mentioned above, businesses dealing with increasing masses of data are looking for a distributed computing analytic solution that provides comprehensive administration and management, easy deployment, and support for effective business continuity and high availability.

Today, commercial open-source models that incorporate MapReduce along with a built-in framework and infrastructure offer another means for avoiding the learning curve and burdens associated with Apache Hadoop deployment.

These commercial players ease skillset acquisition by providing key management tools that interface with Hadoop processes. The value of technical support, services, and training cannot be overstated when it comes to Hadoop implementation.

Commercial vendors offer a means by which these high-level analysis tools can be accessed and used by a wide variety of users, not just those with engineering or BI capabilities. They provide the support that ensures Hadoop users can undertake complex data analysis projects.

As open-source tools proliferate and their increasing importance to big data analytics continues to grow, a need for streamlined administration and support will expand as well. While commercial Hadoop providers offer the necessary support, there is no alternative to learning its platform-specific language.

Adequate knowledge of MapReduce represents an intrinsic component to working with Hadoop. Moreover, in order for users to install, configure, and use the code, thorough training is fundamental.

Hadoop integration with current BI analytics remains a key goal along with the development of analytic tools that can be employed by a wide range of users. Commercial vendors, such as Cloudera, Hortonworks, and MapR, may eventually provide the necessary connectivity between common BI analysis methodology and NoSQL.

Since Apache Hadoop, as a stand-alone, open-source deployment, doesn't contain internal manageability controls or high-level performance monitors, Cloudera offers a number of management tools that make analysis easier to implement for a range of users.

Cloudera's proprietary Hadoop software, Cloudera's Distribution Including Apache Hadoop (CDH), is in its fourth generation (CDH4). It includes a hot failover for the metadata function, NameNode. This is a critical contribution since NameNode is considered a single point of failure, essentially an Achilles heel for Hadoop. The latest version of the software offers a comprehensive package: HA, improved security, and Cloudera Manager for end-to-end Hadoop administration as well as long-term support.

Since part of the promise of big data requires getting past the hype and understanding appropriate applications of Hadoop, Hortonworks has created the Hortonwork Data Platform, version 1.0, which combines HA and failover requirements using VMware virtualization tools. The software's approach relies on running NameNode and Hadoop's JobTracker on virtual machines (VMs). This aspect helps to double up Hadoop's fault tolerance through the automation of VM replacement for failed servers. The software also includes a GUI for dataset integration and for composing workflows as well as HCatalog that enables connectivity with RDBMS products.

MapR has chosen to solve the data volume issue via its replacement of Hadoop's HDFS with a derivative of the UNIX-based file system, NFS. This helps to do away with the NameNode function altogether as a single point of failure. By swapping out HDFS, the company's proprietary components claim to offer improved HA as well as higher scalability and performance.

Commercial Hadoop providers play a critical role in enabling wider platform adoption, and their support services allow the technology to be accessed by those organizations that might otherwise have difficulties around implementation. While these companies represent key players in the ongoing commercialization of Hadoop, they also offer an important function through their training and certification courses—a value that cannot be understated.

Conclusion

Today, IT organizations and independent users must carefully strategize their approach to dealing with big data to avoid being overrun with data that has no intrinsic value due to the lack of adequate processing tools. Even more importantly, these users need to acknowledge that the right analytic tools, such as Apache Hadoop, present a serious challenge to adoption due simply to the rigorous learning curve.

To truly realize the promise of Apache Hadoop and its distributed set of resources for big data analysis, businesses and end-users need to expand their approach by relying on the wealth of resources currently available: access to professional training, commercial platform implementation, and utilizing third-party service providers, such as Cloudera.

It's becoming clear that the open-source Apache Hadoop platform changes the economics and dynamics of large-scale data analytics due to its scalability, cost effectiveness, flexibility, and built-in fault tolerance. It makes possible the massive parallel computing that today's data analysis requires.

However, the proper skillset training will be necessary to achieve large-scale data analysis. That's why commercial providers of open-source tools for Apache Hadoop offer such great value to companies.

These integrated management features enable the platform to be implemented by a wide range of users at all levels of skill expertise. Organizations can then make appropriate business decisions based on the large amounts of data they accrue by accessing the power of a relatively low-cost, highly scalable infrastructure such as Hadoop to tackle the challenges of big data.

Learn More

To learn more about how you can improve productivity, enhance efficiency, and sharpen your competitive edge, Global Knowledge suggests the following courses:

- Cloudera Administrator Training for Apache Hadoop
- Cloudera Developer Training for Apache Hadoop
- Data Science and Big Data Analytics

Visit **www.globalknowledge.com** or call **1-800-COURSES (1-800-268-7737)** to speak with a Global Knowledge training advisor.

About the Author

Kerry Doyle (MA, MSr, CPL) writes for a diverse group of companies based in technology, business and higher education. As an educator, former editor at PCComputing, reporter for PCWeek Magazine and editor at ZDNet/CNet.com, he specializes in computing trends vital to IT professionals, from virtualization and open source to disaster recovery and network storage.