# : What technologies are used for managing big data?

by Daniel J. Power

Editor, DSSResources.COM

Data-driven DSS and predictive analytics frequently need to access and process very large data sets to support decision-making by managers and customers. One way to provide this capability is with the open source Hadoop ecosystem initially developed by Yahoo!. Google innovated in developing these post relational database management systems and it continues to support its proprietary software ecosystem derived from Big Table. For example, Google BigQuery is an enterprise data warehouse with SQL queries. Other technologies for managing big data include cloud services, virtualization and distributed parallel processing (cf., Ashutosh, 2012; Mitchell, 2014). The open source Hadoop ecosystem is the most widely used technology and the focus of this discussion.

Apache Hadoop is a Java based framework for processing, storing and querying large amounts of data distributed on clusters of commodity hardware. Hadoop is a top level Apache project initiated by Yahoo! in 2006. The Hadoop project (http://hadoop.apache.org/) develops open-source software for reliable, scalable, distributed computing.

According to the project webpage, the Apache Hadoop software library is "a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage."

Technically, Hadoop consists of two key services: 1) reliable data storage using the Hadoop Distributed File System (HDFS) and 2) high-performance parallel data processing using a technique called MapReduce (http://www.cloudera.com/what-is-hadoop/).

Hadoop is a family of open-source products and technologies under the Apache Software Foundation (ASF). The Apache Hadoop library includes: the Hadoop Distributed File System (HDFS), MapReduce, Hive, Hbase, Pig, Zookeeper, Flume, Sqoop, Oozie, Hue, and other applications. You can combine these applications in various ways, but HDFS and MapReduce (perhaps with Hbase and Hive) are a useful technology stack for applications in business intelligence, data warehousing, and analytics.

The Hadoop file system (HDSF) is a distributed file system. It hides the complexity of distributed storage and redundancy from the programmer (cf., Vogel, 2010). From a more technical perspective Hive provides data summarization and ad hoc querying. Pig is a high-level data-flow language for parallel computing. Mahout is a machine learning and data mining library. Hadoop has many subprojects.

The colorfully named applications Hive and Pig are the two key components of the Hadoop ecosystem. Hive is a Data Warehousing package that interacts with the Hadoop file system to help analyze vast amounts of data. Hive is mainly developed for users who are comfortable using an SQL like language called HiveQL. Pig provides query like capabilities and better control of data flow processes like extract, transform and load. Pig was developed at Yahoo! while Hive was developed at Facebook. DeZyre.com (2016) has an excellent article comparing Hive and Pig.

According to the Yahoo! Hadoop tutorial, "Performing large-scale computation is difficult. To work with this volume of data requires distributing parts of the problem to multiple machines to handle in parallel. Whenever multiple machines are used in cooperation with one another, the probability of failures rises. ... What makes Hadoop unique is its simplified programming model which allows the user to quickly write and test distributed systems, and its efficient, automatic distribution of data and work across machines and in turn utilizing the underlying parallelism of the CPU cores. In a Hadoop cluster, data is distributed to all the nodes of the cluster as it is being loaded in. The Hadoop Distributed File System (HDFS) will split large data files into chunks which are managed by different nodes in the cluster. In addition to this each chunk is replicated across several machines, so that a single machine failure does not result in any data being unavailable."

In an interview with Doug Cutting, creator of the Hadoop framework, Jaikumar Vijayan for Computerworld (11/7/2011) asked him how he would describe Hadoop to a CIO or a CFO.Cutting explained "At a really simple level it lets you affordably save and process vastly more data than you could before. With more data and the ability to process it, companies can see more, they can learn more, they can do more. [With Hadoop] you can start to do all sorts of analyses that just weren't practical before. You can start to look at patterns over years, over seasons, across demographics. You have enough data to fill in patterns and make predictions and decide, 'How should we price things?' and 'What should we be selling now?' and 'How should we advertise?' It is not only about having data for longer periods but also richer data about any given period, as well."

Hadoop uses a computing strategy of "moving computation to the data to achieve high data locality which in turn results in high performance".

In an InformationWeek cover story on Hadoop, Doug Henschen (2011) concludes "Once Hadoop is proven and mission critical, as it is at AOL, its use will be as routine and accepted as SQL and relational databases are today. It's the right tool for the job when scalability, flexibility, and affordability really matter. That's what all the Hadoopla is about (p. 26)."

Hadoop doesn't meet all the needs associated with managing and using big data.  Babcock (2012) notes "Yahoo isn't using Hadoop as a stand-alone system. Rather, it serves as an information foundation for an Oracle database system, which pulls presorted, indexed data out and feeds it into a Microsoft SQL Server cube for highly detailed analysis. The resulting data is displayed in either Tableau or Microstrategy visualization systems to Yahoo business analysts. They in turn use it to advise advertisers how their campaigns are faring soon after launch."

In 2012, Yahoo! was storing and managing 140 petabytes of data in Hadoop. Hadoop keeps all data sets in triplicate so more than 400 petabytes of storage were needed. The volume of big data is very large and growing.  The social media and search companies have lead the way in managing big data.

**References**

Ashutosh, A., "Best Practices For Managing Big Data," Forbes, July 5, 2012 at URL http://www.forbes.com/sites/ciocentral/2012/07/05/best-practices-for-managing-big-data/#5ef13bafef02.

Babcock, C., "Yahoo And Hadoop: In It For The Long Term," InformationWeek, June 15, 2012 at http://www.informationweek.com/database/yahoo-and-hadoop-in-it-for-the-long-term/d/d-id/1104866

DeZyre, "Difference between Pig and Hive-The Two Key Components of Hadoop Ecosystem," October 15,  2014 (latest update made on October 31, 2016) at URL https://www.dezyre.com/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79.

Harris, D., "The history of Hadoop: From 4 nodes to the future of data," Gigaom,
March 4, 2013 at URL
https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/

Henschen, D., "Why all the Hadoopla?" InformationWeek, November 14, 2011, issue 1,316, pp. 19-26.

Russom, P., "Busting 10 Myths about Hadoop," http://tdwi.org, March 20, 2012, at URL http://tdwi.org/articles/2012/03/20/Busting-10-Hadoop-Myths.aspx

Mitchell, R.L., "8 big trends in big data analytics," Computerworld, October 23, 2014 at URL http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html.

Power, D., "What is Hadoop?" DSS News, Vol. 12, No. 23, November 13, 2011 at URL http://dssresources.com/newsletters/306.php .

Vijayan, J., "Q&A: Hadoop creator expects surge in interest to continue: interview with Doug Cutting," http://www.computerworld.com, November 7, 2011.

Vogel, L., "Apache Hadoop-Tutorial," v. 3, 03.04.2010 at URL http://www.vogella.de/articles/ApacheHadoop/article.html .

Yahoo! Hadoop Tutorial at http://developer.yahoo.com/hadoop/tutorial/index.html

Hadoop at Yahoo! at https://developer.yahoo.com/hadoop/

Author: Daniel Power
Last update: 2016-11-10 05:36