Date: **June 2011**
Prepared by: **Metia**

waterstechnology

The Key to 'Big Data' Mastery

Author: Rob Daly
June 2011

Rob Daly charts the remarkable growth of data volumes in the financial services industry, and looks at the technologies that allow firms to process, analyze and store data quantities that until recently were inconceivable. The amount of data available for financial services—growing at a mind-numbing rate—is both a blessing and a curse in the terms of analyzing and storing it.

During the 2010 Hadoop World Conference, Abhishek Mehta, then a managing director at Bank of America, and now founder of Tresata, cited a Cisco Systems' estimate that by 2013 700 zetabytes (ZBs) of data would be flowing across the Internet. A zetabyte represents 1,000 exabytes (EBs), or 1 million petabytes (PBs). Mehta attributes part of the growth in data volumes to the price of storage, which has become almost free. "A terabyte of storage in 1992 cost $1 million and your first-born child," says Mehta. "Today, you can get the same terabyte for $100."

It's not just that data is overwhelming organizations; firms have a desire to collect and process more data than they have done before," explains Matt Aslett, senior analyst for enterprise software at research firm 451 Group. "Previously, the core database market thought that relational databases could handle everything. People are beginning to think differently and deploy different types of database and storage models depending on the nature of their applications."

Various technologists have dubbed these immense data sets "big data." Aslett says there is a general agreement about what big data is, but there is not a specific definition for it. "The term is applied to data sets that are large, complex and dynamic and for which there is a requirement to capture, manage and process the data set in its entirety, such that it is not possible to process the data using traditional software tools and analytic techniques within tolerable time frames."

The data behind big data in financial services is not just the familiar structured market and reference data, but unstructured data from a variety of new sources.

"The new type of data may include blogs, documents of the visitors to a website, documents that describe market offerings or user-generated content from a customer support or customer engagement site," says Mike Olson, CEO of analytics and data management vendor Cloudera. "It might even include audio recording from phone calls that customers or traders have made."

"There are over 5 billion minutes of voice data generated annually," says Mehta. "There is a ton of useful information we can use to better serve our customers. We did not look at it before. Now we throw it all into Apache Hadoop and layer a large massively parallel database on top of it. So I can enable my data quants who love to program in SQL to manage and ping that data at scale."

metia

However, not all data analytics are up to the task of handling large datasets. Bank of America set up a side-by-side testbed using a platform from an unnnamed but “well-known” data analytics vendor and the open-source Apache Hadoop platform, which is designed especially for big data. In the first test, both platforms processed 90 million rows of data—a total of 15 GB—in nine minutes.

The second test entailed processing 64 GB of data consisting of 540 million rows of data. It took Apache Hadoop nine minutes to finish the test, while the other platform took 45 minutes. The final test involved processing 5.4 billion rows of data—a total of 640 GB. It took Apache Hadoop 28 minutes to complete the test. The other platform could not finish the test, according to Mehta. “We realized there are inflection points where typical data analysis tools break and can’t handle data at scale.”

To handle big data, the industry is turning to a number of available technologies that offer MapReduce functionality, which breaks down queries into massively parallel sub-queries and then reassembles the results for the final answer.

The pace of adoption of this sort of technology has increased in recent years, according to Aslett. Between vendor-developed platforms like ScaleOut Software’s ScaleOut StateServer, which provides MapReduce functionality for in-memory data, and open-source platforms like Apache Hadoop, there are no typical deployments that will get these technologies into the datacenters, according to David Brinker, COO of ScaleOut Software.

Some firms may see a good return on investment (ROI) when it comes to backtesting trading strategies. “The idea of a quant who can quickly run multiple strategies against an in-memory store of stock history is a use-case that gets a lot of interest,” says Brinker. “Churning through 10 years of history for 4,000 stock ticker symbols is compute-intensive.”

Other compute-intensive calculations like Value-at-Risk (VaR), Monte Carlo simulations and portfolio analysis also can also benefit from MapReduce processing.

MapReduce can even handle the more mundane functions like identifying and updating a large number of objects that are stored in cache. “If you have trade data in memory, you can use MapReduce to market those trades as today’s trades then go back and use parallel queries or operations on those objects,” says William Bain, founder and CEO of ScaleOut Software.

Although MapReduce can trace its history back to developments in the supercomputing industry in the 1990s, it received a major boost when the open-source community launched the Hadoop project under the auspices of the Apache Software Foundation in 2006. The project resulted from an academic paper published by Google in 2004 on how the search engine giant handled its own data management issues.

“Ultimately, Google and Facebook are trying to solve the same issues: storing vast amount of data from a variety of sources and processing it as economically as possible,” says Aslett.

“You can put months’ and years’ worth of trading data online as well as combine that information with the data about the state of the market generally,” adds Olson. “You can ingest other unstructured data types alongside that structured data, combine them, and then analyze them with the system.”

Like many other open-source projects, the development of Apache Hadoop moves forward by consensus, which can take more time to add features to the current code build than its proprietary counterparts, such as in-memory processing.

“Hadoop does a lot of logging on disk, but it is possible that you can do logging in-memory and reduce the time it takes to do the MapReduce tasks,” says Prasenjit Sakar, master inventor, manager storage analytics and resiliency at IBM. “Say one MapReduce function ends and dumps its data onto a disk and then a second MapReduce function begins and reaches out to the results of the first MapReduce task. What if we could optimize it by having the results of the first MapReduce task go directly to memory? It might take Hadoop a while to reach this functionality, but it will eventually get there.”

While Hadoop users wait for these functionalities to be added to the base code, a number of independent software vendors (ISVs) have stepped up to provide bridging applications.

Cloudera, which contributes resources to the Apache Hadoop project, offers a suite of open-source technologies, which it bundles with the Apache Hadoop release.

"The entire package is 100 percent open-source and free to download," says Olson. "We've integrated it, tested it and built easy-to-use installers to deliver the package to the market."

To generate revenue, Cloudera also offers its proprietary analytics and management platform for Apache Hadoop, which "basically allows IT staffers who have competency as Oracle database administrators or as Microsoft Certified Engineers to administer tens, hundreds or thousands of nodes in a Hadoop cluster," says Olson.

The tools are written in a few programming languages, including Java, and come with a browser-based client interface and a backend that runs on an Apache Hadoop cluster.

Other soon-to-be available tools include the eponymous Hadapt, which will create a hybrid Apache-Hadoop–relational-database environment.

"What we have done with Hadapt is make Hadoop more relational, and relational databases more Hadoop-friendly," says Justin Borgman, CEO of Hadapt. "We've taken the relational benefits—like the ability to use a SQL interface and the ability to use column-store technology—that have emerged in the past five years or so." Hadapt will also include built-in workload balancing and fault tolerance to deal with queries spread across multiple servers or a cloud environment.

Go Big or Go Home

Cloudera's Olson estimates that the number of Apache Hadoop clusters in production globally are in the thousands rather than the millions. "We expect to see adoption across all the verticals and geographies," Olson says.

Most firms are doing development and testing in order to figure out what benefits that they can gain from an implementation along with other distributed file systems, adds 451 Group's Aslett. "There isn't a chance for firms to get locked into one of these deployments yet, since they're only just dipping their toes into the water." Olson sees this trend reflected in the recent deals that have been signed by Cloudera clients. "We're not doing deals worth tens of millions of dollars for an enterprise-wide deployment," he says. "What we do expect over the course of the next few years is that Hadoop will grow as a platform that will rival the size of relational databases, but that will take a little while."

Currently, the typical implementation starts with the company's tiger team. Once the immediate benefits are seen, the IT team tends to get an evangelical furor, says Olson.

However, Mehta believes that to gain the necessary traction for Apache Hadoop, organizations have to think big—really big.

"We couldn't put five boxes under our desk and claim that it works," he explains. To prove that Apache Hadoop ecosystem could be a game-changer, Mehta's organization needed to build it to scale.

"You have to be ambitious, buy into the fact that this is going to change the business model and disrupt existing economic models," says Mehta. "Hadoop today is like Linux was 20 years ago."

He suggests that IT organizations start framing the adoption of Apache Hadoop more as a commercial opportunity rather than solving a technology problem. "We have a problem to solve and it can't be solved by existing technology and the path that we are going to go forward with will be a new technology platform," he says. "Being sandwiched between the credit crisis and the new regulations, there are enough problems. It's not difficult to find them."