InfoWorld

September 14, 2010

GET TECHNOLOGY RIGHT®

The big promise of Big Data: What you need to know today

Hadoop and other tools can unlock critical insights from unfathomable volumes of corporate and external data

In the never-ending quest for a competitive advantage, organizations are turning to large repositories of corporate and external data to uncover trends, statistics, and other actionable information to help decide on their next move. Those data sets, along with their associated tools, platforms, and analytics, are often referred to as "Big Data," a term that is gaining popularity among technologists and executives alike.

Although decision-makers have realized there's value in Big Data, getting to that value has remained elusive in most businesses. That's where IT can help, creating services that empower researchers to delve through large data stores to perform analytics and discover important trends. In other words, IT will prove to be the catalyst that delivers on the promise of Big Data.

Big Data has already proved its importance and value in several areas. Organizations such as the U.S. National Oceanic and Atmospheric Administration (NOAA), U.S. National Aeronautics and Space Administration (NASA), several pharmaceutical companies, and numerous energy companies have amassed huge amounts of data and now leverage Big Data technologies on a daily basis to extract value from them.

NOAA uses Big Data approaches to aid in climate, ecosystem, weather, and commercial research, while NASA uses Big Data for aeronautical and other research. Pharmaceutical companies and energy companies have leveraged Big Data for more tangible results, such as drug testing and geophysical analysis. The New York Times has used Big Data tools for text analysis and Web mining, while Disney uses them to correlate and understand customer behavior across its stores, theme parks, and Web properties.

Big Data plays another role in today's businesses: Large organizations increasingly face the need to maintain massive amounts of structured and unstructured data — from transaction information in data warehouses to employee tweets, from supplier records to regulatory filings — to comply with government regulations. That need has been driven even more by recent court cases that have encouraged companies to keep large quantities of documents, email messages, and other electronic communications such as instant messaging and IP telephony that may be required for e-discovery if they face litigation.

Perhaps the biggest challenge facing those pursuing Big Data is getting a platform that can store and access all the current and future information and make it available online for analysis cost-effectively. That means a highly scalable platform. Such platforms consist of storage technologies, query languages, analytics tools, content analysis tools, and transport infrastructures — there are many moving parts for IT to deploy and look after.

There are many proprietary and open source resources for these tools, often from startups but also from established cloud technology companies such as Amazon.com and Google — in fact, use of the cloud helps solve the Big Data scalability issue, both for data storage and computational capability. However, Big Data does not necessarily have to be a "roll your own" type of deployment. Large vendors such as IBM and EMC offer tools for Big Data projects, though their costs can be high and hard to justify.

Hadoop: The core of most Big Data efforts

In the open source realm, the big name is Hadoop, a project administered by the Apache Software Foundation that consists of Google-derived technologies for building a platform to consolidate, combine, and understand data.

Technically, Hadoop consists of two key services: reliable data storage using the Hadoop Distributed File System (HDFS) and high-performance parallel data processing using a technique called MapReduce. The goal of those services is to provide a foundation where the fast, reliable analysis of both structured and complex data becomes a reality. In many cases, enterprises deploy Hadoop alongside their legacy IT systems, which allows them to combine old and new data sets in powerful new ways. Hadoop allows enterprises to easily explore complex data using custom analyses tailored to their information and questions.

Hadoop runs on a collection of commodity, shared-nothing servers. You can add or remove servers in a Hadoop cluster at will; the system detects and compensates for hardware or system problems on any server. Hadoop, in other words, is self-healing. It can deliver data — and run large-scale, high-performance processing jobs — in spite of system changes or failures.

Although Hadoop provides a platform for data storage and parallel processing, the real value comes from add-ons, cross-integration, and custom implementations of the technology. To that end, Hadoop offers subprojects, which add functionality and new capabilities to the platform:

- Hadoop Common: The common utilities that support the other Hadoop subprojects.
- Chukwa: A data collection system for managing large distributed systems.
- HBase: A scalable, distributed database that supports structured data storage for large tables.
- HDFS: A distributed file system that provides high throughput access to application data.

- Hive: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- MapReduce: A software framework for distributed processing of large data sets on compute clusters.
- Pig: A high-level data-flow language and execution framework for parallel computation.
- ZooKeeper: A high-performance coordination service for distributed applications.

Most implementations of a Hadoop platform will include at least some of these subprojects, as they are often necessary for exploiting Big Data. For example, most organizations will choose to use HDFS as the primary distributed file system and HBase as a database, which can store billions of rows of data. And the use of MapReduce is almost a given since its engine brings speed and agility to the Hadoop platform.

With MapReduce, developers can create programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers. The MapReduce framework is broken down into two functional areas: Map, a function that parcels out work to different nodes in the distributed cluster, and Reduce, a function that collates the work and resolves the results into a single value.

One of MapReduce's primary advantages is that it is fault-tolerant, which it accomplishes by monitoring each node in the cluster; each node is expected to report back periodically with completed work and status updates. If a node remains silent for longer than the expected interval, a master node makes note and reassigns the work to other nodes.

Building on Hadoop

In addition to many open source support tools such as Clojure and Thrift, dozens of commercial options exist as well, though many are built using Hadoop as the foundation. The PricewaterhouseCoopers Center for Technology and Innovation has published an in-depth guide to the Big Data building blocks and how they relate to both IT deployment and business usage.

One example is Datameer, which provides a platform to collect and read different kinds of large data stores, put them into a Hadoop framework, and then provide tools for analysis of this data. Basically, Datameer seeks to hide the complexity of Hadoop and provide analysis tools on top of it. Datameer's sweet spot is data sources that exceed 10TB, the size at which Datameer says companies begin to struggle with using traditional technologies for data analysis.

Other commercial vendors offering similar approaches to Big Data analytics include Appistry, Cloudera, Drawn to Scale HQ, Goto Metrics, Karmasphere, and Talend. And the three main database vendors -- IBM, Microsoft, and Oracle -- all support Hadoop interaction, though in different ways. The open source BI vendor Pentaho also supports Hadoop.

Big Data fits businesses of all sizes

Big Data is not just all about size; it is also about performance, whatever the dimensions of the data set. That matters for immediate analytics, such as assessing a customer's behavior on a website to better understand what support they need or product they seek, or figuring out implications of current weather and other conditions on delivery routing and scheduling.

That is where server clusters, high-performance file systems, and parallel processing come into play. In the past, those technologies were too expensive for all but the largest businesses. Today, virtualization and commodity hardware have reduced the costs significantly, making Big Data available to small and medium-size businesses.

Those smaller businesses also have another path to Big Data analytics: the cloud. Cloud services for Big Data are popping up, offering platforms and tools to perform analytics quickly and efficiently.

But do smaller businesses really need access to Big Data? Simply put, yes. All companies have Big Data whether they realize it or not. For example, most online businesses collect large volumes of data from their log files and clickstream data. For companies that don't have such data streams, storing gigabytes not terabytes of information, Big Data tools let them tap into the vast trove of publicly available data sources.

Witness: The World Bank makes its statistical data about the entire world available online, and the Library of Congress is archiving all Twitter data since March 2006. What's more, there are plenty of news and investment data services that offer low-cost access to their information. Big Data techniques can be used to analyze these data sources, in addition to your own data — or both together.

Take, for instance, FlightCaster, a company that offers improved accuracy in predicting flight delays and, in the process, outperforms estimates by the major airlines. FlightCaster mines large amounts of historical data on domestic flights and factors in real-time conditions, as well as other proprietary elements using much of the same (public) data available to the airlines. FlightCaster's secret sauce is its practical understanding of Big Data analytics and the application of the proper tools to calculate the outcome in real time.

As costs fall and companies think of new ways to correlate data, Big Data analytics will become more commonplace, perhaps providing the growth mechanism for a small company to become a large one. Consider that Google, Yahoo, and Facebook were all once small companies that leveraged their data and understanding of the relationships in that data to grow significantly. It's no accident that many of the underpinnings of Big Data came from the methods these very businesses developed. But today, these methods are widely available through Hadoop and other tools for enterprises such as yours.

— Frank J. Ohlhorst