Technology Evaluation Centers

About Big Data

Jorge García - 11/18/2011 10:21:00 AM

There is no general consensus with respect to how big *big data* is—some companies deal with data volumes in the order of terabytes or even petabytes—but not many people will disagree that managing these huge amounts of data represents a challenge. It's fair to say that we're dealing with big data when traditional relational databases and systems are no longer sufficient.

Things as simple as data storage and movement between repositories can have a big impact on the organization. Big data management is more than just working with an enormous data set; it has to do with the complexity of analyzing it and getting the most value from it— competitive advantage, performance improvement, and, of course, profit. Big data requires special strategies and tools, and has to be considered from a broader perspective than mere size.

More than Just Size

Big data has three main features:

- **Volume.** Volume is the first and most notorious feature. It refers to the amount of data to be handled. Many organizations are producing large amounts of data internally, or gathering other large amounts of data from the exterior.
- **Variety.** The variety of data that organizations collect has increased in several ways: there are more internal systems having (primarily structured) data collected from them, and there is the rise of internal and external sources of data from semi- or nonstructured social media sources, such as blogs and tweets, as well as data coming from sensors and even plain-text documents.
- **Velocity.** As with traditional types of solutions (e.g., the data warehouse), latency periods are being reduced. Information is often sensitive and needs to be used and moved according to certain timeframes to obtain the best possible value from it. Real-time or near real-time answers are common needs in modern organizations.

Once it's established that big data is an issue, there are a couple of major aspects to consider. **The complexity of the data** will determine the difficulty of reliably exploiting the information beneath the big data. This in turn will guide an organization in acquiring **the technology to deal with the data**—the combination of hardware and software technologies that make the handling of its big data possible.

Some organizations have realized that relational database management systems (RDBMSs) are not sufficient for managing large and diverse amounts of data, and traditional business intelligence (BI) applications are not powerful enough to unveil the insights in a proper and timely manner. They need to deploy specific technologies to be able to work with big data.

A big data solution provides the technical means to perform operations with high volumes of data in a short period of time, with the ability to treat various types of data from disparate sources.

Why the Hype?

One of the main triggers for the design of new applications and technologies is the inability of common BI deployments to manage both structured and unstructured content. The data extraction process can be especially difficult with large amounts of information.

These new tools are changing the traditional BI data cycle. Data can be collected from its sources and analyzed in a matter of seconds, giving reliable results in a fraction of the time required by a traditional BI deployment, thus reducing data latency and speeding up the decision-making process. Some of the advantages of deploying a big data solution include:

- reducing the decision-making process by reading, analyzing, and giving results faster than traditional solutions
- collecting information, whether structured, semi-structured, or nonstructured, from disparate sources, and being able to manage it

- performing data discovery tasks, allowing you to build test scenarios, which is extremely important for building better analytic solutions and improving existing ones, as well as performing analysis on the go

There's also an economic angle to the big data hype. A corporate data warehouse can rapidly become expensive as the data volume increases. Scaling a data warehouse can be a burden when you're dealing with such volumes. Meanwhile, some big data providers can produce solutions not only that are cheaper from the get go, but which can be escalated, adapted, and modified as required.

Open source solutions—such as NoSQL—have also played an important role in the big data movement, forcing market prices to stay down.

The Players

As with any other segment in the software industry, the big data space is filled with providers that serve different aspects of handling big data. We can distinguish two major categories in the big data space.

Big data management systems are for administering big data volumes.

Big Data File and Database Management Systems

Product	Vendor	Commercial Provider of Related Products
Aster Database	Aster Data (acquired by Teradata)	
Ayrris	Appistry	
Cassandra	Apache Software Foundation (open source)	DataStax
Hadoop	Apache Software Foundation (open source)	Cloudera, Hortonworks, MapR, Microsoft Big Data, IBM InfoSphere BigInsights
Hypertable	Hypertable.org (open source)	
MongoDB	MongoDB.org (open source)	10gen
Riak	Basho	

Big data analytics appliances are products for analyzing large volumes and their information sets.

Big Data Analytics Appliances

Product	Vendor
1010Data DBMS	1010Data
Greenplum Data Computing Appliance (DCA)	EMC
IBM Netezza Analytics	Netezza, an IBM company
Infobright Enterprise Edition	Infobright
Oracle Big Data Appliance	Oracle
ParAccel Analytic Platform	ParAccel
SQL Server R2 Parallel Data Warehouse	Microsoft
Sybase IQ	Sybase, an SAP company
Vectorwise	Actian (formerly Ingres)
Vertica Advanced In-Database Analytics	Vertica, an HP company
WX2	Kognitio

Big data has had a rapid uptake by traditional BI vendors. Some of them offer connectors to big data applications in order to be able to analyze the data. A few of these vendors are Pentaho, Tableau Software, Endeca (acquired by Oracle), Jaspersoft, and MicroStrategy.

Getting Started

Here's a summary of some basic things to take into account when selecting a big data provider:

1. Calculate the challenges and opportunities buried within your data. Determine the most important problem in terms of the management and analysis of your vast amounts of data, and focus on it.
2. Identify your needs clearly. Before starting to explore a list of vendors, evaluate the type of technology and information you will require. Once you start to explore your options, make sure you understand your data problem and what you need in order to solve it.
3. Don't rush; plan. Make sure you are aligning your big data initiative with your corporate goals, and be sure the benefits and risks are clear. Clear the path to success.

A big data solution involves the complete data life cycle, from data collection to its visual representation. The explosion of data within an organization can be the impetus for a big data strategy. Organizations that succeed in the deployment of this sort of solution are the ones that are able to identify the type of data to be managed, the process the data needs to undergo, and the nature of the information to be obtained. Following this path an organization can select and deploy the necessary technology for the best use of its data.