BeyeNETWORK
Global coverage of the business intelligence ecosystem

Big Data and its Impact on Data Warehousing

The "big data" movement has taken the information technology world by storm. Fueled by open source projects emanating from the Apache Foundation, the big data movement offers a cost-effective way for organizations to process and store large volumes of any type of data: structured, semi-structured and unstructured.

BY WAYNE ECKERSON

TechTarget

Despite Problems, Big Data Makes it Huge

THE HYPE AND REALITY of the big data movement is reaching a crescendo. It's clear that Hadoop and NoSQL technologies are gaining a foothold in corporate computing environments. But big data software and computing paradigms are still in their infancy and must clear many hurdles before organizations trust them to handle serious data and application workloads.

Most leading big data vendors now count hundreds of customers. Big data is no longer the province of Internet and media companies with large Web properties; companies in nearly every industry are jumping on the big data bandwagon. These include energy, pharmaceuticals, utilities, telecommunications, insurance, retail, financial services and government.

For example, Vestas Wind Systems, a leading wind turbine maker, uses BigInsights to model larger volumes of weather data so it can pinpoint the optimal placement of wind turbines. And a financial services customer uses Hadoop to improve the accuracy of its fraud models by addressing much larger volumes of transaction data.

BIG DATA DRIVERS

Hadoop clearly fills an unmet need in many organizations. Given its open source roots, Hadoop provides a more cost-effective way to analyze large volumes of data compared with traditional relational database management systems (RDBMSes). It's also better suited to processing unstructured data, such as audio, video or images, and semi-structured data, such as Web server log data for tracking customer behavior on social media sites. For years, leading-edge companies have struggled in vain to figure out an optimal way to analyze this type of data in

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

traditional data warehousing environments, but without much luck.

Finally, Hadoop is a load-and-go environment: Administrators can dump the data into Hadoop without having to convert it into a particular structure. Then users (or data scientists) can analyze the data using whatever tools they want, which today are typically languages, such as Java, Python or Ruby. This type of data management paradigm appeals to application developers and analysts, who often feel straitjacketed by top-down, IT-driven architectures and SQL-based tool sets.

SPEED BUMPS

But Hadoop is not a data management panacea. It's clearly at or near the apogee of its hype cycle right now, and its many warts will disillusion all but bleeding- and leading-edge adopters.

For starters, Hadoop is still wet behind the ears. The Apache Software Foundation just released the equivalent of version 1.0. So there are plenty of basic things missing from the environment—like security, a metadata catalog, data quality, backups and monitoring and control. Moreover, it's a batch processing environment, not terribly efficient in the way it exploits a clustered environment. Hadoop knockoffs, like MapR, which embed proprietary technology underneath Hadoop application programming interfaces claim up to five-fold faster performance on half as many nodes.

To run a Hadoop environment, you need to get software from a mishmash of Apache projects, with razzle-dazzle names like Flume, Sqoop, Ooze, Pig, Hive and ZooKeeper. These independent projects often contain competing functionality, have separate release schedules and aren't always tightly integrated. And each project evolves rapidly. That's why there is a healthy market for Hadoop distributions that package these components into a reasonable set of implementable software.

> To run a Hadoop environment, you need to get software from a mishmash of Apache projects, with razzle-dazzle names like Flume, Sqoop, Ooze, Pig, Hive and ZooKeeper.

But the biggest complaint among big data advocates is the current lack of data scientists to build Hadoop applications. These wunderkinds combine a rare set of skills: statistics and math, data, process and domain knowledge and computer programming. Unfortunately, developers have little data and domain experience and data experts don't know how to program. So there is a severe shortage of talent. Some com-

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

panies are hiring several people with related skills to cobble together one complete "data scientist."

EVOLUTION

One good thing about the big data movement is that it evolves fast. There are Apache projects to address most of the shortcomings of Hadoop. One promising project is Hive, which provides SQL-like access to Hadoop, although it's stuck in a batch processing paradigm. Another is HBase, which overcomes Hadoop's latency issues, but is designed for fast row-based reads and writes to support high-performance transactional applications. Both create table-like structures on top of Hadoop files.

Many commercial vendors have jumped into the fray, marrying proprietary technology with open source software to turn Hadoop into a more corporate-friendly compute environment. Vendors such as Zettaset, EMC Greenplum and Oracle have launched appliances that embed Hadoop with commercial software to offer customers the best of both worlds. Many BI and data integration vendors, such as Talend, now connect to Hadoop and can move data back and forth seamlessly. Some even create and run MapReduce jobs in Hadoop using their standard visual development environments. Even Microsoft has jumped into the fray, offering its Hadoop port of Windows Server, an ODBC-to-Hive driver, and a new JavaScript framework for MapReduce to the Apache Foundation.

> Established software vendors stand to lose significant revenue if Hadoop evolves without them and gains robust data management and analytical functionality that cannibalizes their existing products.

COOPERATION OR COMPETITION?

Although vendors are quick to rally behind big data, there is some measure of desperation in the move. Established software vendors stand to lose significant revenue if Hadoop evolves without them and gains robust data management and analytical functionality that cannibalizes their existing products. They either need to generate sufficient revenue from new big data products or circumscribe Hadoop so that it plays a subservient role to their existing products. Most vendors are hedging their bets and playing both options, especially database vendors who perhaps have the most to lose.

Both sides are playing nice and are eager to partner and work together. Hadoop vendors benefit as more applications run on Hadoop, including traditional products centering on business intelligence, extract, transform and load (ETL) and DBMSes. And commercial vendors benefit if their existing tools have a new source of data to connect to and plumb. It's a big new market whose sweet-tasting honey attracts a hive full of bees.

WHY INVEST IN PROPRIETARY TOOLS?

But customers are already asking whether data warehouses and BI tools will eventually be folded into Hadoop environments or the reverse. Why spend millions of dollars on a new analytical RDBMS if you can do that processing without paying a dime in license costs using Hadoop? Why spend hundreds of thousands of dollars on data integration tools if your data scientists can turn Hadoop into a huge data staging and transformation layer? Why invest in traditional BI and reporting tools if your power users can exploit Hadoop using freely available programs such as Java, Python, Pig, Hive or Hbase?

THE FUTURE IS CLOUDY

Right now, it's too early to divine the future of the big data movement and predict winners and losers. It's possible that in the future all data management and analysis will run entirely on open source platforms and tools. But it's just as likely that commercial vendors will co-opt (or outright buy) open source products and functionality and use them as pipelines to magnify sales of their commercial products.

More than likely, we'll get a mélange of open source and commercial capabilities. After all, 30 years after the mainframe revolution, mainframes are still a mainstay at many corporations. In IT, nothing ever dies; it just finds its niche in an evolutionary ecosystem. ■

CHAPTER 2

Two Markets for Big Data: Comparing Value Propositions

THERE ARE TWO types of big data in the market today. There is open source software, centered largely on Hadoop, which eliminates up-front licensing costs for managing and processing large volumes of data. And then there are new analytical engines, including appliances and column stores, which provide significantly higher price-performance than general-purpose relational databases. Both sets of big data software deliver higher return on investment than previous generations of data management technology, but in vastly different ways.

HADOOP

Free software. Hadoop is an open source distributed file system available through the Apache Software Foundation that is capable of storing and processing large volumes of data in parallel across a grid of commodity servers. Hadoop emanated from large Internet providers, such as Google and Yahoo, which needed a cost-effective way to build search indexes. They knew that traditional relational databases would be prohibitively expensive and technically unwieldy, so they came up with a low-cost alternative they built themselves and eventually gave to the Apache Software Foundation so others could benefit from their innovations.

Today many companies are implementing Hadoop software from Apache as well as third-party providers, such as IBM, Cloudera, Hortonworks and EMC. Developers see Hadoop as a cost-effective way to get their arms around large volumes of data that they›ve never been able to do much with before.

Many companies use Hadoop to store, process and analyze large volumes of Web server log data so they can get a better feel for the browsing

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

and shopping behavior of their online customers. Before, companies outsourced the analysis of their clickstream data or simply let it fall on the floor since they didn›t have a way to process it in a timely and cost-effective way. Companies are also turning to Hadoop to process more structured data to improve analytical models.

Data agnostic. Besides being free to implement, the other major advantage of big data software is that it is data agnostic. It can handle any type of data. Unlike a data warehouse or traditional relational database, Hadoop doesn't require administrators to model or transform data before they load it. With Hadoop, you don't define a structure for the data; you simply load and go. This significantly reduces the cost of preparing data for analysis compared with what happens in a data warehouse. Most experts assert that 60% to 80% of the cost of building a data warehouse, which can run into the tens of millions of dollars, involves extracting, transforming and loading data. Hadoop virtually eliminates this cost.
As a result, many companies are using Hadoop as a general-purpose staging area and archive for all their data. So a telecommunications company can store 12 months of call detail records instead of aggregating that data in the data warehouse and rolling the details to offline storage. With Hadoop, they can keep all their data online and eliminate the cost of data archival systems. They can also let power users query Hadoop data directly if they want to access the raw data or can't wait for the aggregates to be loaded into the data warehouse.

Hidden costs. Of course, nothing in technology is ever free. When it comes to processing data, you either pay the piper up front, as in the data warehousing world, or at query time, as in the Hadoop world. Before querying Hadoop data, a developer needs to understand the structure of the data and all of its anomalies. With a clean, well-understood, homogenous data set, this is not difficult. But most corporate data doesn't fit that description. So a Hadoop developer ends up playing the role of a data warehousing developer at query time, interrogating the data and making sure it's format and content match their expectations. Querying Hadoop today is a "buyer beware" environment.

Moreover, to run big data software, you still need to purchase, install and manage commodity servers (unless you run your big data environment in the cloud, say through Amazon Web Services). While each server may not cost a lot, the price adds up.

But what's more costly is the expertise and software required to administer Hadoop and manage grids of commodity servers. Hadoop is still bleeding-edge technology and few people have the skills or experience to run it efficiently in a production en-

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

vironment. These folks are hard to find, and they don't come cheap. The Apache Software Foundation admits that Hadoop's latest release is equivalent to version 1.0 software. So even the experts have a lot to learn, since the technology is evolving at a rapid pace. But nonetheless, Hadoop and its NoSQL brethren have opened up a vast new frontier for organizations to profit from their data.

ANALYTICAL PLATFORMS

The other type of big data predates Hadoop and NoSQL variants by several years. This version of big data is less a "movement" than an extension of existing relational database technology optimized for query processing. These analytical platforms span a range of technology, from appliances and columnar databases to shared nothing, massively parallel processing databases. The common thread among them is that most are read-only environments that deliver exceptional price-performance compared with general-purpose relational databases originally designed to run transaction processing applications.

Teradata laid the groundwork for the analytical platform market when it launched the first analytical appliance in the early 1980s. Sybase was also an early forerunner, shipping the first columnar database in the mid-1990s. IBM Netezza kicked the current market into high gear in 2003 when it unveiled a popular analytical appliance, and was soon followed by dozens of startups. Recognizing the opportunity, all the big names in software and hardware—Oracle, IBM, Hewlett-Packard, and SAP—subsequently jumped into the market, either by building or buying technology, to provide purpose-built analytical systems to new and existing customers.

Although the pricetag of these systems often exceeds $1 million, customers find that the exceptional price-performance delivers significant business value, in both tangible and intangible form. For example, Virginia-based XO Communications recovered $3 million in lost revenue from a new revenue assurance application it built on an analytical appliance, even before it had paid for the system. It subsequently built or migrated a dozen applications to run on the new purpose-built system, testifying to its value.

Kelley Blue Book in Irvine, Calif., purchased an analytical appliance to run its data warehouse, which was experiencing performance issues, giving the provider of online automobile valuations a competitive edge. For instance, the new system reduces the time needed to process hundreds of millions of automobile valuations from one week to one day. Kelley Blue Book now uses the system to analyze its Web advertising business and deliver dynamic pricing for its Web ads.

Challenges. Given the up-front costs of analytical platforms, organizations usually undertake a thorough evaluation of these systems before jumping aboard.

First, a company must determine whether an analytical platform outperforms its existing data warehouse database to a degree that it warrants migration and retraining costs. This requires a proof of concept in which the customer tests the systems in its own data center using its own data across a range of queries.

The good news is that the new analytical platforms usually deliver jaw-dropping performance for most queries tested. In fact, many customers don't believe the initial results and rerun the queries to make sure that the results are valid.

Second, companies must choose from more than two dozen analytical platforms on the market today. For instance, they must decide whether to purchase an appliance or a software-only system, a columnar database or an MPP database, or an on-premises system or a Web service. Evaluating these options takes time, and many companies create a short list that doesn't always contain comparable products.

Finally, companies must decide what role an analytical platform will play in their data warehousing architectures. Should it serve as the data warehousing platform? If so, does it handle multiple workloads easily or is it a one-trick pony? If the latter, what applications and data sets make sense to offload to the new system? How do you rationalize having two data warehousing environments instead of one?

Today we find that companies which have tapped out their SQL Server or MySQL data warehouses often replace them with analytical platforms to get better performance. But companies that have implemented an enterprise data warehouse on Oracle, Teradata or IBM often find that analytical platforms are best used when they sit alongside the data warehouse so they can handle new applications or existing analytical workloads offloaded to them. This architecture helps organizations avoid a costly upgrade to a data warehousing platform, which might easily exceed the cost of purchasing an analytical platform.

The big data movement consists of two separate but interrelated markets: one for Hadoop and open source data management software and the other for purpose-built SQL databases optimized for query processing. Hadoop avoids most of the up-front licensing and loading costs endemic to traditional relational database systems. But since the technology is still immature, there are hidden costs that have thus far kept many Hadoop implementations experimental in nature. On the other hand, analytical platforms are a more proven technology but impose significant up-front licensing fees and potential migration costs. Companies wading into the waters of the big data stream need to carefully evaluate their options. ■

CHAPTER 3

Categorizing Big Data Processing Systems

FACED WITH AN expanding analytical ecosystem, BI managers need to make many technology choices. Perhaps the most difficult involves selecting a data processing system to power a variety of analytical applications (see Chapter 4, "The New Analytical Ecosystem: Making Way for Big Data").

In the past, these types of decisions revolved around selecting one of a handful of leading relational database management systems to power a data warehouse or data mart. Often, the choice boiled down to internal politics as much as technical functionality.

Today the options aren't as straightforward, although politics may still play a role. Instead of selecting a single data management product, BI managers may need to select multiple platforms to outfit an expanding analytical ecosystem. And rather than evaluating four or five alternatives for each platform, the BI manager is faced with dozens of viable options in each category. The once lazy database market is now a beehive of activity.

Staying abreast of all the new products, partnerships and technological advances is now a full-time job. Industry analysts who make a living sifting through products in emerging markets are needed now more than ever. Most analysts will tell you that the first step in selecting an analytical platform is to understand the broad categories of products in the marketplace, and then make finer distinctions from there (see **FIGURE 1,** page 11).

At a high-level, there are four categories of analytical processing systems that are available today: transactional RDBMSes. The following describes those categories and can be used as a starting point when creating a short list of products during a product evaluation process:

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

FIGURE 1.

Database/Platform Positioning

OLTP databases	Analytical platforms	Hadoop	NoSQL
Oracle, DB2, SQL Server	Netezza, Vertica, Exadata, Teradata appliances	Cloudera, EMC, IBM, Hortonworks	Cassandra, MongoDB, MarkLogic, Aster Data
Transaction systems Enterprise data warehouse hub	Enterprise data warehouse to replace MySQL or SQL Server in fast-growing companies Analytical data marts to offload the DW Free-standing analytical sandboxes (big data, extreme performance)	Online data archive for all data (but mostly unstructured) Staging area to feed the data warehouse Analytical system when you want to query all the raw data (Hbase, Hive) Analytical system when you can't wait until data is modeled and put in the data warehouse (Hbase, Hive)	Key value pair databases for rapid data capture and analysis Document databases for high-performance application transactions Graph systems that capture relationships among entities Search databases for querying structured and unstructured data Hybrid SQL-MapReduce databases

1. Transactional RDBM Systems

Transactional RDBMSes were originally designed to support transaction processing applications, although most have been retrofitted with various types of indexes, join paths and custom SQL bolt-ons to make them more palatable to analytical processing. There are two types of transactional RDBMSes: enterprise and departmental.

- **Enterprise hubs.** The traditional enterprise RDBMSes, such as those from IBM, Oracle and Sybase, are best suited as data warehousing hubs that feed a variety of downstream, end-user-facing systems, but don't handle query traffic directly. Although retrofitted with analytical capabilities, these systems often hit performance and scalability walls when used for query processing along with other workloads and are expensive to upgrade and replace. Thus, many customers now use these "gray-bearded" data warehousing sys-

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

tems as hubs to feed operational data stores, data marts, enterprise reporting systems, analytical sandboxes and various analytical and transactional applications.

- **Departmental marts.** A number of companies use Microsoft SQL Server or MySQL as data marts fed by an enterprise data warehouse or as standalone data warehouses for a business unit or small and medium-sized business (SMB). Like their enterprise brethren, these systems also often hit the wall when usage, data volumes or query complexity increases rapidly. A fast-growing business unit or SMB often replaces these transactional RDBMSes with analytical appliances (see below) which provide the same or greater level of simplicity and ease of management as SQL Server or MySQL.

2. Analytical Platforms

Analytic platforms represent the first wave of big data systems (see Chapter 2, "Two Markets for Big Data: Comparing Value Propositions"). These are purpose-built SQL-based systems designed to provide superior price-performance for analytical workloads compared with transactional RDBMSes. There are many types of analytical platforms. Most are being used as data warehousing replacements or standalone analytical systems.

- **MPP database.** Massively parallel processing (MPP) databases with strong mixed workload utilities make good enterprise data warehouses for analytically minded organizations. Teradata was the first on the block with such a system, but it now has many competitors, including EMC Greenplum and Microsoft's Parallel Data Warehousing option, which are relative upstarts compared to the 30-year old Teradata.

- **Analytical appliance.** These purpose-built analytical systems come as an integrated hardware-software combination tuned for analytical workloads. Analytical appliances come in many shapes, sizes and configurations. Some, like IBM Netezza, EMC Greenplum and Oracle Exadata, are more general-purpose analytical machines that can serve as replacements for most data warehouses. Others, such as those from Teradata, are geared to specific analytical workloads and can deliver extremely fast performance or manage super large data volumes.

- **In-memory systems.** If you are looking for raw performance, there is nothing better than a system that lets you put all your data into memory. These systems will soon become more commonplace, thanks

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

- to SAP, which is betting its business on HANA, an in-memory database for transactional and analytical processing, and is evangelizing the need for in-memory systems. Another contender in this space is Kognitio. Many RDBMSes are better exploiting memory for caching results and processing queries.

- **Columnar.** Columnar databases, such as SAP's Sybase IQ, Hewlett-Packard's Vertica, ParAccel, Infobright, Exasol, Calpont and Sand offer fast performance for many types of queries because of the way these systems store and compress data—by columns instead of rows. Column storage and processing is fast becoming a RDBMS feature rather than a distinct subcategory of products.

3. Hadoop Distributions

Hadoop is an open source software project run within The Apache Software Foundation for processing data-intensive applications in a distributed environment with built-in parallelism and failover. The most important parts of Hadoop are the Hadoop Distributed File System (HDFS), which stores data in files on a cluster of servers, and MapReduce, a programming framework for building parallel applications that run on HDFS. The open source community is building numerous additional components to turn Hadoop into an enterprise-caliber, data processing environment. The collection of these components is called a Hadoop distribution. Leading providers of Hadoop distributions include Cloudera, IBM, EMC, Amazon, Hortonworks and MapR.

Today, in most customer installations, Hadoop serves as a staging area and online archive for unstructured and semi-structured data, as well as an analytical sandbox for data scientists who query Hadoop files directly before the data is aggregated or loaded into the data warehouse. But this could change. Hadoop will play an increasingly important role in the analytical ecosystem at most companies, either working in concert with an enterprise DW or assuming most duties of one.

4. NoSQL Databases

NoSQL—shorthand for "not only SQL"—is the name given to a broad set of databases whose only common thread is that they don't require SQL to process data, although some support both SQL and non-SQL forms of data processing. There are many types of NoSQL databases, and the list grows longer every month. These specialized systems are built using either proprietary and open source components or a mix of both. In most cases, they are designed to overcome the limitations of traditional RDBMes to handle unstructured and semi-structured data. Here's a partial listing of NoSQL systems:

- **Key value pair databases.** These systems store data as a simple record structure consisting of a key and content. These are used for operational applications that involve large volumes of data, flexible data structures and fast transactions. Leading key value pair databases include Cassandra, Hbase and Basho Riak.

- **Document stores.** These systems specialize in storing, parsing and processing application objects, typically using a lightweight structure, such as JSON. Like key value databases, document stores are used for high-volume transaction processing. Leaders here include MongoDB and Couchbase.

- **SQL MapReduce.** These systems allow users to use SQL to invoke MapReduce jobs running inside the database or associated file system. Teradata's Aster Data and EMC Greenplum support these capabilities.

- **Graph systems.** These database store associations among entities, making them popular among social media companies that need to track different connections among people.

- **Unified information access.** These systems, such as those from Attivio, MarkLogic and Splunk, use more of a search storage and query paradigm to query structured and unstructured data.

- **Other.** There are many other NoSQL databases that vary in how they store and process data or the types of applications they are designed to support.

The above four categories represent just the start of a broader categorization of data processing systems geared to analytical workloads. This is a fast-changing field. With the multiplicity of choices available today, BI professionals need to understand the differences between data management offerings so they can position themselves properly in the new analytical ecosystem. ■

CHAPTER 4

The New Analytical Ecosystem: Making Way for Big Data

THE BIG DATA revolution has arrived, and it's transforming long-established data warehousing architectures into vibrant, multifaceted analytical ecosystems.

Gone are the days when all analytical processing first passes through a data warehouse or data mart (or their less sanctified spreadmart or data shadow system brethren). Now data winds its way to users through a plethora of corporate data structures, each tailored to the type of content it contains and the type of user who wants to consume it.

Figure 1 depicts a reference architecture for the new analytical ecosystem that has the fingerprints of big data all over it. The objects in blue represent the traditional data warehousing environment, while those in pink represent new architectural elements made possible by big data technologies; namely Hadoop, NoSQL databases, high-performance analytical engines—analytical appliances, MPP databases, in-memory databases—and interactive, in-memory visualization tools.

Most source data now flows through Hadoop, which primarily acts as a staging area and online archive. This is especially true for semi-structured data, such as log files and machine-generated data, but also for some structured data that companies can't cost-effectively store and process in SQL engines (e.g., call detail records in a telecommunications company). From Hadoop, data is fed into a data warehousing hub, which often distributes data to downstream systems, such as data marts, operational data stores and analytical sandboxes of various types, where users can query the data using familiar SQL-based reporting and analysis tools.

Today data scientists analyze raw data inside Hadoop by writing MapReduce programs in Java and other lan-

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

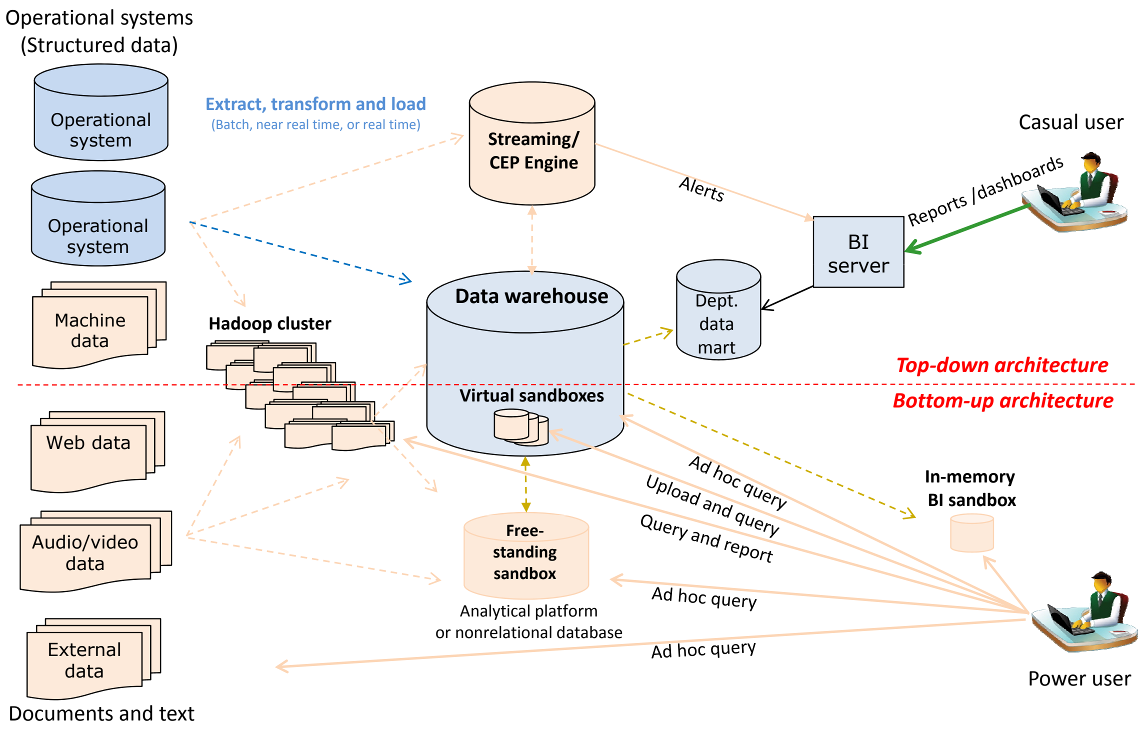
TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

FIGURE 1.

The New Analytical Ecosystem



guages. In the future, users will be able to query and process Hadoop data using SQL-based data integration and query tools.

HARMONIZING OPPOSITES

The big data revolution is not only about analyzing large volumes and new sources of data, it's also about balancing data alignment and consistency with flexible, ad hoc exploration. As such, the new analytical ecosystem features both top-down and bottom-up data flows that meet all business requirements for reporting and analysis.

The top-down world. Here source data is processed, refined and stamped with a predefined data structure—typically a dimensional model—and then consumed by casual users using SQL-based reporting and analysis tools. In this domain, IT developers create data and semantic models so business users can get answers to known questions and executives can track performance of predefined metrics.

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

Design precedes access. The top-down world also takes great pains to align data along conformed dimensions and deliver clean, accurate data. The goal is to deliver a consistent view of the business entities so users can spend their time making decisions instead of arguing about the origins and validity of data artifacts.

The underworld. Creating a uniform view of the business from heterogeneous sets of data is not easy. It takes time, money and patience, often more than most departmental heads and business analysts are willing to tolerate. They often abandon the top-down world for the underworld of spreadmarts and data shadow systems. Using whatever tools are readily available and cheap, these data-hungry users create their own views of the business. Eventually, they spend more time collecting and integrating data than analyzing it, undermining their productivity and a consistent view of business information.

The bottom-up world. The new analytical ecosystem brings these prodigal data users back into the fold. It carves out space within the enterprise environment for true ad hoc exploration and promotes the rapid development of analytical applications using in-memory departmental tools. In a bottom-up environment, users can't anticipate the questions they will ask on a daily or weekly basis or the data they'll need to answer those questions. Often, the data they need doesn't yet exist in the data warehouse.

The new analytical ecosystem supports three analytical sandboxes that enable power users to explore corporate and local data on their own terms: (1) Hadoop, (2) virtual partitions inside a data warehouse and (3) specialized analytical databases that offload data or analytical processing from the data warehouse or handle new untapped sources of data, such as Web server logs or machine data. The new environment also gives department heads the ability to create and use dashboards built with in-memory visualization tools that point both to a corporate data warehouse and other independent sources.

Combining top-down and bottom-up worlds is not easy. BI professionals need to assiduously guard data semantics while opening access to data.

Combining top-down and bottom-up worlds is not easy. BI professionals need to assiduously guard data semantics while opening access to data. For their part, business users need to commit to adhering to corporate data standards in exchange for getting the

DESPITE PROBLEMS, BIG DATA MAKES IT HUGE

TWO MARKETS FOR BIG DATA: COMPARING VALUE PROPOSITIONS

CATEGORIZING BIG DATA PROCESSING SYSTEMS

THE NEW ANALYTICAL ECOSYSTEM: MAKING WAY FOR BIG DATA

keys to the kingdom. To succeed, organizations need robust data governance programs and lots of communication among all parties.

The big data revolution brings major enhancements to the BI landscape. First and foremost, it introduces new technologies, such as Hadoop, that make it possible for organizations to cost-effectively consume and analyze large volumes of semi-structured data. Second, it complements traditional, top-down data-delivery methods with more flexible, bottom-up approaches that promote ad hoc exploration and rapid application development. ■

WAYNE ECKERSON has more than 15 years' experience in data warehousing, business intelligence (BI) and performance management. He has conducted numerous in-depth research studies and wrote the best-selling book Performance Dashboards: Measuring, Monitoring, and Managing Your Business. He is a keynote speaker and blogger and conducts workshops on business analytics, performance dashboards and business intelligence. Eckerson served as director of education and research at The Data Warehousing Institute, where he oversaw the company's content and training programs and chaired its BI Executive Summit.

Eckerson is director of research at TechTarget, where he writes a weekly blog called Wayne's World, which focuses on industry trends and examines best practices in the application of BI. He is also president of BI Leader Consulting and founder of BI Leadership Forum, a network of BI directors who exchange ideas about best practices in BI and educate the larger BI community. Email him at weckerson@techtarget.com.

TechTarget®

Big Data and its Impact on Data Warehousing is is a joint publication of BeyeNETWORK, SearchDataManagement.com and SearchBusinessAnalytics.com.

Hannah Smalltree
Editorial Director

Jason Sparapani
Managing Editor, E-Publications

Wayne Eckerson
Director of Research

Linda Koury
Director of Online Design

Mike Bolduc
Publisher
mbolduc@techtarget.com

Ed Laplante
Director of Sales
elaplante@techtarget.com

About TechTarget: TechTarget publishes media for information technology professionals. More than 100 focused websites enable quick access to a deep store of news, advice and analysis about the technologies, products and processes crucial to your job. Our live and virtual events give you direct access to independent expert commentary and advice. At IT Knowledge Exchange, our social community, you can get advice and share solutions with peers and experts.