BeyeNETWORK
Global coverage of the business intelligence ecosystem

Big Data Analytics:

Profiling the Use of Analytical Platforms in User Organizations

BY WAYNE ECKERSON
Director of Research, Business Applications and Architecture Group, TechTarget, September 2011

Executive Summary

THIS REPORT EXAMINES the rise of "big data" and the use of analytics to mine that data. Companies have been storing and analyzing large volumes of data since the advent of the data warehousing movement in the early 1990s. While terabytes used to be synonymous with big data warehouses, now it's petabytes, and the rate of growth in data volumes continues to escalate as organizations seek to store and analyze greater levels of transaction details, as well as Web- and machine-generated data, to gain a better understanding of customer behavior and drivers.

■ **Analytical platforms.** To keep pace with the desire to store and analyze ever larger volumes of structured data, relational database vendors have delivered specialized analytical platforms that provide dramatically higher levels of price-performance compared with general-purpose relational database management systems (RDBMSs). These analytical platforms come in many shapes and sizes, from software-only databases and analytical appliances to analytical services that run in a third-party hosted environment. Almost three-quarters (72%) of our survey respondents said they have implemented an analytical platform that fits this description.

Companies have been storing and analyzing large volumes of data since the advent of the data warehousing movement in the early 1990s.

In addition, new technologies have emerged to address exploding volumes of complex structured data, including Web traffic, social media content and machine-generated data, such as sensor and Global Positioning System (GPS) data. New nonrelational database vendors combine text indexing and natural language processing techniques with traditional database technology to optimize ad hoc queries against semi-structured data. And many Internet and media companies use new open source

frameworks such as Hadoop and MapReduce to store and process large volumes of structured and unstructured data in batch jobs that run on clusters of commodity servers.

■ **Business users.** In the midst of these platform innovations, business users await tools geared to their information requirements. Casual users—executives, managers, front-line workers—primarily use reports and dashboards that deliver answers to predefined questions. Power users—business analysts, analytical modelers and data scientists—perform ad hoc queries against a variety of sources. Most business intelligence (BI) environments have done a poor job meeting these diverse needs within a single, unified architecture. But this is changing.

Most business intelligence (BI) environments have done a poor job meeting these diverse needs within a single, unified architecture. But this is changing.

■ **Unified architecture.** This report portrays a unified reporting and analysis environment that finally turns power users into first-class corporate citizens and makes unstructured data a legitimate target for ad hoc and batch queries. The new architecture leverages new analytical technology to stage, store and process large volumes of structured and unstructured data, turbo-charge sluggish data warehouses and offload complex analytical queries to dedicated data marts. Besides supporting standard reports and dashboards, it creates a series of analytical sandboxes that enable power users to mix personal and corporate data and run complex analytical queries that fuel the modern-day corporation. ■

Research Background

THE PURPOSE OF this report is to profile the use of analytical platforms in user organizations. It is based on a survey of 302 BI professionals as well as interviews with BI practitioners at user organizations and BI experts at consultancies and software companies.

■ **Survey.** The survey consists of 25 pages of questions (approximately 50 questions) with four branches, one for each analytical platform deployment option: analytical database (software-only), analytical appliance (hardware-software combo), analytical service and file-based analytical system (e.g., Hadoop and NoSQL). Respondents who didn't select an option were passed to a fifth branch where they were asked why they hadn't purchased an analytical platform and whether they planned to do so.

The survey ran from June 22 to August 2, 2011, and was publicized through several channels. The BI Leadership Forum and BeyeNetwork sent several email broadcasts to their lists. I tweeted about the survey and asked followers to retweet the announcement. Several sponsors, including Teradata, Infobright, and ParAccel, notified their customers about the survey through email broadcasts and newsletters.

[This report] is based on a survey of 302 BI professionals as well as interviews with BI practitioners and BI experts.

■ **Respondent profile.** Survey respondents are generally IT managers based in North America who work at large companies in a variety of industries (see **FIGURES 1-4**, page 5). ■

FIGURE 1: *Which best describes your position in BI?*

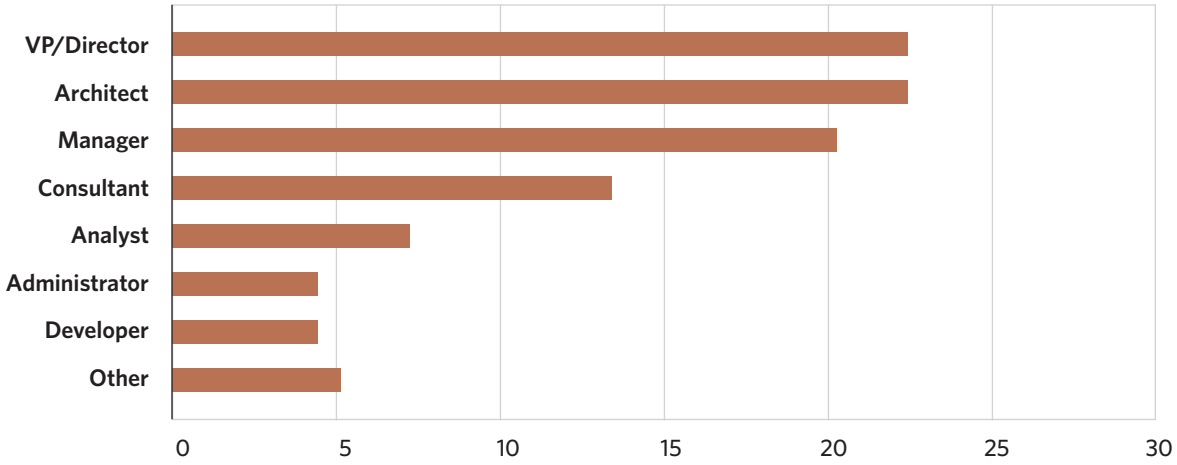


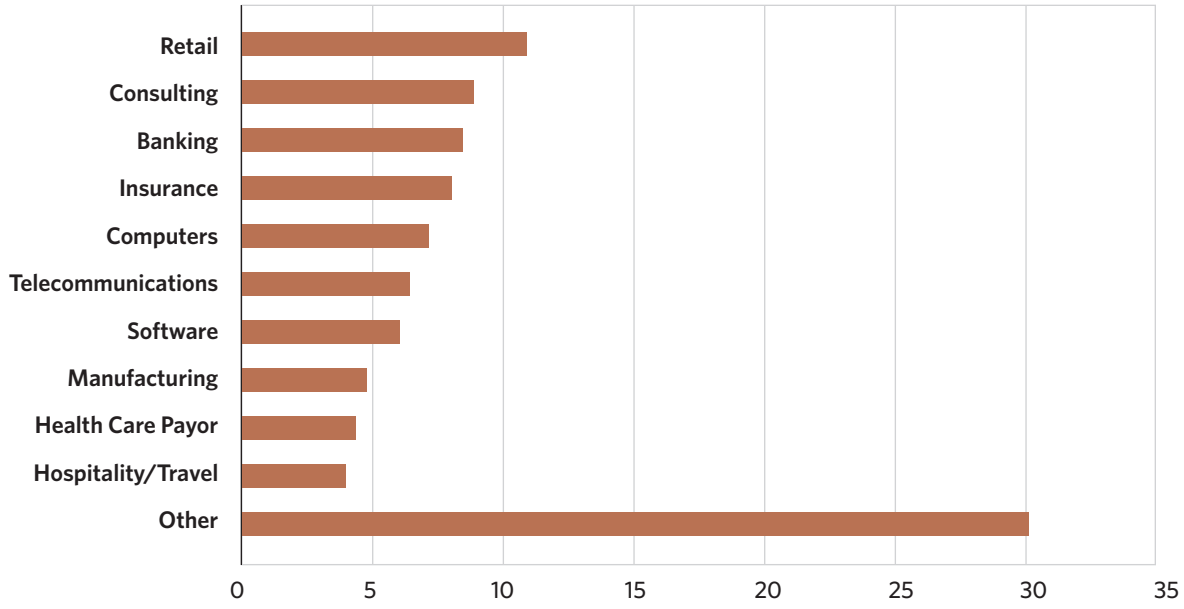
FIGURE 2: *Where are you located?*

North America	66.7%
Europe	16.5%
Other	16.9%

FIGURE 3: *What size is your organization by revenues?*

Large ($1B + revenues)	52.4%
Medium ($50M to $1B)	24.8%
Small (<$50M)	22.8%

FIGURE 4: *In what industry do you work?*



Why Big Data?

THERE HAS BEEN a lot of talk about "big data" in the past year, which I find a bit puzzling. I've been in the data warehousing field for more than 15 years, and data warehousing has always been about big data.

Back in the late 1990s, I attended a ceremony honoring the Terabyte Club, a handful of companies that were storing more than a terabyte of raw data in their data warehouses. Fast-forward more than 10 years and I could now be attending a ceremony for the Petabyte Club. The trajectory of data acquisition and storage for reporting and analytical applications has been steadily expanding for the past 15 years.

So what's new in 2011? Why are we are talking about big data today? There are several reasons:

The growth in data is fueled by largely unstructured data from websites and machine-generated data from an exploding number of sensors.

1. Changing data types. Organizations are capturing different types of data today. Until about five years ago, most data was transactional in nature, consisting of numeric data that fit easily into rows and columns of relational databases. Today, the growth in data is fueled by largely unstructured data from websites (e.g, Web traffic data and social media content) as well as machine-generated data from an exploding number of sensors. Most of the new data is actually semi-structured in format, because it consists of headers followed by text strings. Pure unstructured data, such as audio and video data, has limited textual content and is more difficult to parse and analyze, but it is also growing (see **FIGURE 5**, page 7).

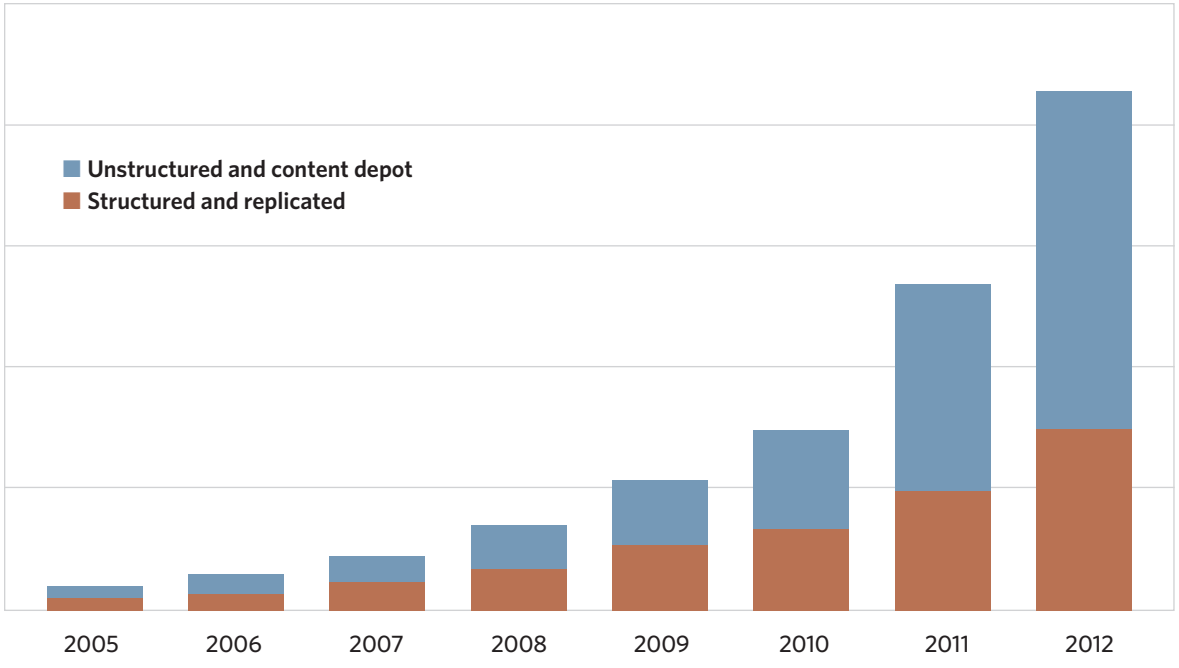
2. Technology advances. Hardware has finally caught up with software. The exponential gains in price-performance exhibited by computer processors, memory and disk storage have finally made it possible to store and analyze

large volumes of data at an affordable price. Database vendors have exploited these advances by developing new high-speed analytical platforms designed to accelerate query performance against large volumes of data, while the open source community has developed Hadoop, a distributed file management system designed to capture, store and analyze large volumes of Web log data, among other things. In other words, organizations are storing and analyzing more data because they can.

Organizations are storing and analyzing more data because they can.

3. Insourcing and outsourcing. Because of the complexity and cost of storing and analyzing Web traffic data, most organizations traditionally outsourced these functions to third-party service bureaus like Omniture. But as the size and importance of corporate e-commerce channels have increased, many are now eager to insource this data to gain greater insights about customers. For example,

FIGURE 5: ***Data growth***



SOURCE: IDC DIGITAL UNIVERSE 2009: WHITE PAPER, SPONSORED BY EMC, 2009.

automobile valuation company Kelley Blue Book is now collecting and storing Web traffic data in-house so it can combine that information with sales and other corporate data to better understand customer behavior, according to Dan Ingle, vice president of analytical insights and technology at the company. At the same time, virtualization technology is beginning to make it attractive for organizations to consider moving large-scale data processing outside their data center walls to private hosted networks or public clouds.

> ***"We are the beginning of an amazing world of data driven applications. It's up to us to shape the world."***
>
> ***—TIM O'REILLY,***
> *founder, O'Reilly Media*

4. Developers discover data. The biggest reason for the popularity of the term *big data* is that Web and application developers have discovered the value of building new data-intensive applications. To application developers, big data is new and exciting. Tim O'Reilly, founder of O'Reilly Media, a longtime high-tech luminary and open source proponent, speaking at Hadoop World in New York in November 2010, said: "We are the beginning of an amazing world of data-driven applications. It's up to us to shape the world." Of course, for those of us who have made their careers in the data world, the new era of "big data" is simply another step in the evolution of data management systems that support reporting and analysis applications. ■

Big Data Analytics: Deriving Value from Big Data

BIG DATA BY itself, regardless of the type, is worthless unless business users do something with it that delivers value to their organizations. That's where analytics comes in. Although organizations have always run reports against data warehouses, most haven't opened these repositories to ad hoc exploration. This is partly because analysis tools are too complex for the average user but also because the repositories often don't contain all the data needed by the power user. But this is changing.

- **BIG VS. SMALL DATA.** A valuable characteristic of "big" data is that it contains more patterns and interesting anomalies than "small" data. Thus, organizations can gain greater value by mining large data volumes than small ones. While users can detect the patterns in small data sets using simple statistical methods, ad hoc query and analysis tools or by eyeballing the data, they need sophisticated techniques to mine big data. Fortunately, these techniques and tools already exist thanks to companies such as SAS Institute and SPSS (now part of IBM) that ship analytical workbenches (i.e., data mining tools). These tools incorporate all kinds of analytical algorithms that have been developed and refined by academic and commercial researchers over the past 40 years.

- **REAL-TIME DATA.** Organizations that accumulate big data recognize quickly that they need to change the way they capture, transform and move data from a nightly batch process to a continuous process using micro batch loads or event-driven updates. This technical constraint pays big business dividends because it makes it possible to deliver critical information to users in near real time. In other words, big data fosters operational analytics by supporting just-in-time information delivery. The market today is witnessing a perfect storm with the convergence of big data, deep analytics and real-time information delivery.

■ **Complex analytics.** In addition, during the past 15 years, the "analytical IQ" of many organizations has evolved from reporting and dashboarding to lightweight analysis conducted with query and online analytical processing (OLAP) tools. Many organizations are now on the verge of upping their analytical IQ by implementing complex analytics against both structured and unstructured data. Complex analytics spans a vast array of techniques and applications. Traditional analytical workbenches from SAS and SPSS create mathematical models of historical data that can be used to predict future behavior. This type of predictive analytics can be used to do everything from delivering highly tailored cross-sell recommendations to predicting failure rates of aircraft engines. In addition, organizations are now applying a variety of complex analytics to Web, social media and other forms of complex structured data that are hard to do with traditional SQL-based tools, including path analysis, graph analysis, link analysis, fuzzy matching and so on.

> ***Analytics increases corporate intelligence. ... and is the only true source of sustainable advantage.***

Organizations are now recruiting analysts who know how to wield these analytical tools to unearth the hidden value in big data. They are hiring analytical modelers who know how to use data mining workbenches, as well as data scientists, application developers with process and data knowledge who write programming code to run against large Hadoop clusters.

■ **Sustainable advantage.** At the same time, executives have recognized the power of analytics to deliver a competitive advantage, thanks to the pioneering work of thought leaders such as Tom Davenport, who co-wrote the book *Competing on Analytics: The New Science of Winning*. In fact, forward-thinking executives recognize that analytics may be the only true source of sustainable advantage since it empowers employees at all levels of an organization with information to help them make smarter decisions. In essence, analytics increases corporate intelligence, which is something you can never package or systematize and competitors can't duplicate. In short, many organizations have laid the groundwork to reap the benefits of "big data analytics."

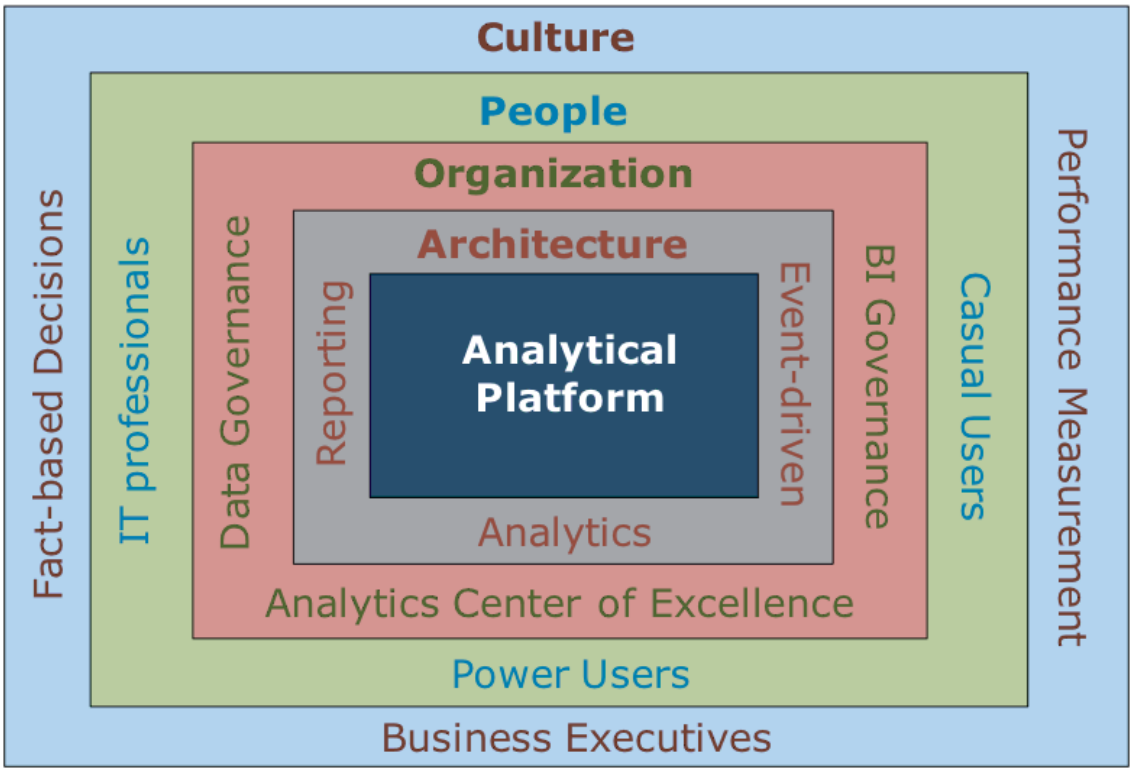
A FRAMEWORK FOR SUCCESS

However, the road to big data analytics is not easy and success is not guaranteed. Analytical champions are still rare. That's because succeeding with big data analytics requires the right culture, people, organization, architecture and technology (see **Figure 6**).

1. The right culture. Analytical organizations are championed by executives who believe in making fact-based decisions or validating intuition with data. These executives create a culture of performance measurement in which individuals and groups are held accountable for the outcomes of predefined metrics aligned with strategic objectives. These leaders recruit other executives who believe in the power of data and are willing to invest money and their own time to create a learning organization that runs by the numbers and uses analytical techniques to exploit big data.

2. The right people. You can't do big data analytics without power users, or

Figure 6: ***Big data analytics framework***



more specifically, business analysts, analytical modelers and data scientists. These folks possess a rare combination of skills and knowledge: They have a deep understanding of business processes and the data that sits behind those processes and are skillful in the use of various analytical tools, including Excel, SQL, analytical workbenches and coding languages. They are highly motivated, critical thinkers who command an above-average salary and exhibit a passion for success and deliver outsized value to the organization.

3. The right organization. Historically, analysts with the aforementioned skills were pooled in pockets of an organization hired by department heads. But analytical champions create a shared service organization (i.e., an analytical center of excellence) that makes analytics a pervasive competence. Analysts are still assigned to specific departments and processes, but they are also part of a central organization that provides collaboration, camaraderie and a career path for analysts. At the same time, the director maintains a close relationship with the data warehousing team (if he doesn't own the function outright) to ensure that business analysts have open access to the data they need to do their jobs. Data is fuel for a business analyst or data scientist.

4. The right architecture. The data warehousing team plays a critical role in delivering deep analytics. It needs to establish an architecture that ensures the delivery of high-quality, secure, consistent information while providing open access to those who need it. Threading this needle takes wisdom, a good deal of political astuteness and a BI-savvy data architecture team. The architecture itself must be able to consume large volumes of structured and unstructured data and make it available to different classes of users via a variety of tools (see "Architecture for Big Data Analytics" below).

5. Analytical platform. At the heart of an analytical infrastructure is an analytical platform, the underlying data management system that consumes, integrates and provides user access to information for reporting and analysis activities. Today, many vendors, including most of the sponsors of this report, provide specialized analytical platforms that provide dramatically better query performance than existing systems. There are many types of analytical platforms sold by dozens of vendors (see "Types of Analytical Platforms" below). ■

Architecture for Big Data Analytics

IF BIG DATA is simply a continuation of longstanding data trends, does it change the way organizations architect and deploy data warehousing environments? Big data analytics doesn't change data warehousing or BI architectures; it simply supplements them with new technologies and access methods better tailored to meeting the information requirements of business analysts and data scientists.

■ **TOP DOWN.** For the past 15 years, BI teams have built data warehouses that serve the information needs of casual users (e.g., executives, managers and front-line staff.) These top-down, report-driven environments require developers to know in advance what kinds of questions casual users want to ask and which metrics they want to monitor. With requirements in hand, developers create a data warehouse model, build extract, transform and load (ETL) routines to move data from source systems to the data warehouse, and then create reports and dashboards to query the data warehouse (see **FIGURE 7**, page 14).

Whether by choice or not, power users who operate in an exclusively top-down BI environment are largely left to fend for themselves, using spreadsheets, desktop databases, SQL and data-mining workbenches. Business analysts generally find BI tools too inflexible and data warehousing data too limited. At best, they might use BI tools as glorified extract engines to dump data into Microsoft Excel, Access or some other analytical environment. The upshot is that these analysts and data scientists generally spend an inordinate amount of time preparing data instead of analyzing it and create hundreds if not thousands of data silos that wreak havoc on information consistency from a corporate perspective.

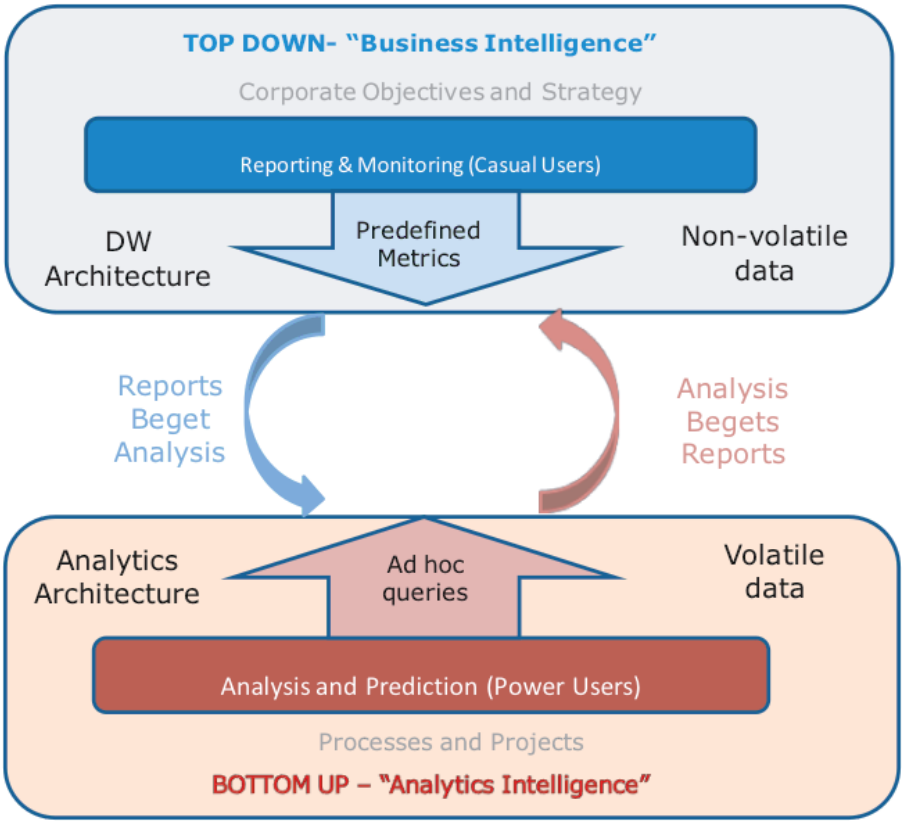
■ **BOTTOM-UP.** Business analysts and data scientists need a different type of analytical environment, one that caters to their needs. This is a bottom-up environment that fosters ad hoc exploration of any data source, both inside and outside corporate boundaries, and minimizes the need for analysts to create data silos. Here, business analysts don't know what questions they need to

answer in advance because they are usually responding to emergency requests from executives and managers who need information to address new and unanticipated events in the marketplace. Rather than focus on goals and metrics, business analysts spend most of their time engaged in ad hoc projects, or they work closely with business managers to optimize existing processes.

As you can see, there is a world of difference between a top-down and bottom-up BI environment. Many organizations have tried to support both types of processing within a single BI environment. But that no longer works in the age of big data analytics. Forward-thinking companies are expanding their data warehousing architectures and data governance programs to better balance the dynamic between top-down and bottom-up requirements. (See *Analytic Architectures: Approaches to Supporting Analytics Users and Workloads*, a 40-page report by Wayne Eckerson, available for free download.)

FIGURE 7: ***Top-down vs. bottom-up BI***

Top-down and bottom-up BI environments are distinct, but complementary, environments, but most organizations try to shoehorn both into a single architecture.

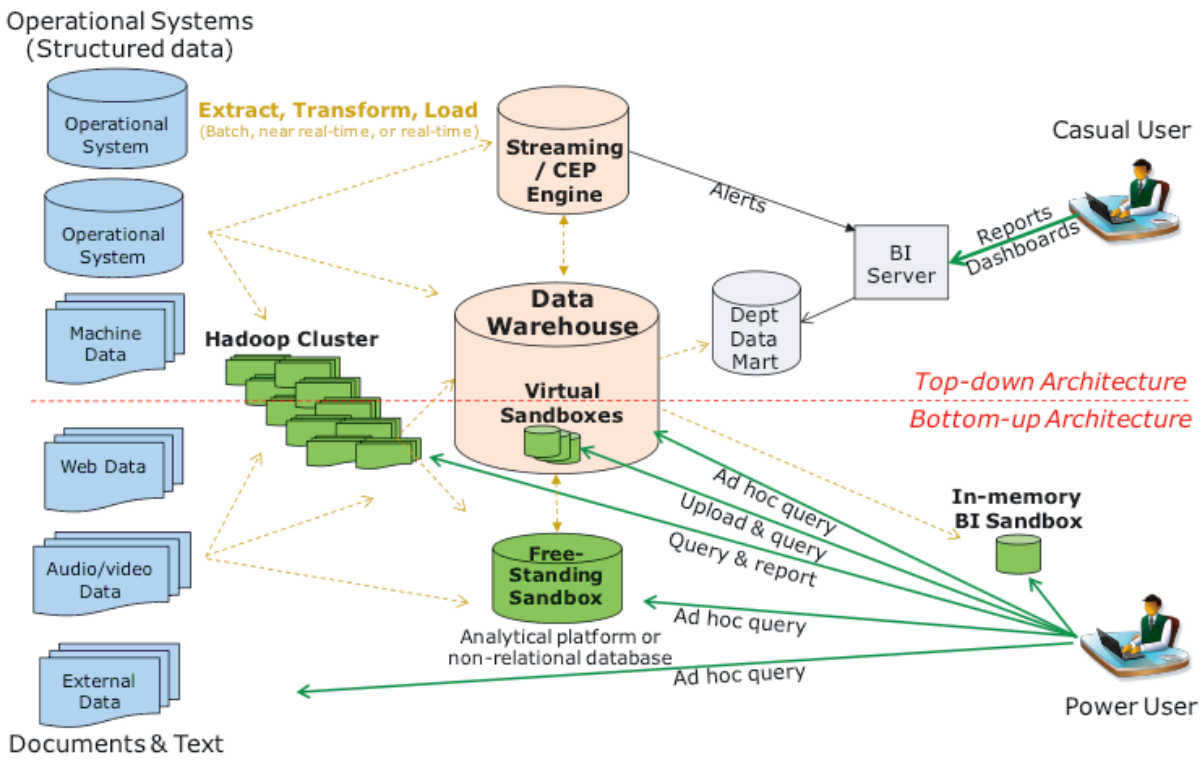


NEXT-GENERATION BI ARCHITECTURE

FIGURE 8 represents the next-generation BI architecture, which blends elements from top-down and bottom-up BI into a single cohesive environment that adequately supports both casual and power users. The top half of the diagram represents the classic top-down, data warehousing architecture that primarily delivers interactive reports and dashboards to casual users (although the streaming/complex event processing (CEP) engine is new.) The bottom half of the diagram adds new architectural elements and data sources that better accommodate the needs of business analysts and data scientists and make them full-fledged members of the corporate data environment.

FIGURE 8: *The new BI architecture*

The next-generation BI architecture is more analytical, giving power users greater options to access and mix corporate data with their own data via various types of analytical sandboxes. It also brings unstructured and semi-structured data fully into the mix using Hadoop and nonrelational databases.



SERVER ENVIRONMENT

■ **Hadoop.** The biggest change in the new BI architecture is that the data warehouse is no longer the centerpiece. It now shares the spotlight with systems that manage structured and unstructured data. The most popular among these is Hadoop, an open source software framework for building data-intensive applications. Following the example of Internet pioneers, such as Google, Amazon and Yahoo, many companies now use Hadoop to store, manage and process large volumes of Web data.

Hadoop runs on the Hadoop Distributed File System (HDFS), a distributed file system that scales out on commodity servers. Since Hadoop is file-based, developers don't need to create a data model to store or process data, which makes Hadoop ideal for managing semi-structured Web data, which comes in many shapes and sizes. But because it is "schema-less," Hadoop can be used to store and process any kind of data, including structured transactional data and unstructured audio and video data. However, the biggest advantage of Hadoop right now is that it's open source, which means that the up-front costs of implementing a system to process large volumes of data are lower than for commercial systems. However, Hadoop does require companies to purchase and manage dozens, if not hundreds, of servers and train developers and administrators to use this new technology.

The biggest change in the new BI architecture is that the data warehouse is no longer the centerpiece.

■ **Data warehouse integration.** Today some companies use Hadoop as a staging area for unstructured and semi-structured data (e.g., Web traffic) before loading it into a data warehouse. These companies keep the atomic data in Hadoop and push lightly summarized data sets to the data warehouse or nonrelational systems for reporting and analysis. However, some companies let power users with appropriate skills query raw data in Hadoop.

For example, LiveRail, an online video advertising service provider, follows the lead of most Internet providers and uses both Hadoop and a data warehouse to support a range of analytical needs. LiveRail keeps its raw Web data from video campaigns in Hadoop and summarized data about those campaigns in Infobright, a commercial, open source columnar database. Business

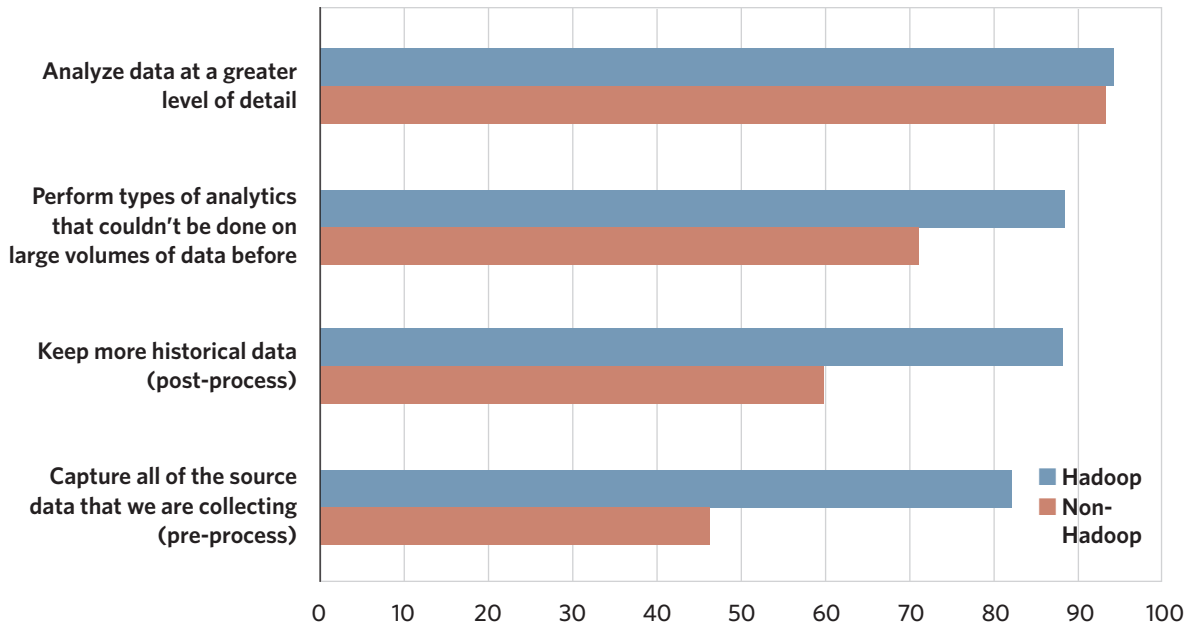
users, including LiveRail's customers who want to check on the performance of their campaigns, run ad hoc queries and reports against Infobright, while developers who need to access the raw data use Hive to schedule and run reports against Hadoop.

"Since Hadoop isn't interactive, we needed a fast database that scales to support our ad hoc environment, which is why we chose Infobright," said Andrei Dunca, chief technology officer at LiveRail.

■ **Use cases for Hadoop.** According to a new report by Ventana Research titled *Hadoop and Information Management: Benchmarking the Challenge of Enormous Volumes of Data*, Hadoop is more likely to be used than traditional data management systems for three purposes: to "perform types of analytics that couldn't be done on large volumes of data before," "capture all the source data we are collecting (pre-process)" and "keep more historical data (post-process)." (See **Figure 9**.)

Once data lands in Hadoop, whether it's Web data or not, organizations have several options:

Figure 9: ***Role of Hadoop***



SOURCE: *HADOOP AND INFORMATION MANAGEMENT: BENCHMARKING THE CHALLENGE OF ENORMOUS VOLUMES OF DATA: EXECUTIVE SUMMARY,* VENTANA RESEARCH, JUNE 23, 2011.

- **Create an online archive.** With Hadoop, organizations don't have to delete or ship the data to offline storage; they can keep it online indefinitely by adding commodity servers to meet storage and processing requirements. Hadoop becomes a low-cost alternative for meeting online archival requirements.

- **Feed the data warehouse.** Organizations can also use Hadoop to parse, integrate and aggregate large volumes of Web or other types of data and then ship it to the data warehouse, where both casual and power users can query and analyze the data using familiar BI tools. Here, Hadoop becomes an ETL tool for processing large volumes of Web data before it lands in the corporate data warehouse.

- **Support analytics.** The big data crowd (i.e., Internet developers) views Hadoop primarily as an analytical engine for running analytical computations against large volumes of data. To query Hadoop, analysts currently need to write programs in Java or other languages and understand MapReduce, a framework for writing distributed (or parallel) applications. The advantage here is that analysts aren't restricted by SQL when formulating queries. SQL does not support many types of analytics, especially those that involve inter-row calculations, which are common in Web traffic analysis. The disadvantage is that Hadoop is batch-oriented and not conducive to iterative querying.

- **Run reports.** Hadoop's batch-orientation, however, makes it suitable for executing regularly scheduled reports. Rather than running reports against summary data, organizations can now run them against raw data, guaranteeing the most accurate results.

■ **Nonrelational databases.** While Hadoop has received a lot of press attention lately, it's not the only game in town for storing and managing semi-structured data. In fact, an emerging and diverse set of products goes one step further than Hadoop and stores both structured and unstructured data within a single index. These so-called nonrelational databases (depicted in Figure 8 supporting a "free-standing sandbox") typically extract entities from documents, files and other databases using natural language processing techniques and index them as key value pairs for quick retrieval using a document-

centric query language such as XQuery. As a result, these products can give users one place to go to query both structured and unstructured data.

This style of analysis, which some call "unified information access," exhibits many search-like characteristics. But instead of returning a list of links, the systems return qualified data sets or reports in response to user queries. And unlike Hadoop, the systems are interactive, allowing users to submit queries in an iterative fashion so they can understand trends and issues.

These nonrelational systems complement Hadoop, an enterprise data warehouse or both. For example, organizations might use Hadoop to transcribe audio files and then load the transcriptions into a nonrelational database for analysis. Or they might replicate sales and customer data from a data warehouse and combine it with Web data in a nonrelational database so power users can find correlations between Web traffic and customer orders without bogging down performance of the data warehouse with complex queries. This type of unified information access is critical in a growing number of applications.

Nonrelational systems can store both structured and unstructured data within a single index, giving users one place to query any type of data.

For example, an oil and gas company uses MarkLogic to track the location of ships at sea. The MarkLogic Server stores data from GPS, news feeds, weather data, commodity prices, among other things, and surfaces all this data on a map that users can query. For example, a user might ask, "Show me all the ships within this polygon (i.e., geographic area) that are carrying this type of oil and have changed course since leaving the port of origin." The application then displays the results on the map.

■ **DATA WAREHOUSE HUBS.** While Hadoop and nonrelational systems primarily manage semi-structured and unstructured data, the data warehouse manages structured data from run-the-business operational systems. Except for Teradata shops, many companies increasingly use data warehouses running on traditional relational databases as hubs to feed other systems and applications rather than to host reporting and analysis applications.

For example, Dow Chemical, which maintains a large SAP Business Ware-

house (BW) data warehouse, now runs all queries against virtual cubes that run in memory using SAP BW Accelerator. "By running our cubes in memory, we've de-bottlenecked our data warehouse," said Mike Masciandaro, director of business intelligence at Dow. "Now our data warehouse primarily manages batch loads to stage data." Likewise, Blue Cross Blue Shield of Kansas City has transformed its IBM DB2 data warehouse into a hub that feeds transaction and analytical systems and implemented a Teradata Data Warehouse Appliance 2650 to handle all reports and queries and support a self-service BI environment.

Many companies increasingly use their data warehouse as hubs to feed other systems and applications rather than as targets for reporting and analysis applications.

■ **ANALYTICAL SANDBOXES.** In keeping with its role as a hub, an enterprise data warehouse at many organizations now distributes data to analytical sandboxes that are designed to wean business analysts and data scientists off data shadow systems and make them full-fledged consumers of the corporate data infrastructure. There are four types of analytical sandboxes:

› **Hadoop.** Hadoop can be considered an analytical sandbox for Web data that developers with appropriate skills can access to run complex queries and calculations. Rather than analyze summarized or transformed data in a data warehouse, developers can run calculations and models against the raw, atomic data.

› **Virtual DW sandbox.** A virtual data warehouse sandbox is a partition, or set of tables, inside the data warehouse, dedicated to individual analysts. Rather than create a spreadmart, analysts upload their data into a partition and combine it with data from the data warehouse that is either "pushed" to the partition by the BI team using ETL processes or "pulled" by analysts through queries. The BI team carefully allocates compute resources so analysts have enough horsepower to run ad hoc and complex queries without interfering with other workloads running on the data warehouse.

- **Free-standing sandbox.** A free-standing sandbox is a separate system from the data warehouse with its own server, storage and database that is designed to support complex, analytical queries. In some cases, it is an analytical platform that runs complex queries against a replica of the data warehouse. In other cases, it runs on a nonrelational database and contains an entirely new set of data (e.g., Web logs or sensor data) that either doesn't fit in the data warehouse because of space constraints or is processed more efficiently in a nontraditional platform. Occasionally, it mixes both corporate and departmental data into a local data mart that can run on-premises or off-site in a hosted environment. In all cases, it provides a dedicated environment for a targeted group of analytically minded users.

- **In-memory BI sandbox.** Some deskop BI tools, such as QlikView or PowerPivot, maintain a local data store in memory to support interactive dashboards or ad hoc queries. These sandboxes are popular among analysts, because they generally let them pull data from any source, quickly link data sets, run super-fast queries against data held in memory, and visually interact with the results, all without much or any IT intervention. Also, some server-based environments, such as SAP HANA, store all data in memory, accelerating queries for all types of BI users.

■ **Streaming/CEP Engine.** The top-down environment picks one important new architectural feature, streaming and CEP engines. Designed to support continuous intelligence, CEP engines are designed to ingest large volumes of discrete events in real-time, calculate or correlate those events, enrich them with historical data if needed, and apply rules that notify users when specific types of activity or anomalies occur. For example, these engines are ideal for detecting fraud in a stream of thousands of transactions per second.

These rules-driven systems are like intelligent sensors that organizations can attach to streams of transaction data to watch for meaningful combinations of events or trends. In essence, CEP systems are sophisticated notification systems designed to monitor real-time events. They are ideal for monitoring continuous operations, such as supply chains, transportation operations, factory floors, casinos, hospital emergency rooms, Web-based gaming systems and customer contact centers.

Streaming engines are similar to CEP engines but are designed to handle

enormous volumes of a single discrete event type, such as a sensor data generated by a pipeline or medical device. Streaming engines typically ingest an order of magnitude more events per second than CEP engines but typically only pull data from a single source. However, streaming and CEP engines are merging in functionality as vendors seek to offer one-stop shopping for continuous intelligence capabilities.

CLIENT ENVIRONMENT

■ **CASUAL USERS.** The front-end of the new BI architecture remains relatively unchanged for casual users, who continue to use reports and dashboards running against dependent data marts (either logical or physical) fed by a data warehouse. This environment typically meets 60% to 80% of their information needs, which can be defined up-front through requirements-gathering exercises. Predefined reports and dashboards are designed to answer questions tailored to individual roles within the organization.

However, meeting the ad hoc needs of casual users continues to be a problem. Interactive reports and dashboards help to some degree, but casual users today still rely on the IT department or "super users"—tech-savvy business colleagues—to create ad hoc reports and views on their behalf. Search-based exploration tools that allow users to type queries in plain English and refine their search using facets or categories offer significant promise but are not yet mainstream technology.

Search-based exploration tools that allow users to type queries in plain English and refine their search using facets or categories offer significant promise but are not yet mainstream technology.

One new addition to the casual user environment are dashboards powered by streaming/CEP engines. While these operational dashboards are primarily used by operational analysts and workers, many executives and managers are keen to keep their fingers on the pulse of their companies' core processes by accessing these "twinkling" dashboards directly or, more commonly, receiving alerts from these systems inside existing BI environments.

■ **Power users.** The biggest change in the new analytical BI architecture is how it accommodates the information needs of power users. It gives power users many new options for consuming corporate data rather than creating countless spreadmarts. A power user is a person whose job is to crunch data on a daily basis to generate insights and plans. Power users include business analysts (e.g., Excel jockeys), analytical modelers (e.g., SAS programmers and statisticians) and data scientists (e.g., application developers with business process and database expertise.) Power users have five options as depicted in **Figure 8**:

- **Query a virtual sandbox.** Rather than spend valuable time creating spreadmarts, power users can leverage the processing power and data of the data warehouse by using a virtual sandbox. Here, they can upload their own data to the data warehouse, mix it with corporate data and perform their analyses. However, if they want to share something they've built in their sandboxes, they need to hand over their analyses to the BI team to turn it into production applications.

- **Query a free-standing sandbox.** Power users can also query a free-standing sandbox created for their benefit. In most cases, the system is tuned to support ad hoc queries and analytical modeling activities against a replica of the data warehouse or another data set designed for power users.

- **Query a BI sandbox.** Power users can download data from a data warehouse or other source into a local BI tool and interact with the data in memory at the speed of thought. While these sandboxes have the potential to become spreadmarts, new analytical tools usually bake in server environments that encourage, if not require, power users to publish their analyses to an IT-controlled environment. Many also are starting to come up with collaboration capabilities that encourage reuse and minimize the proliferation of data silos.

- **Query the data warehouse.** Some BI teams give permission to a handful of trusted power users to directly query the data warehouse or DW staging area. This requires that analysts have a deep understanding of the raw data and advanced knowledge of SQL to avoid creating runaway queries or generating incorrect results.

› **Query Hadoop.** If power users want to analyze big data in its raw or lightly aggregated form, they can query Hadoop directly by writing MapReduce code in a variety of languages. However, power users must know how to write parallel queries and interrogate the structure of the data prior to querying it since Hadoop data is schema-less. Vendors are also beginning to ship BI tools that access Hadoop through Hive or Hbase and return data sets to the BI tool.

■ **Data integration.** The new BI architecture also places a premium on managing and manipulating data flows between systems. This calls for a versatile set of data integration tools that can access any type of data (e.g., structured, semi-structured and unstructured), load it into any target (e.g., Hadoop, data warehouse or in-memory database), navigate data sources that exist on-premises or in the cloud, work in batch and real time, and handle both small and large volumes of data.

Data integration tools for Hadoop are in their infancy but evolving fast. The open source community has developed Flume, a scalable distributed file system that collects, aggregates and loads data into the HDFS. But longtime data integration vendors, such as Informatica, are also converting their visual design tools to interoperate with Hadoop. That way, ETL developers can use familiar tools to extract, load, parse, integrate, cleanse and match data in Hadoop by generating MapReduce code under the covers.

Data integration products that run on both relational and nonrelational platforms and maintain a consistent set of metadata across both environments will reduce overall training and maintenance costs.

In this respect, Hadoop is both another data source for ETL tools as well as a new data processing engine geared to handling semi-structured and unstructured data. Data integration products that run on both relational and nonrelational platforms and maintain a consistent set of metadata across both environments will reduce overall training and maintenance costs.

■ **Analytical services.** Although it doesn't happen often, a growing number of

companies are outsourcing analytical applications to a third-party provider and occasionally an entire data warehouse. The most popular applications to outsource to a third party are test, development and prototyping applications. Often, IT administrators will provision servers on demand from a public cloud to support these types of applications. However, if an organization wants a permanent online presence to support an analytical sandbox or data warehouse, it will subscribe to a private hosted service, which provides higher levels of guaranteed availability and performance compared with a public cloud.

In either case, the motivation to implement an analytical service is straightforward: An analytical service requires minimal IT involvement and up-front capital, making it easy, quick and painless to get up and running quickly. Ironically, few companies think of outsourcing their analytical environments when exploring options.

■ **Dollar General.** That was the case with Dollar General, a discount retailer that wanted to purchase an analytical system to supplement its Oracle data warehouse, which could not store atomic-level point-of-sale (POS) data from its 9,500 stores nationwide. With a reference from a consumer products partner, Dollar General decided to implement an analytical platform from services provider, 1010data. The product, which is accessed via a Web browser, offers an Excel-like interface that provides native support for time-series data and analytical functions. Within five weeks, Dollar General was running daily reports against atomic-level POS data, according to Sandy Steier, executive vice president and co-founder of 1010data.

A year later, Dollar General decided to replace its Oracle data warehouse and conducted a proof of concept with several leading analytical platform providers. 1010data, which participated in the Bake-Off, demonstrated superior performance and now runs Dollar General's entire data warehouse.

And while Dollar General didn't set out to purchase an analytical service, that proved a smart move. Besides quick deployment times and reduced internal maintenance costs, the analytical service made it easier for Dollar General to open up its data warehouse to suppliers, which now use it to track sales and make recommendations for product placement and promotions, Steier said. ■

Platforms for Running Big Data Analytics

SINCE THE BEGINNING of the data warehousing movement in the early 1990s, organizations have used general-purpose data management systems to implement data warehouses and, occasionally, multidimensional databases (i.e., "cubes") to support subject-specific data marts, especially for financial analytics. General-purpose data management systems were designed for transaction processing (i.e., rapid, secure, synchronized updates against small data sets) and only later modified to handle analytical processing (i.e., complex queries against large data sets.) In contrast, analytical platforms focus entirely on analytical processing at the expense of transaction processing.[1]

Analytical platforms focus entirely on analytical processing at the expense of transaction processing.

■ **THE ANALYTICAL PLATFORM MOVEMENT.** In 2002, Netezza (now owned by IBM), introduced a specialized analytical appliance, a tightly integrated, hardware-software database management system designed explicitly to run ad hoc queries at blindingly fast speeds. Netezza's success spawned a host of competitors, and there are now more than two dozen players in the market. The value of this new analytical technology didn't escape the notice of the world's biggest software vendors, each of whom has made a major investment in the space, either through organic development or an acquisition (see **TABLE 1**).

To be accurate, Netezza wasn't the first mover in the market, but it came along at the right time. In the mid-2000s, many corporate BI teams were

[1]Like most things, there are exceptions to this rule. For example, Oracle Exadata runs on Oracle 10g and, as such, it supports both transactional and analytical processing, often with superior performance in both realms compared with standard Oracle 10g installations.

TABLE 1: Types of analytical platforms

(Companies in parentheses recently acquired the preceding product or company)

TECHNOLOGY	DESCRIPTION	VENDOR/PRODUCT
Massively parallel processing analytic databases	Row-based databases designed to scale out on a cluster of commodity servers and run complex queries in parallel against large volumes of data.	Teradata Active Data Warehouse, Greenplum (EMC), Microsoft Parallel Data Warehouse, Aster Data (Teradata), Kognitio, Dataupia
Columnar databases	Database management systems that store data in columns, not rows, and support high data compression ratios.	ParAccel, Infobright, Sand Technology, Sybase IQ (SAP), Vertica (Hewlett-Packard), 1010data, Exasol, Calpont
Analytical appliances	Preconfigured hardware-software systems designed for query processing and analytics that require little tuning.	Netezza (IBM), Teradata Appli-ances, Oracle Exadata, Greenplum Data Computing Appliance (EMC)
Analytical bundles	Predefined hardware and software configurations that are certified to meet specific performance criteria, but customers must purchase and configure themselves.	IBM SmartAnalytics, Microsoft FastTrack
In-memory databases	Systems that load data into memory to execute complex queries.	SAP HANA, Cognos TM1 (IBM), QlikView, Membase
Distributed file-based systems	Distributed file systems designed for storing, indexing, manipulating and querying large volumes of un-structured and semi-structured data.	Hadoop (Apache, Cloudera, MapR, IBM, HortonWorks), Apache Hive, Apache Pig
Analytical services	Analytical platforms delivered as hosted or public-cloud-based services.	1010data, Kognitio
Nonrelational	Nonrelational databases optimized for querying unstructured data as well as structured data.	MarkLogic Server, MongoDB, Splunk, Attivio, Endeca, Apache Cassandra, Apache Hbase
CEP/streaming engines	Ingest, filter, calculate and correlate large volumes of discrete events and apply rules that trigger alerts when conditions are met.	IBM, Tibco, Streambase, Sybase (Aleri), Opalma, Vitria, Informatica

struggling to deliver reasonable query performance using general-purpose data management systems as the volumes of data and numbers of users and applications running against their data warehouses exploded. Netezza offered a convenient way to offload long-running or complex queries from data warehouses and satisfy the needs of business analysts. Yet long before Netezza shipped its first box, in the 1980s, Teradata delivered the first massively parallel database management system geared to analytical processing, and Sybase shipped the first columnar database in the 1990s. Both of these products now have thousands of customers and, in this respect, can be considered front-runners in the so-called analytical platform market.

Today, the technology behind analytical platforms is diverse: appliances, columnar databases, in memory databases, massively parallel processing (MPP) databases, file-based systems, nonrelational databases and analytical services. What they all have in common, however, is that they provide significant improvements in price-performance, availability, load times and manageability compared with general-purpose relational database management systems. Every analytical platform customer I've interviewed has cited order-of-magnitude performance gains that most initially don't believe. "The performance is blinding; just amazing," says Masciandaro of Dow.

■ **Analytical Techniques.** Analytical platforms offer superior price-performance for many reasons. And while product architectures vary considerably, most support the following characteristics:

› **MPP.** Most analytical platforms spread data across multiple nodes, each containing its own CPU, memory and storage and connected to a high-speed backplane. When a user submits a query or runs an application, the "shared nothing" system divides the work across the nodes, each of which process the query on its piece of the data and ship the results to a master node that assembles the final result and sends it to the user. MPP systems are highly scalable, since you simply add nodes to increase processing power. And if the nodes run on commodity servers, as many MPP systems today do, then this scalability is more cost-effective than MPP systems running on proprietary hardware or symmetric multiprocessing systems, which require big, expensive boxes to scale.

› **Balanced configurations.** Analytical platforms optimize the configuration

of CPU, memory and disk for query processing rather than transaction processing. Analytical appliances essentially "hard wire" this configuration into the system and don't let customers change it, whereas analytical bundles or analytical databases (i.e., software-only solutions) allow customers to configure the underlying hardware to match unique application requirements. Analytical appliances offer convenience and ease of use while analytical databases offer flexibility.

- **Storage-level processing.** Netezza's big innovation was to move some database functions, specifically data filtering functions, into the storage system using field programmable gate arrays. This storage-level filtering reduces the amount of data that the DBMS has to process, which significantly increases query performance. Many vendors have followed suit, moving various databases functions into hardware. In fact, Kickfire (purchased by Teradata) runs all SQL functions on a chip to accelerate query processing.

- **Columnar storage and compression.** Many vendors have followed the lead of Sybase, Sand Technology, ParAccel, and other columnar pioneers, by storing data in columns, not rows. Since most queries ask for a subset of columns in a row (i.e., the "where" clause) rather than all rows, storing data in columns minimizes the amount of data that needs to be retrieved from disk and processed by the database, accelerating query performance. In addition, since data elements in many columns are repeated (e.g., "male" and "female" in the gender field), column-store systems can eliminate duplicates and compress data volumes significantly, sometimes as much as 10:1. This enables more data to fit into memory, which speeds processing and minimizes the amount of disk required to store data, making the systems more cost-effective.

- **Memory.** Many analytical platforms make liberal use of memory caches to speed query processing. Some products, such as SAP HANA and QlikTech's QlikView, store all data in-memory, while others store recently queried results in a smart cache so others who need to retrieve the same data can pull it from memory rather than from disk. Given the growing affordability of memory and the widespread deployment of 64-bit operating systems, which lift constraints on the amount of data that can be held

in memory, many analytical platforms are expanding their memory footprints to speed query processing.

- **Query optimizer.** Analytical platform vendors invest a lot of time and money researching ways to enhance their query optimizers to handle various workloads. A good query optimizer is perhaps the biggest contributor to query performance. In this respect, the older vendors with established products have an edge.

- **Plug-in analytics.** True to their name, many analytical platforms offer built-in support for complex analytics. This includes complex SQL, such as correlated subqueries, as well as procedural code implemented as plug-ins to the database. Some vendors offer a library of analytical routines, from fuzzy matching algorithms to market-basket calculations. Some, like Aster Data (now owned by Teradata), provide native support for MapReduce programs that are called using SQL.

■ **Hadoop and NoSQL.** Some may argue whether Hadoop and the nonrelational databases are analytical platforms. While they don't store data in rows and columns, both are well-suited to process large volumes of data for analytical purposes. And most use an MPP architecture that scales out on commodity servers. And some, such as MarkLogic, are full-fledged databases that support transactional integrity.

Hadoop in particular differs significantly from most analytical platforms. As a batch system, it's not focused on optimizing query performance like other analytical platforms, and thus, does not implement many of the characteristics in the above bulleted list. However, Hadoop's biggest value is that it's open source and so can process large volumes of data in a cost-effective way. And like many nonrelational systems, it is schema-less, giving administrators greater flexibility to change data structures without having to spend weeks or months rewriting a data model. ■

Profiling the Use of Analytical Platforms

NOW THAT WE understand the business context for analytical platforms, the technical architecture in which they run and their technical characteristics, we can profile their use in user organizations. To do this, I conducted a survey of BI professionals and asked them to describe their use of analytical platforms from a business and technical perspective. The survey provided respondents with the following definition of an analytical platform:

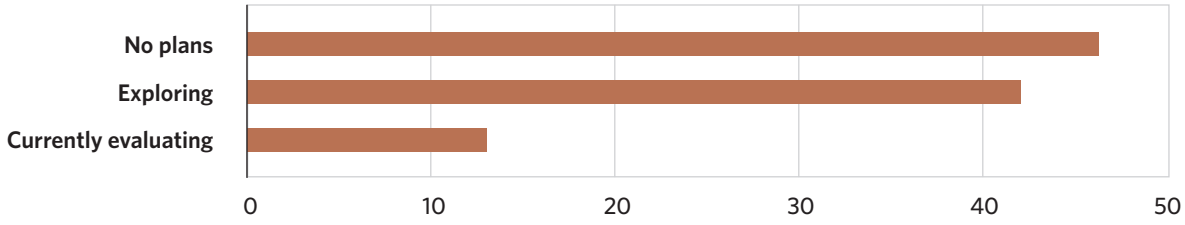
> *An analytical platform is a data management system optimized for query processing and analytics that provides superior price-performance and availability compared with general purpose database management systems.*

Given this definition, almost three-quarters (72%) of our survey respondents said that they had purchased or implemented an analytical database.

While the growth of the analytical platform market has been strong, this 72% figure seems a tad high, given that a majority of analytical database products have been on the market for less than five years. Upon closer investigation, despite our definition, a sizable number of respondents when asked to name their analytical platform identified a general-purpose database, in particular Microsoft SQL Server and Oracle (non-Exadata). Regardless, the data still shows that many companies are turning to specialized analytical platforms to better meet their analytical requirements.

■ **Non-customers.** Among respondents that haven't purchased an analytical platform, 46% have no plans to do so, 42% are exploring the idea and just 12% are currently evaluating vendors. On the whole, about 75% of respondents will have an analytical platform in the near future (see **Figure 10**, page 33).

FIGURE 10: ***Do you plan to purchase or implement an analytical platform?***
(Asked of respondents who don't yet have an analytical platform)



DEPLOYMENT OPTIONS

Our survey grouped analytical platforms into four major categories to make it easier to compare and contrast various product offerings:

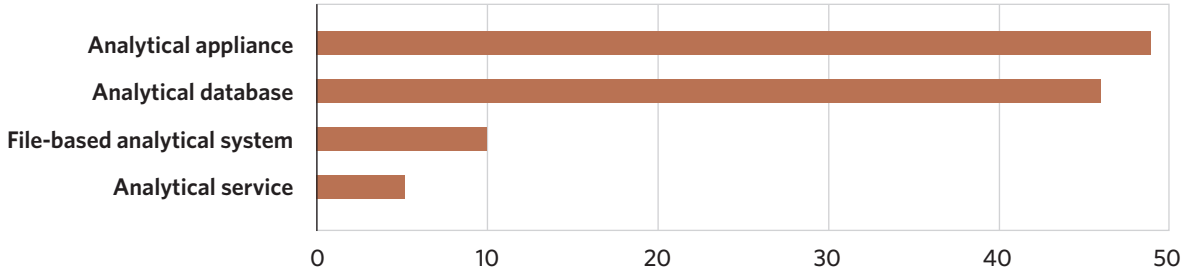
1. Analytical databases: They can be described as software-only analytical platforms that run on a variety of hardware that customers purchase. Customers install, configure and tune software, including the analytical database, before they can use the analytical system. Most MPP analytical databases, columnar databases and in-memory databases listed in **TABLE 1** qualify as analytical databases.

2. Analytical appliances: These are hardware-software combinations designed to support ad hoc queries and other types of analytical processing. Analytical appliances tightly integrate the hardware and software, often using proprietary components, to optimize performance and minimize the need for tuning. Analytical bundles, which consist of standalone hardware and software products that a vendor ships as a package, also qualify as analytical appliances. Bundles give administrators more flexibility to tune the system but sacrifice deployment speed and manageability.

3. Analytical services: Rather than deploy an analytical platform in a customer's data center, an analytical service enables customers to house the system in an off-site hosted environment or public cloud. This eliminates up-front capital expenditures and lessens maintenance.

4. File-based analytical system: This generally refers to Hadoop, but we also

FIGURE 11: ***Analytical platform deployment options: Which have you purchased and implemented?***



lumped NoSQL or nonrelational systems into this category, although it's not entirely accurate, since nonrelational systems are databases. However, since both are used to store and analyze large volumes of unstructured data and don't' require an up-front schema design, they share more similarities than differences.

Given these categories, most analytical platform customers have either purchased or implemented analytical databases (46%) or analytical appliances (49%). Many fewer have implemented a file-based analytical system (10%) or analytical service (5%). (See **FIGURE 11**.)

Looking under the covers, analytical database customers are most likely to have purchased Microsoft SQL Server or Oracle, while appliance customers have purchased Teradata Active DW, a Teradata Appliance, or Netezza. Analytical services customers subscribed to a host of different vendors, while customers of file-based analytical systems were most likely to purchase a Hadoop distribution from Cloudera, Apache or EMC Greenplum.

DEPLOYMENT STATUS

Drilling into each category further, we find that most of the respondents who have purchased an analytical platform of some type have also deployed the system. Roughly three-quarters of customers with analytical databases (73%) and slightly more customers of analytical appliances (80%) have deployed their systems. Not surprisingly, 100% of analytical services customers have deployed their systems, but only 33% of customers with file-based analytical systems have implemented theirs (see **TABLE 2**, page 34).

TABLE 2: Status of deployment options

	ANALYTICAL DATABASE	ANALYTICAL APPLIANCE	ANALYTICAL SERVICE	FILE-BASED ANALYTICAL SYSTEM
Percentage deployed	72%	81%	100%	33%
Average years deployed	4.0	4.9	3	1.3

With an analytical service, you simply create a data model (which in many cases is optional) and load your data either by using the Internet or shipping a disk to the provider, and the provider takes care of the rest. Thus, it's much easier to deploy an analytical service than the other options, accounting for the 100% deployment figure in **TABLE 2**.

Analytical appliances generally take less time to deploy than analytical databases, which may account for the slightly higher deployment percentage. Analytical databases require customers to purchase and install hardware, which may take many months and require multiple sign-offs from the IT, legal and purchasing departments. Despite overwhelming press coverage of Hadoop, few companies have implemented the system. Among those that have, most are largely experimenting, which explains the low deployment percentage compared with the other options.

The figures for "average years deployed" tell a similar story. As the new kid on the block, Hadoop systems have only been deployed for an average of 1.3 years, followed by analytical services, which have been deployed an average of three years. In contrast, analytical appliances have been deployed for 4.9 years and analytical databases for 4.0 years.

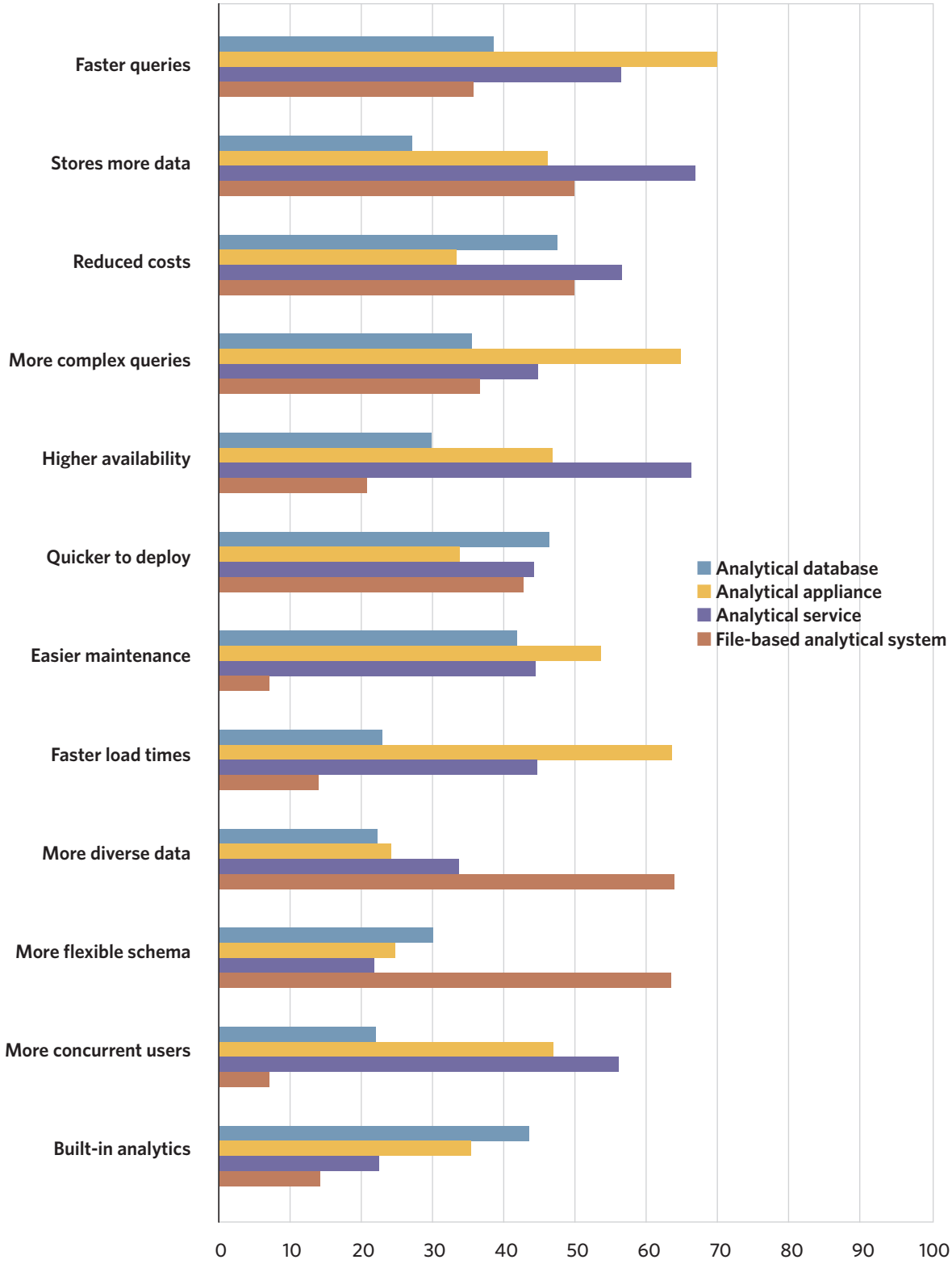
TECHNICAL DRIVERS

When examining the business requirements driving purchases of analytical platforms overall, three percolate to the top: "faster queries," "storing more data" and "reduced costs." These requirements are followed by "more complex queries," "higher availability" and "quicker to deploy." This ranking is based on summing the percentages of all four deployment options for each requirement (see **FIGURE 12**, page 35).

More important, Figure 12 shows that customers purchase each deployment

FIGURE 12: ***Business requirements by deployment option***

(Sorted from most to least for the percentage total of all four deployment options)



option for different reasons. Analytical database customers value "quicker to deploy" (46%), "built-in analytics" (43%) and "easier maintenance" (41%) more than other requirements, while analytical service customers favor "storing more data" (67%), "high availability" (67%), "reduced costs" (56%) and "more concurrent users" (56%). Not surprisingly, customers with file-based systems look for the ability to support "more diverse data" (64%) and "more flexible schemas" (64%), two hallmarks of a Hadoop/NoSQL offering.

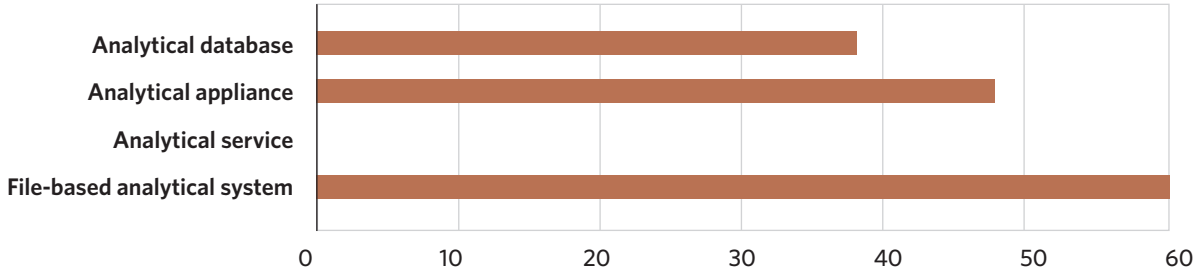
Analytical appliance customers had the most emphatic requirements. Almost two-thirds value faster queries (70%), more complex queries (64%) and faster load times (63%), suggesting that analytical appliance customers seek to offload complex ad hoc queries from data warehouses.

Analytical appliance customers had the most emphatic requirements. Almost two-thirds value faster queries, more complex queries and faster load times.

This is exactly the reason that Blue Cross Blue Shield of Kansas City purchased a Teradata Data Warehouse Appliance 2650 and MicroStrategy analysis tools. The company plans to push about 1 TB of data from its IBM DB2 data warehouse to the Teradata appliance to support a self-service BI environment for about 300 users. "We want executives and managers to be able to get data and make decisions without depending on the IT department," says Darren Taylor, vice president of enterprise analytics and data management. Taylor said the key requirements for the system were query performance and the ability to support complex analytical models and advanced visualization techniques, which will be embedded in the self-service solution.

Many companies also offload analytical processing to analytical databases. For example, a large U.S. retailer recently offloaded complex analytical queries from its maxed-out Teradata data warehouse to ParAccel, a high-performance columnar database. The company chose a software-only system so it could implement the database in a private cloud and spawn new instances in response to user demand, a key requirement that an analytical appliance does not support. The customer also implemented a direct connection between the two systems using Teradata's parallel FastExport wire protocol, eliminating the need for the customer to expand its ETL footprint, saving considerable time

FIGURE 13: ***Were you explicitly looking for [this deployment option]?***
(Percentages based on respondents who answered "Yes")



and money.

"Agility and interoperability with existing technologies were key drivers for the customer," said Rick Glick, vice president of customer and partner development at ParAccel.

■ **SELECTION BY CATEGORY.** We also asked respondents if they were looking for a specific deployment option when evaluating products (see **FIGURE 13**). Except for customers of file-based systems, most customers investigated products irrespective of its technology category. For example, Blue Cross Blue Shield of Kansas City looked at three columnar databases (i.e., software-only) and an appliance before making a decision. Interestingly, no analytical service customers intended to subscribe to a service prior to evaluating products. That's because many analytical service customers subscribe to such services on a temporary basis, either to test or prototype a system or to wait until the IT department readies the hardware to house the system. Some of these customers continue with the services, recognizing that they provide a more cost-effective test and development environment than an in-house system.

BUSINESS APPLICATIONS

When push comes to shove, the value of an analytical platform is judged not by its technical merits, but by the business applications it supports or makes possible. The most popular business applications running on analytical platforms are customer analytics, followed by management reports, financial analytics, data integration, executive dashboards and risk analytics. This ranking is

based on summing the percentages of all four deployment options for each requirement (see **Figure 14**).

Figure 14: ***Business applications by deployment option***

(Sorted from most to least for the percentage total of all four deployment options)

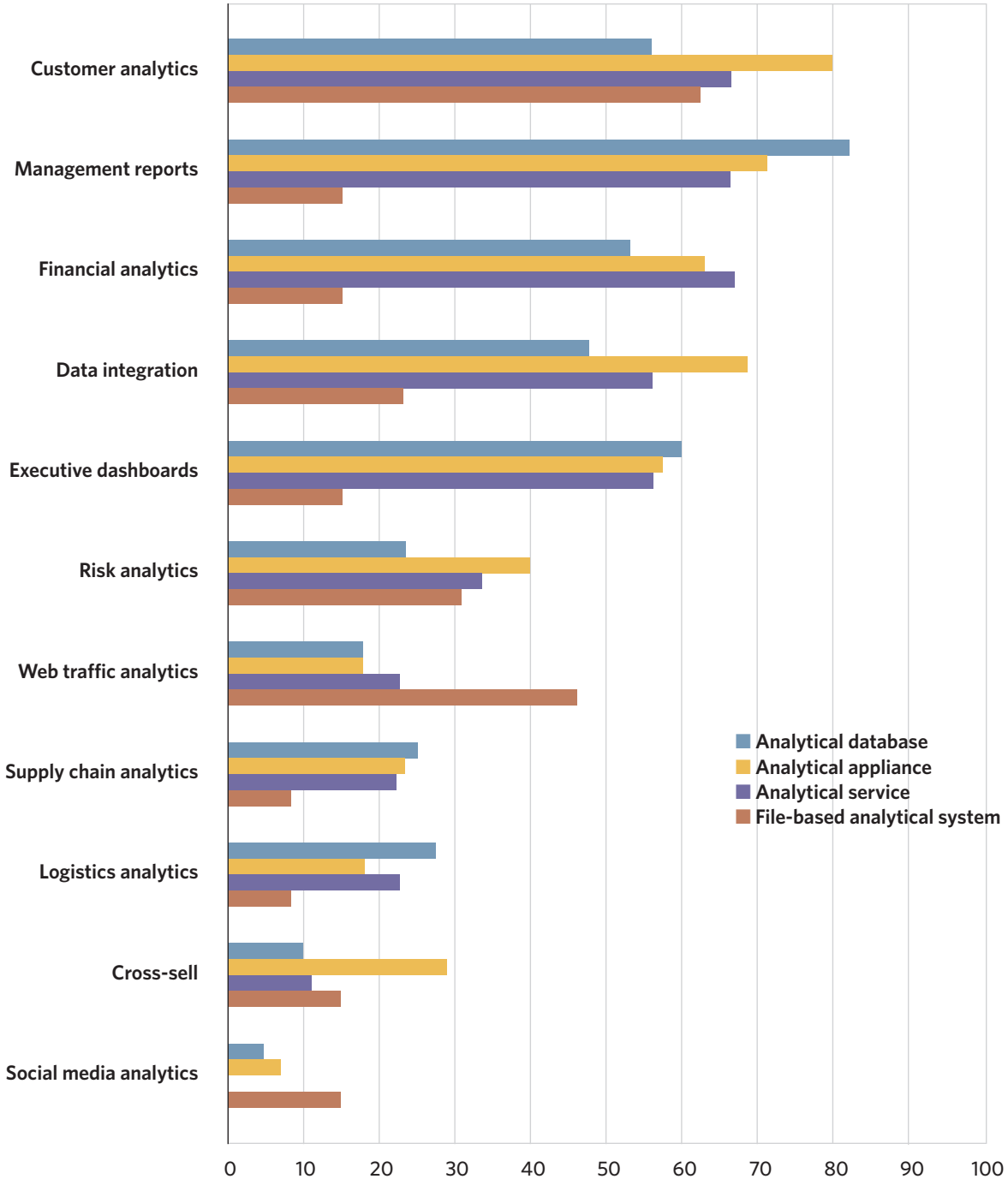


FIGURE 15: ***ROI by deployment option***

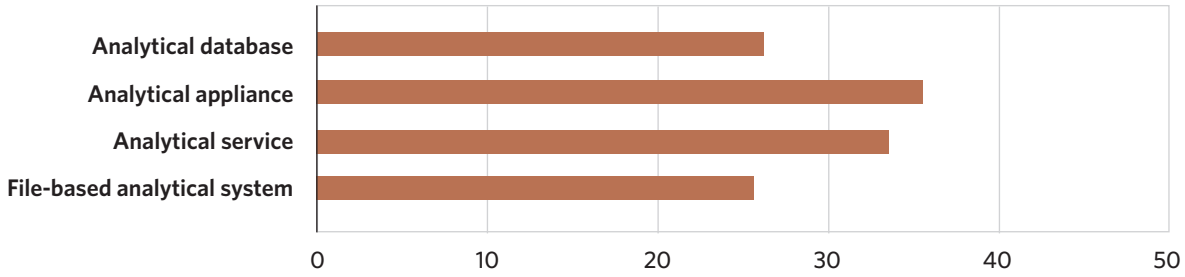


Figure 14 also exposes stark differences in the business applications supported by each deployment option. For example, an analytical appliance is more likely to be used for customer analytics (80%), risk analytics (40%), and cross-sell recommendations (29%) than analytical databases, which are more likely to be used for management reports (82%) and executive dashboards (60%). Thus, analytical databases are more likely to be used for traditional top-down reporting, while analytical appliances are used for bottom-up analytics. This contrast makes sense when you remember that many of our analytical database users are customers of Microsoft SQL Server and Oracle 10, which are best-suited to reporting, not analytics. **FIGURE 14** also shows that file-based systems are twice as likely to be used for Web traffic analysis (46%) and social media analysis (15%) than the other options.

■ **ROI.** Not surprisingly, given its emphasis on analytics versus reporting, analytical appliances (35%) have a higher ROI than analytical databases (26%), with analytical services close behind at 33%. Given their newness, file-based systems delivered a surprisingly strong 25% ROI, but that's probably because most file-based systems are open source and don't require an up-front investment in software (see **FIGURE 15**).

TECHNICAL ATTRIBUTES

■ **Applications and users.** When examining business attributes of each deployment option, it's clear that analytical appliances support far more applications and users than analytical services, analytical databases or file-based systems. The analytical appliance figure is perhaps skewed by the high

TABLE 3: Applications and users

	ANALYTICAL DATABASE	ANALYTICAL APPLIANCE	ANALYTICAL SERVICE	FILE-BASED ANALYTICAL SYSTEM
Average number of applications	5.9	11.3	8.0	4.9
Average number of concurrent users	47.1	81.4	27.5	27.8

number of Teradata Active DW customers who responded to the survey. Teradata Active EDW is geared to supporting multiple workloads, serving as a data warehouse, data mart and operational data store. In addition, its customers have used the product for many years, and the longer a product is used, the more applications it tends to support (see **TABLE 3**).

■ **Data volumes.** Analytical appliances and file-based systems are neck and neck in terms of the amount of data they store. More than 40% of both sets of customers use the systems to store between 10 TB and 100 TB of data, and more than 14% of both options store over 100 TB. In contrast, 40% of analytical database customers have less than 1 TB of data (see **FIGURE 16**, page 41).

Analytical appliances and file-based systems areneck and neck in terms of the amount of data they store.

■ **Types of data.** Not surprisingly, the analytical database and analytical appliance, both of which rely on relational technology, primarily hold structured data (90% and 95% respectively). In contrast, analytical services and file-based analytical systems hold a more balanced mix of data types. More than three-quarters (78%) of analytical services customers manage structured data, while 67% manage semi-structured data and 33% unstructured data. In contrast, file-based systems have more semi-structured data (73%) than either structured (67%) or unstructured (33%). This high percentage reflects the trend of companies insourcing Web data from service bureaus so they can

combine it with other corporate data, such as sales and orders, and derive more value from it (see **FIGURE 17**).

FIGURE 16: ***Volume of raw data***

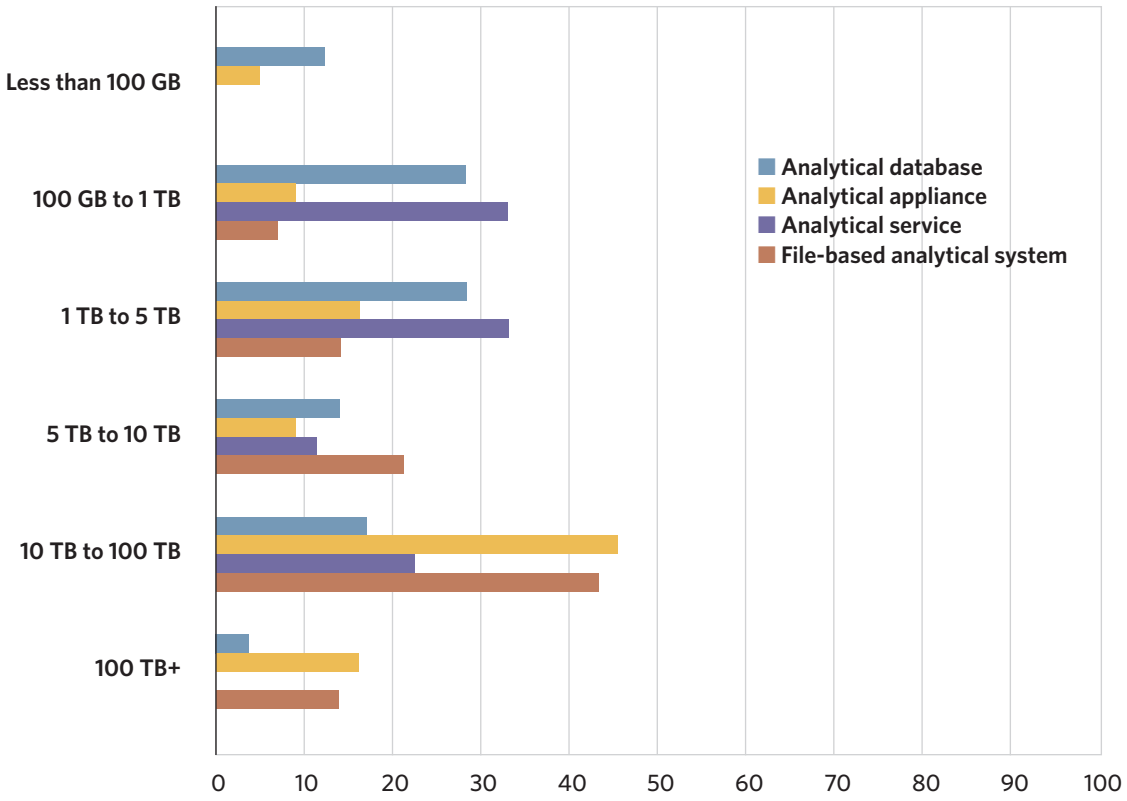
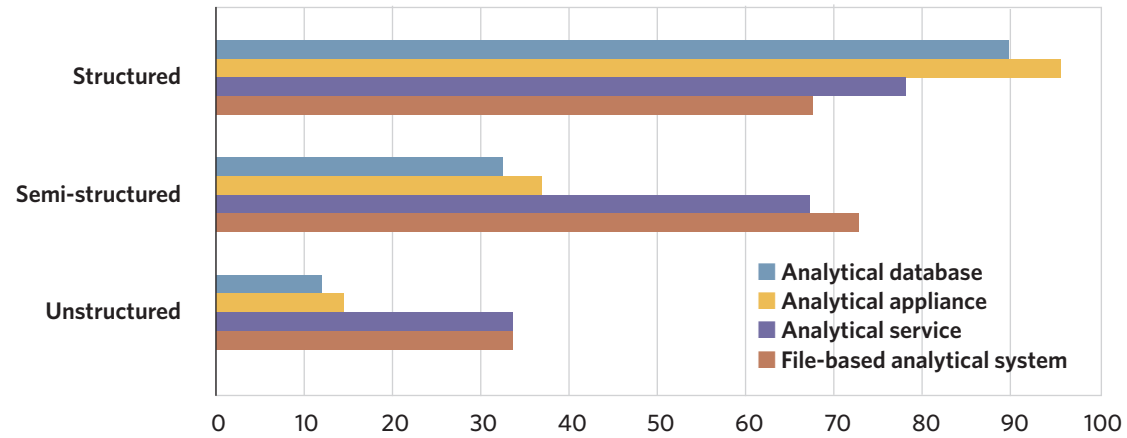


FIGURE 17: ***Types of data***

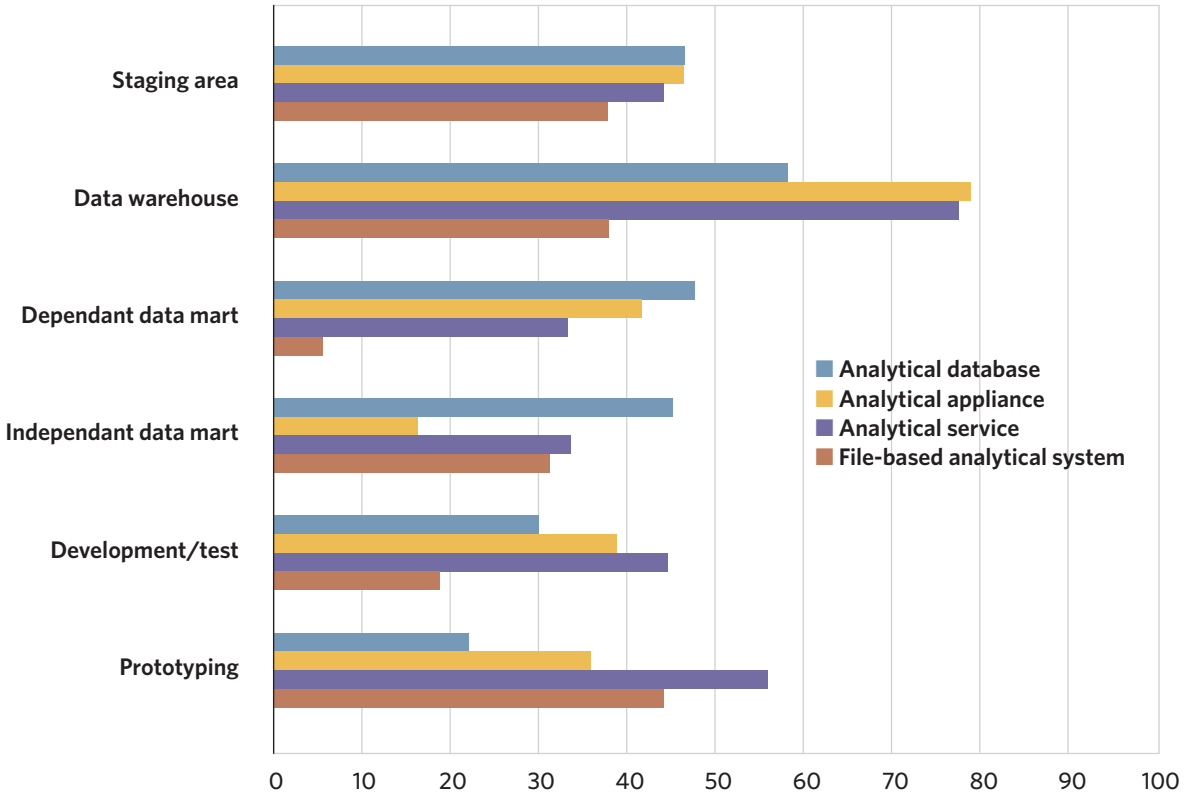


ARCHITECTURE

Architectural roles are fairly consistent across deployment options. The noticeable exception is that analytical appliances and analytical services are much more likely to be used for data warehouses than the other options. Also, analytical services and file-based systems are more likely to be used for prototyping than other systems, and analytical databases are more likely to be used as independent data marts (see **FIGURE 18**).

In addition, the most prominent use of file-based systems is for "prototyping" (44%), followed by "staging area" (38%) and "data warehouse" (38%). Currently, Hadoop is in its early days, and many companies are experimenting with the technology, which explains the high percentage of prototyping applications. But it's often used to stage and process Web traffic so companies can summarize and transfer the data into the data warehouse for analysis. But some companies aren't moving this data into data warehouses; they are simply leaving it in Hadoop and allowing data scientists to query this "data

FIGURE 18: ***Architecture by deployment option***



warehouse" of semi-structured or unstructured data.

I'm surprised by the 79% of analytical appliance customers who are using the system as a data warehouse. I believe this reflects the large percentage of Teradata Active DW customers who took the survey. But it's not dissimilar from 2010 survey results that showed that 68% of companies were using analytical platforms in general to power their data warehouses. What I've discovered anecdotally is that companies that power their data warehouse with Microsoft SQL Server are more likely to replace the product with an analytical platform than augment it. Conversely, companies with scalable data warehouse RDBMSs are more likely to augment the RDBMS with an analytical platform than replace it.

I'm also surprised that analytical databases are the leading platform for dependent and independent data marts, with 47% and 45% of customers selecting these architectural roles respectively. This undoubtedly reflects the large number of Microsoft SQL Server customers who took the survey, but it's also not too far out of line with our 2010 survey results.

Companies that power their data warehouse with Microsoft SQL Server are more likely to replace the product with an analytical platform than augment it. Conversely, companies with scalable data warehouse RDBMSs are more likely to augment the RDBMS with an analytical platform than replace it.

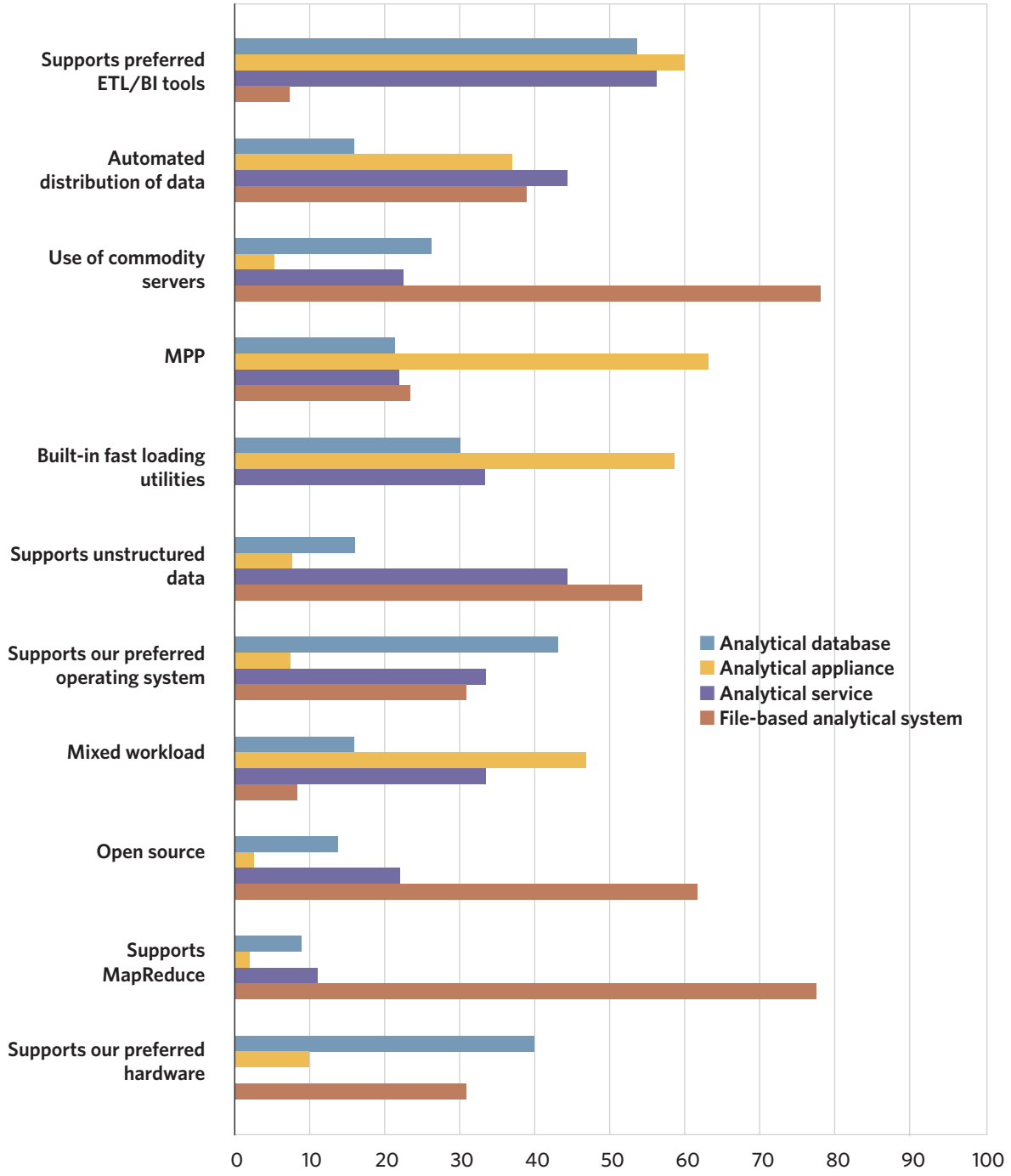
TECHNICAL REQUIREMENTS

The technical requirements for selecting products varied widely by deployment option. **Figure 19** (see page 44) ranks the requirements by sum of the percentages across all four deployment options. This shows that the top technical requirements are "supports our preferred ETL/BI tools," "automated distribution of data" and "use of commodity servers." This is followed by "MPP," "built-in fast loading," "supports unstructured data," "supports our preferred operating system" and "mixed workload."

What's striking is the variation in support for these requirements by deployment option. For example, interoperability with existing BI and ETL tools is a

critical requirement for all options except the file-based system. This makes sense, since most Hadoop developers write custom code in Java, Perl or some other language to construct queries rather than use packaged BI tools.

FIGURE 19: ***Technical requirements***



However, BI and ETL vendors are extending their products to interoperate with Hadoop, so this will undoubtedly change, since it's often easier to use tools than write code.

Another variation is that file-based customers are much more interested in "commodity servers," "open source" and "MapReduce" than customers of other deployment options. This makes sense, since all three requirements are critical elements of a Hadoop ecosystem. In contrast, analytical appliances are concerned with MPP, fast-loading utilities and mixed workload functionality. This aligns with the predominant needs of Teradata Active DW customers, who constituted a large portion of the analytical appliance respondents.

It came as no surprise that support and service are more critical for customers of analytical services.

VENDORS

When asked why they selected their chosen vendors, respondents were mostly likely to say, "Met more of our requirements," followed by "successful proof of concept," "liked and trusted the vendor" and "support and service." Interestingly, pricing was ranked fifth on the list, followed by customer references and vendor incumbency (see **FIGURE 20**, page 46).

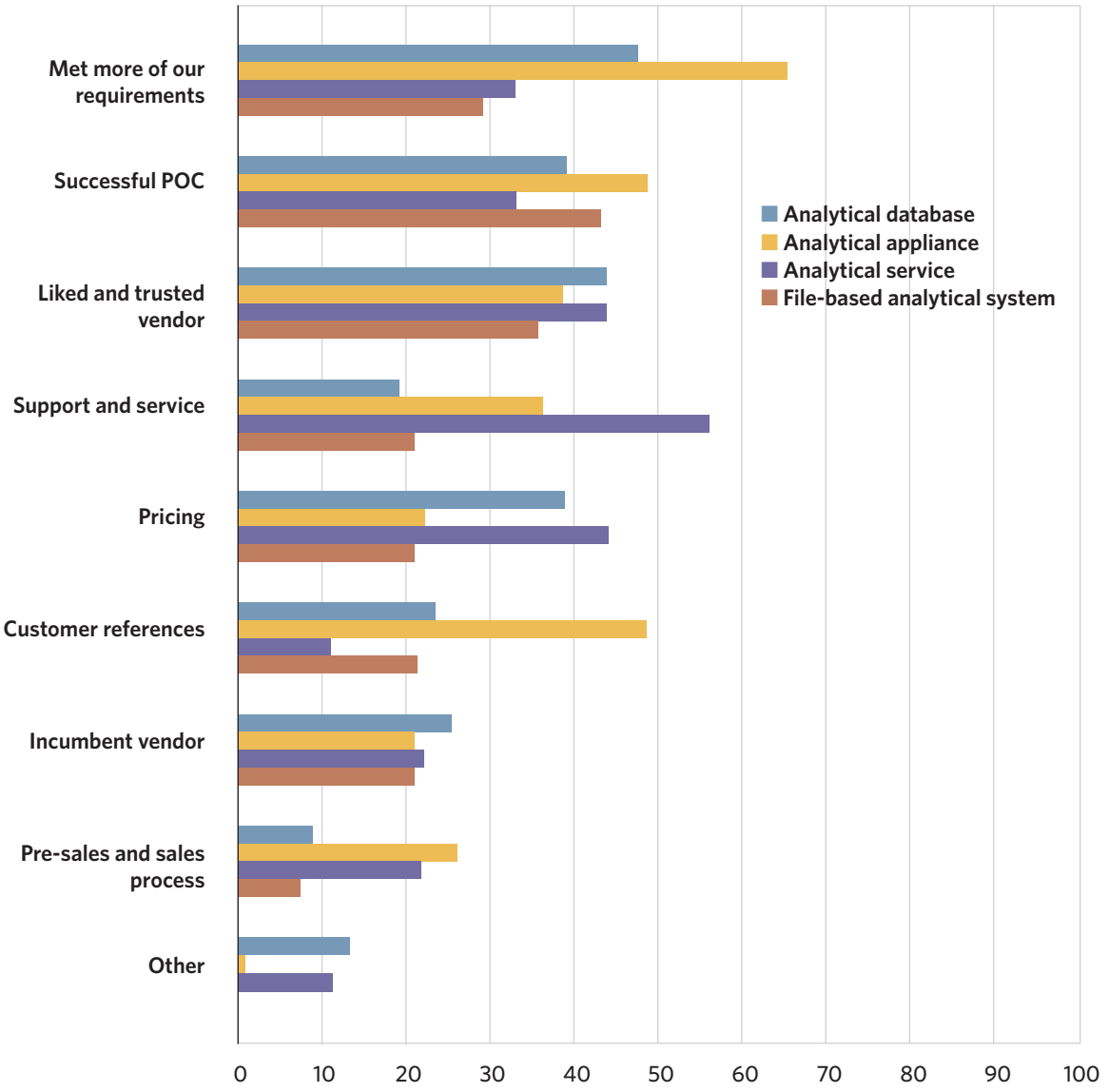
A vendor's ability to meet more of a customer's requirements was more important for analytical appliance customers than customers of other deployment options. This is for two reasons: (1) Many analytical appliances are new products from startup vendors (except Teradata Active DW), so customers need to make doubly sure that the product meets their requirements, since these vendors have less of a track record and (2) the customer is spending a significant amount of money on the product, and it's playing a central role in the data warehousing architecture. (This is also true for Teradata Active DW customers.)

It came as no surprise that support and service are more critical for customers of analytical services, who are totally dependent on the quality and responsiveness of the vendor to meet their needs. Pricing is also clearly a bigger issue for analytical service customers than others, since price (or more likely lack of an up-front capital investment) is a major inducement to hand

over responsibility for corporate data to a third party.

■ **Incumbency.** Interestingly, incumbency can cut both ways. Blue Cross Blue Shield of Kansas City decided not to examine an appliance product from its incumbent data warehouse vendor because it didn't want to expand its relationship with that vendor. At the same time, it considered a columnar database from Sybase because it had an existing relationship with the vendor in another area of the business. ■

Figure 20: ***Vendor selection criteria***



Recommendations

TO ADDRESS THE information needs of the modern corporation, organizations should take the following steps:

1 **Support both top-down and bottom-up business requirements.** For too long, organizations have tried to shoehorn all types of users into a single information architecture. That has never worked. Organizations need to recognize that casual users, who represent a majority of their business users, primarily need top-down, interactive reports and dashboards, while power users need ad hoc exploratory tools and environments. Balancing these polar opposite requirements in a single architecture requires new thinking.

2 **Implement a new BI architecture.** The BI architecture of the future incorporates traditional data warehousing technologies to handle detailed transactional data and file-based and nonrelational systems to handle unstructured and semi-structured data. The key is to integrate these systems into a unified architecture that enables casual and power users to query, report and analyze any type of data in a relatively seamless manner. This unified information access is the hallmark of the next generation BI architecture. More immediately, companies are using Hadoop to preprocess unstructured data so that it can be loaded and integrated with other corporate data for reporting and analysis. This allows BI and ETL users to use familiar tools to query and analyze data.

3 **Create analytical sandboxes.** The new BI architecture brings power users more fully into the corporate information architecture by creating analytical sandboxes that enable them to mix personal and corporate data and run complex, ad hoc queries with minimal restrictions. Types of analytical sandboxes include (1) virtual sandboxes running as partitions within a data warehouse, (2) a free-standing data mart running a replica of the data warehouse or other data not available in the data warehouse and powered by

an analytical platform or nonrelational database, (3) an in-memory BI tool that runs on an analyst's desktop or a corporate server and (4) a Hadoop cluster that stores atomic-level unstructured or semi-structured data.

4 **Implement analytical platforms that meet business and technical requirements.** Today, organizations implement analytical platforms for various reasons. For example, analytical appliances are fast to deploy and easy to maintain and make good replacements for Microsoft SQL Server or Oracle data warehouses that have run out of gas and are ideal as free-standing data marts that offload complex queries from large, maxed-out data warehousing hubs. Analytical databases, as software-only solutions, run on a variety of hardware platforms and are good for organizations that want to tune database performance for specific workloads or run the RDBMS software on a virtualized private cloud. Analytical services are great for development, test and prototyping applications as well as for organizations that don't have an IT department or want to outsource data center operations or get up and running very quickly. File-based analytical systems and nonrelational databases are ideal for processing large volumes of Web traffic and other log-based or machine-generated data. Organizations need to carefully evaluate the type and capabilities of the analytical platform they need before purchasing and deploying a system. ■

ABOUT THE AUTHOR

Wayne Eckerson has been a thought leader in the data warehousing, business intelligence and performance management fields since 1995. He has conducted numerous in-depth research studies and is the author of the best-selling book *Performance Dashboards: Measuring, Monitoring, and Managing Your Business*. He is a noted keynote speaker and blogger and he consults and conducts workshops on business analytics, performance dashboards and business intelligence (BI), among other topics. For many years, Eckerson served as director of education and research at The Data Warehousing Institute (TDWI), where he oversaw the company's content and training programs and chaired its BI Executive Summit.

Eckerson is currently director of research at TechTarget, where he writes a popular weekly blog called Wayne's World, which focuses on industry trends and examines best practices in the application of business intelligence. (See www.b-eye-network.com/blogs/eckerson.) Wayne is also president of BI Leader Consulting (www.bileader.com) and founder of BI Leadership Forum (www.bileadership.com), a network of BI directors who meet regularly to exchange ideas about best practices in BI and educate the larger BI community. He can be reached at weckerson@techtarget.com.

- **The Post-Relational Reality Sets In: 2011 Survey on Unstructured Data**
- **Leading Analyst Predicts Big Changes from Big Data: Exclusive Interview Recording**
- **Addressing the Challenges of Unstructured Information with Purpose-built Technology**

About MarkLogic:
MarkLogic empowers organizations to make high stakes decisions on Big Data in real time. Customers trust MarkLogic for mission critical applications that drive revenue and growth through Big Data Analytics enabled by MarkLogic products, services, and partners. MarkLogic is a fast growing enterprise software company that has been providing solutions to the public sector and Global 1000 for nearly a decade. Operating at petabyte scale, MarkLogic Server is a next generation database for unstructured information that allows customers to outflank their competition by consistently getting to better decisions faster.

MarkLogic is headquartered in Silicon Valley with field offices in Austin, Boston, Frankfurt, London, Tokyo, New York, and Washington D.C.