Exploiting Big Data: Strategies for Integrating with Hadoop to Deliver Business Insights

by Wayne Eckerson

Published: June 1, 2012

In this research report, Wayne Eckerson describes the emerging big data ecosystem in which Hadoop and analytical databases play an increasingly important role.

Research Background

The purpose of this report is to describe the emerging big data ecosystem in which Hadoop and analytical databases play an increasingly important role. It will then examine ways that traditional BI, extract, transform and load (ETL) and database vendors integrate with Hadoop. Finally, it will speculate on how Hadoop and analytical ecosystems will evolve and affect traditional BI approaches, technologies and vendors.

The research is based on interviews with BI practitioners, briefings with BI providers, including sponsors of this report, and a survey of BI professionals. The five-minute survey was promoted to the BI Leadership Forum, an online group of about 1,000 BI directors and managers, and my 2,000-plus Twitter followers in April 2012. The survey was started by 170 people and completed by 152 people for an 89.4% completion rate. Survey results are based on 158 respondents who completed the survey and indicated their positions as "BI or IT professional," "BI sponsor or user" or "BI consultant." Responses from people who selected the position of "BI vendor" or "Other" or who didn't complete the survey were excluded from the results.

Among this set of qualified respondents, most are BI or IT professionals (75%) from large companies with more than $1 billion in annual revenues (53%). The industries with the highest percentage of respondents are consulting (14%) and banking (11%) (see Figures 1, 2 and 3).

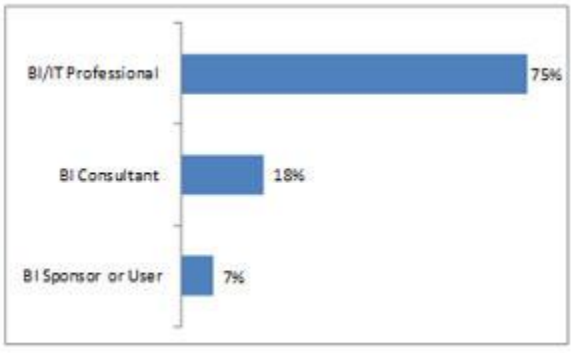

Figure 1: Demographics: Respondent Profile
Based on 158 respondents (BI Leadership Forum, January 2012, www.bileadership.com)

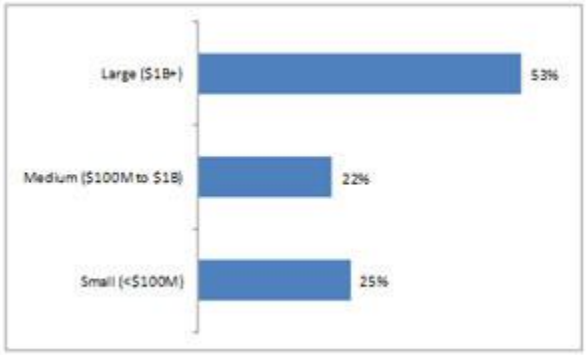

Figure 2: Demographics – Company Size
Based on 158 respondents (BI Leadership Forum, January 2012, www.bileadership.com)

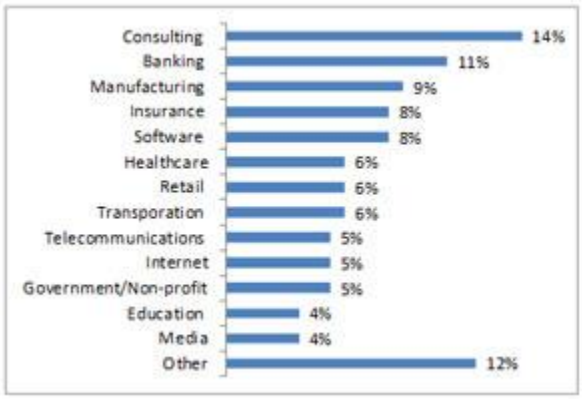

Figure 3: Demographics – Industries
Based on 158 respondents (BI Leadership Forum, January 2012, www.bileadership.com)

The New Analytical Ecosystem

There's a revolution afoot in the way organizations architect the flow of data to support reporting and analysis applications. The ringleader of the revolution is big data. There are two types of big data products in the market today. There is open source software, centered largely on on Hadoop, which eliminates up-front licensing costs for managing and processing large volumes of data. And 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 returns on investment than previous generations of

data management technology, but in vastly different ways. Organizations are gradually implementing both types of systems, among others, to create a new analytical ecosystem.

Understanding Hadoop

For many, big data has become synonymous with Hadoop—an open source framework for parallel computing that runs on a distributed file system (for example, the Hadoop Distributed File System, or HDFS.) For the first time, Hadoop enables organizations to cost-effectively store all their data as well as multi-structured data, such as Web server logs, sensor data, email and extensible markup language (XML) data. Because multi-structured data consists of 80% of all data by most estimates, the advent of Hadoop heralds the dawn of a new age of data processing in which organizations can pluck the needle out of the data haystack, which consists of terabytes, if not petabytes, of information.

Benefits. Hadoop, an open source project hosted by the Apache Software Foundation, brings three major benefits to the world of analytical processing:

1. **Low cost.** Because Hadoop is open source software that runs on commodity servers, it radically alters the financial equation for storing and processing large volumes of data—at least in terms of up-front licensing costs. With Hadoop, organizations can finally store all the data they generate in its raw form without having to justify its business value up front. This creates a low-cost staging and refining area and fosters greater data exploration and reuse.

2. **Load and go.** Compared with relational databases, Hadoop does not require developers to convert data to a specific format and schema (e.g., fields with fixed data types, lengths, labels and relationships) to load and store it. Rather, Hadoop is a load-and-go environment that handles any data format and speeds load cycles, which is critical when ingesting terabytes of data.

3. **Procedural code.** Finally, Hadoop MapReduce enables developers to use Java, Python or other programming languages to process the data. This means developers don't have to learn SQL, which is not as expressive as procedural code for certain types of analytical work, such as data mining operations or inter-row calculations. MapReduce is also implemented in a number of analytical databases to augment the familiarity and enterprise adoption of SQL.

Given these benefits, the data management industry is understandably abuzz about big data. Organizations are now exploiting creative ways to use Hadoop. For example, Vestas Wind Systems, a wind turbine maker, uses Hadoop 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.

Drawbacks. But Hadoop is not a data management panacea. It's essentially a 1.0 product that is missing many critical ingredients of an industrial-proof data processing environment—robust security, a universal metadata catalog, backup software, rich management utilities and in-

memory parallel pipelining. Moreover, Hadoop is a batch-processing environment that doesn't support speed-of-thought, iterative querying. And although higher-level languages exist, Hadoop today requires data scientists who know how to write Java or other programs to run data processing jobs, and these folks are in scarce supply.

To run a production-caliber Hadoop environment, you need to get software from a mishmash of Apache projects, which have colorful names like Flume, Sqoop, Oozie, Pig, Hive and ZooKeeper. These independent projects often have competing functionality and separate release schedules, and they 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. It's also why the jury is still out on the total cost of ownership for Hadoop. Today the cost of putting Hadoop into production is high, although this will improve as the environment stabilizes and administrative and management utilities are introduced and mature.

Many BI vendors—including database, data integration and reporting and analysis vendors—are working to integrate with Hadoop, making it easier to access, manipulate and query Hadoop data than is currently possible with native Apache software. The rapid merging of traditional BI and Hadoop environments creates a new analytical ecosystem that expands the ways organizations can effectively exploit data for business gain.

Adoption rates. Leading-edge companies from a variety of industries have already implemented Hadoop, and many more are considering it. A survey of the BI Leadership Forum conducted in April 2012 shows that 10% of organizations are implementing Hadoop or have already put it into production, while 20% are experimenting with it in-house and 32% are considering it. Not bad for a 1.0 technology. On the downside, 38% have no plans to use Hadoop (see Figure 4).

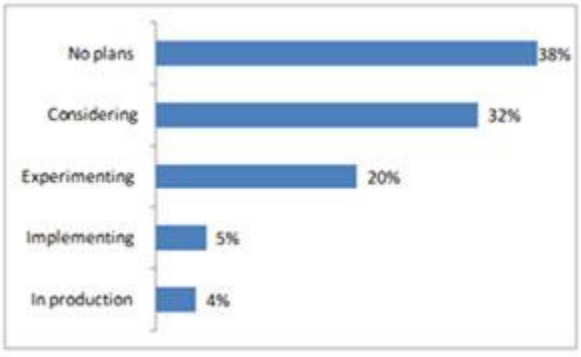

Figure 4: Adoption of Hadoop
Based on 158 respondents (BI Leadership Forum, January 2012, www.bileadership.com)

Many BI directors have heard the buzz around Hadoop and are starting to explore its value. Most are focused on use cases that Hadoop can support, ranging from technical workloads such as data processing, reporting and data mining to business applications such as fraud detection, risk modeling, churn analysis, sentiment analysis and cross-selling. But most haven't sufficiently examined how Hadoop fits into and extends their current BI architecture.

Understanding Analytical Databases

The other type of big data predates Hadoop by several years. It is less a "movement" than an extension of existing relational database technology optimized for query processing or advanced analytics. These analytical platforms span a range of technology, from appliances and columnar databases to shared nothing, massively parallel processing (MPP) databases. And they have unique extensions that in some cases include a MapReduce processing framework for advanced analytics. 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. Oracle launched the Exadata appliance in 2008, and it has been one of the company's most successful products ever. Aster Data (now part of Teradata) tightly integrated MapReduce with SQL processing. Startup ParAccel, which offers a software-only analytical database, has also achieved considerable traction, especially among financial services and other companies that require the highest level of performance for complex queries and analytical workloads. And several BI vendors, notably MicroStrategy, Oracle and Tableau, have been architecturally designed to exploit these big data sources.

Benefits. Although the price tag of analytical databases can often exceed $1 million including hardware, customers find that they are well worth the cost. Some use the technology to offload complex analytical workloads from existing data warehouses and avoid costly upgrades and time-consuming rewrites of existing applications. Others use SQL-MapReduce capabilities embedded in some analytical databases to analyze multi-structured data at scale and increase the accuracy of customer behavior models. But most use analytical databases to replatform aging data warehouses in an effort to eliminate performance bottlenecks and run a backlog of critical applications.

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. Gilt Groupe, an online luxury retailer, uses SQL-MapReduce processing to tie together clickstream information with email logs, ad viewing information, Twitter and operational information to discover consumer preferences and improve the shopping experience. Australian Finance Group, the largest provider of mortgage services in Australia, implemented an analytical appliance to accelerate query and log-in performance of a critical application used by its mortgage brokers, achieving a 42% return on investment in three years.

Challenges. Given the up-front costs of analytical databases, organizations usually do a thorough evaluation of these systems before jumping on board. First, companies determine whether an analytical database outperforms their existing data warehouse databases to a degree that warrants migration and retraining costs. This requires a proof of concept in which a 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 databases 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 databases 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-based cloud service through vendors, such as Amazon Web Services. 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 and scale linearly to handle the growth in numbers of concurrent users? Or is it best used as an analytical sandbox? If so, how well does it handle complex queries? Does it have a built-in library of analytical algorithms and an extensibility framework to add new ones? What kinds of data make sense to offload to the new system? How do you rationalize having two data warehousing environments instead of one?

Today, small and medium-sized companies that have tapped out their Microsoft SQL Server or MySQL data warehouses usually replace them with analytical appliances to get better performance. But large companies that have invested millions of dollars in a data warehouse typically augment the data warehouse with outboard, purpose-built systems designed to handle unique workloads or types of data. Here the data warehouse becomes the hub that feeds multiple downstream systems with clean, integrated data. This hub-and-spoke architecture helps organizations avoid a costly upgrade to their legacy data warehouses, which might exceed the cost of a new analytical database.

The New Analytical Ecosystem

Figure 5 depicts a reference architecture for a 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 new big data and analytical technologies.

By extending the traditional BI architecture with Hadoop, analytical databases and other analytical components, organizations gain the best of top-down reporting and bottom-up analytics in a single ecosystem. Traditionally, reports and analyses run in separate environments—reports using data warehouses and analyses using Excel, Microsoft Access and renegade data marts. But the new analytical ecosystem brings business analysts into the mainstream, enabling them to conduct freeform analyses inside the corporate data infrastructure using a variety of analytical sandboxes.

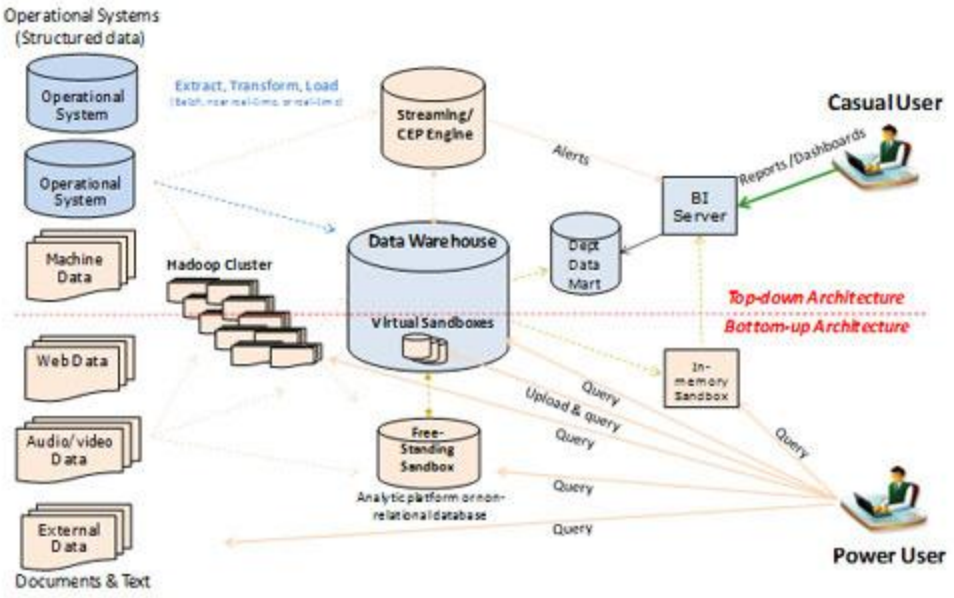

Figure 5: The New Analytical Ecosystem

Top-down processing. The top-down world of reporting and dashboard (depicted in the top half of Figure 5) delivers answers to questions that users define in advance. BI developers gather requirements, stamp questions into a multidimensional data model, find relevant data, map it to the target model and build interactive reports that answer the predefined questions with some options for further exploration. Although the top-down world generally processes data in large batches at night, it can also handle continuous data using an operationalized data warehousing environment or a complex event processing (CEP) engine when ultra low-latency analytics is required. CEP systems examine patterns within high-volume event streams and trigger alerts and workflows when predefined thresholds are exceeded.

Bottom-up processing. The bottom-up world of analysis (depicted in the lower half of Figure 5) answers new and unanticipated questions using ad hoc queries, visual exploration and predictive analytical tools. In the new analytical ecosystem, business analysts have many options for exploring and analyzing data. While they used to be left to their own devices to collect, standardize and integrate data, now they can access raw data in Hadoop or standardized, corporate data in the data warehouse and mix it with their own data. They can then model and explore data however they want in officially sanctioned sandboxes without inciting the ire of corporate IT managers. When they're ready to share insights with others, they publish their analyses to a corporate server and turn over responsibility to the enterprise BI team to produce the output.

Analytical sandboxes. New analytical sandboxes let power users explore a combination of enterprise and local data in a safe haven that won't interfere with top-down workloads. Different types of analysts use different sandboxes. Data scientists query raw, multi-structured data in Hadoop using Java, Python, Perl, Hive or Pig. Business analysts query virtual sandboxes in the data warehouse or free-standing analytical appliances designed for complex query processing

using SQL, built-in analytical algorithms or a combination of SQL and MapReduce capabilities. Business analysts and superusers explore subsets of data using in-memory visualization tools and immediately publish their findings as dashboards for departmental colleagues to consume.

Data flows. In the new ecosystem, a lot of the source data flows through Hadoop, which acts as a staging area and online archive for all of an organization's data. This is especially true for multi-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, much of the data is fed into a data warehousing hub, which often distributes data to downstream systems, such as data marts, operational data stores, analytical appliances and in-memory visualization applications, where users can query the data using familiar SQL-based reporting and analysis tools. Some data is also fed directly into free-standing analytical databases and appliances. To build data flows, organizations use graphical data integration and data quality tools that hide the technical complexity of Hadoop and MapReduce code generation. 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 multi-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.

But 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 analysts and data scientists need to commit to adhering to corporate data standards in exchange for getting the keys to the kingdom. To succeed, organizations need robust data governance programs and lots of communication among all parties.

Big Data Trends

According to surveys of the BI Leadership Forum, user organizations have invested significantly in analytical databases while early adopters are taking the plunge with Hadoop.

Almost half of respondents to a 2011 BI Leadership Forum survey said they have either implemented an analytical database (i.e., a software-only system) or an analytical appliance (i.e., a hardware-software combination). In contrast, only 11% have implemented Hadoop (see Figure 6).

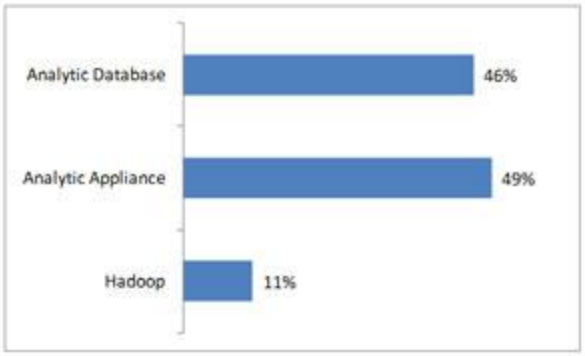

Figure 6: Deployment Trends
Based on 302 respondents (BI Leadership Forum, April 2011, www.bileadership.com)

Companies have deployed all three types of platforms in a variety of ways. One significant standout is that 79% of companies that have deployed analytical appliances have implemented them to run data warehouses. In most cases, this is to replace a legacy relational database or a SQL Server or MySQL implementation that ran out of gas. For instance, many Oracle customers have migrated their Oracle database licenses to the Oracle Exadata Database Machine. Hadoop is also more likely to be used as a prototyping area than anything else, which reflects the fact that most companies today are experimenting with Hadoop rather than putting it in production. Analytical databases tend to be used as data marts, both dependent and independent, more than analytical appliances. This reflects their use as standalone systems to handle unique workloads outside of the data warehouse (see Figure 7).

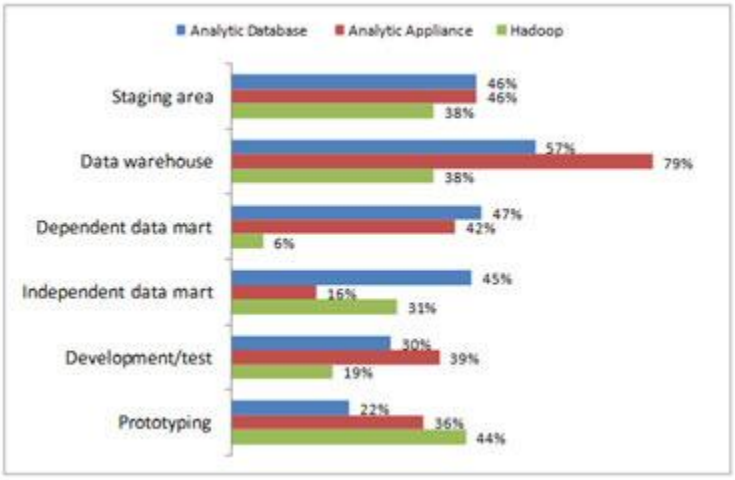

Figure 7: Use Cases
Based on 302 respondents (BI Leadership Forum, April 2011, www.bileadership.com)

Hadoop Attracting Interest

Although the analytical ecosystem depicted in Figure 5 portrays Hadoop as a staging area for enterprise data warehouses and other downstream analytical systems, some big data and BI

professionals think Hadoop may attract a growing percentage of analytical queries in the future.

Netflix, for example, uses Hadoop to stage all its operational data before aggregating and loading it into its Teradata data warehouse, where users can access it using MicroStrategy and other SQL-based tools, according to Kurt Brown, director of business intelligence platforms at the company. Business analysts can also query the raw data in Hadoop using Java, MicroStrategy (through free-form HiveQL) or other languages if they need access to the atomic-level data or have an urgent request for information that can't wait for the data to be loaded into the data warehouse.

Like Netflix, almost all organizations today that have implemented Hadoop are using it as a staging area (92%) to land and transform data, presumably for loading into a downstream data warehouse, where users can access and analyze it using standard BI tools. They are also using it as an online archive (92%) so they never have to delete data or summarize it and move it offline where it is less accessible (see Figure 8). In some cases, organizations move data from their data warehouse into Hadoop for safekeeping instead of into an archival system.

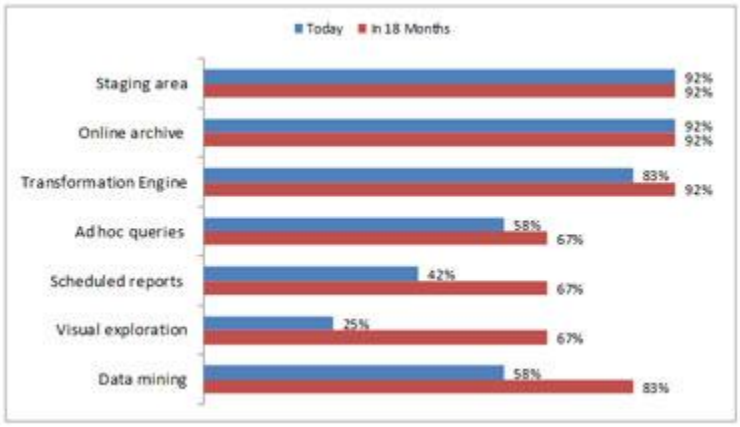

Figure 8: What Analytical Functions Does Hadoop Support in Your Organization? Based on respondents who have implemented Hadoop (BI Leadership Forum, April 2012, www.bileadership.com)

Far fewer organizations are using Hadoop for running queries or reports or exploring or mining data. But this will change in the next 18 months, according to the survey. The percentage of companies that plan to use Hadoop for ad hoc queries will jump from 58% to 67%; reporting will climb from 42% to 67% and data mining from 58% to 83%. But the biggest growth will come in visual exploration, from 25% to 67%, as companies use visualization tools such as Tableau to explore data in Hadoop.

Hadoop data. Organizations are using Hadoop to store and process all kinds of data, not just multi-structured data. In fact, almost all respondents are using Hadoop today (92%) to store transaction data, and that figure will climb to 100% in 18 months. In fact, Netflix lands all its operational data from its consumer websites and networking streaming devices into Hadoop. Very little operational data gets into Netflix's data warehouse before first passing through Hadoop.

After transaction data, Hadoop is most often used to store Web and systems logs (67% each), followed by social media and semi-structured data (58% each). Some companies are using Hadoop to store documents (33%), email (25%) and sensor data (17%), but few are using it to store audio or video data yet.

What's interesting is that these same companies plan to expand the types of data they store in Hadoop over the next 18 months. The fastest growing data types stored in Hadoop among early adopters will be audio and video data, sensor data and email (42% each) and documents (50%). More companies also plan on capturing social media data in Hadoop (75%) as well as Web logs (75%), transaction data (100%) and semi-structured data (67%) (see Figure 9).

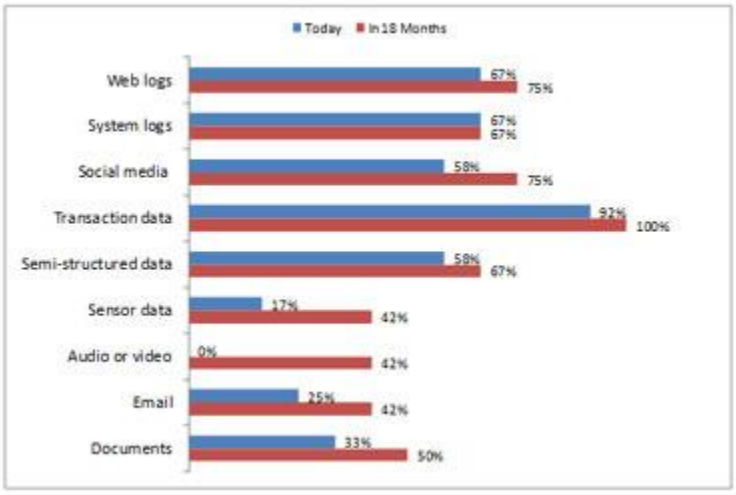

Figure 9: What Data Does Hadoop Support?
Based on respondents who have implemented Hadoop (The BI Leadership Forum, April 2012, www.bileadership.com)

Despite some of the rhetoric coming out of the Hadoop community, early Hadoop adopters see the technology as serving a complementary role to the data warehouse. Most are using Hadoop to handle new workloads (66.7%) that don't run in their data warehouse or they are offloading existing workloads that are straining the current data warehouse (50%). Another third (33%) use Hadoop to share existing data warehousing workloads, while another quarter (25%) use it to share new workloads (see Figure 10).

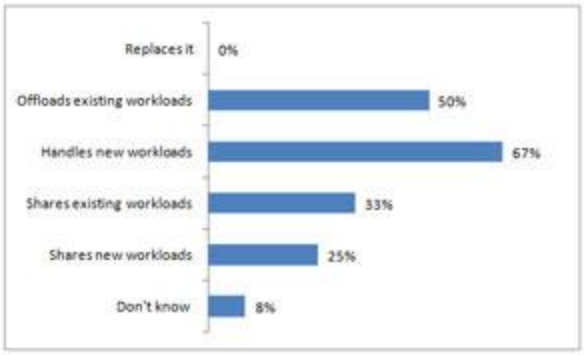

Figure 10: Impact of Hadoop on Your Data Warehouse
Based on respondents who have implemented Hadoop (The BI Leadership Forum, April 2012, www.bileadership.com)

Non-Implementers

Companies that have yet to implement Hadoop are fairly bullish about their plans for the technology. "For us, Hadoop is a given," said Mike King, technical fellow at FedEx Services, who develops database architectures for the shipping and logistics company. "We're looking at big data with an eye towards [implementing] a full analytics stack. We see Hadoop as part of the puzzle and believe it's a good place to land raw data sources for analysis."

While FedEx has a robust data warehouse, it is used primarily to run reporting and dashboarding applications on structured data sets, according to King. Hadoop will support new data sources, such as server log data, system performance data, Web clickstream data and reference data—including addresses, shipments, invoices and claims. And King expects people will analyze much of this data in place, using MapReduce jobs to query the data.

Like FedEx, 40% of companies plan to implement Hadoop within 12 months, and another 22% will do so within two years, according to the survey. About a third (30%) are not sure of their plans, and surprisingly, no one said they will "never" implement Hadoop (see Figure 11).

Clearly, Hadoop has fixed itself in the current imaginations of BI professionals who don't want to get left out in the cold—it's one of the biggest new developments in data management in about five years. (The last big trend was the advent of analytical appliances and platforms, which are still in the early adoption phase.)

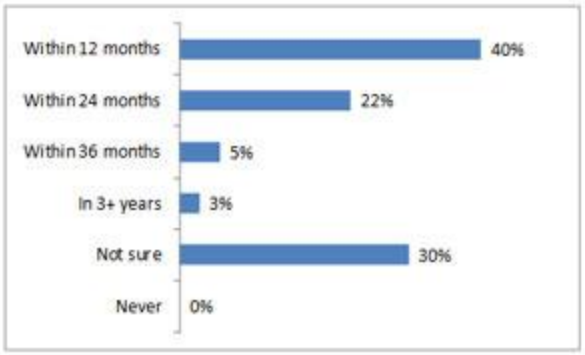

Figure 11: Expected Adoption Rate of Hadoop by Non-Implementers
Based on 76 respondents who have not yet implemented Hadoop (The BI Leadership Forum, April 2012, www.bileadership.com)

Hadoop functionality. Companies that have yet to implement Hadoop have a different idea about how they'll use the new technology. The largest percentage of potential Hadoop users plans to use it for data mining (57%), followed by ad hoc queries (45%) (see Figure 9). A smaller percentage sees Hadoop as a staging area (37%), online archive (23%) or transformation engine (39%), which differs sharply from organizations that have already implemented Hadoop (see Figure 12).

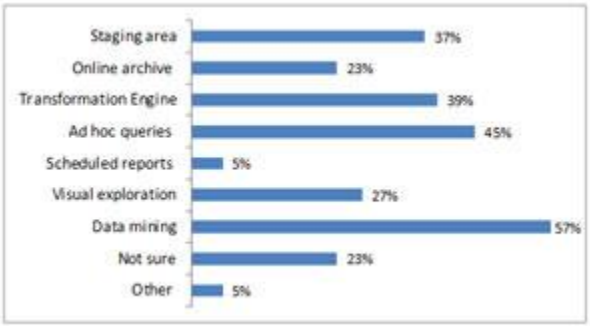

Figure 12: Expected Use of Hadoop by Non-Implementers
Based on 76 respondents who have not yet implemented Hadoop (The BI Leadership Forum, April 2012, www.bileadership.com)

Much of the disconnect between current and future adopters of Hadoop comes from the hype emanating out of the big data community, which generally portrays Hadoop as an analytical system for multi-structured data instead of a staging and pre-processing area. Analytics is the promise of Hadoop, but this won't happen until there is a viable market of easy-to-use graphical tools for accessing and analyzing Hadoop data. The BI industry has spent the last 20 years developing such tools for SQL databases, and it's only now getting the formula right. Fortunately, Hadoop vendors will be able to piggyback on the lessons learned by BI vendors, if not the actual tools and techniques themselves (see "Integrating with Hadoop").
Hadoop data. Non-implementers also share a similar profile as Hadoop implementers when it comes to the kind of data they expect to store in Hadoop with one exception: Non-implementers

are less bullish about storing transaction data in Hadoop (see Figure 13). In fact, only 47% said they plan to store transaction data in Hadoop; that's compared with 100% of implementers. For non-implementers, Hadoop is synonymous with non-structured data.

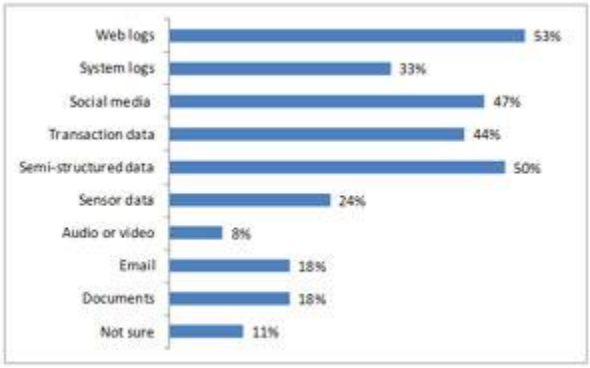

Figure 13: Data That Non-Implementers Will Store in Hadoop

Impact on data warehouses. Non-implementers share much the same profile in the way Hadoop will affect their existing data warehouses. First, no one plans to replace a data warehouse with Hadoop (see Figure 11). The largest percentage of respondents plans to use it to handle new workloads (58%), followed by sharing new workloads (37%), offload existing workloads (27%) and share existing workloads (24%) (see Figure 14).

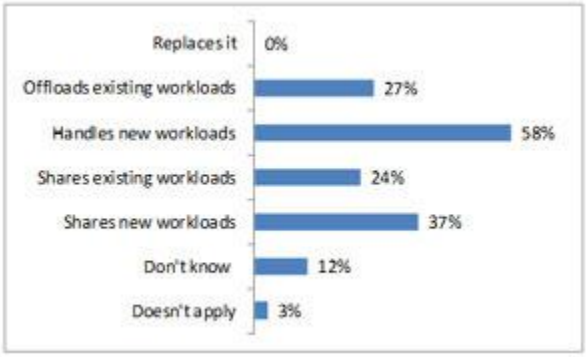

Figure 14: The Impact of Hadoop on Data Warehouses of Non-Implementers
Based on 76 respondents who have not yet implemented Hadoop (The BI Leadership Forum, April 2012, www.bileadership.com)

A majority of non-implementers view Hadoop as a means to analyze new multi-structured data in its native form or process that data and load it into their data warehouses for analysis. One respondent wrote, "I need more 'gory processing' of semi-structured and unstructured data in an MPP environment prior to getting that content linked up to our relational data [in the data warehouse.]" Another said, "Hadoop will augment our current business intelligence and data warehousing environment with large unstructured data sets, mostly social networking logs. It will in no way replace our existing data warehouse and data marts or ETL environments."

Hadoop Hype

Science experiment. For some, Hadoop is a curiosity they feel obligated to explore because of all the hype. “We’re looking at Hadoop to see what all the excitement is about and whether it offers any benefits for us,” said John Rome, deputy chief information officer and BI strategist at Arizona State University. “All we hear about is big data, big data, big data.”

Rome said Hadoop might be an appropriate place to house the university’s voluminous Web log, swipe card and possibly social media data as well as data from its learning management system, which generates 4 million records daily. But he wonders whether Hadoop would be overkill for these applications since it’s a good possibility IT could flow this data directly into the university’s data warehouse. “We look forward to applying these emerging tools to administrative tasks and gain a better understanding of how we can help those at ASU who are working on big data challenges and research problems,” he said.

A senior vice president of analytics at a major online publishing company concurs with Rome. “I am not convinced Hadoop is the right solution for the problem they are trying to solve; however, our team is taking a look. My personal belief is that the team partially just wants to play with Hadoop because it is a hot technology.”
Nonetheless, both managers plan to evaluate Hadoop in the near future. Rome plans to carve out time this summer or fall to evaluate Hadoop, while the analytics manager is awaiting the results of his team’s experimental usage of Hadoop.

Integrating with Hadoop

Vendors of all stripes are working to integrate their products with Hadoop so users can do any kind of workload with any tools they want. In many respects, there seems to be a race between members of the Apache Software Foundation and established BI and data warehousing (DW) vendors to see who can wrap Hadoop with the enterprise and analytical functionality it currently lacks. Both sides are reaching out to the other to improve interoperability and integration between Hadoop and established BI and DW environments.

Quest for Interoperability

Apache analytical projects. The Apache Software Foundation has multiple projects to address Hadoop shortcomings in the arena of reporting and analytics. One promising project is Hive, a higher-level data access language that generates MapReduce jobs and is geared to BI developers and analysts. Hive provides SQL-like access to Hadoop, although it does nothing to overcome its 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. Another high-level scripting language is Pig Latin, which like Hive runs on MapReduce, but is better suited to building complex, parallelized data flows, so it’s something to which ETL programmers might gravitate.

Another emerging Apache project that addresses the lack of built-in metadata in Hadoop is

HCatalog, which takes Hive's metadata and makes it more broadly available across the Hadoop ecosystem, including Pig, MapReduce and HBase. Soon SQL vendors will be able to query HCatalog with MapReduce, Web and Java Database Connectivity (JDBC) interfaces to determine the structure of Hadoop data before launching a query. Armed with this metadata, SQL products will be able to issue federated queries against SQL and Hadoop data sources and return a unified result set without users having to know where data lies or how to access it.

On the data-mining front, Apache Mahout is a set of data-mining algorithms for clustering, classification and batch-based collaborative filtering that run on MapReduce. Finally, Apache Sqoop is a project to accelerate bulk loads between Hadoop and relational databases data ingestion, and Apache Flume streams large volumes of log data from multiple sources into Hadoop.

These projects are still in the early stages, although many leading-edge adopters have implemented them and are providing ample feedback, if not code, to the Apache community to make these software environments enterprise-ready. They are assisted by many commercial open source vendors that contribute most of the code to Apache Hadoop projects and want to create a rich, enterprise-ready ecosystem.

BI and DW outreach. In addition, many BI and DW vendors have either extended their tool sets to connect to or interoperate with Hadoop. For example, many vendors now connect to Hadoop and can move data back and forth seamlessly. Some even enable BI and ETL developers to create and run MapReduce jobs in Hadoop using familiar graphical development environments. Some startup companies go a step further and build BI and ETL products that run natively on Hadoop. Many database vendors are shipping appliances that bundle relational databases, Hadoop and other relevant software in an attempt to offer customers the best of both worlds.

Most of these Apache projects are led by commercial open source vendors, typically the Hadoop distribution vendors, which strive to enrich their distribution and make it more enterprise-ready than the next one. By working under the Apache Software Foundation umbrella, they accelerate the time to market of these features and ensure that compatibility will be maintained throughout the process.
Types of integration. Since many organizations plan to add Hadoop to their existing BI environments, it's important to understand the degree to which BI and DW products can interoperate with Hadoop.

There are four types of integration, each with its own advantages and disadvantages:

Connectivity. Prebuilt data connectors enable developers to view and move data in bulk between the two environments.
Hybrid systems. These products blend SQL and MapReduce functionality into a single data processing environment.
Native Hadoop. These run natively on Hadoop but are accessed via a graphical user interface that hides the complexities of the underlying processing environment.
Interoperability. Application programming interfaces and metadata catalogs enable developers in one environment to design and dynamically execute native jobs in another environment.
Type 1. Connectivity

A data connector enables developers to view data in a remote system and download selected data to their local applications. Since Hadoop stores data in a distributed file system, the data connector can simply be a file system viewer that connects to Hadoop's name node, which contains a list of all files in the cluster. The connector also usually executes Hadoop file system commands to retrieve specified files distributed across the cluster and move them to the requesting system.

Many data connectors connect via an application programming interface (API), like Open Database Connectivity (ODBC), and JDBC drivers, which link client applications to relational databases. With an API approach, the connector enables users to invoke functions on the remote system, which is the second level of integration (see "Interoperablity"). Most vendors don't charge for connectors.

For example, ParAccel offers a bidirectional parallel connector for Hadoop that enables developers to issue SQL calls that invoke appropriate prewritten MapReduce functions in Hadoop and transfer files to ParAccel for further processing. Similarly, Oracle offers its Direct Connector for Hadoop Distributed File System, which queries HDFS data using SQL and loads it into the Oracle Database or joins it with data already stored there.

Pros and cons. The benefit of data connectors is that they are relatively easy to build and use. The downsides are that they require moving entire data sets, which in a big data environment is not ideal. If the data volumes are sizeable, the download may take a long time and create network bottlenecks. Administrators who want to move data between Hadoop and relational database management systems will use Sqoop or Flume or some proprietary bulk load utility instead.

Type 2. Hybrid Systems

Hybrid systems create a single environment that supports both Hadoop and SQL data and functions. There are no remote function calls because all processing happens locally. The single system runs both a relational database and MapReduce and supports SQL and file system storage.

Hybrid systems vary considerably in architecture. Some embed relational databases inside Hadoop, while others embed Hadoop and a relational database inside a single data warehousing appliance.

For instance, the Teradata Aster MapReduce platform embeds a MapReduce engine inside a relational database. Administrators load structured and multi-structured data into the platform for data discovery. Teradata's patented Aster Data SQL-MapReduce framework enable users to invoke and execute MapReduce jobs that run inside the Aster platform using standard SQL and BI tools. As a result, SQL-MapReduce enables organizations to store all their data—both structured and multi-structured—in a single place and query it using standard SQL or a variety of programming languages with MapReduce. This strategy effectively puts Hadoop and unstructured data into the familiar land of SQL-based processing.

Hadapt, on the other hand, takes the opposite approach in its cloud-based product, which is set to

launch this summer. Hadapt installs a Postgres relational database on every node in a Hadoop cluster and then converts SQL calls to MapReduce, where needed, to process both sets of data at once. Like Aster SQL-MapReduce, Hadapt lets users store all their data in one place and use a single query paradigm to access all data. These specialized environments are ideal for supporting applications, such as customer analytics, that require querying both structured and unstructured data to get a complete picture of customer behavior.

Pros and cons. The benefits of a hybrid system are straightforward: (1) It creates a unified environment for all types of data, (2) you buy and support one system instead of two, (3) you can issue a single query that runs against any data type and (4) all processing runs locally, providing fast performance. The major downside of such systems is they are somewhat redundant if you already manage an enterprise data warehouse or Hadoop environment or both. As such, hybrid systems make good analytical sandboxes or discovery platforms to handle new applications that require a synthesis of data types and analytical techniques.

Type 3. Native Hadoop

Some vendors, including a bevy of startups and some established BI and DW players, are aggressively courting the Hadoop market by building applications that use MapReduce as the underlying execution engine.

As a new platform, Hadoop needs to encourage developers to write applications that run on top of it. Some stalwart BI and data integration veterans have taken up the challenge and have built new applications from scratch or are porting existing applications to run on Hadoop. In most cases, the new applications provide a graphical interface that hides the underlying processing from end users, who don't know where the application runs or the nature of its execution engine. All they know is that they can query Hadoop data using a point-and-click interface.

For example, startup Datameer offers a native Hadoop application for creating dashboards. The tool builds a meta catalog of Hadoop data that it imports using 20-plus connectors to a variety of sources, including most relational databases, NoSQL databases and social media feeds, such as Twitter.

Developers use the catalog to populate a Web-based spreadsheet with a subset of data that they use to design the dashboard. Once they're satisfied with the design, they schedule a job to create the dashboard by running it against the entire data set in Hadoop.

Some established BI and DW vendors are contemplating shipping native Hadoop versions of their products. For instance, one leading data integration vendor plans this year to ship a native Hadoop version of its ETL tool that will convert existing mappings to run as MapReduce jobs on Hadoop. This capability is part of a wider set of functionality that pushes down processing to the most efficient platform in the BI and DW ecosystem.

Pros and cons. The benefit of this native approach is that it provides seamless access to Hadoop data and optimizes MapReduce functions. Presumably, the new tool would better stay abreast of new developments in the Hadoop world and optimize MapReduce processing better than SQL vendors. The downside is that Hadoop MapReduce is a batch environment that does not support iterative queries, at least today. Users need to design the entire dashboard before they hit the execute button, since there is sizable latency when running MapReduce jobs. Also, the

application would need to make ODBC or JDBC calls via MapReduce to access SQL data in relational databases when joining this data with Hadoop data or use data services to retrieve this data dynamically.

Type 4. Interoperability

As mentioned above, interoperability takes a data connector one step further and lets users execute native functions on the remote system, often using a graphical interface that doesn't require users to write code. This is application-level interoperability using custom or standard APIs and a metadata catalog rather than data-level connectivity that links directly to a database or file system. Obviously, writing code to an API requires more time and testing than does creating a data connector.

For example, Tableau uses Cloudera's ODBC driver for Hive to query Hadoop data directly. The driver shields users from having to write HiveQL statements. Since Hive does not support iterative queries, Tableau users can view, connect and download the data they want to an in-memory cache that users can query in an ad hoc manner. MicroStrategy also uses Cloudera's ODBC driver but goes further by embedding Hive metadata into its semantic layer (for example, Projects) and Query Builder, dynamically generating HiveQL under the covers. It also lets developers execute predefined HiveQL and Pig scripts using the free-form query capability.

Hortonworks, a startup that provides a software distribution of Apache Hadoop along with training, service and support, establishes deep levels of integration with established BI and DW products. Its developers, most of whom helped build and manage Yahoo's 42,000-node Hadoop cluster, contribute the majority of code to Apache Hadoop. One of the key pieces of Apache Hadoop that it is building is HCatalog, a SQL-like metadata catalog accessible via Hive, MapReduce and various REST-based APIs. Talend, an open source ETL vendor, already uses HCatalog to design MapReduce workflows and transformations. HCatalog is likely to become Hadoop's de facto metadata repository and key piece in any interoperability framework.

Pros and cons. In the case of Hadoop, the benefit of an application-level interface is significant. Users of SQL-based tools can issue a query against Hadoop and retrieve just the data they want without having to download a much larger data set. In other words, they can use SQL to locate Hadoop data and invoke a MapReduce job to process or query that data. In addition, rather than execute an import-export function, users query Hadoop on demand as part of a single SQL query, which may also join the Hadoop data with other data to fulfill the SQL request. One downside: these APIs are not bidirectional—in other words, a Hadoop developer would need to work with a SQL API to query data from a relational database as part of a MapReduce job.

Future of Hadoop

Cooperation or competition? Vendors have been quick to jump on the big data bandwagon largely because it opens up a new market for their database, data integration and BI products, but also because Hadoop represents a potential threat. Established software vendors stand to lose significant revenue if Hadoop evolves without them or gains robust data management and analytical functionality that reduces the appeal of their existing products. Most are hedging their bets, keeping their friends close and their enemies closer. They are working to interoperate with Hadoop, in case it becomes a major new data source, and ensure it remains subservient to their

existing products.

Both sides are playing nice and are eager to partner and work together. Hadoop vendors benefit as more applications run on Hadoop, including BI, ETL and relational database management system (RDBMS) products. 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 honey attracts a hive full of bees.

Who wins? Some customers are already asking whether data warehouses and BI tools will eventually be folded into Hadoop environments or the reverse. As one survey respondent put it, "We are still evaluating between two schools of thought: pull from all sources including Hadoop clusters into the RDBMS or pull from all, including the RDBMS, into the Hadoop cluster."

It's no one wonder BI professionals are confused. In the past year, they have been bombarded with big data hype, some of which makes it sounds like the days of the data warehouse and relational database are numbered. They think, "Why spend millions of dollars on a new analytical RDBMS if we can do that processing without paying a dime in license costs using Hadoop?" Or "Why spend hundreds of thousands of dollars on data integration tools if our data scientists can turn Hadoop into a huge data staging and transformation layer?" or "Why invest in traditional BI and reporting tools if our power users can exploit Hadoop using freely available programs, such as Java, Python, Pig, Hive or Hbase?"

Given the hype it shouldn't be surprising that some user organizations are opting for Hadoop when a NoSQL database or more traditional BI solution would suffice. For example, several customers of Splunk, a NoSQL vendor that applies search technology to log data, planned to implement Hadoop until they came across Splunk. Even though Splunk is commercial software, the customers found the software much easier to implement and use than Hadoop, making its total cost of ownership much lower. Also, a customer that has large volumes of XML data was considering Hadoop but recently discovered MarkLogic, another NoSQL database vendor that uses an XML schema to process data. It is now investigating whether MarkLogic is a better fit than Hadoop.

The future is cloudy. Right now, it's too early to divine the future of the big data movement and name 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. Since Hadoop vendors are venture-backed, it's likely we'll discover the answer soon enough as investors get restless and push for a sale. In this case, a commercial vendor will gain access to the major Hadoop distributions and more than likely pursue a hybrid strategy.

More than likely, we'll get a mélange of open source and commercial capabilities that interoperate to create the big data ecosystem as described in section one of this report. After all, 30 years after the mainframe revolution, mainframes are still an integral component of many corporate data processing architectures. And multidimensional databases didn't replace relational databases as the building blocks of data warehousing environments; rather, multidimensional

databases became an external data mart or cubing technology geared to intensive, multidimensional analytical workloads. Our corporate computing environments have an amazing ability to ingest new technologies and find an appropriate niche for each within the whole. In IT, nothing ever dies; it just finds its niche in an evolutionary ecosystem.

Accordingly, Hadoop will most likely serve as a staging area and pre-processing environment for multi-structured data before it is loaded into analytical databases. Like any staging area, some users with requisite skills and permission will query and report off this layer, but the majority of users will wait until the data is safely integrated, aggregated, cleansed and loaded into high-performance analytical databases, where it can be queried and analyzed with SQL-based tools.

Recommendations

Implement an analytical database. If you haven't already, it's time to implement an analytical database, either to augment or replace your data warehousing engine or create an analytical sandbox or discovery platform for business analysts or both. As part of the process, identify a critical application for the new system that can kick-start adoption and build support among executives to expand the analytics program.

Role of multi-structured data. Identify multi-structured data in your organization and how harvesting it might provide business value. Talk to business-side people and ask what they would do if they could query this multi-structured data and what it would be worth to them.

Investigate Hadoop. Examine Hadoop and monitor its evolution. Understand the pros and cons of using Hadoop to manage your multi-structured data versus other approaches, such as analytical databases, text mining and NoSQL databases. Identify a critical application for Hadoop, such as improving the accuracy of fraud or cross-sell models, to justify the commitment of time and money and keep it from being labeled a "science experiment."

Investigate how to integrate with Hadoop. Understand the four types of integration with Hadoop from SQL environments. Identify which levels you need for the various Hadoop applications you have identified.

Understand tradeoffs between Hadoop and analytical databases. When you should process data in Hadoop versus new analytical databases is not always clear. Obviously, Hadoop is geared more toward unstructured data because it's based on a file system, while analytical databases are geared more toward structured data. But that doesn't mean you can't do any kind of processing in either environment. It simply depends on data volumes, internal expertise and the nature of the consuming application. On the whole, it's best to experiment before moving workloads to any new environment.

About the Author

Wayne Eckerson - Wayne 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.

Wayne 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. He can be reached at weckerson@techtarget.com.