TDWI RESEARCH

TDWI BEST PRACTICES REPORT

FOURTH QUARTER 2013

MANAGING BIG DATA

By Philip Russom

CO-SPONSORED BY

sas

tdwi.org

tdwi

MANAGING BIG DATA

By Philip Russom

Table of Contents

About the Author

PHILIP RUSSOM, Ph.D., is a well-known figure in data warehousing and business intelligence, having published more than 500 research reports, magazine articles, opinion columns, speeches, Webinars, and more. Today, he's the TDWI Research Director for Data Management at The Data Warehousing Institute (TDWI), where he oversees many of the company's research-oriented publications, services, and events. Before joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and BI consultant and was a contributing editor with leading IT magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org, @prussom on Twitter, and at linkedin.com/in/philiprussom.

About TDWI

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence and data warehousing industry. TDWI is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence and data warehousing solutions. TDWI also fosters the advancement of business intelligence and data warehousing research and contributes to knowledge transfer and the professional development of its members. TDWI offers a worldwide membership program, five major educational conferences, topical educational seminars, role-based training, onsite courses, certification, solution provider partnerships, an awards program for best practices, live Webinars, resourceful publications, an in-depth research program, and a comprehensive website, tdwi.org.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies and is supplemented by surveys of business intelligence professionals.

To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving business intelligence problems or an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical business intelligence issues. Please contact TDWI Research Director Philip Russom (prussom@tdwi.org) to suggest a topic that meets these requirements.

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who agreed to our requests for phone interviews. Second, our report sponsors, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: Jennifer Agee, Michael Boyda, and Denelle Hanlon.

Sponsors

Cloudera, Dell Software, Oracle, Pentaho, SAP, and SAS sponsored the research for this report.

Research Methodology and Demographics

Report Scope. The purpose of this report is to accelerate users' understanding of data management in the age of big data. TDWI assumes that most of the best-known practices for data management still apply to managing big data—albeit with adjustments. A number of new best practices for managing big data are also emerging. Many of these are described in this report.

Terminology. For the purposes of this report, big data is characterized as very large data sets (multi-terabyte or larger) that usually consist of a wide range of data types (relational, text, multi-structured data, etc.) from numerous sources, including relatively new ones (Web applications, machines, sensors, and social media).

Survey Methodology. In May 2013, TDWI sent an invitation via e-mail to the data management professionals in its database, asking them to complete an Internet-based survey. The invitation was also posted in Web pages, newsletters, and publications from TDWI and other firms. The survey drew responses from 693 survey respondents. From these, we excluded incomplete responses and respondents who identified themselves as academics or vendor employees. The resulting completed responses of 461 respondents form the core data sample for this report. Due to branching in the survey, some questions were answered by only 189 respondents who have experience managing big data.

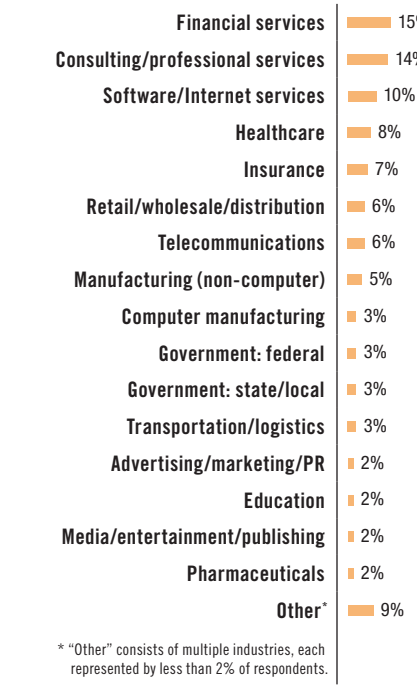
Research Methods. In addition to the survey, TDWI Research conducted many telephone interviews with technical users, business sponsors, and recognized data management experts. TDWI also received product briefings from vendors that offer products and services related to the best practices of managing big data.

Survey Demographics. The majority of survey respondents are IT professionals (67%), whereas the others are consultants (21%) and business sponsors or users (12%). We asked consultants to fill out the survey with a recent client in mind.

The financial services industry (15%) dominates the respondent population, followed by consulting (14%), software/Internet (10%), healthcare (8%), insurance (7%), and other industries. Most survey respondents reside in the U.S. (48%), Europe (20%), or Asia (11%). Respondents are somewhat evenly distributed across all sizes of companies and other organizations.

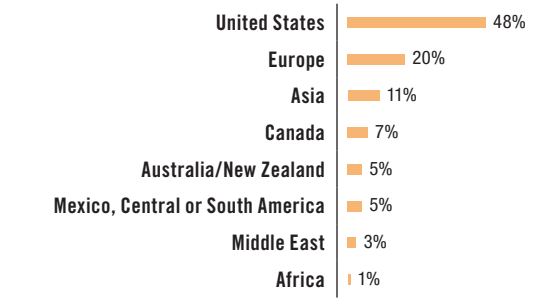
Position

Industry

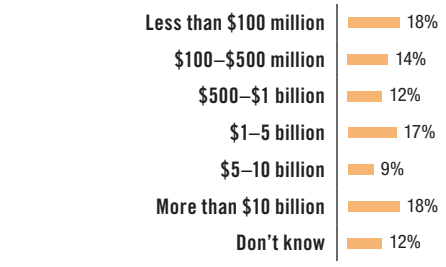


* "Other" consists of multiple industries, each represented by less than 2% of respondents.

Geography



Company Size by Revenue



Based on 461 survey respondents.

Executive Summary

Getting business value from big data starts with managing big data appropriately.

The emerging phenomenon called *big data* is forcing numerous changes in businesses and other organizations. Many struggle just to manage the massive data sets and non-traditional data structures that are typical of big data. Others are managing big data by extending their data management skills and their portfolios of data management software. This empowers them to automate more business processes, operate closer to real time, and through analytics, learn valuable new facts about business operations, customers, partners, and so on.

Managing big data brings together old and new technologies and practices.

The result is **big data management (BDM)**, an amalgam of old and new best practices, skills, teams, data types, and home-grown or vendor-built functionality. All of these are expanding and realigning so that businesses can fully leverage big data, not merely manage it. At the same time, big data must eventually find a permanent place in enterprise data management.

BDM is well worth doing because managing big data leads to a number of benefits. According to this report's survey, the business and technology tasks that improve most are analytic insights, the completeness of analytic data sets, business value drawn from big data, and all sales and marketing activities. BDM also has challenges, and common barriers include low organizational maturity relative to big data, weak business support, and the need to learn new technology approaches.

Half of organizations manage big data today, some with the full range of big data types.

Despite the newness of big data, half of organizations surveyed are actively managing big data today. For a quarter of organizations, big data mostly takes the form of the relational and structured data that comes from traditional applications, whereas another quarter manages traditional data along with big data from new sources such as Web servers, machines, sensors, customer interactions, and social media.

Big data, big decision: Manage it with existing systems or with a new one built for big data?

A quarter of surveyed organizations have managed to scale up preexisting applications and databases to handle burgeoning volumes of relational big data. Another quarter has gone out on the leading edge by acquiring new data management platforms that are purpose-built for managing and analyzing multi-structured big data. Many more are evaluating such big data platforms now, creating a brisk market of vendor products and services for managing big data.

Expect growing use of Hadoop, CEP, NoSQL, in-memory, and in-database technologies.

According to the survey, the Hadoop Distributed File System (HDFS), MapReduce, and various Hadoop tools will be the software products most aggressively adopted for BDM in the next three years. Others include complex event processing (for streaming big data), NoSQL databases (for schema-free big data), in-memory databases (for real-time analytic processing of big data), private clouds, in-database analytics, and grid computing.

Successfully managing big data demands new skills, and perhaps, new hires.

Organizations are adjusting their technical best practices to accommodate BDM. Most are schooled in extract, transform, and load (ETL) in support of data warehousing (DW) and reporting. Preparing big data for analytics is similar, but different. Organizations are retraining existing personnel, augmenting their teams with consultants, and hiring new personnel. The focus is on data analysts, data scientists, and data architects who can develop the applications for data exploration and discovery analytics that organizations need for getting value from big data.

This report accelerates users' understanding of the many options that are available for big data management (BDM), including old, new, and upcoming options. The report brings readers up to date so they can make intelligent decisions about which tools, techniques, and team structures to apply to their next-generation solutions for BDM.

Introduction to Big Data Management

Definitions of Data Management and Big Data

Big data management is about two things—big data and data management—plus how the two work together to achieve business and technology goals. To get us all on the same page, let's start with definitions of both, and then bring them together.

Big data is about size first, then diverse data types and streaming.

The base definition of "big data." Big data is first and foremost about data volume, namely large data sets measured in tens of terabytes, or sometimes in hundreds of terabytes or petabytes. Before the term *big data* became common parlance, we talked about very large databases (VLDBs). VLDBs usually contain exclusively structured data, managed in a database management system (DBMS). In many organizations, big data and its management follow the VLDB paradigm.

The extended definition of big data. In addition to very large data sets, big data can also be an eclectic mix of structured data (relational data), unstructured data (human language text), semi-structured data (RFID, XML), and streaming data (from machines, sensors, Web applications, and social media). In this report, the term *multi-structured data* refers to data sets or data environments that include a mix of these data types and structures.[1]

Data management includes many user disciplines and vendor tool types.

Data management. This involves the collection and storage of data, plus its processing and delivery—whether traditional data, new big data, or both. Processing can be extensive, especially when data is repurposed for a use differing from that of its origin (as is common in business intelligence [BI], data warehousing [DW], and analytics). Data management is a broad practice that encompasses a number of data disciplines, including data warehousing, data integration, data quality, data governance, content management, event processing, database administration, and so on.

BDM applies old and new skills and tools to managing big data.

Big data management (BDM). This is where data management disciplines, tools, and platforms (both old and new) are applied to the management of big data (in the base definition or the extended one). Traditional data and new big data can be quite different in terms of content, structure, and intended use, and each category has many variations within it. To accommodate this diversity, software solutions for BDM tend to include multiple types of data management tools and platforms, as well as diverse user skills and practices.

Usage Rates for Big Data Management Today

Big data management is important to companies and other organizations that have big data to manage, but big data is still relatively new, so how many organizations actually have it? To quantify this issue, the survey for this report presented the above definitions of big data and data management, and then asked: "Does your organization have big data, as defined above?" (See Figure 1.)

Most organizations manage some form of big data today.

Big data regularly focuses on structured data. The survey bears this out; a quarter of respondents (26%) reported that their big data is mostly structured, as in the base definition above.

The full range of big data is being managed by real-world organizations today. Thirty-one percent of respondents reported that their big data is fairly diverse, as in the extended definition.

BDM is mainstream, not a minority practice.

Most organizations are managing some definition of big data today (57%). This is true if you combine responses for the base definition of big data (26%) and the extended one (31%). This indicates that big data and its management have crossed into mainstream usage.

[1] For more detailed definitions of big data, see the TDWI Best Practices Report *Big Data Analytics* (September 2011), online at tdwi.org/bpreports.

For many organizations, big data has not yet arrived. More than a third of respondents (38%) said they don't yet have big data, in any definition. Oddly enough, the survey population for this report has unusually large percentages of respondents from the two kinds of organizations that are most prone to big data: namely, midsize-to-large Internet firms and very large corporations (with $10 billion of annual revenue or greater). Without these concentrations in the survey population, the percent of organizations without big data would no doubt have been higher, but still less than a majority.

Does your organization have big data, as defined above?

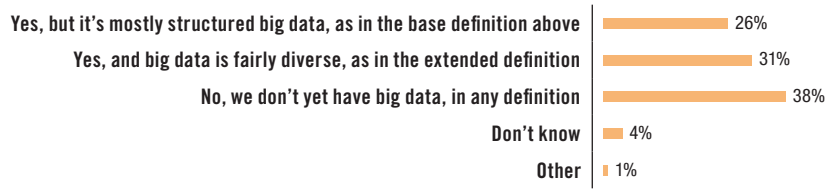


Figure 1. *Based on 461 respondents.*

Why Big Data and Data Management are Colliding

Structured data from applications is a common form of big data, although it's not new.

Remember the buzz term *eBusiness*, which was trendy during the 1990s? eBusiness was about automating as many organizational processes as possible with software. In that spirit, many companies and other organizations deployed applications in great numbers, starting in the mid-'90s. Today, we no longer use the term eBusiness because we assume that an organization of any size or complexity should have numerous applications for the sake of efficiency and competitiveness. A consequence of the post-eBusiness era is that many organizations now have massive volumes of application data to manage and to leverage for business value. Although organizations have the skills for structured data (which is what comes out of most operational applications), today's unprecedented data volume and speed of generation make big data management a challenge.

Big data can be industry specific, such as unstructured text in insurance, healthcare, and government.

Big data comes from many sources, in many formats. Some industries have large, valuable stores of unstructured data, typically in the form of human language text. For example, the claims process in insurance generates many textual descriptions of accidents and other losses, plus the related people, locations, and events. Most insurance companies process this unstructured big data using technologies for natural language processing (NLP), often in the form of text analytics. The output from NLP may feed into older applications for risk and fraud analytics or actuarial calculations, which benefit from the larger data sample provided via NLP.

Sensor data and other machine data are new and large, and they enable new applications.

Sensors are coming online in great numbers as a significant source for big data. For example, robots have been in use for years in manufacturing, but now they have additional sensors so they can perform quality assurance as well as assembly. For decades, mechanical gauges have been common in many industries (such as chemicals and utilities), but now the gauges are replaced by digital sensors to provide real-time monitoring and analysis. GPS and RFID signals now emanate from mobile devices and assets—ranging from smartphones to trucks to shipping pallets—so all these can be tracked and controlled precisely.

BDM's toughest challenges simultaneously involve scale, speed, and diverse data types.

The swelling swarm of sensors worldwide (plus the extended "Internet of Things") produces outrageous volumes of machine data, which is further compounded by the fact that many sensor sources stream big data in real time and have schema-free data formats. On the bleeding edge, the most extreme challenge to data management is to capture and manage large volumes of data that is arriving continuously in real time and in multi-structured formats.

Business and Technology Drivers behind Big Data Management

Big data just gets bigger. It's important to beef up data management infrastructure and skills as early as possible. Otherwise, an organization can get so far behind from a technology viewpoint that it's difficult to catch up. From a business viewpoint, delaying the leverage of big data delays the business value. Similarly, capacity planning is more important than ever, and should be adjusted to accommodate the logarithmic increases typical of big data.

Embrace big data ASAP to keep pace with its growth, get a business return, and fold it into enterprise data architecture.

Resistance is futile: big data will be assimilated into enterprise data. You have to start somewhere, even if it's a data management silo devoted to one form of big data. Typical silos manage Web logs, sensor and machine data logs, and persisted data streams. Yet, it's also important to determine how each form of big data will eventually fit into an overall architecture for enterprise data.

Leverage big data, don't just manage it. It costs money to collect and store big data, so don't let it be a cost center. Look for ways to get business value from big data. As you select data platforms for managing big data, consider low-cost new ones and open source.

Get business value from big data by analyzing it and by combining it with traditional data.

Advanced analytics is the primary path to business value from big data. This fact is so apparent that there's even a name for it: *big data analytics*. In many ways, the current uptick in advanced analytics among user organizations is driven by the availability of new big data, plus the new business facts and insights that can be learned from its study.

Joining big data with traditional data is another path to value. For example, so-called 360-degree views of customers and other business entities are more complete and bigger when based on both traditional enterprise data and big data. In fact, some sources of big data come from new customer touchpoints (mobile apps, social media) and so belong in your customer view.

Use big data to create new applications and extend older ones.

Big data can enable new applications. For example, in recent years, a number of trucking companies and railroads have added multiple sensors to each of their fleet vehicles and train cars. The big data that streams from sensors enables companies to more efficiently manage mobile assets, deliver products to customers more predictably, identify noncompliant operations, and spot vehicles that need maintenance.

Big data can extend older applications. This includes any application that relies on a 360-degree view, as mentioned above. Big data can also beef up the data samples parsed by many analytic applications, especially those for fraud, risk, and customer segmentation.

> **USER STORY SMALL MESSAGES ABOUT CONSUMER BEHAVIOR ADD UP TO BIG DATA.**
>
> "We provide a comprehensive set of marketing services, applications, and data solutions for our clients," said Leo D. Davis III, vice president of technology services at Epsilon. "Many of our solutions leverage big data, much of it coming and going in real time.
>
> "For example, with a typical e-mail campaign, we can identify when a consumer engages with a marketing e-mail. We may then update the consumer's profile and create follow-up messages based on their level of engagement. Likewise, if the consumer clicks a URL that's in the e-mail, Epsilon can identify an even greater level of interest in the message or brand from that consumer.
>
> "Other behaviors, such as visiting a Web page, redeeming a coupon, watching a video, checking an account balance, or seeking online help with a product are also valuable insights that can inform follow-up marketing actions. All these real-time events can be combined with aggregated big data and other information about consumers to create and maintain comprehensive, up-to-date marketing profiles of customers. In turn, these profiles can enable our clients to send highly personalized and relevant marketing communications in real time for greater returns and better customer experiences."

Problems and Opportunities for Big Data Management

User perception says that BDM is mostly an opportunity today, not a problem.

In recent years, TDWI has seen many organizations adopt new vendor platforms and user best practices that enabled them to overcome some of the performance issues with big data that dogged them for years, especially data volume scalability and real-time data processing. With that progress in mind, this report's survey asked: "Is the management of big data mostly a problem or mostly an opportunity?" (See Figure 2.)

The vast majority consider BDM an opportunity (89%). Conventional wisdom today says that big data enables data exploration and predictive analytics to discover new facts about customers, markets, partners, costs, and operations.

A tiny minority consider BDM a problem (11%). No doubt, big data presents technical challenges due to its size, speed, and diversity. Data volume alone is a showstopper for a few organizations.

Is the management of big data mostly a problem or mostly an opportunity?

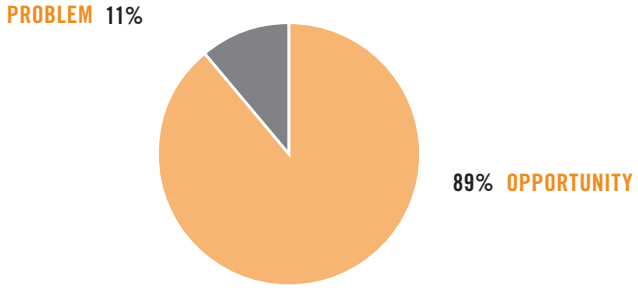


Figure 2. *Based on 461 respondents.*

Benefits of Big Data Management

To determine the most compelling reasons for managing big data, the survey asked respondents: "If your organization were to successfully manage and leverage big data, which business and technology tasks would improve?" (See Figure 3.)

The key beneficiaries of BDM are analytics, data sets, business value, and sales and marketing activities.

All things analytic stand to gain from big data management. At the top of the list, survey respondents selected data analytics (61%) more than any other answer. This isn't surprising considering that big data and analytics go together naturally, as noted earlier. According to survey respondents, common analytic applications can benefit from BDM, including fraud detection (21%) and risk quantification (16%).

Analytic data sets would benefit from big data management. Big data's large data samples and diverse range of data sources can lead to broader data sourcing for analytics (32%), more data for data warehousing (24%), and improved data staging for data warehousing (23%). In turn, the data sets improved via BDM can enable better information exploration (39%) and the development of new data-driven applications (22%).

BDM delivers value to the business. This is borne out in the ranking of survey responses, which place near the top of the list business value from big data (33%) and numerous and accurate business insights (34%). Similar benefits of BDM include business optimization (28%), addressing new business requirements (22%), and understanding business change (22%).

Sales and marketing activities improve with BDM. These include the recognition of sales and market opportunities (28%), definitions of churn and other customer behaviors (18%), better targeted social influencer marketing (16%), and understanding consumer behavior via clickstreams (16%), as well as related analytic applications such as customer-base segmentation (27%) and sentiment analytics and trending (24%).

If your organization were to successfully manage and leverage big data, which business and technology tasks would improve? Select seven or fewer.

Figure 3. *Based on 2,583 responses from 461 respondents; 5.6 responses per respondent, on average.*

Barriers to Big Data Management

Big data management has benefits, as we just saw. Yet, it also has barriers. To get a sense of which problems are more likely than others, this report's survey asked respondents: "What problems hinder the successful management of big data in your organization?" (See Figure 4.)

The most common barriers to BDM are low maturity, weak business support, and embracing new design paradigms.

Being new to big data and its management is the biggest challenge users face. When an organization is new to big data, it typically has (relative to managing big data) inadequate staffing or skills (40%), inadequate data management infrastructure (23%), and immaturity with new data types and sources (22%). The cure is to dive in with training and new hires (or consultants, more likely), then work through the learning curve, as with any new project type.

Serious BDM efforts are unlikely without proper business support. It's difficult for any new project type to get off the ground when it lacks governance or stewardship (33%), business sponsorship (33%), or a compelling business case (27%).

Solution design and architecture can be challenging, but not a showstopper. It takes time and angst to work through data integration complexity (30%) and the architecture of a big data management system (25%), but it's doable for teams with solid data management experience. In a related issue, it's difficult to determine big data's role in enterprise data architecture if you don't have one (25%). This is one reason why many BDM solutions are silos.

Some problems aren't much of a problem at all. A few issues ranked so low in the survey that we should consider them non-issues, namely loading large data sets (13%), fast processing of queries (9%), scalability with big data (7%), and network bandwidth (4%).

What problems hinder the successful management of big data in your organization? Select seven or fewer.

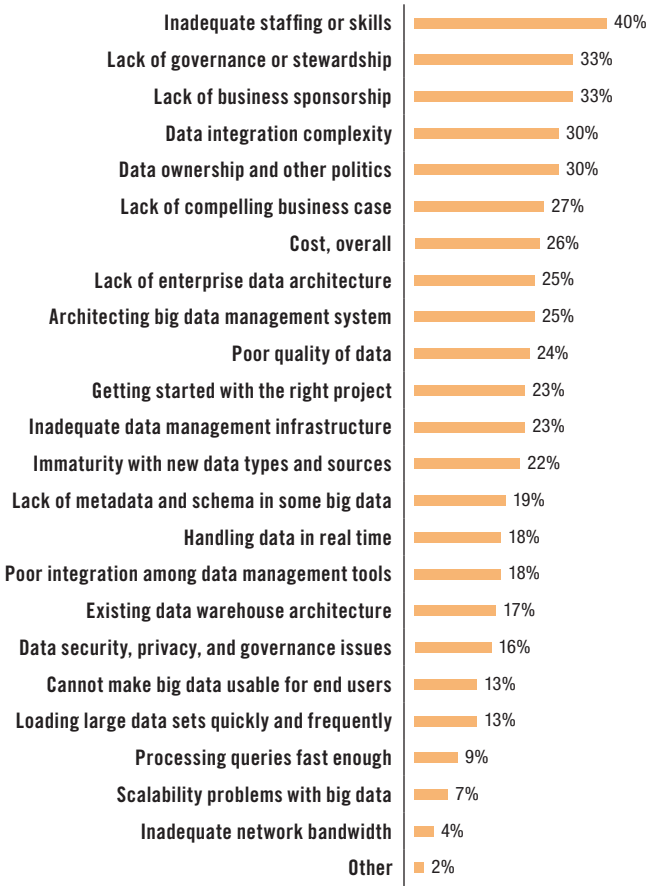


Figure 4. *Based on 2,287 responses from 461 respondents; 5 responses per respondent, on average.*

USER STORY SENSORS ON MOBILE ASSETS CAN IMPROVE BUSINESS EFFICIENCY, SAFETY, AND ASSET MAINTENANCE.

"There are thousands of locomotives and railcars in our network, all with RFID tags placed in multiple locations within a train," said Jerry D. Ward, an applications team member at Norfolk Southern. "We can collect information, primarily about car locations, safety issues, and maintenance issues. We get data every time a locomotive or railcar passes a milepost on a track.

"Right now, we just capture and store the information. But engineers are starting to use the data to develop predictive analyses. For example, if the temperature of a mechanical component rises to a certain temperature, a maintenance ticket is issued to prevent a part failure. Data from tags and other sensors about friction levels, temperatures around the cabin, oil level, pressure in the brake line—all these trigger preventative maintenance, which improves safety and on-time service."

The State of Big Data Management

Status of Implementations for Big Data Management

A number of user organizations are actively managing big data today, as seen in survey results. However, do they manage big data with a dedicated BDM solution, as opposed to extending existing data management platforms? To quantify these issues, this report's survey asked: What's the status of big data management in your organization today? (See Figure 5.) The survey also asked: When do you expect to have a big data management solution in production? (See Figure 6.)

BDM is a minority practice today, but will be a majority practice within three years.

Dedicated BDM solutions are quite rare, for the moment. Only 10% of respondents report having deployed a special solution for managing big data today. Most of these are very new (7%), whereas a few are relatively mature (3%), as seen in Figure 5. This is consistent with the 11% of respondents who already have a BDM solution in production, as seen in Figure 6.

In the short term, the number of deployed BDM solutions will double. Another 10% of respondents say they have a BDM solution in development as a committed project, as seen in Figure 5. This is consistent with the 10% who say they will deploy a dedicated BDM solution within six months, as seen in Figure 6.

Half of surveyed organizations plan to bring a BDM solution online within three years. In addition to the 10% over six months just noted, more solutions will come online in 12 months (20%), 24 months (19%), and 36 months (12%). If users' plans pan out, dedicated BDM solutions will jump from rare to mainstream within three years. But note that users' plans are by no means certain, because many projects are still in the prototyping or discussion stage (20% and 37%, respectively, in Figure 5).

Few organizations don't need a special solution for managing big data. Just a quarter report no plans at present for such a solution (23% in Figure 5); even fewer say they'll never deploy a BDM solution (6% in Figure 6).

What's the status of BDM in your organization today?

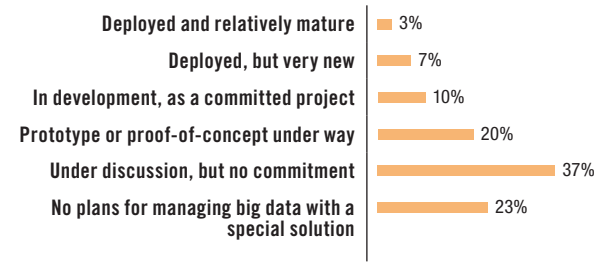


Figure 5. *Based on 461 respondents.*

When do you expect to have a BDM solution in production?

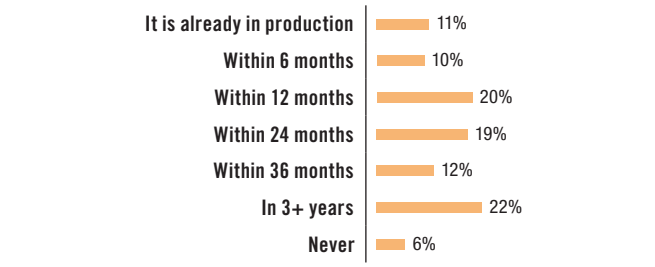


Figure 6. *Based on 461 respondents.*

USER STORY MANAGING BIG DATA CAN REVEAL SECURITY AND COMPLIANCE BREACHES.

"Our business unit for corporate security was the first part of the company to start managing big data for analytic purposes. That's because security was a prominent pain point," said a data specialist interviewed by TDWI for this report. "The security team today uses a vendor distribution of the Hadoop Distributed File System to collect massive data sets that come largely from application logs and system logs. They also have some open source analytic tools from Apache. By studying Hadoop data, they can spot unauthorized or suspicious accesses to systems. For example, one thing they do is to look at log events for server reboots and crashes, then correlate those with IP addresses and other log information to determine whether each was purely an internal event or externally driven. These analyses have really tightened up our security, risk reduction, and compliance."

Strategies for Managing Big Data

Different organizations take different technology approaches to managing big data. On one hand, a "fork in the road" decision is whether to manage big data in existing data management platforms or to deploy one or more dedicated solutions just for managing big data. On the other hand, some organizations don't have or say they don't need a strategy for managing big data. (See Figure 7.)

If you're committed to big data, you need a strategy for managing it.

Half of organizations have a strategy for managing big data. This is true whether the strategy involves deploying new data management systems specifically for big data (20%) or extending existing systems to accommodate big data (31%). One survey respondent selected "other" and added the comment: "Our big data strategy is a core competency for our business."

The other half doesn't have a strategy, for various reasons. Some don't have a strategy because they're not committed to big data (15%). "The business value is questionable," said one respondent. Others lack a strategy for managing big data, as yet, even though they know they need one (30%). "Once our POC completes, strategy can be defined."

A lack of maturity can prevent a strategy from coalescing. One survey respondent added the comment: "We don't know enough yet to determine a strategy." Another commented: "Our data management is in a nascent stage. [It] needs to mature before a strategy becomes clear."

BDM strategy can rely on existing platforms, additional ones, or both.

As with many strategies, hybrids can be useful. According to one respondent: "[We'll use] a blend of extending existing [platforms] and deploying new [ones] in a hybrid mode." Another echoed that strategy, but turned it into an evolutionary process: "[We'll] extend existing systems now, and add new and better systems later."

Strategy should be part business, part technology. Ideally, BDM strategy should start with upper management, who determines that big data and its management supports business goals enough that the business should in turn support big data management. Without this business strategy in place first, technology strategies for BDM are putting the cart before the horse.

Which of the following best describes your organization's strategy for managing big data?

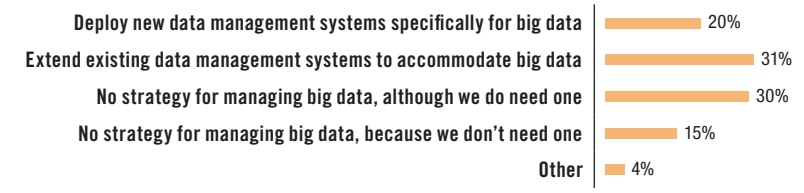


Figure 7. *Based on 461 respondents.*

The Success of Big Data Management

BDM success is about hitting both technology and business goals.

Managing big data successfully on a technology level is one thing. Managing big data so that it supports business goals successfully is a different matter. For example, the benefits of BDM noted in the discussion of Figure 3 include business goals such as more numerous and accurate business insights, greater business value from big data, and business optimization.

To estimate metrics for these measures of success, this report's survey asked two related questions: "How successful has your organization been with the technical management of big data? How successful has big data management been in terms of supporting business goals?" (See Figures 8 and 9.) Note that these questions were answered by a subset of 188 survey respondents (which is 41% of the total respondents) who claim they've managed one or more forms of big data. Hence, their responses are strongly credible, as they are based on direct, hands-on experience.

Big data management (BDM) is moderately successful for both technology and business. A clear majority of respondents feel BDM (which they've done hands-on) is moderately successful on both technology and business levels (65% in Figure 8 and 64% in Figure 9, respectively). This is good news, considering that BDM is a relatively new practice. It also suggests that BDM can balance both technology and business goals.

Few consider BDM to be highly successful. This is the case for both technology (11%) and business (12%). No doubt, BDM will mature into higher levels of success.

Roughly a quarter of respondents consider BDM to be not very successful. Again, this is true for both technology (24%) and business (24%). The lack of success in some organizations may be due to the newness of BDM. At this point in BDM, we've mostly seen organizations' first attempts and early implementation stages; as these mature, success ratings will likely improve.

How successful has your organization been with the technical management of big data?

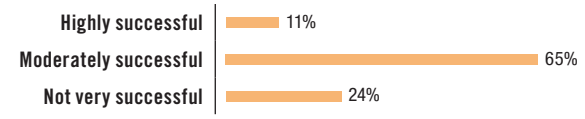


Figure 8. *Based on 188 respondents who have experience managing big data.*

How successful has big data management been in terms of supporting business goals?

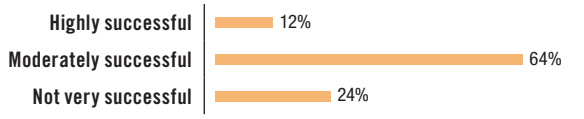


Figure 9. *Based on 188 respondents who have experience managing big data.*

USER STORY DON'T LET BIG DATA TAKE OVER IT SYSTEMS THAT HAVE OTHER PRIORITIES.

According to a data architect interviewed for this report: "We're collecting big data today, but we haven't done much with it yet. Most of our storage today is on relational databases, with a little VSAM [virtual access storage method] and other legacy databases. And we keep an archive dump in ASCII format. All these and other systems have scaled up to growing data volumes very well, although that's not what transactional systems like this were designed for.

"As big data volumes accelerate, we fear we'll end up squandering the capacity of our mission-critical transactional systems on big data. To avoid that, we've decided to look for an additional system that can specialize in managing big data so we can keep older systems focused on what they do best. We'd also like to manage as much big data as possible in one place, so it's more easily shared across teams; so we move less big data around; and so we have a single version of the truth for analytics. So far, all that points to Hadoop as the additional system, and we've just started testing it."

Organizational Practices for Big Data Management

The survey responses discussed in this section of the report come from a subset of survey respondents who report that they have experience managing big data. Based on direct, hands-on experience, their responses provide a credible glimpse into emerging best practices for big data management. As with the total survey population, this subset is dominated by BI/DW professionals and their bias is reflected in survey results.

Big Data Ownership and Sponsorship

As we'll see in subsequent sections of this report, big data management is performed by a variety of specialists on a variety of teams. These specialists and teams come together in a variety of ways. To get us started sorting out how they all work together, this report's survey asked: Who provides your primary big data environment?" (See Figure 10.)

Many departments and groups have their own big data platforms.

Data warehouse group (52%). Big data and analytics go hand in hand, as we've already seen. Big data can also be a source for reporting and performance management. So it makes sense that a data warehouse environment would include its own specialized platform for big data management, typically as a complement to and extension of the core warehouse and other data platforms in that environment. (This report will later drill into how data warehouse architectures are evolving to accommodate big data.)

Central IT (45%). In many organizations, IT provides infrastructure (but not applications), and it's up to independent application teams to build and/or deploy applications atop that infrastructure. In such organizations, IT infrastructure is expanding to include big data platforms, similar to how IT has for years provided data storage via SAN (storage area network) and NAS (network attached storage) systems. This is a good approach for organizations that hope to avoid the big data silos that result when departments deploy their own platforms for big data management.

One trend is toward big data platforms supplied by IT as infrastructure.

Miscellaneous departments (63%). This could be a business unit (23%), research or analysis group (21%), or an application group (19%). This makes sense (if we ignore the silo problem), because big data and its analysis are often a departmental affair.

For example, TDWI interviewed a data management professional who works in the quality assurance department of a consumer electronics firm. He manages a few billion quality and test records collected from sensors on the robots that assemble products. The data is analyzed for product, supply chain, and efficiency improvements. Although big in the extreme, this data is purely a departmental asset, managed on a departmental system.

Third parties (8%). A few respondents added comments to the effect that their firm has outsourced its entire data center (including all big data and its management) to a third party. Others have just outsourced big data, typically to a cloud provider.

Who provides your primary big data environment? Select up to three.

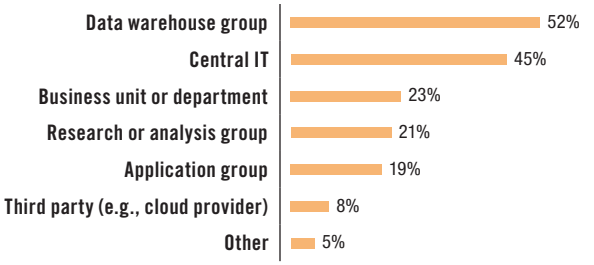


Figure 10. *Based on 323 responses from 186 respondents who have experience managing big data; 1.7 responses per respondent, on average.*

Job Titles and Team Structures for Big Data Management

The types of people who manage and use big data in their work are surprisingly diverse. To form an inventory of these people, this report's survey told respondents: "Enter the titles of people who manage big data in your organization." Respondents' responses were manually normalized into standard job titles. The result is a rather long list of diverse job titles. No single team structure arises from the list, but subsets within the list suggest that teams could be focused around data, applications, IT, or business-domain functions, as well as combinations. (See Figure 11.)

Data architects. Architects of various types are prominent in big data management, especially those focused on data, as seen in the job titles data architect (16%) and data warehouse architect (3%). The related job title ETL developer (2%) also applies here.

Many diverse job titles and teams are already involved with managing big data.

Data analysts. Various types of analysts are also prominent in BDM, especially those with a data focus, as in the data analyst (10%) and business analyst (6%). Related to data analysis, the relatively new job title "data scientist" fared well in the survey (6%).

Managers and directors. Data management requires IT management, as seen amply in the job titles BI/DW manager (8%), IT director/manager (5%), project/product manager (3%), and CIO or similar executive (2%). In a similar vein, business managers who are also data stewards (3%) govern big data, as they would any data.

Technical administrators. Managing big data requires administration, as seen in the titles database administrator (DBA, 7%) and systems analyst/architect (6%).

Applications specialists. When an application generates or handles big data, application specialists must manage big data. In other words, big data management is not just for data specialists. In fact, there's already a tradition of Web application teams managing big data, and that's exactly where Hadoop and other technologies and practices for big data came from. In the survey, we see this in job titles such as application developer (6%) and application architect (5%).

Domain experts. Many end users (who are business people or other non-IT personnel) need to manage big data because it's integral to their work. For example, as with any data, marketers manage big data in support of customer analytics, customer segmentation, profitability analytics, customer profiling, and so on. Many scientists design new products and services based on big data, as with pharmaceutical researchers and healthcare providers.

Enter the job titles of people who manage big data in your organization.

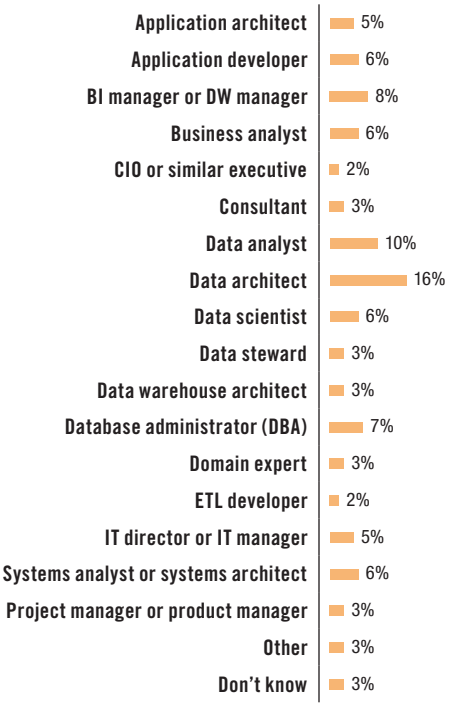


Figure 11. *Based on 297 responses from 166 respondents who have experience managing big data; 1.8 responses per respondent, on average.*

Collaborative Practices around Big Data Management

Remember the definition of *data management* (DM) presented early in this report? It states that DM is a broad practice that encompasses a number of disciplines, including data warehousing, data integration, data quality, data governance, content management, event processing, database administration, and so on. It's often the case that preparing any data (including big data) for use with different departments and applications requires a combination of DM tools, techniques, and teams. Hence, the collaboration of multiple DM teams and the integration of their tools and platforms can be critical success factors for big data management.

BI/DW and related disciplines have the strongest involvement with BDM.

To determine which DM disciplines are involved with BDM (and hence which need to be coordinated and integrated), the survey asked: "Which data management disciplines and teams are involved in managing big data?" Responses to the question were structured, such that respondents ranked a DM discipline as strongly involved, moderately involved, or not involved. The survey results charted in Figure 12 are sorted by rankings for "strongly involved."

Business intelligence and data warehousing. These are the most common disciplines for strong involvement with managing big data, along with the closely related field of data integration. Note that these disciplines are rarely absent.

Administrators and architects. As seen earlier in the list of BDM job titles, data architects are strongly involved with designing big data environments, just as DBAs are integral to managing them.

Quality and governance. It's a red flag that data quality and data governance are only moderately involved with big data management at present. As with all enterprise data, big data has problems and opportunities concerning quality and governance.

Application disciplines. Disciplines far removed from BI/DW and somewhat removed from data are not often strongly involved, but instead are more often moderately involved or not involved. This includes (at the bottom of the chart) content management, application integration, and application teams.

Which data management disciplines and teams are involved in managing big data? Select one answer per row.

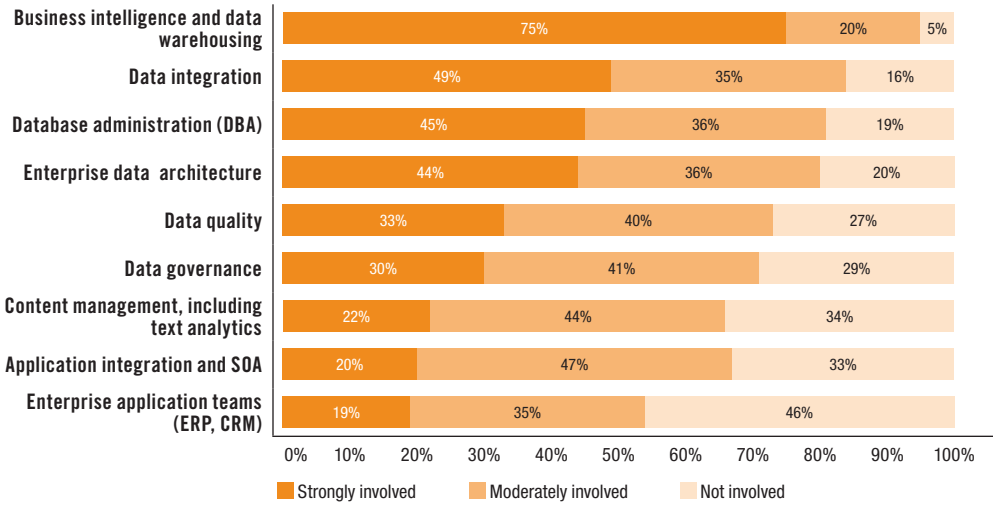


Figure 12. *Based on 189 respondents who have experience managing big data. Sorted by "strongly involved."*

EXPERT COMMENT MANY FORMS OF ANALYTICS APPLY TO BIG DATA.

"The ability to analyze big data provides unique opportunities for organizations in terms of the kinds of analysis they can do," said Fern Halper, TDWI Research director for advanced analytics. "For example, instead of being limited to certain kinds of data, organizations can start to incorporate text and geospatial data and other forms of unstructured data into their analysis. However, traditional business intelligence products weren't designed to handle this kind of data. This data can come from untrusted sources (think social media), it can be inaccurate, incomplete (think of a broken sensor), and real time. The algorithms you might have used on your desktop often need to be refactored to run on a big data infrastructure. Newer kinds of analytics such as text analytics will have to be utilized to extract data from text.

"Users should start with a business reason for analytics, then determine which form of analytics they need, and finally decide how data should be collected, managed, and processed for the chosen form of analytics. Organizations using big data analytics must make sure they have the skills to deal with the nuances of analyzing big data and the governance in place to deal with managing it. Starting with a proof of concept is generally a best practice."

Technical Practices for Big Data Management

BDM for Many Different Data Types and Structures

Structured data is still in the lead, with semi-structured data in hot pursuit.

Structured data retains its hegemony, even with diverse big data. At 88%, structured data is by far the most managed data type today, according to the survey (see Figure 13). In fact, we can safely assume that most of this structured data is actually relational, meaning that relational data is still very prominent. In turn, that means that DBMSs, SQL, and other tool types and technologies for relational data are important to managing big data.

Semi-structured data is the most prominent secondary data format. A number of data formats include a mixture of structured data, hierarchies, text, and so on. Common examples include documents that adhere to standards for XML, JSON, and RSS. Coincidentally, these documents are often used as formats for messages and events, so they may also be considered event data, which (along with semi-structured data) also ranked highly in the survey as a prominent secondary data format for big data.

Web data ranked low; social data ranked high.

Web data ranked surprisingly low. Web servers and Web applications have been with us for almost 20 years, and Web data is a common source of big data today. So it's surprising that almost half of survey respondents (45%) don't manage Web logs and clickstreams at all.

Social media data ranked surprisingly high. Social media Web sites are only a few years old, and it's only in the last three years that user organizations have started to collect social data for study. So it's surprising that so many organizations surveyed are already managing social data.

Text ranked low; sensor data ranked high.

Unstructured data still eludes many organizations. All forms of unstructured data require highly specialized technologies and skills, which may explain why approximately half of organizations surveyed still don't manage unstructured big data in the form of human language or audio/video (45%), personal productivity files (43%), or e-mail (53%).

Sensor data, machine data, and geospatial data have all arrived. Approximately half of user organizations surveyed are managing and leveraging these data formats today, and they ranked well as somewhat prominent secondary data formats.

Scientific data and surveillance data are lagging behind at the moment. Most user organizations are not capturing and storing this data today (in the 72% to 80% range).

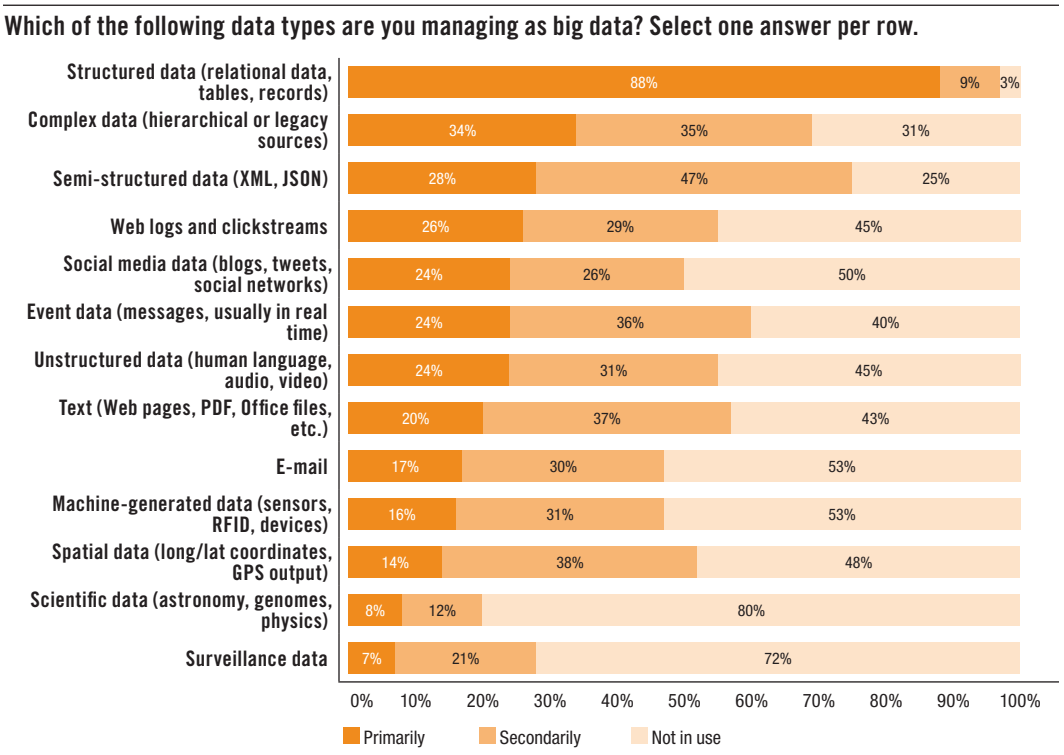


Figure 13. *Based on 188 respondents who have experience managing big data. Sorted by the "primarily" column.*

Storage Strategies for Big Data Management

An obvious question for data specialists who manage big data is: Where do you put it all? To get at the heart of the matter, the survey asked: "What physical storage media do you use for managing big data?" Responses to the question were structured such that respondents ranked each storage medium as primary, secondary, or not in use. The survey results charted in Figure 14 are sorted by rankings for primary media.

Most big data is stored on traditional drives, but solid-state drives and in-memory functions are gaining.

Disk drives are the primary medium for storing big data. This is not a surprise, since traditional disk drives are by far the most common storage medium in use today. They have been for years, and will remain so for years.

Solid-state drives (SSDs) are an important secondary medium for big data. Survey results show that SSDs already have a firm foothold; over half of users use them today. Traditional, spinning hard drives have gotten larger, cheaper, and more reliable—but no faster. Based on Flash memory, with no moving mechanical parts, SSDs provide a speedy—albeit expensive—alternative. A number of data warehouse appliances already incorporate SSDs because they are significantly faster than traditional drives when it comes to queries.

An emerging best practice is to configure a server or appliance with a mix of drive types. The bulk of big data is stored on commodity-priced traditional drives, while "hot data" (which is being accessed frequently by high-value applications and therefore demands the best query performance) is moved to high-priced SSDs.

In-memory functions for big data become progressively more feasible in terms of data volume and cost. Dropping server memory prices and 64-bit computing have enabled an influx of in-memory databases for analytics. For example, the database that's in memory may be a cube for OLAP, a table of metrics for performance management, analytic models that are re-scored in real time, or an entire relational database to serve both OLTP and OLAP functions. In-memory databases are very fast because they avoid most I/O operations. A few survey respondents entered comments explaining that in-memory data management is even faster than SSDs for queries and some forms of data processing.

Off-premises storage is a viable option for big data management. A surprisingly high 40% of survey respondents report using off-premises systems for storing and managing data, including hosted data centers and clouds.

Traditional backup media play a diminishing role with big data. Backing up hundreds of terabytes of big data is not all that feasible, and restoring it is even less so. As backup becomes less feasible, organizations are moving to alternate strategies such as data management platforms that are highly available and purposefully redundant with storage.

What physical storage media do you use for managing big data? Select one answer per row.

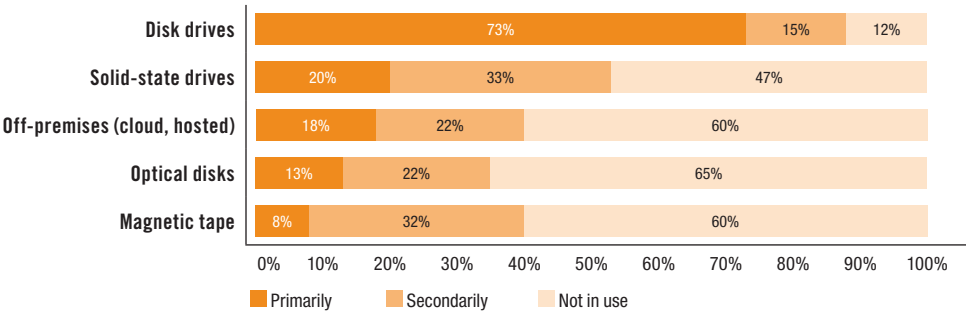


Figure 14. *Based on 189 respondents who have experience managing big data.*

Volumes of Big Data Being Managed

Everyone wants to know: How big is big data? What's the volume managed today? How will that change in the future? To quantify these issues, this report's survey asked: "What's the approximate total volume of big data (by any definition of big data) that your organization manages, both today and in three years?" (See Figure 15.)

10 to 99 terabytes is the big data norm today.

Many organizations have broken the 10-terabyte barrier. In fact, the 10-to-99 TB range received more survey responses than other ranges, indicating that it's the norm for today's big data volumes. Within three years, 100 TB will become the norm.

Smaller data sets will become less common as they grow into larger ones. In forecasting big data volumes for three years from now, survey respondents project far fewer data sets in the 1 TB and 1-to-9 TB ranges. This is natural as big data repositories mature into greater volume. TDWI surveys on big data analytics (mid-2011) and high-performance data warehousing (mid-2012) showed near-identical declines in sub-10-TB data volumes.[2]

Many firms anticipate breaking the one-petabyte barrier within three years.

Conversely, very large data sets are rare today, but will become more numerous. Looking at data volumes in the 100 TB and greater range, many more organizations will manage big data volumes in this range within three years (51%) as compared to today (28%). Furthermore, almost a quarter of users surveyed (23%) anticipate breaking the one-petabyte barrier within three years.

[2] Access the TDWI Best Practices Reports that discuss these surveys and related points at tdwi.org/bpreports.

What's the approximate total volume of big data (by any definition of big data) that your organization manages, both today and in three years? Select one answer per row.

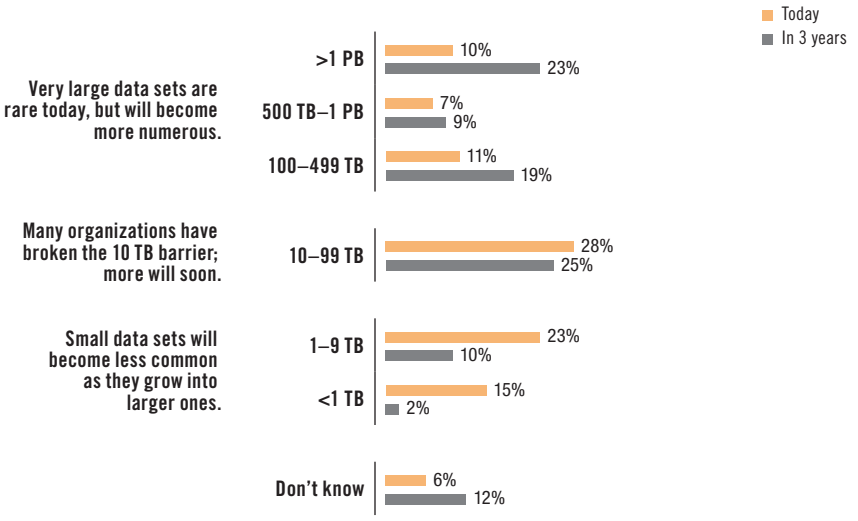


Figure 15. *Based on 188 respondents who have experience managing big data.*

EXPERT COMMENT ENGINEER BIG DATA INTO AN ENTERPRISE INFORMATION ARCHITECTURE.

"The promise of big data analytics reflects a common trend of data reusability for alternate purposes, however massive the scale might be," said David Loshin, the president of Knowledge Integrity, Inc.

"One risk is falling into the old pattern of segregating the application and its data from the rest of the organization, creating islands of data. Even an environment that supports bleeding-edge technologies can be prone to implementing analytical applications within a virtual silo.

"…Big data analytics cannot operate within a vacuum; the results need to be integrated with existing reports, customer profiles, data warehouses, and other aspects of the traditional BI stack. Maintaining fitness for purposes and sustainability implies a more governed set of best practices for data management with respect to data utility across the enterprise. Use big data (as well as other types of analytics), but do it in alignment with the existing BI strategy. Plan your design and development to use existing methods for defined data models already present in the organization's information architecture."[3]

Analytic Practices and Big Data Management

Organizations want more business value from big data, and analytics is an important route to value. Conventional wisdom now says that big data should be leveraged via advanced analytics rather than treated as a cost center. In fact, that's why big data and analytics go together. The catch, however, is that big data must be managed so it's in the proper structure and condition for the data exploration and discovery analytics that most user organizations want. Let's look at a few of the emerging practices for managing and preparing big data for analytic applications.

[3] For more insights from David Loshin (www.knowledge-integrity.com), see the TDWI Checklist Report this expert comment comes from: *Strategic Planning for Big Data Management*, May 2012, available at tdwi.org/checklists.

Approaches to Managing Big Data for Analytics

Most of the time, the overarching goal of BDM for advanced analytics is to maintain a large data store of fairly "raw" data—that is, source data straight out of source systems, with little or no alteration. There are good reasons for this:

For open-ended, discovery-oriented analytics, manage big data in its original form.

Most applications of advanced analytic methods enable discovery. This is especially true of data mining technologies and similar clustering or correlation algorithms. The user is usually looking for facts about the business or related entities (customers, partners) that were previously unknown. Or they may be looking for fraud, new customer segments, hidden costs, correlations among people, affinities among products, and so on. Such discoveries often depend on highly detailed source data. If the source data is merged, transformed, or standardized, the details can be obscured or lost. This is why all data management (including BDM) must maintain source data in its original state when the data will apply to discovery analytics.

Often, an analytic discovery is expressed as a unique ad hoc data set in a unique model. For example, a business analyst, data analyst, data scientist, or similar user may use a query tool or hand-coded SQL to run iterative ad hoc queries against structured and relational big data in search of a special data subset. A unique data set can reveal important insights such as a new customer segment, widgets from select suppliers that tend to fail in certain field conditions, or a class of transactions with an anomaly that suggests fraud. In this SQL-based analytic method, SQL transforms data, standardizes data, and models the result set. Hence, there's no need to preprocess data up front via ETL, data quality functions, or data modeling. In fact, such a priori processing could impose data standards and data models on the data that would get in the way of the desired discovery. Again, it's best to manage most big data in its original form so it can be repurposed at analysis time, thus enabling analysts to go any direction the discovery mission suggests.

Reporting and analytics are different; managing data for each is, too.

Managing big data for analytics is not the same as managing DW data for reporting. In fact, the two are almost opposites, as you probably surmised from the last two paragraphs. For example, reporting is about seeing the latest values of the numbers that you track over time via a report. Obviously, you know the report, the business entities it represents, and the data warehouse that feeds the report. An analysis is more about discovering variables you don't know, based on data that you probably don't know very well. Also, a report requires a solid audit trail, so its data must be managed with well-documented metadata and possibly master data, too. Since most analyses have no expectation of an audit trail, there's no need to manage one. That's just a sampling of the differences. The point is to embrace BDM for analytics as a unique practice that doesn't follow all the strict rules we're taught for reporting and data warehousing.

Big data needs data standards, but different ones compared to enterprise data.

Classic practices and standards for enterprise data still apply to big data, but only late in the analytic process. After an analyst gets an epiphany from the big data being explored, the analyst shares the discovery with a few business people and other analysts. But don't stop there. The analyst should also take that epiphany to the BI/DW team so they can determine whether the findings should be operationalized in metrics, reports, and data structures for the warehouse. After all, analytics can be a first step for BI/DW development. At this point, the BI/DW team applies its best practices and standards for ETL, data quality, and data modeling, as they take what started as analytic big data and transform it into enterprise data suited to reuse via the data warehouse.[4]

When you can, bring the analytic algorithm to big data, not the other way around. There's a long tradition of extracting data from a warehouse and other data stores, to be moved to a platform (and sometimes transformed into a new model) that's conducive to how an analytic tool works. Many tools for data mining and statistical analysis demanded this for years. This is a problem with the massive volumes

[4] For more details, see the TDWI Checklist Report *Data Requirements for Advanced Analytics*, September 2009, available at tdwi.org/checklists.

of big data, which are not trivial to move and transform just for one run of one analytic algorithm. Luckily, some analytic tool vendors and database management system vendors have partnered to enable in-database analytics. That means one or more analytic algorithms can run internally within the DBMS and process data there, without the need to move the data out of the DBMS. Hadoop technologies were built under this assumption; analytic and data processing tools (such as MapReduce, HBase, and Hive) can be layered over the Hadoop Distributed File System to process data in place without moving it to another platform. Furthermore, some vendors have versions of their analytic tools that can run as a stored procedure or user-defined function within another vendor's DBMS.

Diverse big data is subject to diverse processing, which may require multiple platforms. In general, manage big data on as few data management platform types as possible. This minimizes data movement, as well as avoids data synchronization and silo problems that work against the "single version of the truth." However, there are ample exceptions to this rule. As you expand into multiple types of analytics with multiple structures of big data, you will inevitably spawn many different types of data workloads. Because no single platform available today runs all workloads equally well, most DW and analytic environments are trending toward a multi-platform environment, as explained in a later section of this report.

BDM for Streaming and Other Real-Time Big Data

The pace of business continues to accelerate, such that speedy decisions based on fresh information have become a competitive advantage. This is true of many modern business practices, from operational business intelligence to business performance management, and from just-in-time inventory to facility monitoring.

Some forms of big data stream in real time. Streaming data shoots out of a number of systems in real time, including Web servers, robots and other machinery, sensors, social media, supply chains, RSS feeds, events, transactions, and customer interactions. This is one of the reasons big data is getting so big; high counts of small messages or streaming events can add up to big data. All these sources produce valuable information that should be captured and analyzed. Some of the information in a stream merits analytic processing in real time.

Streaming big data is easy to capture, but tough to process in real time.

Managing streaming big data in real time requires specialized technology. A successful tool must do several things with streaming data. It must capture each event from a stream, separate events of interest from noise events, make correlations with other streams and databases (especially data warehouses), react to some events in real time, and store most events for offline analytics. The technology in use must have high performance for each of these atomic units of work, so that the aggregate performance of the overall system is fast and scalable.

Real-time functionality at this extreme level is beyond the tools and platforms found in the average BI/DW technology stack, and even beyond the capabilities of most platforms designed for managing big data. For that reason, organizations wishing to get the most out of streaming big data typically complement their existing systems with a tool for complex event processing (CEP), which can be programmed to spot opportunities and problems in streaming data in real time.

CEP makes analytic correlations across multiple data sets. Most analytic methods involve correlations, associations, and relationships that can only be discovered when data comes from multiple sources. CEP excels with multi-data-source correlations, even when the sources are an eclectic mix of traditional enterprise sources (enterprise applications, relational databases) and new ones (streaming data, machine data, social media data, and NoSQL data platforms such as Hadoop). That's a long

list of old and new data sources, so a mature CEP platform must have the appropriate interfaces to support them all, along with the data models and standards that many of these sources assume.

CEP's extreme real-time functionality enables new business practices.

CEP enables immediate business action based on the analysis of streaming big data. For example, CEP empowers you to:

- Understand customer behavior so you can improve the customer experience as it's happening.
- Monitor and maintain the availability, performance, and capacity of interconnected infrastructures such as utility grids, computer networks, and manufacturing facilities.
- Identify compliance and security breaches, then halt and correct them immediately.
- Spot and stop fraudulent activity, even as fraud is being perpetrated.
- Evaluate sales performance in real time and take instant measures to achieve quotas.

Multi-Platform Architectures for BDM and Analytics

Analytics, big data, real time, and unstructured data present new DW workloads.

Workload-centric DW architecture. One way to measure a data warehouse's architecture is to count the number of workloads it supports. TDWI Research estimates that a little over half of data warehouses support only the most common workloads, namely those for standard reports, performance management, and online analytic processing (OLAP). The other half also supports workloads for advanced analytics, detailed source data, and real-time data feeds. The trend is toward the latter. In other words, the number and diversity of DW workloads is increasing due to organizations embracing big data, multi-structured data, real-time or streaming data, and data management and processing for advanced analytics.

Diversification of DW workloads leads to distributed architectures for DWs.

Distributed DW architecture. The issue in a multi-workload environment is whether a single-platform data warehouse can be designed and optimized such that all workloads run optimally, even when concurrent. More and more DW teams are concluding that a single-platform DW is no longer desirable. Instead, they maintain a core DW platform for traditional workloads (reports, performance management, and OLAP), but offload other workloads to other platforms. For example, data and processing for SQL-based analytics are regularly offloaded to DW appliances and columnar DBMSs. A few teams offload workloads for big data and advanced analytics to HDFS, MapReduce, and similar platforms. The result is a strong trend toward distributed DW architectures, where many areas of the logical DW architecture are physically deployed on standalone platforms instead of the core DW platform.

A distributed DW architecture is both good and bad. It's good if your fidelity to business requirements and DW performance leads you to deploy another data platform in your DW environment, and the new platform integrates well with others in the distributed architecture. But it's bad when disconnected systems proliferate uncontrolled, like the errant data marts we all fear. So far, the new generation of analytic databases and data management platforms are controlled by users far better than the marts of yore, but you still have to be diligent to avoid abuses.

Rearrange the acronym from EDW to DWE for "data warehouse environment," meaning multi-platform DW.

From the EDW to the multi-platform DWE. A consequence of the workload-centric approach is a trend away from the single-platform monolith of the enterprise data warehouse (EDW) toward a physically distributed data warehouse environment (DWE). A modern DWE consists of multiple platform types, ranging from the traditional warehouse (and its satellite systems for marts and ODSs) to new platforms such as DW appliances, columnar DBMSs, noSQL databases, MapReduce tools, and HDFS. In other words, users' portfolios of tools for BI/DW and related disciplines are diversifying aggressively. The multi-platform approach adds more complexity to the DW environment, but

BI/DW professionals have always managed complex technology stacks successfully. The upside is that users love the high performance and solid information outcomes that they get from workload-tuned platforms.

The survey results in Figure 16 indicate the usage of just a few of the many data platforms now available for a modern data warehouse environment.

Relational databases still rule DWEs, for the moment.

Relational database management systems still dominate among DW data platforms, whether SMP or MPP. For this survey population, the MPP computing architecture has pulled ahead of SMP; with other populations tested by TDWI, SMP had a slight lead. The overall trend is definitely toward MPP because of its obvious advantages for data operations, as compared to SMP, which is still the preferred architecture for operational and transactional applications.

Columnar databases and data (warehouse) appliances have recently become ubiquitous in DWEs. They're a natural fit, because both are built from the bottom up as SQL-based query machines, clearly with SQL-based forms of advanced analytics in mind. Both make good sense as easily deployed standalone platforms for extending a DWE, offloading SQL-based analytics, and managing forms of big data that are not appropriate to the core DW (especially detailed source data for analytics).

Non-SQL, non-relational platforms are coming on strong, and we need them.

It's not just SQL: Non-SQL data platforms have a good foothold in DWEs. On the one hand, SQL is back with a vengeance in the form of SQL-based analytics. On the other, many forms of analytics that users want do not require SQL, and they scale and perform well with other approaches. Furthermore, many forms of big data are schema-free, and hence are more appropriately managed on a non-SQL platform than on a relational DBMS. NoSQL DBMSs registered moderately in the survey, but HDFS is clearly being used as a primary data platform, despite zero usage by half of respondents.

What types of database management systems (DBMSs) and other data platforms are you using for big data management? Select one answer per row.

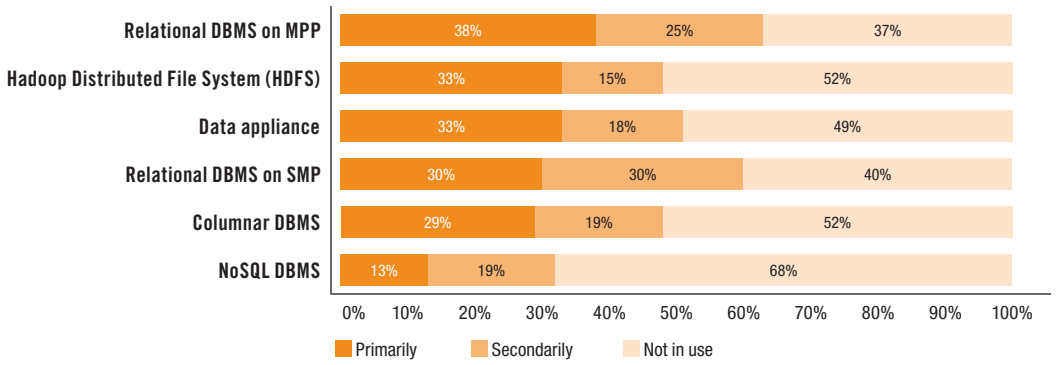


Figure 16. *Based on 189 respondents who have experience managing big data.*

USER STORY ONE BIG REPOSITORY FOR BIG DATA IS AN ARCHITECTURAL CHOICE.

"My team is very interested in big data because our telephone network generates tens of terabytes of detail records—CDRs and IPDRs—every year," said Ranko Petrovic, a data architect at Canadian telco Telus. "My team doesn't do billing, although CDRs and IPDRs assist with that. Instead, we're focused on network performance and capacity planning, which is mostly based on the analysis of big data consisting of CDRs and IPDRs.

> "What we're trying to do with our big data initiative (although we don't have anything in production yet) is to collect big data in a way that multiple teams can use. In our view, it would be a waste of resources if each team had its own copy. Plus, it's too easy for multiple repositories to present conflicting information, and a single big data repository is best if you want a complete view of customers, the network, and products.
>
> "In support of our fledgling big data initiative, we have drafted a reference document that describes our big data architecture. Assuming that the reference doc gets approved, it will set high-level guidelines for Telus' architects and development teams to follow when developing big data solutions. Development teams should comply with the reference big data architecture, and they should report deviations to the architecture team."

Future Trends in User Practices and Vendor Tools for Big Data Management

Good news: There are many options for managing big data.

Bad news: it's hard to know them all and select the best one.

By now, you've probably noticed that there are many different options that you can select for your solutions in big data management (BDM). An option can be many things, including vendor tool types and tool features, as well as users' techniques and methodologies. Regardless of what project stage you're in with BDM, knowing the available options is foundational to making good decisions about approaches to take and software or hardware products to evaluate.

To quantify these issues and to draw the big picture of available options, TDWI presented survey respondents with a long list of options for BDM. (See Figure 17.) The list includes options that have arrived fairly recently (Hadoop, MapReduce, event processing), have been around for a few years but are just now experiencing broad adoption (in-database analytics, in-memory DBMSs, clouds), or have been around for years and are firmly established (metadata management, data federation, appliances, columnar DBMSs). The list is a catalog of available options for BDM, and responses to survey questions indicate what combinations of user designs, platform types, and tool functions users are employing today, as well as which they anticipate using in a few years. From this information, we can quantify trends and project future directions for managing big data. We can also deduce priorities that can guide users in planning their future efforts in big data management.

Concerning the list of BDM options presented in the survey, TDWI asked respondents: "Of the following list of techniques, tool types, tool features, and user practices, which is your organization using for big data management?" Each row (representing a BDM option) presented three multiple-choice answers:

1. Using today; will keep using
2. Will use within three years
3. No plans to use

These survey responses are charted on the left side of Figure 17.

Survey responses quantify the use of BDM options today and predict their future increase or decline.

The pairs of compound horizontal bars on the right side of Figure 17 paint a slightly different picture for option usage. The "potential growth" bars calculate the per-option difference between responses for "using today" and "will use"; this metric provides an indication of how much the usage of a BDM option will increase or decrease. An option's "commitment" value is the percentage of survey respondents who did not select "no plans for using"; this metric provides an indication of how many organizations are committed to using that option, whether today, within three years, or both.

Note that the BDM options listed in Figure 17 are a subset of the options listed in the survey. These are the top 20 options in terms of potential growth. There isn't space to chart all 40 or so options listed in the survey. The focus here is on those that are set for the highest adoption rate in the near future, so only the top 10 options by growth are of interest to this discussion.

Of the following list of techniques, tool types, tool features, and user practices, which is your organization using for big data management?

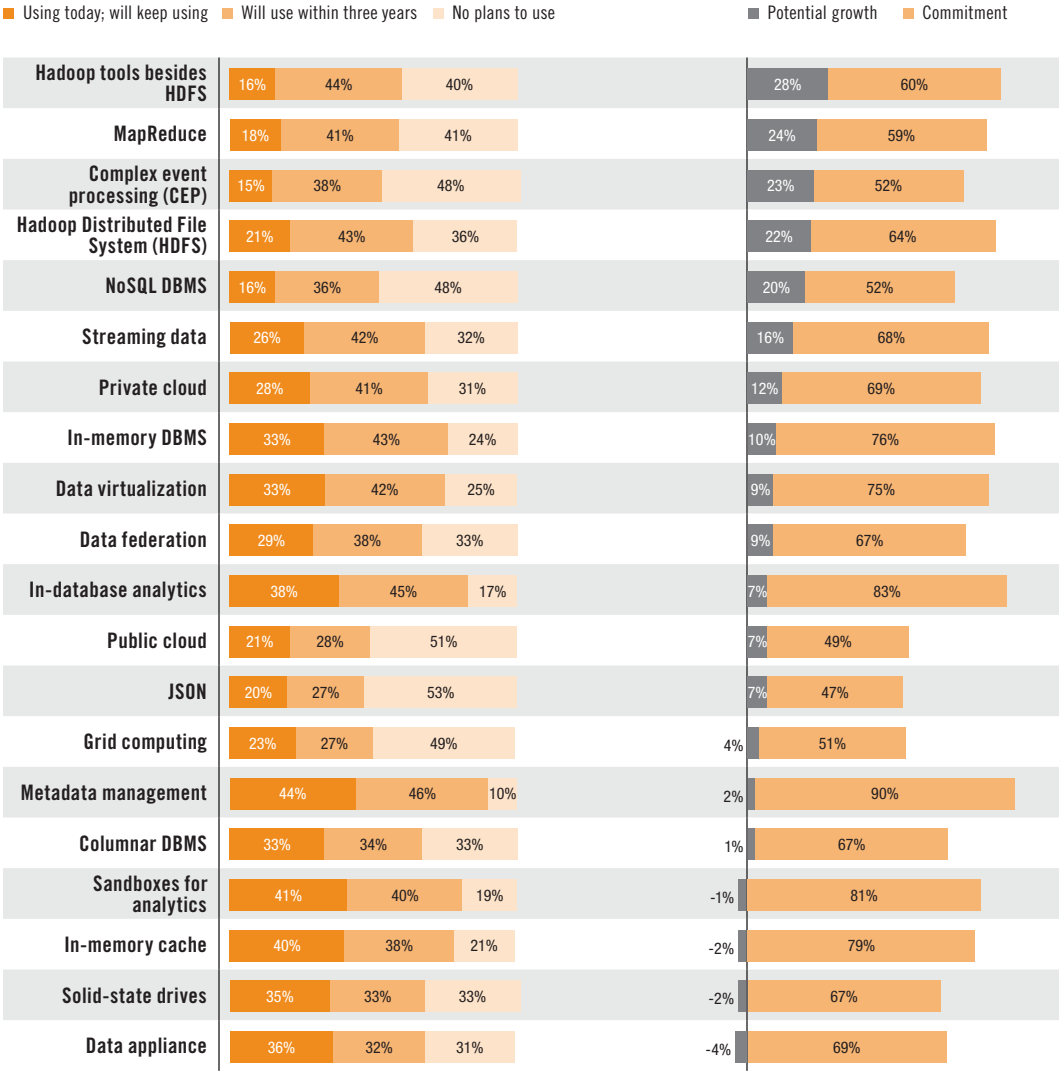


Figure 17. *The number of respondents varies per answer, from 357 to 395. The charts are sorted by the "potential growth" column. Only the top 20 multiple-choice answers are charted here, to focus on the fastest-growing BDM options. Due to rounding, not all groups of percentages on the left side of the figure sum to 100.*

Potential Growth versus Commitment for BDM Options

Figure 17 reveals several interesting things about the use of tools and techniques for BDM. For example, all of Figure 17 is sorted by the "potential growth" column, in descending order. In this sort order, "Hadoop tools besides HDFS" appears at the top of the chart, because—with a growth

projection of 28% (greater than other options)—this tool category exhibits the highest potential for growth among the options listed. In the "commitment" column, we see that 60% of survey respondents have committed to implementing Hadoop tools, whether today or within three years. The prediction deduced from these data points is that future BDM solutions will dramatically increase their usage of Hadoop tools.

Commitment and potential growth are two indicators for quantifying the future of BDM options.

From this, we see that there are two forces at work in Figure 17, as well as in the planning processes of user organizations.

- **Potential growth.** The potential growth value is the product of "will use" minus "using today," and the delta provides a rough indicator for the growth or decline of usage of options for BDM over the next three years. The charted numbers are positive or negative. Note that a negative number indicates that the number of new deployments of an option may decline or remain flat instead of grow. A positive number indicates growth (in the sense of new deployments), and the size of the number suggests a growth rate.
- **Commitment.** The "commitment" value represents the percentage of survey respondents who did *not* select "no plans to use." Note that the measure of commitment is cumulative, in that the commitment may be realized via use today or in the near future.
- **Balance of commitment and potential growth.** To get a complete picture, it's important to look at the metrics for both growth and commitment. For example, some features or techniques may have significant growth rates, but within a weakly committed segment of the BDM user community (CEP, NoSQL DBMSs, public cloud). Or they could have low growth rates (even flat or declining rates), despite being strongly committed through common use today (data appliances, analytic sandboxes). Options seeing the greatest activity in the near future will most likely be those with strong ratings for both growth and commitment (Hadoop, MapReduce, streaming data).

To visualize the balance of growth and commitment, Figure 18 plots the "potential growth" and "commitment" numbers from Figure 17 as opposing axes. BDM options are plotted in terms of growing or declining usage (*x*-axis) and narrow or broad commitment (*y*-axis).

Trends for BDM Options

Rates of growth and commitment identify four groups of options for BDM.

Figures 17 and 18 reveal that most BDM options will experience some level of growth in the near future. The figures also indicate which options will grow the most versus those that will stagnate or decline. Four groups of options cluster together based on intersections of growth and commitment. (See the groups circled, numbered, and labeled in Figure 18.) Furthermore, the groups are indicative of trends in BDM and related disciplines.

Group 1—Moderate commitment, moderate-to-strong potential growth

Options that have the highest probability of altering best practices for BDM are those with a moderate or strong potential growth (according to survey results), coupled with a moderate or strong organizational commitment. Group 1 meets those requirements, and it includes tool types and techniques that TDWI has seen adopted aggressively in recent years. In many ways, Group 1 is the epitome of big data management because of its mix of leading-edge options supported by real-world organizational commitment. Furthermore, today's strongest trends in data management, real-time operation, multi-structured data, and analytics are apparent in Group 1.

Real-time operation. As discussed earlier in this report, the incremental movement toward real-time operation is the most influential trend today in BI, DW, data management, and analytics. As

examples, real-time practices such as operational BI and operational analytics require very fresh data that's collected, processed, and delivered in or near real time. For this purpose, real-time data integration into and out of EDWs has become commonplace. A few organizations capture streaming big data and analyze the stream in real time or close to it; examples of applications include financial trading systems, business activity monitoring, utility grid monitoring, e-commerce product recommendations, and facility monitoring and surveillance. As a method for processing multiple data streams and correlating them with other data sources, complex event processing has arrived recently to handle streaming data. The survey results reveal that CEP and streaming data are among the fastest-growing techniques used in big data management.[5]

Real time is the most influential BI trend, and it makes tough demands of BDM.

Options for High-Performance Data Warehousing Plotted by Growth and Commitment

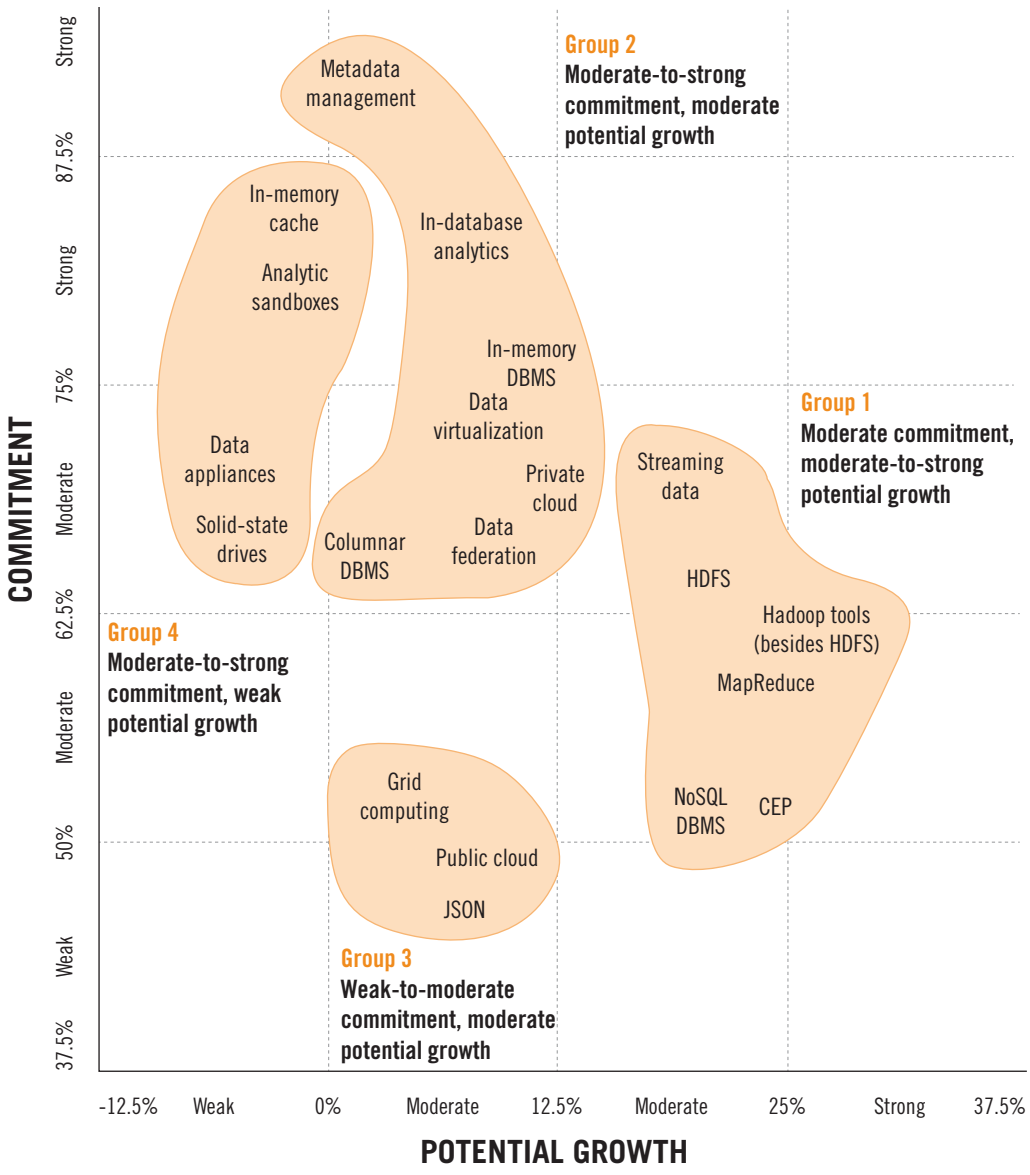


Figure 18. *Based on the "commitment" and "potential growth" columns of Figure 17. Plots are approximate.*

[5] For an introduction to new practices for real-time big data, see the TDWI Checklist Report *Operational Intelligence: Real-Time Business Analytics from Big Data*, August 2013, available online at tdwi.org/checklists.

Interest is high in distributed file systems and distributed analytic processing.

Hadoop Distributed File System (HDFS). At the moment, users' interest in HDFS is extremely high (hence, the moderate potential growth in Figure 18). In this survey, HDFS has risen into the moderate commitment category (higher than in earlier TDWI surveys), which indicates that HDFS adoption by users is well under way.

Interest is high because of Hadoop's good reputation for managing and processing the extremes of big data for data integration, data warehousing, and analytics, all at a low price. HDFS clusters are known to scale out to hundreds of nodes that scale up to handle hundreds of terabytes of file-based data. Furthermore, as a data-type-agnostic file system, HDFS manages a very wide range of file-based data that is structured, unstructured, semi-structured, or a mix. HDFS's clustered architecture—with other Hadoop products layered above HDFS—provides a very scalable and relatively high-performance platform for a wide range of data-intense applications.[6]

MapReduce. Parallel processing—usually in the form of massively parallel processing (MPP)—is a key enabler for high performance with data-intense operations. That's what MapReduce is all about. Open source MapReduce is an execution engine that can give multithreaded parallelism to hand-coded routines written in a wide variety of programming languages. A typical analytic application is to write analytic logic in Java, Pig, or R routines, then have MapReduce execute the routines using parallel processing to access the massive file and data repositories managed by an HDFS cluster. When open source MapReduce is deployed as a layer atop HDFS this way, the result is a high-performance analytic application that scales to massive data collections.

NoSQL DBMSs. A NoSQL database is a DBMS that manages data that's not relational and thus the DBMS need not or cannot support SQL. Although not relational, NoSQL data may be structured according to other schema such as records or value pairs, or the data may have no fixed schema. NoSQL databases are categorized according to schema and storage strategies such as document stores, key-value stores, BigTable implementations, and graph databases.

A consequence of big data is increasing volumes of non-relational, schema-free data. To manage and analyze non-relational big data, a few organizations are embracing NoSQL databases, as well as similar non-DBMS data platforms such as HDFS. This makes sense when the majority of data types analyzed are not relational and converting them to relational structures is not practical. In other situations, a NoSQL approach is useful when modeling and indexing data in relational schema would inhibit the discovery mission that many modern analytic projects are all about. In this report's survey, NoSQL DBMSs rose into the moderate commitment category, indicating their new popularity.

Group 2 – Moderate-to-strong commitment, moderate potential growth

Similar to Group 1, Group 2 is poised for moderate growth. However, Group 2 today has a moderate-to-strong commitment in the user marketplace. That's because most of the options in Group 2 have been available for years and have seen brisk adoption recently (namely, in-database analytics, in-memory analytics, private clouds, columnar DBMSs, metadata management, and data federation and virtualization). These are all compelling and proven technologies that are getting a lot of attention, so we should expect to see more of these options implemented in upcoming years.

In-database analytics takes processing to the data, instead of vice versa.

In-database analytics. Analytic processing based on data mining, statistical analysis, and other non-relational approaches has for many years required that the data be located in a special database or a flat file. Moving large volumes of data for advanced forms of analytics is time-consuming—and therefore antithetical to real time. As big data volumes soar, moving big data becomes ever less tenable. Hence, a new trend takes analytic algorithms to the data instead of vice versa. That's exactly

[6] For more information about Hadoop's role in data management and data warehouses, see the TDWI Research reports *Integrating Hadoop into Business Intelligence and Data Warehousing* (April 2013, tdwi.org/bpreports) and *Where Hadoop Fits in Your Data Warehouse Architecture* (July 2013, tdwi.org/checklists).

what in-database analytics does. It assists speed and scale for large-volume analytics, and, depending on how you set it up, it also reduces architectural complexity. Users are aggressively adopting this kind of analytics, which is why in-database analytics already has a strong commitment and should experience good growth.

Reducing disk I/O increases the performance of data-intense processes.

In-memory databases. One way to get high performance (in the sense of fast data access) from a database is to manage it in server memory, thereby eliminating disk I/O and other bottlenecks. For several years now, TDWI has seen consistent adoption of in-memory databases among its members and other organizations. An in-memory database can serve many purposes, but in BI they usually support real-time dashboards for operational BI, and the database usually stores metrics and KPIs—sometimes OLAP cubes. We're now seeing a similar trend among users in the adoption of in-memory databases for advanced analytics, typically to speed the scoring of analytic models. In a related trend, leading vendors now offer data warehouse appliances and similar configurations with solid-state drives (see Group 4 in Figure 18). I/O-bound operations with these are very fast compared to traditional drives.

A cloud can be central to a data management strategy.

Private clouds. According to TDWI Technology Surveys, users have a clear preference for private clouds over public ones due to concerns about data security, privacy, and governance with public clouds. Even so, user organizations are adopting a mix of cloud types, and freely combining them with traditional on-premises platforms. For many, a cloud is a data management strategy due to its fluid allocation and reapportionment of virtualized system resources. For example, an analytic database is set up in a cloud as an "analytic sandbox" (in Group 4 in Figure 18) to accommodate the large and fluctuating volumes of data (and to isolate the unpredictable ad hoc query workloads) that are part and parcel of the work of business analysts and other power users.

Data federation and virtualization. Data federation fetches fresh data in near real time from source systems only when an application, report, or user asks for it. This reduces the overhead of continuous data capture and makes federation a viable approach to collecting and managing data in real time. Federation has been around for years in low-end forms such as distributed queries and materialized views. Modern tools, however, provide superior design and maintenance functions for federation (plus higher performance and scalability) that make it far more compelling as a feature, even for relational big data. Federation is also more compelling as it merges with data virtualization. These advances help explain why data federation and virtualization have already achieved a solid commitment among users and should see continued growth.

Columnar DBMSs. A columnar DBMS is a relational database management system that relies on a column-oriented data store, unlike the row-oriented stores of most relational DBMSs. In a column store, data is stored by table columns. The close proximity of related data in storage speeds up the DBMS's retrieval of data for a specific column. The DBMS also creates statistics and lists about the content of columns; it heavily compresses data, which also speeds up columnar queries. Hence, a columnar DBMS provides high performance and scalability for complex queries against large data sets, and that's why the columnar DBMS has become a popular platform for managing and analyzing big data using SQL-based analytic methods.[7]

Group 3 – Weak-to-moderate commitment, moderate potential growth

Group 3 is a mixed bag of options, namely grid computing, public clouds, and JSON. Although commitment to these is low today, all are poised for moderate growth. Expect to see more of them in the near future.

[7] For more information about columnar databases and other analytic databases, see the TDWI Checklist Report *Analytic Databases for Big Data* (October 2012), available at tdwi.org/checklists.

Group 4 – Moderate-to-strong commitment, weak potential growth

The adoption of a few mature and saturated BDM options will slow down.

Group 4 includes some of the most common options in use today for BDM and other data management strategies, namely data warehouse appliances and analytic sandboxes. If these are so popular, why does the survey show them in decline or suffering flat growth?

Sometimes an established tool type or user technique reaches a saturation point because it's been deployed by most of the organizations that need it, in most of the situations where it's needed. After this point, deployments receive maintenance but little or no new development. This is most likely the case with the BDM options of Group 4.

> **EXPERT COMMENT GETTING VALUE FROM BIG DATA'S DIVERSITY DEMANDS DIVERSE DATA MANAGEMENT PLATFORMS, INTEGRATED ARCHITECTURALLY IN A HYBRID DATA ECOSYSTEM.**
>
> "We're currently experiencing a confluence of paradigm shifts that are driven by a maturing data consumer community, new technologies for data management and data processing, changes in the economics of data management platforms, and unprecedented volumes of valuable new data types," said Shawn Rogers, a vice president at IT analyst firm Enterprise Management Associates. "These trends are driving us toward a hybrid data ecosystem, which is essentially an integrated environment consisting of different types of data management platforms and tools, ranging from traditional databases and data appliances to columnar databases, NoSQL databases, and Hadoop.
>
> "Modern organizations need this diversity of platforms so they can match the growing number of data workloads to platforms that are optimized for such workloads. Otherwise, technical users are frustrated as they attempt to manage and process data in new ways—for advanced analytics, broader 360-degree views, real-time reactions to streaming data, lower TCO—with platforms that weren't designed for these new requirements. In short, getting full business value and technical efficiency from new big data opportunities demands a hybrid data ecosystem."

Vendor Platforms and Tools for Managing Big Data

Since the firms that sponsored this report are all good examples of software and hardware vendors that offer tools, platforms, and professional services for managing big data, let's take a brief look at the product portfolio of each. The sponsors form a representative sample of the vendor community, but their offerings illustrate different approaches to big data management.[8]

Cloudera

Cloudera is a leader in enterprise analytic data management powered by Apache Hadoop and the top contributor to the Hadoop open source community. Cloudera's 100% open source distribution, CDH, is comprehensive and widely deployed, and it includes: Cloudera Impala for interactive SQL of data in HDFS and HBase through popular BI tools; Cloudera Search to enable non-technical users to intuitively explore Hadoop data; plus enterprise capabilities such as Sentry for fine-grained, role-based access control and native high availability for extreme fault tolerance. Cloudera Enterprise, available as a commercial subscription, couples that platform with Cloudera's Support and a suite of system and data management software built for the enterprise, including: Cloudera Manager to simplify and reduce the cost of Hadoop configuration, deployment, upgrades, and administration; Cloudera Navigator for audit and access control of Hadoop data; and optional Support for Impala, Search, and HBase. Cloudera also offers consulting services and a broad array of Hadoop training and certification programs. More than 700 partners across hardware, software, and services have teamed with Cloudera to ensure maximum integration with customers' existing investments.

[8] The vendors and products mentioned here are representative, and the list is not intended to be comprehensive.

Leading companies in every industry run Cloudera in production, including over 65% of the *Fortune* 500 in finance, telecommunications, retail, Internet, insurance, energy, healthcare, biopharmaceuticals, networking, and media, plus three of the largest intelligence agencies and two of the top three defense agencies.

Dell Software

For years, Dell Software has been acquiring and building software tools (including the acquisition of Quest Software and its database development and administration tool, Toad) with the goal of assembling a comprehensive portfolio of IT administration tools for securing and managing networks, applications, systems, endpoints, devices, and data. Within that portfolio, Dell Software now offers a range of tools specifically for data management, with a focus on big data and analytics. For example, Toad Data Point provides IT users with data provisioning and administrative functions for most traditional databases and packaged applications, plus new big data platforms such as Hadoop, MongoDB, Cassandra, SimpleDB, and Azure. Toad Business Intelligence Suite combines Toad Data Point, Toad Decision Point for data visualizations, and a server component called Toad Intelligence Central for collaboration, to provide a fully integrated self-service BI solution that works alongside corporate BI systems. Shareplex supports Oracle-to-Oracle replication, as well as Oracle-to-Hadoop replication. Kitenga Analytics Suite is a big data analytics tool that combines search, visualization, and analytics into a platform that enables rapid transformation of diverse unstructured data into actionable insights. Dell Boomi is a cloud-based data integration service that connects any combination of hosted and on-premises applications. Combine all the new Dell Software capabilities with Dell's traditional Hadoop and hardware options, and Dell is well positioned to provide a true end-to-end big data analytics solution.

Oracle

Oracle has recently made a substantial investment in new products and platforms engineered to provide all aspects of big data management—especially scalability and real-time operation—specifically for advanced analytics with big data. For example, the Oracle Big Data Appliance integrates and optimizes all the hardware and software components needed to build comprehensive analytic applications. The Big Data Appliance includes Cloudera's distribution of Hadoop, Cloudera Manager, a distribution of R, Oracle Linux, Oracle NoSQL Database, and the Oracle HotSpot Java Virtual Machine. Oracle Big Data Connectors provides a gateway from Hadoop and NoSQL Database to Oracle Database 11*g* and 12*c*. As another example, the Oracle Exalytics In-Memory Machine provides high-performance, in-memory analytics, as required of growing practices such as business performance management, operational BI, and operational analytics. Other platforms conducive to high-performance big data management and analytics include the Oracle Exadata Database Machine and the Oracle Exalogic Elastic Cloud.

Pentaho

Pentaho, Inc., is well known for its unified, open, embeddable, pluggable Business Analytics platform that tightly couples both data integration (DI) and business intelligence (BI). As we moved into the age of big data analytics, Pentaho evolved Pentaho Data Integration (PDI) (an enterprise-class, graphical ETL tool) to include support for multiple layers and types of Hadoop, NoSQL, and analytic appliance environments, and added its Visual MapReduce tool that eliminates the need for the complex coding normally required. Pentaho evolved its Business Analytics Suite, as well. Pentaho Business Analytics now also includes visualization tools, data mining and predictive algorithms, plus an analytic modeling workbench, in addition to its traditional BI reporting and dashboard tools. Pentaho supports the entire big data analytics process. Pentaho makes this unified approach a practical reality by generating data integration logic and predictive models in Hadoop-friendly Java. And its engine can be run directly in the Hadoop cluster without generating code. This means that Pentaho-based solutions are easy to embed in Hadoop, and a new adaptive big data layer from Pentaho ensures portability of Pentaho solutions across Hadoop distributions from Cloudera, Hortonworks, MapR, and Intel, as well as Cassandra, MongoDB, and Splunk.

SAP SAP provides a comprehensive set of solutions for big data, including analytic applications, rapid deployment solutions, BI and advanced analytic tools, analytic databases, data warehousing solutions, and information management tools. Furthermore, SAP enables its customers to integrate Hadoop into their existing BI, advanced analytic, and data warehousing environments in multiple ways, giving customers the ability to tailor Hadoop to their needs. Many customers are deploying Hadoop alongside SAP HANA, an in-memory database used for real-time analytics and other applications. Customers can use SAP Data Services to search and load data from HDFS or Hive into SAP HANA or SAP Sybase IQ, or they can use SAP HANA smart data access to push queries into Hive Hadoop (or other data sources) and achieve data virtualization with virtual tables in SAP HANA. To accommodate the extremes of big data management, version 16 of SAP Sybase IQ software recently achieved a Guinness World Record for loading and indexing big data, achieving an audited result of 34.3 terabytes per hour. Streaming and CEP are ably served by SAP Sybase ESP. Furthermore, SAP BusinessObjects BI, SAP Visual Intelligence, and SAP Predictive Analysis users can query Hive environments, giving business analysts the ability to explore Hadoop data directly. New in-memory spatial and enhanced natural language capabilities in SAP HANA enable organizations to uncover richer and more meaningful signals from business and geospatial data. Finally, the completion of the integration of Sybase data management with SAP HANA will further transform customers' end-to-end data management landscape.

SAS SAS is a leader in analytic solutions that provide business performance improvements. SAS looks at the big data challenge holistically and promotes a lifecycle methodology that orchestrates data management, preparation, and exploration to model development and governance. Recently, SAS launched its High-Performance Analytics product that supports the Hadoop approach to data filing, query processing, and management. For example, SAS Visual Analytics allows users to access the shared memory of the Hadoop cluster to provide an in-memory exploratory analytics platform, enabling users to visualize and assess massive amounts of structured data and text in the Hadoop ecosystem. SAS High-Performance Analytics is an in-memory suite of analytic solutions that empowers analysts to develop predictive models (including those using text) and deliver insights in minutes, without data movement outside the Hadoop cluster. From a data management perspective, SAS provides access to Hadoop (via HiveQL), Oracle Exadata, SAP HANA, Teradata, Vertica, and IBM PureData for Analytics (formerly known as Netezza) using SAS/ACCESS software. Users can also apply existing MapReduce, Pig, or Hive code from within the SAS environment, supporting Cloudera and similar systems. SAS Data Federation Server provides an intuitive, graphical interface to integrate and transform data to and from Hadoop and other sources. Finally, the new SAS Event Stream Processing Engine implements a form of complex event processing to manage streaming big data and execute related high-volume, real-time tasks such as risk management and anti-fraud analytics.

Top 10 Priorities for Big Data Management

In closing, let's summarize the report by listing the top 10 priorities for managing big data, with a few comments about why each is important. Think of the priorities as requirements, rules, or recommendations that can guide organizations into successful strategies for big data management.

1. **Demand business value from big data.** The first step toward value is to manage big data. If you don't capture, store, process, and deliver big data, you won't have it to repurpose for valuable applications in customer intelligence, operational efficiency, business monitoring, and so forth. Eighty-nine percent of survey respondents say BDM is an opportunity—but only if you seize it. Know the common paths to business value and follow them. The primary path to business value from big data is discovery analytics. A second path joins new big data with older enterprise data to extend complete views of customers and other business entities. A third path taps streaming big data to enlighten and accelerate time-sensitive business processes.
2. **Use big data to create new applications and extend old ones.** As noted earlier, streaming big data can enable business monitoring applications, and big data can expand the data samples that data mining and statistical analysis applications depend on for accurate actuarial calculations and customer segments or profiles.
3. **Get training (and maybe new staff) for big data management.** The focus should be on training and hiring data analysts, data scientists, and data architects who can develop the applications for data exploration, discovery analytics, and real-time monitoring that organizations need if they're to get full value from big data. When in doubt, hire and train data specialists to manage big data, not application specialists. Most BI/DW professionals are already cross trained in many data disciplines; cross train them more.
4. **Collaboration is key to all data management, especially when managing big data.** Due to big data's diversity, diverse technology teams will need to play coordinated roles. As a business asset, big data should be managed for broad access and leveraged by multiple business units and stakeholders. It takes a lot of collaboration to be sure everyone knows their role and has their needs met.
5. **Beware the proliferation of siloed big data repositories.** After all, the goal is to integrate big data into your well-integrated enterprise data and BI/DW environments, not proliferate 21st-century spreadmarts. Someone (probably not you) should decide whether big data platforms will be departmentally owned (as a lot of analytic applications are) or shared enterprise infrastructure supplied by central IT (similar to how IT provides SAN/NAS storage, servers, the network, etc.).

   Even if big data management begins in a silo (and you do have to start somewhere), make integration with other enterprise data management systems a second-phase priority. To make the integration happen, look for big data platforms (both open source and vendor built) that enable the integration points discussed elsewhere in this report.
6. **Define places for big data in architectures for data warehousing and enterprise data management.** For example, an obvious place to start is to rethink the data staging area within your data warehouse. That's where big data enters a data warehouse environment and where it is usually stored and processed. Consider moving your data staging area to a standalone big data management platform (on Hadoop, a columnar DBMS, or an appliance) outside the core data warehouse. Architecture can enable or inhibit critical next-generation BDM functions such as extreme scalability, complete views, unforeseen forms of analytics, big data as an enterprise asset, and real-time operation.

7. **Reevaluate your current portfolio of data platforms and data management tools.** For one thing, big data management is, more and more, a multi-platform solution (as are most data warehouse architectures), so you should expect to further diversify your software portfolio accordingly to accommodate big data fully. For another thing, survey data shows that the software types poised for the most brisk new adoption in the next three years are Hadoop (including HDFS, MapReduce, and miscellaneous Hadoop tools) and complex event processing (for streaming real-time big data). After those come NoSQL DBMSs, private clouds, and data virtualization/federation. If you're like most organizations surveyed, all these have a potential use for your BDM solution, so you should educate yourself about them, then evaluate the ones that come closest to your BDM requirements.
8. **Select data platforms that have special support for big data.** There are many types to consider, including relational DBMSs, columnar DBMSs, data appliances, and other engineered systems, as well as Hadoop, NoSQL DBMSs, and the many other options listed earlier in this report. However, matching your BDM requirements to vendor products isn't always about the entire platform; it sometimes comes down to specific functions such as in-memory processing, in-database analytics, complex event processing, MapReduce, and robust interfaces to other data platforms.
9. **Embrace all formats of big data, not just relational big data.** Create a plan for your BDM maturation process. You have to start somewhere, so start with relational data, then move on to to other structured data, such as log files that have a recurring record structure. Carefully select a beachhead for unstructured data, such as text analytics applied to call center text in support of sentiment analysis. Look for mission-critical data that's semi-structured, as in the XML documents your procurement department is exchanging with partnering companies. Then continue down the line of big data types.
10. **Develop and apply a technology strategy for big data management.** The strategy needs to spell out a wide range of road maps, standards, preferences, and stewardship guidelines, depending on what your organization needs and has a culture for. For example, you could lay out a road map for maturing from structured to semi-structured to unstructured data, as noted above. Since big data comes in many forms, you need to determine the preferred platforms and interfaces for capturing, storing, and processing each form. You should design a workflow for managing big data forms in their original state, plus processing that into other states for customer intelligence, data warehousing, discovery analytics, and so on. Big data isn't the storage problem it used to be, but you still have to plan capacity carefully, as well as related issues such as the acquisition and upgrade of data management platforms. Assuming you have an enterprise-scope data strategy and data architecture, you need to determine the many places diverse forms of big data should take in those. Finally, all the above must be supported by influential business sponsors (through stewardship and governance) so that big data management aligns with business goals for maximum business value.

www.sas.com

Upgraded SAS® Data Management delivers big data payoff

SAS adds in-database data quality, improved event-stream processing, access to more data sources including Hadoop.

Today's big data mania reflects a harsh reality: companies have huge amounts of data, but struggle to turn it into value. SAS, a leader in data quality and data integration, is offering new technologies and upgrades to its SAS® Data Management family of software to help organizations meet this big data challenge.

The new release helps customers address high-priority items first, with one platform that provides modules for data access, data quality, event stream processing, data integration, and more. This modular approach helps companies expand and adapt their data management environment as their needs evolve.

Enhancements to SAS Data Management include new in-database data quality capabilities, improved event-stream processing, and new access engines to more easily use data from Hadoop, Amazon Redshift, and other sources.

NEW IN-DATABASE DATA QUALITY

This update introduces in-database data quality to SAS Data Management. Cleansing within the database means less data movement, dramatically reducing the time to transform raw data into clean, usable information. Ultimately that means better, faster business decisions and applications.

The new SAS Data Quality Accelerator for Teradata is the only in-database data-cleansing technology for the Teradata platform. By executing data quality functions directly in Teradata, companies improve processing speeds, reduce IT workload and, most importantly, provide high-quality data for a host of downstream business uses.

STREAMING DATA IN REAL TIME

SAS Event Stream Processing Engine seamlessly brings streaming data—from trading activities, transaction systems, smart grids and location-based devices—into data management processes for high-volume, real-time tasks.

Now with an improved interface, SAS Event Stream Processing Engine lets companies access and analyze data as it happens, foregoing intelligence latency. Companies such as electric utilities, which continually collect high volumes of data on customer power use from smart meters, use real-time data to anticipate and react to sudden spikes in demand. And financial services firms use event-stream processing to evaluate trades as they happen, finding risky or fraudulent activity.

NEW DATA SOURCES

Data comes not only in many sizes, but from many sources, including the cloud, Hadoop clusters, streaming transactional data, on-site databases, mainframes, in-memory databases, and more. Regardless of format or source, SAS Data Management can natively access, cleanse, and enrich data so it is a valid data source for operational systems and business analytic applications.

With this upgrade, SAS makes it even easier to manage data from HP Vertica, Hortonworks, RedShift, and PostgreSQL. With direct data connectivity and integration across these data sources, customers can edit, move, cleanse, and utilize data regardless of source, platform, size, or variety.

TDWI RESEARCH

TDWI Research provides research and advice for business intelligence and data warehousing professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence and data warehousing solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

555 S Renton Village Place, Ste 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org