TDWI RESEARCH

SECOND QUARTER 2015

TDWI BEST PRACTICES REPORT

Hadoop for the Enterprise:

Making Data Management Massively Scalable, Agile, Feature-Rich, and Cost-Effective

By Philip Russom

Co-sponsored by

pentaho®

tdwi.org

tdwi
Advancing all things data.

BEST PRACTICES REPORT

SECOND QUARTER 2015

Hadoop for the Enterprise:

Making Data Management Massively Scalable, Agile, Feature-Rich, and Cost-Effective

By Philip Russom

Table of Contents

About the Author

PHILIP RUSSOM is a well-known figure in data warehousing and business intelligence, having published over five hundred research reports, magazine articles, opinion columns, speeches, and Webinars. Today, he's the Research Director for Data Management at 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 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 on LinkedIn at linkedin.com/in/philiprussom.

About TDWI

For 20 years, TDWI has provided individuals and teams with a comprehensive portfolio of business and technical education and research about all things data. The in-depth, best-practices-based knowledge TDWI offers can be quickly applied to develop world-class talent across your organization's business and IT functions to enhance analytical, data-driven decision making and performance. TDWI advances the art and science of realizing business value from data by providing an objective forum where industry experts, solution providers, and practitioners can explore and enhance data competencies, practices, and technologies. TDWI offers five major conferences as well as topical seminars, onsite education, membership, certification, live Webinars, resourceful publications, industry news, and in-depth research. See tdwi.org.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence (BI) 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 BI professionals.

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

Sponsors

Actian Corporation, Cloudera, EXASOL, IBM, MapR Technologies, MarkLogic, Pentaho, SAS, Talend, and Trillium Software sponsored the research for this report.

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 responded 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: Michael Boyda, Roxanne Cooke, Marie Gipson, James Powell, and Denelle Hanlon.

Research Methodology and Demographics

Report Scope. TDWI believes that Hadoop usage will become mainstream in coming years, and—more to the point—Hadoop will serve whole enterprises, not just a handful of users with niche applications in a limited number of industries. This report explains how Hadoop and its uses are evolving to enable enterprise-grade deployments that serve a broadening list of use cases, user constituencies, and organizational profiles.

Terminology. In this report, the term "Hadoop" refers to the growing ecosystem of open source and commercial software platforms, tools, and maintenance available from the Apache Software Foundation (open source) and several software vendor firms.

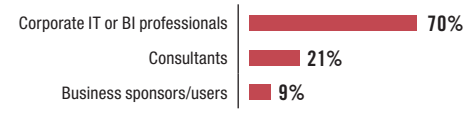
Survey Methodology. In November 2014, TDWI sent an invitation via e-mail to the data management professionals in its database, asking them to complete an Internet-based survey. The invitation was also distributed via websites, newsletters, and publications from TDWI and other firms. The survey drew responses from 334 survey respondents. From these, we excluded incomplete responses and respondents who identified themselves as academics or vendor employees. The resulting completed responses of 247 respondents form the core data sample for this report. Due to branching in the survey, some questions were answered only by 100 respondents who have hands-on Hadoop experience.

Research Methods. In addition to the survey, TDWI Research conducted 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 we discuss in the report.

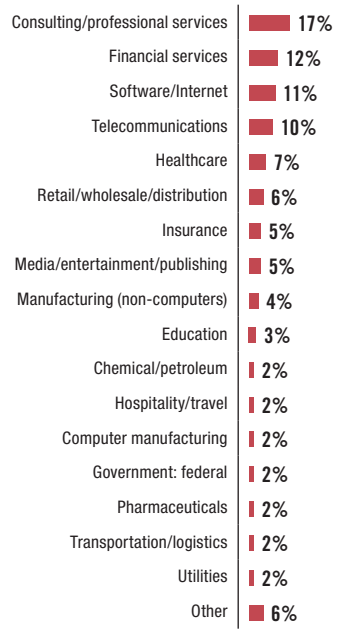
Survey Demographics. The majority of survey respondents are IT or BI/DW professionals (70%), whereas the others are consultants (21%) and business sponsors or users (9%). We asked consultants to fill out the survey with a recent client in mind.

The consulting industry (17%) dominates the respondent population, followed by financial services (12%), software/Internet (11%), telecommunications (10%), and other industries. Most survey respondents reside in the U.S. (47%), Europe (18%) or Asia (15%). Respondents are fairly evenly distributed across all sizes of organization.

Position

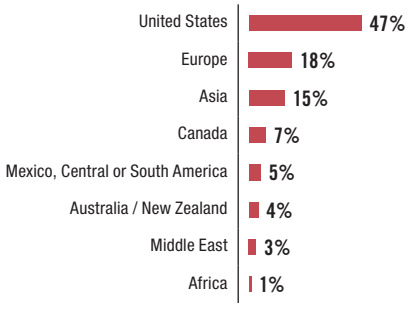


Industry

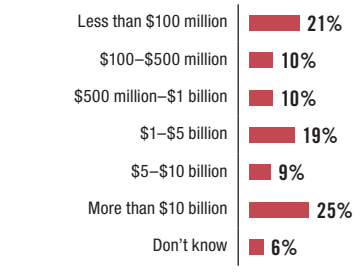


("Other" consists of multiple industries, each represented by less than 2% of respondents.)

Region



Company Size by Revenue



Based on 247 survey respondents.

Executive Summary

Hadoop is evolving to serve mainstream enterprises.

Hadoop began its journey by proving its worth as a Spartan but highly scalable data platform for reporting and analytics in Internet firms and other digital organizations. The journey is now taking Hadoop into a wider range of industries, use cases, and types of organization. Hadoop is again challenged to prove its worth, this time by satisfying the stringent requirements that traditional IT departments and business units demand of their platforms for enterprise data and business applications.

Technical users want Hadoop to scale, extend older systems, and leverage exotic data.

Hadoop for the enterprise is driven by several rising needs. On a technology level, many organizations need data platforms to scale up to handle exploding data volumes. They also need a scalable extension for existing IT systems in warehousing, archiving, and content management. Others need to finally get BI value out of non-structured data. Hadoop fits the bill for all these needs.

Business people want Hadoop for value from big data and for insights from analytics.

On a business level, everyone wants to get business value and other organizational advantages out of big data instead of merely managing it as a cost center. Analytics has arisen as the primary path to business value from big data, and that's why the two come together in the term "big data analytics." Hadoop is not just a storage platform for big data; it's also a computational platform for business analytics. This makes Hadoop ideal for firms that wish to compete on analytics, as well as retain customers, grow accounts, and improve operational excellence via analytics.

Hadoop in production is up 60% in two years.

For these and other reasons, Hadoop adoption is accelerating. TDWI survey results show that Hadoop clusters in production are up 60% in two years. Almost half of respondents have new Hadoop clusters in development, and these will come online within 12 months. At this rate, 60% of users surveyed will have Hadoop in production by 2016, a giant step forward.

Hadoop is an opportunity for innovation.

Adoption is accelerating because most users (89%) consider Hadoop an opportunity for innovation According to this report's survey, Hadoop's leading benefits include improvements to analytics, data warehousing, data scalability, and the handling of exotic data types, in that order. Leading barriers are inadequate technical skills, weak business support, security issues, and weak open source tools. All these barriers (and others) are being corrected by user best practices and advancements from both open source and vendor communities.

As Hadoop goes enterprise in scope, ownership, staffing, development methods, and economics shift.

As Hadoop broadens across the enterprise, its ownership is shifting from departments and application teams to central IT. This makes sense when IT provides Hadoop clusters as shared enterprise infrastructure. The people working on these clusters are most often data scientists, data architects, data analysts, and developers. These people are rare and expensive on the job market, so most organizations train existing employees in Hadoop skills instead of hiring them.

Best practices for enterprise Hadoop are coalescing. Developers employ a mix of programming and high-level tools, though they prefer the latter. Most clusters are on premises today but going to clouds soon. Developers complain of poor SQL and relational functions on and off Hadoop today, but vendors and open source contributors are working aggressively on improvements.

Hubs and archives will see big growth atop Hadoop.

The leading future use cases for enterprise Hadoop (according to survey respondents) are enterprise data hubs, archives, and BI/DW. Half of respondents expect to improve existing Hadoop clusters by integrating them with data quality and master data management tools.

This report accelerates users' understanding of the many new products, technologies, and best practices that have emerged recently around Hadoop. It will also help readers map newly available options to real-world use cases, with a focus on mainstream enterprise uses, while respecting tried-and-true IT practices and delivering maximum business value.

An Introduction to Hadoop for the Enterprise

Hadoop is rising to the challenges of broad enterprise use by successfully evolving to serve enterprise IT—as well as multiple departments and application teams—in mainstream industries. As we'll see in this report, the evolution of Hadoop (and users' best practices for it) is driven by new business and technology requirements, compelling use cases, diversification beyond analytics, scalability issues, and economic concerns.

The Hadoop Ecosystem Continues to Expand

The Hadoop ecosystem includes products from the open-source and vendor communities, plus best practices.

Open Source Software (OSS). Apache Hadoop is an open-source software project administered by the Apache Software Foundation (ASF). *Hadoop* is the brand name Apache and its open source community have given to a family of related open source products and technologies. The Hadoop family includes several products, including the Hadoop Distributed File System (HDFS), MapReduce, Hive, Hbase, and Pig.[1]

The Hadoop Ecosystem. HDFS and other Hadoop products are available from Apache and several software vendors. The number of vendor products that integrate with Hadoop family products increases almost daily.

User Applications of Hadoop. TDWI sees Hadoop products used in a variety of data-driven use cases. Hadoop usage is well established for advanced analytics, data visualization, and data warehousing; newer use cases are arising for data integration, data archiving, and content management. TDWI expects this trend to continue such that Hadoop will soon mature into broad enterprise-scale use across multiple departments and use cases.

In this report, "Hadoop" means the entire Hadoop family of products, regardless of their open source or vendor origins. This is how users use the name, and it makes sense because users usually apply many Hadoop products in an integrated fashion along with vendor products from outside open source.

Multiple Business and Technology Drivers Point to Hadoop

From Hadoop, users need scalability, low cost, analytics, and support for unstructured data.

Organizations need scalability for all data. This report contains several user stories and other quotations from surveys and interviews. As you read the quotes, note that users rarely talk about big data or massive scalability for one data domain, one database, or one source. Almost all talk about the challenges of scaling all enterprise data, whether it is new big data or traditional enterprise data. Hadoop has established itself as an enterprise-scope data management platform for multiple data types and domains, and new best practices are also established for these, as seen in Hadoop-based data lakes and enterprise data hubs.

Enterprises need to change the economics of all data. As one user put it, "The high-end relational databases are really expensive in configurations big enough to deal with big data. Data warehouse appliances are almost as expensive. We all need a more economical platform, which is the main reason we're all considering Hadoop." These economic concepts also apply to other enterprise use cases outside data warehousing.

Hadoop can extend the life of other platforms and wring value from big data.

Existing data platforms need greater capacity and life span. A common use case is to extend a data warehouse environment by integrating Hadoop into the environment. In upcoming use cases, Hadoop extends systems for content management, records management, content archives, and data archives. In these configurations, Hadoop doesn't replace existing systems. Instead, it offloads those

[1] Hadoop's open source family and extended ecosystem are described later in the trend section of this report. That section also discusses recent and upcoming changes in the ecosystem. Readers new to Hadoop may wish to read other TDWI reports first, especially *Integrating Hadoop into Business Intelligence and Data Warehousing* and *Managing Big Data*, available for free download at www.tdwi.org/bpreports.

systems and provides economical components, so that each system has greater capacity and a longer life span yet at a lower cost, which in turn means more value for the enterprise over the long haul.

Organizations need to wring business value from big data. Capturing and storing very large datasets in Hadoop is not enough. Organizations should actively explore and process big data for maximum organizational advantage, often to achieve better customer intelligence, improve processes, or detect fraud. More organizations are turning to Hadoop as a cost-effective platform for collecting diverse data prior to leveraging it.

Firms are eager to compete on analytics. Hadoop isn't just a data platform for managing large data sets. It's also a computational platform that enables advanced forms of analytics, such as those based on data mining, statistical analysis, text analytics, graph, and ad hoc algorithmic approaches. This is why Hadoop appeals to for-profit enterprises that now seek to compete on analytics. Other organizations apply Hadoop's analytic power to determine the root cause of customer churn, modernize actuarial calculations, or improve patient outcomes in healthcare, among other uses.

Organizations need better value from multi-structured data. Early adopters have shown that Hadoop excels at storing, managing, and processing unstructured data (e.g., human language text), semi-structured data (XML and JSON documents), and data with evolving schema (some sensor, log, and social data). For organizations with these forms of data in large quantities (e.g., healthcare, government, insurance, and supply chain industries), Hadoop can make the analysis of it more affordable, scalable, and valuable.

Enterprises are moving closer to real-time operations. Early versions of Hadoop had little or no real-time capabilities. This situation has improved considerably with the introduction of open source projects for capturing and analyzing streaming data (Samza, Spark, Storm) and for low-latency query responses (Drill, Impala). Hadoop is becoming a preferred real-time platform because of its low cost (as compared to most commercial real-time platforms) and its massive storage capabilities (volumes of data that stream add up in a hurry).

USER STORY HADOOP CONTINUES TO BE THE SUBJECT OF SERIOUS USER EVALUATIONS

"We don't have a history of using open source, so IT is a bit uncomfortable with Hadoop," said a senior IT director at a global pharmaceutical firm. "We are therefore focusing on vendor distributions of Hadoop in our ongoing proof of concept (POC) evaluations. We are more comfortable that we'll get the support, security, consulting, and administrative tools we need for the effective use of Hadoop, as well as assistance in integrating and configuring with our analytics tools and cloud providers. These aren't available from the open source community, which is why we believe a vendor distribution is the way to go.

"We need to land data, prepare it, and serve it up to users more efficiently than we can currently in our data warehouse. Obviously, we don't have to use Hadoop for that. However, we like Hadoop's commodity-priced hardware and its elasticity, especially in the cloud. Furthermore, every time we extend our appliance-based data warehouse it costs another half-million bucks, which limits the number and size of extensions we can do. We'd like to migrate half of the data warehouse's data to Hadoop, which would free up capacity for new reporting solutions without having to buy another expensive appliance extension. New analytics—especially advanced analytics running against newer, larger data sets—would run mostly on Hadoop.

"At this point, I don't believe Hadoop will completely replace a relational data warehouse, but I do believe we can reduce our footprint on expensive relational databases by migrating some data to Hadoop. That would make our data warehouse platform more affordable and free up capacity for growth, which in turn makes it look more valuable from an economic perspective."

Hadoop's Applications Are Compelling

This report's survey asked participants to name the most useful applications of Hadoop if their organization were to implement it. (See Figure 1.) Data warehouse (DW) and business intelligence (BI) use cases percolated to the top of the sort order in Figure 1. This is no surprise because DW/BI uses cases for Hadoop are well established. However, the prominence of non-DW/BI applications (e.g., archiving, content management, and operational applications) shows that these are becoming more common among Hadoop users, and they are gaining mind share among organizations contemplating Hadoop. Again, this is a sign that Hadoop usage is diversifying across enterprises.

Hadoop's uses in BI, DW, analytics, and data management are well established.

Data warehouse extensions. Among TDWI members, Hadoop regularly appears as a complementary extension of a data warehouse (46%) when warehouse data that doesn't necessarily require the warehouse is migrated to Hadoop. A similar extension migrates data staging and data landing functions (39%) to Hadoop. Fork-lifting operational data stores (ODSs) (17%) to Hadoop is a trend that TDWI has just started seeing.[2]

Analytics and BI. Some of the hottest BI user practices of recent years involve data exploration and discovery (46%), which are critical to learning new facts about a business as well as getting to know new big data and its potential business value. To enable the broadest possible exploration, some users are co-locating numerous large datasets on Hadoop. Data exploration is usually the first step in an analytics project, so it's a fortuitous coincidence that Hadoop is also a capable computational platform and sandbox for advanced analytics (33%).

Data lakes, enterprise hubs, and data archiving are growing use cases for Hadoop.

Lakes and hubs. Data lakes (36%) and enterprise data hubs (28%) are two of the hottest best practices on Hadoop today. Both involve loading multiple massive datasets into Hadoop (easily reaching petabyte scale) with little or no preparation of the data. That way, data ingestion is fast, simple, and inexpensive. To make up for the lack of data modeling and data transformation, both lakes and hubs usually rely on data federation and virtualization techniques. These layer logical data structures over Hadoop that can perform data aggregation and transformation on the fly. Definitions of lakes and hubs vary, but the hub usually involves more data preparation than the lake, similar to how some operational data stores transform data slightly to make it more queryable but without losing original source details.

Data archiving. For legal, audit, and compliance reasons, many organizations are modernizing their enterprise data archiving facilities. The goal is two-fold:

- Manage and document archived data so its integrity and lineage are credible in an audit or compliance investigation
- Manage the data on a platform that is easily queried and searched (without time-consuming restoration) so that the business gets more operational use, BI use, and business value out of the archive

Users are finding that Hadoop has favorable economics and scalability for modern queryable archives, whether involving non-traditional data (Web, machine, sensor, social; 36%) or traditional enterprise data (19%).[3]

Miscellaneous. A few minority use cases for Hadoop are on the horizon, and will probably gain greater adoption by users in coming years, including content, document, and records management (17%) and operational application support (11%).

[2] Hadoop's many roles in DW and BI environments are described in the TDWI Best Practices Report *Integrating Hadoop into Business Intelligence and Data Warehousing*, available at tdwi.org/bpreports.

[3] For a description of modern data archive techniques, see the TDWI Checklist Report *Active Data Archiving for Big Data, Compliance, and Analytics*, available at tdwi.org/checklists.

In your perception, what would be the most useful applications of HDFS if your organization were to implement it? Select four or fewer.

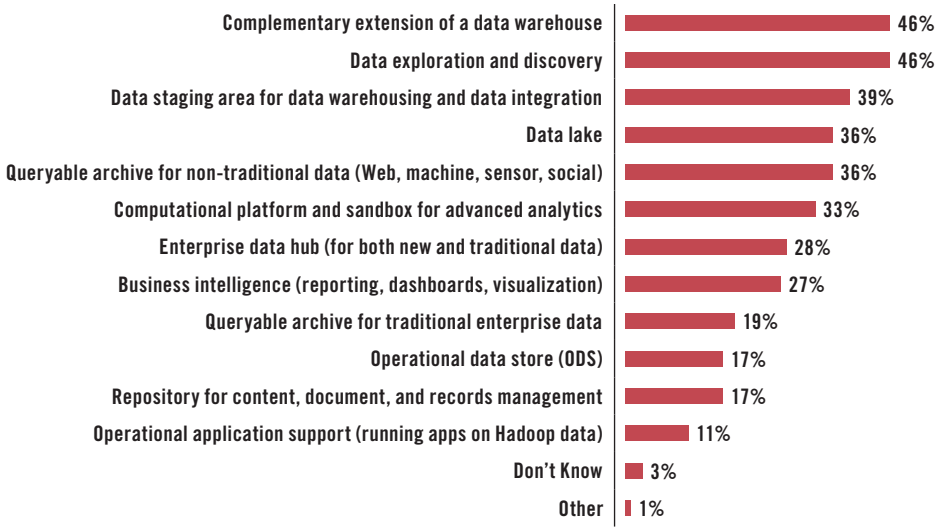


Figure 1. *Based on 743 responses from 207 respondents. 3.6 responses per respondent on average.*

USER STORY HADOOP IS ENABLING MODERN DATA ARCHIVE PRACTICES

"Our clients are big banks, and most are committed to Hadoop, often across multiple departments and divisions," said a data strategy manager at a firm that provides processing services for the financial services industry. "I think Hadoop adoption in my industry is driven by the need to retain sensitive data over the long term, including transactional data, e-mails, and customer communications of certain classes. For those use cases, we found that Hadoop scales up cost effectively, with the level of data integrity and security we need.

"I personally have in-house experience with data archiving and record management applications, which is our primary use of Hadoop. For us, archives are fully modern in that they are online and easily accessed in near real time, unlike the offline backups of the past. Likewise, we follow modern practices in archiving in that we prepare data before committing it to the archive on Hadoop, so the data is easier to access and better suited to complex search and query. Plus, we keep data lineage information, so the archived data is fully documented and trustworthy in audits, investigations, and legal activities.

"To get even more business value out of our live archive on Hadoop, we also use it as an analytic platform. For example, many of our operational reports are built with Hive, and some of our data visualizations and customer analytics are based on HBase data. Our next step is to conduct tests to see if Hadoop will be an appropriate data platform for transactional systems."

Hadoop Can Play a Role in Any Data Strategy

Survey responses suggest that Hadoop can play a role in enterprise data strategies, but it's not a very high priority at the moment. (See Figure 2.) Why?

Hadoop can contribute to data warehouse architectures and enterprisewide data strategies.

Hadoop has a large "foot in the door" at many organizations, and (as we've seen elsewhere in this report) Hadoop usage is proliferating across enterprises. Yet, Hadoop is still so new that its footprint is much smaller than that of other deployed data platforms. A user interviewed for this report put it succinctly: "Relational databases continue to be the most common type of data platform we have, by far. If you ask central IT what our data management strategy is based on, they'll probably say our big investment in network-attached storage subsystems. We keep finding more uses for Hadoop, but it'll be years before it's as high a priority or as big a footprint as older, incumbent data systems."

TDWI has seen Hadoop come out of nowhere to play constructive roles in the systems architectures and data strategies of data warehouse environments. Likewise, TDWI suspects Hadoop will soon be a common contributor to large-scale enterprise data strategies.

How important is Hadoop for the success of your organization's data strategy?

Figure 2. *Based on 247 respondents.*

A Diverse Range of Users Consider Hadoop Important

Survey respondents are unanimous in their zeal for Hadoop. We wanted to know why. To get their unvarnished opinions, the survey asked an open-ended question: "In your own words, why is implementing Hadoop important (or not important)?" The respondents' comments reveal a number of use cases, needs, and trends, as seen in the representative excerpts reproduced in Figure 3. Note that the users quoted work in many industries and locations. Hadoop is certainly top of mind for data professionals and their business sponsors in many contexts worldwide.

In your own words, why is implementing Hadoop important (or not important)?

Hadoop is critical for data scalability, cost containment, modernization, multi-structured data, and many use cases.

- "Cost savings. Linear scalability. Evaluate 'the hype' practically. Complement BI."
  —BI architect, telecom, Europe
- "Hadoop extends traditional data warehouse infrastructure, to support more data and speed up time to market."
  —BI lead, financial services, U.S.
- "Offers aggregation, analysis, and presentation options for non-structured and semi-structured data."
  —Program director, consulting, U.S.
- "Enables us to archive and query 10+ years of data for hundreds of clients."
  —Data architect, healthcare, U.S.

- "Reduces cost of data. New ability to query big data sets. Supply chain improvements. Predictive analytics."
  —Vice president, food and beverage, Asia
- "Our existing infrastructure cannot handle the 10-fold increase in data volumes."
  —Data strategy manager, hospitality, U.S.
- "We need to do ad hoc analysis with a greater volume/variety of data captured."
  —Data architect, insurance, Canada
- "Hadoop is useful for processing read-intensive models, but it is seen as a complement to our MPP databases."
  —Data architect, media/entertainment, Australia
- "Low-cost commodity hardware. Data lake approach is very attractive. Complements relational databases."
  —Practice manager, transportation, U.S.
- "It's important to realize the potential of big data and to explore new business opportunities."
  —Data specialist, consulting, Asia

Figure 3. *Based on 183 respondents.*

Although the vast majority of survey respondents are gung ho for Hadoop, a few expressed reasonable doubts and concerns about Hadoop's use cases, capabilities, and hype.

For example, A U.S. manager of data and analytics in the petroleum industry noted, "We have found no use case for Hadoop that our current MPP cannot handle at this time." The head of BI for a hospitality firm in Europe pointed out that "Hadoop can't support decision making, which requires the transformation of large data sets."

The principal architect of a software/Internet firm in the U.S. felt that Hadoop is mostly hype. "There are some uses cases, but most organizations can get away without using Hadoop."

Hadoop Adoption Is Accelerating

Many organizations are planning new Hadoop systems, and few will skip Hadoop.

In late 2014, our survey asked a question that was first asked in a TDWI survey conducted in late 2012, namely: "When do you expect to have HDFS in production?" Comparing the results of the two surveys shows that Hadoop adoption is up. (See Figure 4.)

Clusters in production are up 60%. In the 2014 survey, 16% of respondents report having HDFS already in production, which is a 60% gain over the 2012 survey. That's an impressive gain for a two-year period. HDFS is still being used by a minority of users and organizations; but if these gains continue, it will become a majority practice within five years.

Many more clusters will come online soon. A substantial gain is seen in the number of respondents expecting to bring HDFS clusters into production within 12 months, up from 28% in 2012 to 44% in 2014. If users' plans pan out, 60% of respondents will be in production by the first quarter of 2016, which is a huge leap from the current 16%.

Very few organizations have ruled out Hadoop. The percentage of respondents expecting to "never" deploy HDFS dropped from 27% in 2012 to 6% in 2014. This suggests a higher level of commitment from the user community than Hadoop has enjoyed in the past.

When do you expect to have HDFS in production?

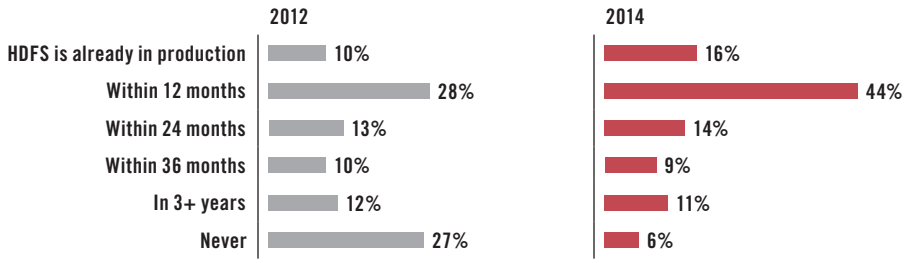


Figure 4. *Based on 263 respondents in 2012 and on 247 respondents in 2014.*

USER STORY A NUMBER OF FIRMS ARE PLANNING THEIR FIRST EFFORTS WITH HADOOP

"Our enterprise data architecture group just did a study of Hadoop, and we figure we're two or three years away from implementing Hadoop," said a data warehouse architect at a global insurance company. "The study determined that our first use cases for Hadoop should be extending existing risk analytics, improving the bottom line, mining social data, and bringing in more data from third parties. Other potential use cases will be fraud detection, mining data about annuities and indices, and cross-selling and up-selling. Our industry is risk-averse by nature, so a pure open source solution is not an option; our study determined that we should use a Hadoop distribution from a vendor, which would be lower risk, if it includes maintenance, additional tools, and consulting services."

Hadoop's Benefits and Barriers

Hadoop: Problem or Opportunity?

Hadoop is an opportunity for innovation.

As Hadoop proliferates across mainstream industries and the departments of individual enterprises, it is being embraced mostly by technology and business people new to the Hadoop ecosystem. These people must scramble to quickly learn Hadoop's infrastructure and development requirements, then push the envelope over time to expand and fine tune their best practices and applications. To test whether Hadoop is worth the effort and risk, this report's survey asked: "Is Hadoop a problem or an opportunity?" (See Figure 5.)

The vast majority (89%) consider Hadoop an opportunity. In fact, a growing number of users rely on Hadoop to spur enterprise business and technology innovations. For example, Hadoop can be the co-location point for multiple multi-terabyte datasets, which enables data exploration and discovery of an unprecedented scope, which in turn reveals both top- and bottom-line opportunities. Some Hadoop-driven innovations are incremental, such as expanding the data samples of existing customer analytics, which takes the firm to a higher level of customer service and account growth. Hadoop's ability to economically manage streaming data and exotic data types provides visibility into business processes and entities that were previously dark, thereby enabling new insights and management practices for running the business.

A small minority (11%) considers Hadoop a problem. The many organizations that have put Hadoop into production attest to the fact that once an organization identifies a compelling use case for Hadoop, it can learn Hadoop within weeks or months, then develop an application that provides benefits. As we'll see later in this report, users consider their lack of skills and experience with Hadoop to be the most likely barrier to implementing Hadoop—but it doesn't stop them.

Is Hadoop a problem or an opportunity?

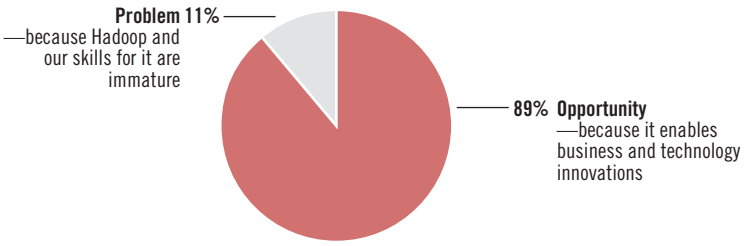


Figure 5. *Based on 247 respondents.*

Benefits of Hadoop

In the perceptions of survey respondents, Hadoop has its benefits. (See Figure 6.)

Hadoop's leading benefits concern analytics, scalability, DW, and new data types.

Advanced analytics. Hadoop supports advanced analytics, based on techniques for data mining, statistics, complex SQL, and so on (48%). This includes the exploratory analytics with big data (44%) that many organizations need to do today. It also includes related disciplines such as information exploration and discovery (36%) and data visualization (22%).

Data warehousing and integration. Many users feel Hadoop complements a data warehouse well (37%), is a big data source for analytics (45%), and is a computational platform for transforming data (26%).

Data Scalability. With Hadoop, users feel they can capture more data than in the past (28%). This is the perception, whether use cases involve analytics, warehouses, or active data archiving (26%). Technology and economics intersect because users feel they can achieve extreme scalability (28%) while running on low-cost hardware and software (34%).

New and exotic data types. Hadoop helps organizations get business value from data that is new to them or previously unmanageable, simply because Hadoop supports widely diverse data and file types (22%). In particular, Hadoop is adept with schema-free data staging (32%) and machine data from robots, sensors, meters, and other devices (17%).

Business applications. Although survey respondents are focused on technology tools and platforms, they also recognize that Hadoop contributes to a number of business applications and activities, including sentiment analytics (22%), understanding consumer behavior via clickstreams (20%), more numerous business insights (16%), recognition of sales and market opportunities (15%), fraud detection (14%), and greater ROI for big data (13%).[4]

[4] For more details about making a business case for Hadoop, see the TDWI Checklist report *Hadoop: Revealing Its True Value for BI*, available for free download on tdwi.org/checklists.

If your organization were to implement Hadoop technologies, which business processes, data, and applications would most likely benefit? Select eight or fewer.

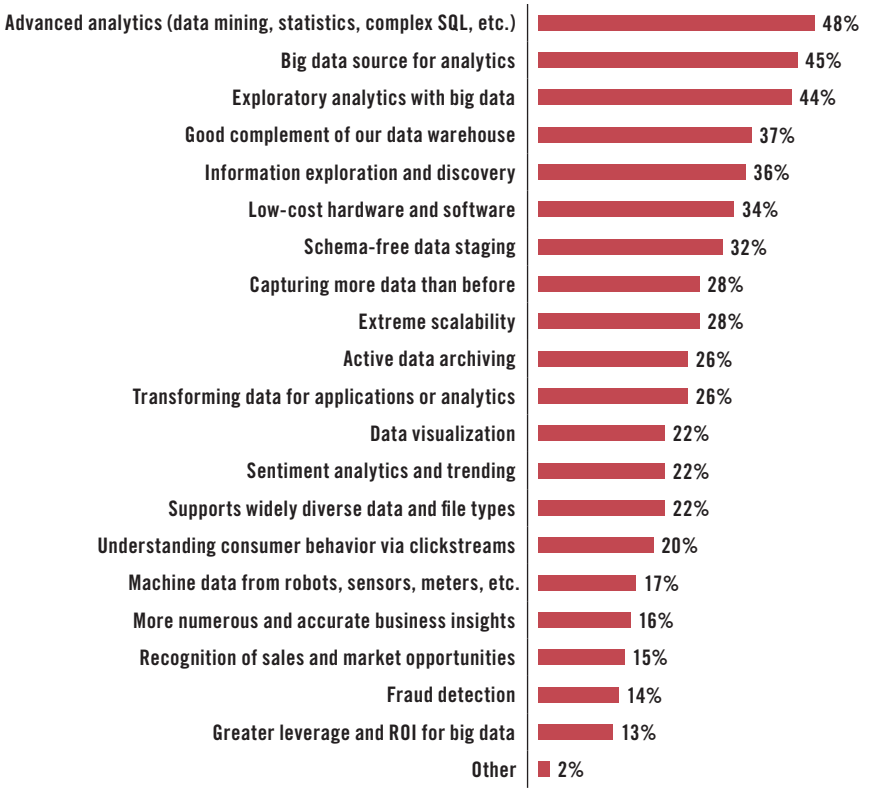


Figure 6. *Based on 1,348 responses from 247 respondents. 5.5 responses per respondent on average.*

Barriers to Hadoop

Hadoop has its benefits, as we just saw, but it also has potential barriers, according to survey results. (See Figure 7.)

Users are concerned about their weak skills and business support.

Skills gap. According to survey respondents, the leading barriers to Hadoop implementation are inadequate skills or the difficulty of finding skilled staff (42%). This is natural because Hadoop is still quite new, but it's not a show stopper. Determined users tend to learn Hadoop on their own without looking to hire rare personnel who have Hadoop experience.

Weak business support. As with any technical implementation, success is unlikely when there's a lack of compelling business case (31%), a lack of business sponsorship (29%), or a lack of data governance (29%).

Users are also concerned about Hadoop's weak support for security, metadata, SQL, and developers.

Security concerns. Business and technology people are concerned about securing Hadoop data (29%) the same way they're concerned about securing other new enterprise platforms, such as clouds, appliances, and NoSQL databases. TDWI feels that as familiarity with these platforms grows, so will users' confidence in their security measures. To help Hadoop users achieve enterprise-class manageability (17%), a number of Hadoop distributions and other tools from software vendors include additional data security, masking, and encryption capabilities that go far beyond the mere access authorization built into purely open source Hadoop.

Data management hurdles. Enterprise data management professionals used to mature relational database management systems are often deterred by Hadoop's lack of metadata management (28%), immature support for ANSI-standard SQL (19%), and limited interoperability with existing systems or tools (19%). However, as these professionals work with Hadoop and understand its methods, they see that metadata is managed at run time (not *a priori*), and Hadoop has its own approach to queries and relational data structures (as seen in Hive, HBase, Drill, Impala, etc.). These methods have advantages with the unstructured and schema-free data that Hadoop typically manages, so users adopt them. Furthermore, Hadoop's approaches to metadata, SQL, and standard interfaces improve regularly.

Tool deficiencies. Open source and vendor-supplied development tools are evolving quickly, but developing solutions with Hadoop still requires excessive hand coding (27%) in languages that few data management professionals know, such as Java, R, Hive, and Pig. This is a skills issue, but it also slows down development (16%). This situation will improve as software tools build out high-level language support (10%).

Containing costs. Survey respondents and interviewees alike feel confident that Hadoop's low costs are real and attainable. Yet, they still have small concerns about the cost of staffing (25%), implementing (22%), and operating (9%) Hadoop.

What are the most likely barriers to implementing Hadoop technologies in your organization? Select eight or fewer.

Figure 7. *Based on 1,162 responses from 247 respondents. 4.7 responses per respondent on average.*

EXPERT COMMENT THE DATA IS THE WAREHOUSE, REGARDLESS OF THE PLATFORM TYPE THAT MANAGES THE DATA.

According to a data warehouse professional interviewed for this report, "What is a data warehouse? Our view is that the data is the warehouse, and our data just happens to be managed with a relational database today. Our data could be managed on a non-relational platform, and it would still be a warehouse.

"For example, over the years we've migrated our warehouse several times. Twice we've migrated from one vendor brand of database to another. We migrated from 16-bit platforms to 32-bit ones and then 64-bit. We migrated from SMP computing architectures to platforms based on MPP. We've migrated data more times than we can remember as we've moved data among the core warehouse, data marts, operational data stores, file systems, and appliances. Now we're migrating some of the data from those platforms to Hadoop. Even if we migrated to Hadoop all the data in our extended data warehouse environment, the data warehouse would still live on. It would simply be managed on Hadoop instead of one of the many other data platforms we've used.

"The idea that Hadoop would replace a warehouse is misguided because the data and its platform are two non-equivalent layers of the data warehouse architecture. It's more to the point to conjecture that Hadoop might replace an equivalent data platform, such as a relational database management system."

Organizational Practices for Hadoop

Ownership of Hadoop

Hadoop environments may be owned and primarily used by a number of organizational units. In enterprises with multiple Hadoop environments, ownership can be quite diverse. (See Figure 8.)

Hadoop as an analytic platform may be owned by teams for DW, operational research, or departmental BI.

Central IT (41%). As Hadoop broadens into enterprise usage, it is becoming more often the property of central IT instead of individual programs and departments. This is natural because centralized IT is primarily a provider of infrastructure, namely networks, data storage, and server resources. Hadoop is becoming just another infrastructure provided by central IT, with a wide range of applications tapped into it, not just analytic ones.

Data warehouse group (28%). In organizations intent on integrating Hadoop into the practices and infrastructure for BI, DW, DI, and analytics, it makes sense that the DW group or an equivalent BI team deploy and maintain its own Hadoop environment.

Business unit or department (7%). The earliest Hadoop implementations were almost exclusively departmental, supporting only the analytic applications of that department. TDWI saw this proclivity fade as Hadoop was picked up by enterprise BI and DW teams. Now that mainstream industries want Hadoop as IT infrastructure, departmental ownership is far less common in such organizations.

Other owners of Hadoop include research teams (which support product developers who depend heavily on data), third parties (cloud providers, data center outsourcing), and application groups.

Who owns and maintains your Hadoop environment? Select only one.

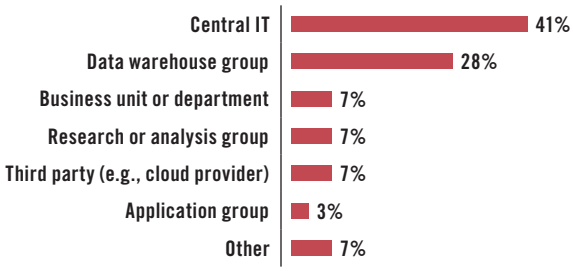


Figure 8. *Based on 100 respondents with Hadoop experience.*

Job Titles for Hadoop Workers

As we saw in our discussion of potential barriers to Hadoop success, many users are concerned about hiring and training the appropriate skills. One way to approach a better understanding of skills is to look at the job titles of people who regularly work with Hadoop. In that spirit, this report's survey asked respondents who have Hadoop experience to enter the job titles of Hadoop workers. (See Figure 9.)

Hadoop workers are often data scientists, architects, developers, and analysts.

Data Scientist. TDWI has seen the job title "data scientist" appear out of nowhere a few years ago and rise to fair prominence today. It's the most common title reported by the current survey, although it was in fourth place when TDWI last asked this question in late 2012. In most cases, a data scientist is essentially a cross-trained BI/DW professional who can also program advanced analytic algorithms in a variety of languages, plus design complex multi-platform data and systems architectures. That's a long and daunting list of desirable technical skills, which makes the data scientist a high-payroll employee. Because analytics in a Hadoop environment requires the entire list, the data scientist is a natural choice (albeit an expensive one) for working with Hadoop.

Architect. It's interesting that the word *architect* appears in many job titles. These include architects for data, BI, applications, and systems. Most architects (regardless of type) guide designs, set standards, and manage developers. Architects are most likely providing management or governance for Hadoop because Hadoop has an impact on data, applications, and system architectures.

Developer. Similar to *architect*, this word appeared in many job titles. Again, there's a distinction between application developers and data (or DW/BI) developers. Application developers may be there to satisfy Hadoop's need for hand-coded solutions, regardless of the type of solution. The data and BI developers obviously bring their analytics expertise to Hadoop-based solutions.

Analyst. Among other things, Hadoop is a powerful computation platform for analytics, which is why data analysts (8%) commonly work with it, as do others who rely on analytics in their work such as engineers (7%) and consultants (3%).

Miscellaneous. The survey uncovered some new job titles, namely big data specialist (8%) and Hadoop specialist (3%).

Enter the job titles of people who design and execute applications using Hadoop technologies.

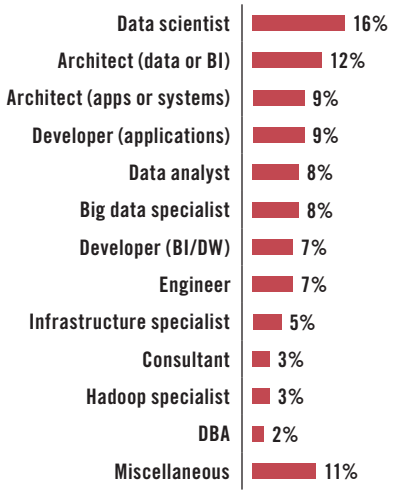


Figure 9. *Based on 110 responses from 76 respondents who have experience with Hadoop. 1.5 responses per respondent on average.*

Staffing Hadoop

Multiple staffing strategies are available and they can be combined.

An ideal strategy for staffing Hadoop would be to hire multiple data scientists. The challenge for that strategy is that data scientists are rare and expensive. In fact, people with any Hadoop work experience are likewise rare, and (because of demand) most are employed currently without need for a new job. Hence, hiring new employees with Hadoop experience (41% in Figure 10) is a possibility but introduces challenges.

Recall that most BI/DW teams are heavily cross-trained; each employee has skills in multiple areas, from report design to ETL development to data modeling to data discovery. Having a pool of cross-trained team members makes it easy for a manager to assign people to projects as they arise, and each team member can "pinch hit" for others when necessary. Most BI/DW professionals prefer the diverse work assignments that cross-training enables. Hence, in the tradition of BI/DW cross-training, most teams are training existing employees for Hadoop skills (73%) instead of hiring.

Another tradition is to rely on consultants when implementing something entirely new to the organization or that involves time-consuming and risky system integration. Hadoop fits both of these criteria, which is why many BI/DW teams are depending on consultants for Hadoop skills (36%). Likewise, many teams get their training and knowledge transfer from experienced consultants.

Finally, more than one staffing strategy can be appropriate, and the three discussed here can be combined differently at different life cycle stages. For example, you might start with consultants when Hadoop enters an enterprise, followed by some cross-training, then add more full-time employees as Hadoop matures into enterprisewide use.

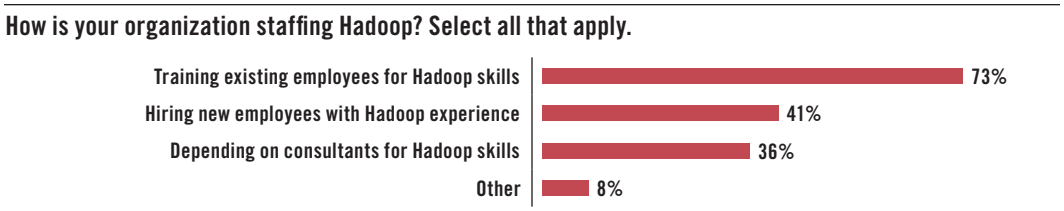


Figure 10. *Based on 157 responses from 99 respondents. 1.6 responses per respondent, on average.*

USER STORY IN RARE CASES, HADOOP MIGHT REPLACE A RELATIONAL DATA WAREHOUSE PLATFORM. According to a data manager in medical research: "Our bespoke data warehouse isn't really a data warehouse. It's just an archive of relational data from a short list of packaged applications, along with lots of logs from miscellaneous applications. This is all we need, and we have no plans for a more sophisticated warehouse. All of the warehouse data has rather simple data models, and our tests have shown that the data is easily managed and queried in Hadoop, using mostly Hive and HBase, with a little Pig and MapReduce. We're planning to migrate the warehouse to Hadoop to get it off of the highly expensive relational database it's on today. Again, our tests show that this should be a simple 'forklift.' In other words, we'll copy data from point A to point B, then spend a couple of weeks tweaking to get the performance optimization and integration with reporting tools that we need. If this works out as well as our testing suggests, we will soon decommission the relational platform that the warehouse runs on today."

Best Practices for Enterprise Hadoop

The survey responses discussed in this section of the report come from a subset of 100 respondents who report that they have experience with Hadoop.[5] As with the total survey population, this subset is dominated by BI/DW professionals. Based on direct hands-on experience with both Hadoop and BI/DW systems, their responses provide a credible glimpse into emerging best practices for implementing Hadoop and integrating it with other enterprise systems.

Securing Hadoop

Users need multiple security methods for Hadoop, not just Kerberos.

Security in purely open source Hadoop is limited to file-permission checks and access control based on Kerberos mechanisms, which ensure that clients connecting to a particular Hadoop service have the necessary, pre-configured permissions. This kind of authorization is important, but it's only one approach to security, whereas mature enterprise IT teams tend to prefer multiple approaches for security to attain desirable redundancy and to simply have more options for diverse situations. Given that security is the number one potential barrier to Hadoop adoption (see our discussion of Figure 7), Hadoop adoption at the enterprise level could be hamstrung unless users get a more comprehensive array of security functionality. To supply this demand, a number of vendor and open source efforts are available or coming.

Roles and directories are priorities for Hadoop security.

Directory-based security. Many IT organizations have standardized on role-based and directory-based approaches to security, especially LDAP and Active Directory, so they need to see these supported in Hadoop (or in tools sitting atop Hadoop). Likewise, such approaches enable single sign-on, which is a requirement for some users. Furthermore, Active Directory makes security manageable and scalable on Hadoop in that administrators can define user-based security in one place, and it will spread throughout all layers of Hadoop.

[5] Not all of the 100 respondents answered every survey question, so the number of respondents varies per question.

Apache Sentry is an open source project incubating under the guidance of the Apache Software Foundation, with contributions from both the open source community and vendors. Sentry is a system for enforcing fine-grained, role-based authorization to data and metadata stored on a Hadoop cluster. Therefore, it addresses many users' real-world needs.

Granular security. In general, HDFS is deployed and treated by tools as one big data volume due to limitations in the current software. Various approaches to data volumes in Hadoop (including virtual approaches) are emerging, such that security can be applied per volume in a more granular manner.

Other areas for improvement. Eventually, users will also demand encryption and (possibly) data masking capabilities for Hadoop; if these appear, they'll need to be high-performance and scalable to be practical with the very large datasets of the Hadoop world. As Hadoop security options increase in number, application vendors will need to support them natively. Finally, add-on products that provide additional security measures are available for Hadoop from a few third-party vendors today, especially vendors known for their Hadoop distributions. In fact, enhancements for security (and related areas such as cluster administration) are good reasons to use a vendor's distribution instead of open source Hadoop direct from Apache.

Data Architecture Issues

Hadoop is progressively integrated into complex multi-platform environments.

As noted earlier (in our discussion of Figure 8), early deployments of Hadoop were mostly silos, supporting a few (but very large) analytic applications, typically at a departmental level. However, the trend today is toward Hadoop clusters that source data for many tools and applications, and therefore must integrate and interoperate with many types of systems. Hadoop's movement from departmental silo to integrated enterprise system has a number of architectural ramifications.

For example, Hadoop is progressively becoming a permanent fixture in data warehouse environments. The appearance of Hadoop usually coincides with the movement of certain datasets from the warehouse to Hadoop and other platforms. Data staging structures, operational data stores (ODSs), and stores of extracted source data are usually the first to migrate to Hadoop or other non-warehouse platforms.

Hadoop affects the architectures of the data ecosystems with which it's integrated.

It's possible to migrate such data without altering the data warehouse's logical architecture, which defines data structures without pinning them to specific physical locations. Even so, savvy users leverage data migration projects to make improvements to data and its architecture, under the assumption that we should improve data, not just move it. However, introducing new data platforms such as Hadoop (but also columnar databases, appliances, and NoSQL platforms) changes the topology of software and hardware servers, which constitute the systems architecture of the physical warehouse. You can see that Hadoop's arrival in a data ecosystem will naturally entail adjustments to that ecosystem's architectural layers.

Note that the data warehouse is—in many ways—a microcosm of modern hybrid data environments seen elsewhere in the enterprise, including the multi-platform environments for operational applications, transactional applications, data archives, and content management. Hadoop likewise forces changes in those complex architectures.

USER STORY RELATIONAL WAREHOUSES, HADOOP, AND OTHER PLATFORMS COEXIST IN MODERN HYBRID DATA ARCHITECTURES.

"Like a lot of organizations, we spent several years building up a report-oriented data warehouse on a mature relational database platform," said Dirk Garner, the director of data technology innovation at Macys.com, "but new business requirements and opportunities arose, so we diversified the platform types in the warehouse environment. For example, we deployed a standalone columnar database, which is good for the intense ad hoc queries that SQL-based analytics requires. As another example, we've deployed a vendor distribution of Hadoop, which is ideal for managing the mountains of credit-card transactions our e-commerce application generates. There are other platforms for specialized applications in real-time, analytics, graphs, and so on. The result is a hybrid data architecture, where multiple types of data platforms and data structures coexist.

"The upside is that we have the right platform for storing or processing data of diverse schema, formats, origins, and uses. The downside is that we have many standalone platform types from many vendors, and data is strewn across them all. We feel that the complexity of this hybrid data environment is worth integrating and managing so that individual workloads and users get the best results possible.

"To be sure we stay ahead of the complexity and keep the diverse platforms integrated, we are going deeper into techniques for data federation, data virtualization, and the logical data warehouse. These techniques have enabled us to establish a common centralized toolset for consuming all data, and the toolset's virtual semantic layer accesses and presents data from many diverse platform types, as if they are in a single database. This, in turn, reduces data redundancy and data movement among platforms, while increasing our ability to govern and standardize data."

Tool Types for Hadoop Development

Recall from our discussion of Figure 7 that the hand coding often required of development with pure open source Hadoop is one of the barriers to Hadoop adoption. To quantify what Hadoop users are doing in this realm, our survey asked: "Today, are you hand coding Hadoop or using vendor-supplied Hadoop tools and frameworks?" (See Figure 11.)

Hadoop users depend on many types of tools for development, even more for deployment.

Roughly a quarter of Hadoop users surveyed (23%) are hand coding most of their solutions. Languages include Java and R. Pig is a higher-level open source tool that simplifies programming for MapReduce; Pig generates Java that's optimized for MapReduce. As Hadoop's support for ANSI SQL improves, hand-coded SQL will probably become common.

A small minority of Hadoop users (14%) use vendor-supplied Hadoop development tools exclusively. These include Spark, Impala, Solr, Drill, and several tools from the IBM InfoSphere BigInsights family.

A mix of tool types is the norm (58%) for Hadoop development. Besides the tools already mentioned, TDWI has found traditional tools for data modeling, programming, utilities, and technical project management applied to Hadoop development. Similarly, traditional tools for reporting, analytics, and data integration regularly access Hadoop data.

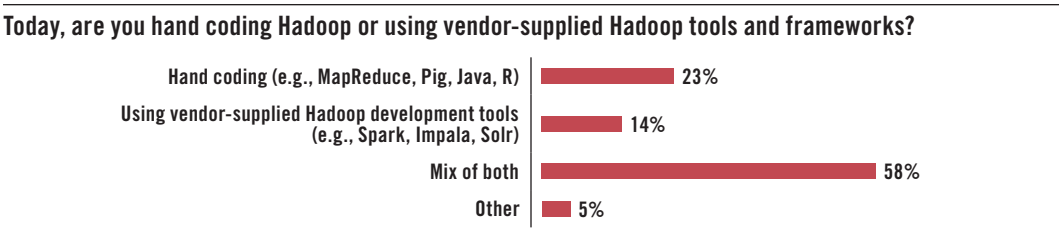


Figure 11. *Based on 99 respondents.*

HDFS Clusters and Nodes

Median enterprises have only four HDFS clusters.

Number of HDFS clusters per enterprise. One way to measure Hadoop adoption is to count the HDFS clusters per enterprise. When asked how many HDFS clusters are in production, 82 survey respondents (who report having Hadoop experience) gave answers ranging from one to 100. (See Figure 12.) Most responses were in the single digits, and only 13 out of 82 respondents reported 20 or more clusters. This drove the average number of HDFS clusters down to 10 and the median down to 4.

Note that ownership of Hadoop products can vary, as discussed earlier, thereby affecting the number of HDFS clusters. Sometimes central IT provides a single, very large HDFS cluster for shared use by departments across an enterprise. At the other extreme, departments and development teams may have their own, sometimes a cluster per analytic application.

The size of HDFS clusters is up, with 250 nodes on average and a median at 32 nodes.

Nodes per HDFS cluster. We can also measure HDFS cluster maturity by counting the nodes in the average cluster. When asked how many nodes are in the HDFS cluster they most often used, respondents replied in the range of one to 5,000. (See Figure 13.) Most responses were in double digits, and only 5 out of 77 respondents reported 1,000 or more nodes. That comes to 250 nodes per production cluster on average, with the median at 32. If we exclude the outliers (where the number of nodes exceeds 1,000), the average comes down to 92 nodes, with a median at 22. With or without outliers, the node count is up considerably from the last time TDWI asked this question in late 2012, which revealed 45 nodes on average, with a median at 12. Clearly, individual Hadoop clusters are maturing into larger sizes.

To put these numbers in context, note that TDWI conference presenters from large Internet firms have spoken about HDFS clusters with approximately 1,000 nodes. However, speakers discussing mature HDFS usage in data warehousing usually have 50 to 100 nodes. Proof-of-concept, department, and development clusters observed by TDWI typically have no more than eight nodes. Hence, in terms of node count, Hadoop clusters can scale down to departmental size but also scale up to the most demanding enterprise datasets.

Enter a number representing the approximate number of HDFS clusters in production across your enterprise

Range	=	1 to 100
Average	=	10
Median	=	4

Figure 12. *Based on 82 respondents who have experience with Hadoop.*

For the HDFS cluster in production you work with most, enter an integer representing the number of nodes

Range	=	1 to 5,000
Average	=	250
Median	=	32

Figure 13. *Based on 77 respondents who have experience with Hadoop.*

Locating Your Hadoop Cluster

Hadoop is mostly on premises today, but going more off premises soon.

Most Hadoop clusters are located on premises today (51%). (See Figure 14.) In addition, some clusters (10%) run on a private cloud located on premises.

Off-premises clouds are a viable choice for locating Hadoop clusters. This includes third-party cloud providers (9%), plus providers whose infrastructure (or part of it) just happens to be cloud-based, such as managed service providers (5%) and software-as-a-service providers (2%). These and other providers now offer services and infrastructure specifically tuned for Hadoop, which shortens users' time to using Hadoop, avoids risky system integration, and cuts administrative costs. As a variety of cloud types gain prominence as enterprise system platforms, TDWI expects to see more Hadoop implementations on clouds.

Hybrid approaches are also possible. In other words, some users employ both on-premises and off-premises (usually cloud-based) Hadoop (18%).

For the Hadoop implementation you work with most, where is it deployed?

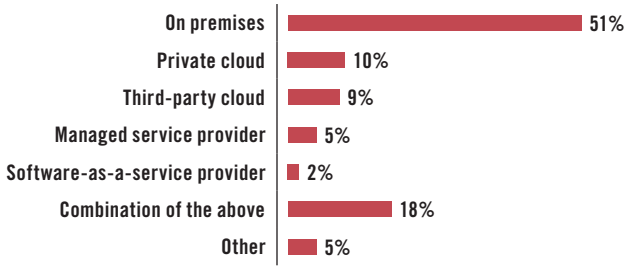


Figure 14. *Based on 100 respondents.*

SQL and other Relational Functions in Hadoop

Key use cases demand better SQL on Hadoop.

As data management professionals have gone deeper into Hadoop usage, many have determined that important use cases require so-called "SQL on Hadoop"—that is, where Hadoop natively supports the execution of ANSI-standard SQL. Desirable use cases include SQL-based analytics and the management of tabular data. One of the hotter trends is to push ETL's transformation processing down into HDFS, following an ELT model often used with relational databases. Transformation logic varies considerably, but much of it is inherently relational because it involves complex table joins and SQL.

As one user put it: "We absolutely feel the need for more and better SQL support in Hadoop, especially in compliance with the ANSI standard. That's because our current skills and tools rely on ANSI SQL. We also need Hadoop's SQL support to be low latency and low cost." The vast majority of survey respondents also feel they must have SQL on Hadoop (69%). (See Figure 15.)

For your organization, how important is "SQL on Hadoop"—that is, Hadoop tools that support ANSI-standard SQL for queries against data managed on Hadoop?

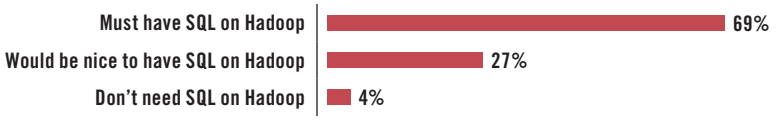


Figure 15. *Based on 99 respondents.*

Hive is useful, but not standard SQL.

Some argue that Hadoop Hive already fulfills SQL requirements, but that's unlikely. Although the Hive Query Language (HiveQL) is very useful, it is not ANSI SQL. Data management professionals who know SQL can learn HiveQL quickly, but that doesn't make their numerous SQL-based tools Hadoop compliant. Furthermore, using HiveQL does not leverage their SQL skills. Luckily, more software vendors are building Hadoop interoperability, Hive and HBase support, and SQL-to-HiveQL translation into their tools for reporting, analytics, and data integration.

SQL off Hadoop is fine for some use cases.

Another argument is that "SQL off Hadoop" is the way to go. In this scenario, data is managed and pre-processed on Hadoop, then moved to relational databases and tools that comply with SQL. On one hand, this scenario fits well with multi-platform data architectures, where we expect data to move to the platform best suited to a specific workload. On the other hand, SQL off Hadoop works against a common goal of Hadoop, which is to offload relational databases.

Relational functions are key to enterprise acceptance of Hadoop.

To exacerbate the situation, Hadoop's support for standard SQL and other relational functions are key to making Hadoop more palatable to traditional enterprise entities, such as central IT, database administrators, and data warehouse professionals. After all, SQL is more than a query language; when coupled with ODBC/JDBS, it's the most common interface today for interoperating among diverse IT systems. Attaining better INSERT and UPDATE functionality is a gateway to more operational and transactional applications with Hadoop datasets.

Users want it all: better relational functionality, both on and off Hadoop.

How do we resolve the argument over SQL on Hadoop versus SQL off Hadoop and related issues? TDWI always prefers that users have diverse options to consider so they have a better chance of finding an approach that fits their current requirements. In that spirit, most users want it all, namely *both* SQL on Hadoop and SQL off Hadoop. They also want improvements to relational functions in Hive and HBase, plus related functions such as metadata management and indexing.

The vendor and open source communities are well down the road to these improvements. Folding HCatalog into Hive was a giant step forward. Open source projects such as Impala and Drill bring better tools and capabilities for queries against Hadoop data, as do a number of vendor-supplied tools.[6] Finally, SQL off Hadoop gets better with each vendor tool release. Again, Hadoop improves almost constantly; query and relational capabilities are the target of many improvements.

[6] For examples of vendor-supplied tools for Hadoop, see the section "Vendor Platforms and Tools in the Hadoop Ecosystem" later in this report.

Data Quality Techniques for Hadoop

As Hadoop goes enterprise scope, data quality and other enterprise standards become more urgent.

Data quality is one of many data management best practices that seem to get short shrift with big data applications, both on and off of Hadoop. Other practices being ignored include data governance, stewardship, and master data and metadata management. That's because most of the early Hadoop applications were about advanced forms of analytics that tolerate data in poor condition (based on mining, clustering, or statistical approaches), and data volumes make it impractical to move data for quality processing. Reporting on Hadoop continues to involve reports that require neither great accuracy nor much of an audit trail. In other words, a Web traffic report on Hadoop is a far cry from the intensely scrutinized, cleansed, and documented financial report that's typical in traditional BI and DW contexts.

Data quality is an established enterprise requirement, and it applies increasingly to Hadoop as Hadoop is used in more mainstream enterprise applications. Hence, with enterprise Hadoop, you ignore data quality at your peril. The catch is that Hadoop best practices are still evolving, and users are rethinking data management as they go. For example, for decades we've been faced with key questions: When do we apply data quality functions: before loading data into a target database or after loading? Do we cleanse whole databases or just a record at a time? Which data quality functions do we apply as applications access data?

In short, Hadoop users tend to go with practices that load data first and improve it only when absolutely necessary. Given the extreme amount of data typically managed by Hadoop, data quality functions down in the cluster, without moving data, have become a firm requirement. Examples include name standardization, de-duplicating customer lists, and extracting semantic meaning from textual data.

Hadoop methods apply data management best practices, but in a different order compared to traditional approaches.

To learn what Hadoop users are doing today, TDWI asked: "Which of the following best describes your primary strategy for improving the quality of data managed on Hadoop?" (See Figure 16.)

Most users have a strategy for improving Hadoop data. Only 22% say they don't have a strategy.

The norm is to ingest data into Hadoop immediately, then improve it later as needed (24%). This is especially true when data is repurposed for analytics (21%). Hadoop data is regularly processed at run time, regardless of application, so it makes sense to follow this paradigm with data quality functions on the fly (14%).

Few users improve data before it enters Hadoop (18%). This is what we've done in data warehousing forever, yet it's a rare practice with Hadoop. In a lot of ways, data management with Hadoop is in reverse order compared to data warehousing. Many of the same data management disciplines and best practices apply to both. However, in data warehousing we apply most of them before loading a target database, whereas Hadoop methods start with data ingestion, then apply data management practices later, often ad hoc at run time. For many seasoned data management veterans, wrapping their heads around Hadoop's unorthodox practices is a bit stupefying at first, but they soon comprehend it and work within the new paradigm.

Which of the following best describes your primary strategy for improving the quality of data managed on Hadoop?

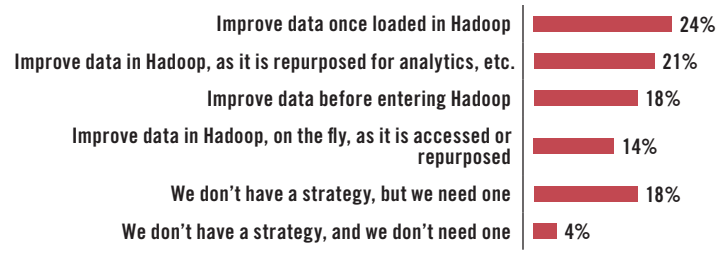


Figure 16. *Based on 99 respondents who have Hadoop experience.*

USER STORY USER EXPECTATIONS FOR THE QUALITY OF HADOOP DATA ARE SOMEWHAT UNREALISTIC AT THIS TIME.

One interviewee summed up the situation thus: "People want Hadoop to work like an RDBMS, but without the time and pain put into data preparation that an RDBMS requires. This is not a realistic expectation, because—for enterprise applications—failing to prepare Hadoop data properly just leads to later problems with data integrity, quality, standardization, and semantics. Even so, there might be some exceptions, especially in analytics, like when you dump unwashed data into a data lake and then improve the data slightly as you explore it and analyze it."

First Impressions of YARN

Enterprise-grade Hadoop came closer to reality in 2013 with the release of Hadoop 2 and its new data operating system called YARN (Yet Another Resource Negotiator). It subsumes MapReduce and adds much of the functionality of a modern operating system to enable a wider range of concurrent, self-managed, interactive, and real-time use cases. This makes Hadoop's new YARN-based architecture far more palatable than Hadoop 1 to a wider range of mainstream industries, user organizations, and IT departments.

YARN is a positive step toward making Hadoop enterprise grade.

This report's survey identified 16 respondents who've been using YARN as part of their Hadoop implementation. To get honest, straightforward opinions, TDWI asked these respondents to briefly explain why and for what purposes they are using YARN.

Several users said they use YARN simply because it's built into the vendor distribution of Hadoop they use. One of the many advantages of a distribution is that users get the latest Hadoop tools thanks to the distribution's maintenance agreement.

A couple of respondents said they upgraded Hadoop to the new version because they needed higher performance for MapReduce. Plus, YARN enables concurrency, so they can now run multiple MapReduce jobs at a time in a self-managed fashion instead of in serial execution.

Other improvements in Hadoop 2 and YARN mentioned by respondents include: greater speed of execution, more reliable high availability, easier administration, and a number of job queue and resource management features.

Trends in Hadoop Implementations

In this section, we examine where Hadoop users are today and where they're going with their applications, platforms, and tools. Respondents reveal what tools, techniques, feature functionality, and use cases they are using today and which they hope to be using within three years. The survey responses examined here come from the subset of 100 respondents who report having experience with Hadoop; hence, the responses are real-world and credible. These predictions are intended to assist Hadoop users—and those contemplating Hadoop usage—with their planning of future implementations or upcoming transitions in existing implementations.

New and Upcoming Use Cases for Hadoop

Expect faster adoption of enterprise data hubs, archives, and BI/DW use cases.

The survey question charted in Figure 17 asked: "What are your organization's top three uses of Hadoop today? In three years, what do you think your organization's top three uses of Hadoop will be?" The rightmost column of the chart quantifies a growth indicator for each use case listed as a choice in the multiple-choice survey question. A growth indicator is calculated by subtracting the value for "today" from that for "in three years." The greater the growth indicator, the more likely that use case will see additional adoption by Hadoop users in coming years.

Enterprise data hubs (EDHs) will experience significant user adoption. In fact, EDHs have the greatest growth indicator in Figure 17 (9%). EDHs are compelling because they strike a practical balance. An EDH provides fast ingestion and uncomplicated management for a wide range of data types and sources while providing additional data preparation and improvement for data, applications, and business processes that demand better data. By definition, an EDH targets broad enterprise use and therefore signals Hadoop's maturation into enterprise usage.

Archives on Hadoop will continue to proliferate. Most of these will follow modern archiving practices that demand an archive be online and accessible via query and search. So far, TDWI regularly finds Hadoop-based archives for non-traditional data and content (Web, machine, sensor, and social data). However, as Hadoop serves more of the enterprise, the significant growth will be with queryable archives for traditional enterprise data (7% growth).

Expect greater adoption of use cases in BI, DW, and analytics. These are already well established and will experience increased user adoption, as seen in their 4% to 5% growth indicators.

Non-BI/DW use cases are coming. So far, Hadoop has mostly been about analytic processing and the data management methods required for advanced analytics. We've already discussed data archiving. Other examples include operational application support and content management (4% growth indicated for each).

Very common use cases today include data lakes, data exploration, and data staging.

Some use cases are so common already that growth will be limited. Ironically, data lakes, data exploration, and data staging have negative growth indicators in the survey responses of Figure 17, despite being the most common Hadoop use cases deployed today. The negative growth indicators do not mean these cases will decline; they are just so proliferated that few existing Hadoop users expect to add them later.

Data lakes, data exploration, and data staging are typical first phase deliverables in a Hadoop implementation because all three enable users to study their big data before making more specific application decisions. This explains why these use cases are so common today, but also why users contemplating a new implementation should consider developing all three—plus plan to go *beyond* all three.

What are your organization's top uses of Hadoop today? In three years, what do you think your organization's top uses of Hadoop will be? Select three or fewer answers for each question.

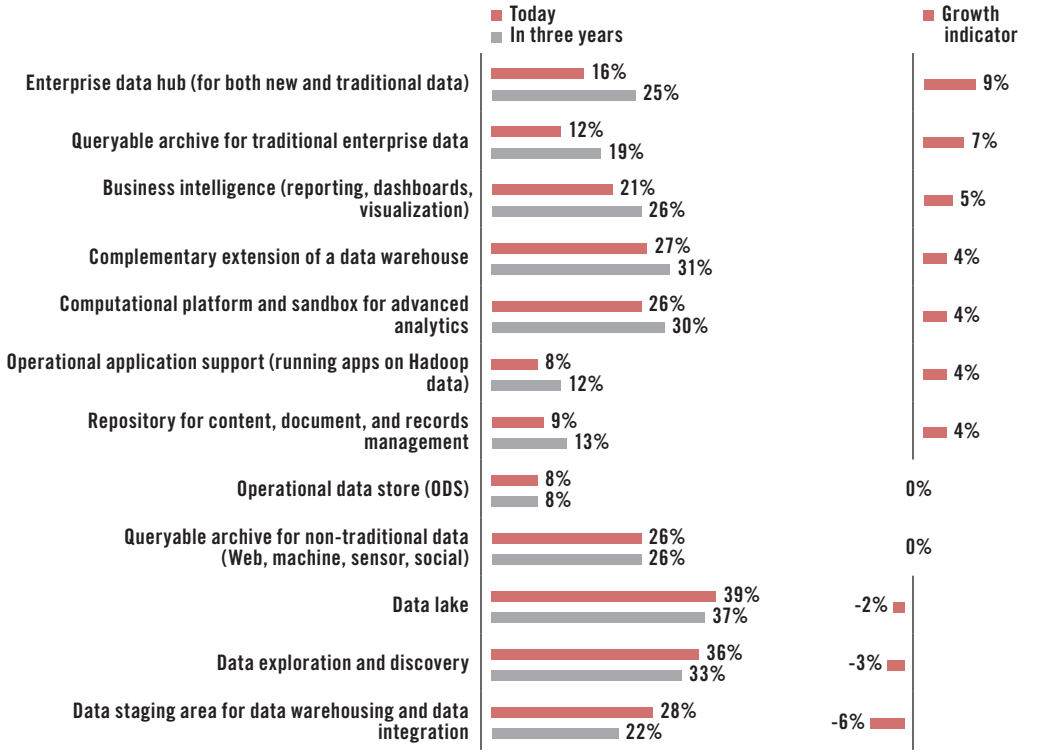


Figure 17. *For "today," based on 262 responses from 100 respondents; 2.6 responses per respondent on average. For "in three years," based on 288 responses from 100 respondents; 2.8 responses per respondent on average. Sorted by "Growth indicator."*

USER STORY HADOOP IS A GOOD PLATFORM FOR DATA-DRIVEN OPERATIONAL APPLICATIONS.

"I believe that Hadoop has matured to a point that people can successfully build large and complex applications atop the platform," said Bob Zurek, SVP of products at global marketing firm Epsilon. "Hadoop has met our scalability requirements for handling large and varied types of data. It is not just for business intelligence and data warehousing. There's a new class of business applications coming that are mostly data-driven in nature, even when there's an analytic component embedded.

"For example, our new cloud-based campaign management platform, Agility Harmony, runs atop a Hadoop cluster in our global data centers. Harmony collects data from a variety of sources and provided by our customers. Our customers can then create advanced segments for targeting their opt-in customers who desire to receive messages across channels including e-mail and SMS. For example, a business may want to send only messages to customers who have a tendency to purchase sports equipment or a certain outerwear. Harmony then takes the customers' interactions (including e-mail clicks and opens, as well as type of device) and enables our customers to gauge the success of their campaigns using a variety of analyses provided by Harmony. Although new to market, Harmony is sending billions of messages on behalf of our customers and this is thanks to the scalability and performance of Hadoop, running across multiple nodes in our data centers."

Tools and Platforms Integrated with Hadoop

BI/DW integration is common today, but DQ, MDM, and machines are absent.

This report's survey asked respondents: "In your organization, with which platform and tool types is Hadoop integrated today? With which will Hadoop be integrated within three years? For which do you have no plans to integrate?" Their responses are charted in Figure 18. The rightmost column of the chart quantifies growth for each tool or platform listed in the multiple-choice survey question. A growth indicator is calculated by subtracting the value for "integrated today" from that for "will integrate within three years."

The tools and platforms most commonly integrated with Hadoop today involve BI/DW. These include tools and platforms for reporting (36%), data warehousing (36%), analytics (32%), data visualization (30%), data integration (29%), and analytic databases (29%). Conversely, very few respondents have "no plans" for integrating Hadoop with these. For many users, Hadoop is both a big data platform for warehousing and a computational platform for analytics.

Expect a huge increase in integration with DQ, MDM, and enterprise applications.

Data quality (DQ) and master data management (MDM) are poised for huge growth. Despite the focus on BI/DW integration, the critical data practices of DQ and MDM are rare with Hadoop today (DQ=11%; MDM=10%). Apparently, Hadoop users are aware of the omission (and the adverse affect on data's condition), and so are planning to integrate these within three years (DQ=55%; MDM=44%). In fact, data quality and master data management have the two greatest growth indicators in this survey (43% and 34%, respectively).

Enterprise applications are rarely integrated with Hadoop today—but that will change. From the bottom of the chart upward, these include applications for supply chain (6% today), ERP (7%), financials (8%), CRM (13%), and operational applications (16%). However, significant percentages of respondents (from 28% to 38%) say they will integrate enterprise applications with Hadoop within three years. Likewise, the growth indicators for these are strong (from 22% to 29%). It's natural that Hadoop will eventually integrate with the usual enterprise applications as it expands its portfolio of enterprise uses.

Machines are rarely integrated with Hadoop today, despite the hype about them. These include machinery such as robots and vehicles (11%) and sensors such as thermometers (13%). Many respondents have no plans to support these (35% and 26%, respectively).

In your organization, with which platform and tool types is Hadoop integrated today? With which will Hadoop be integrated within three years? For which do you have no plans to integrate?

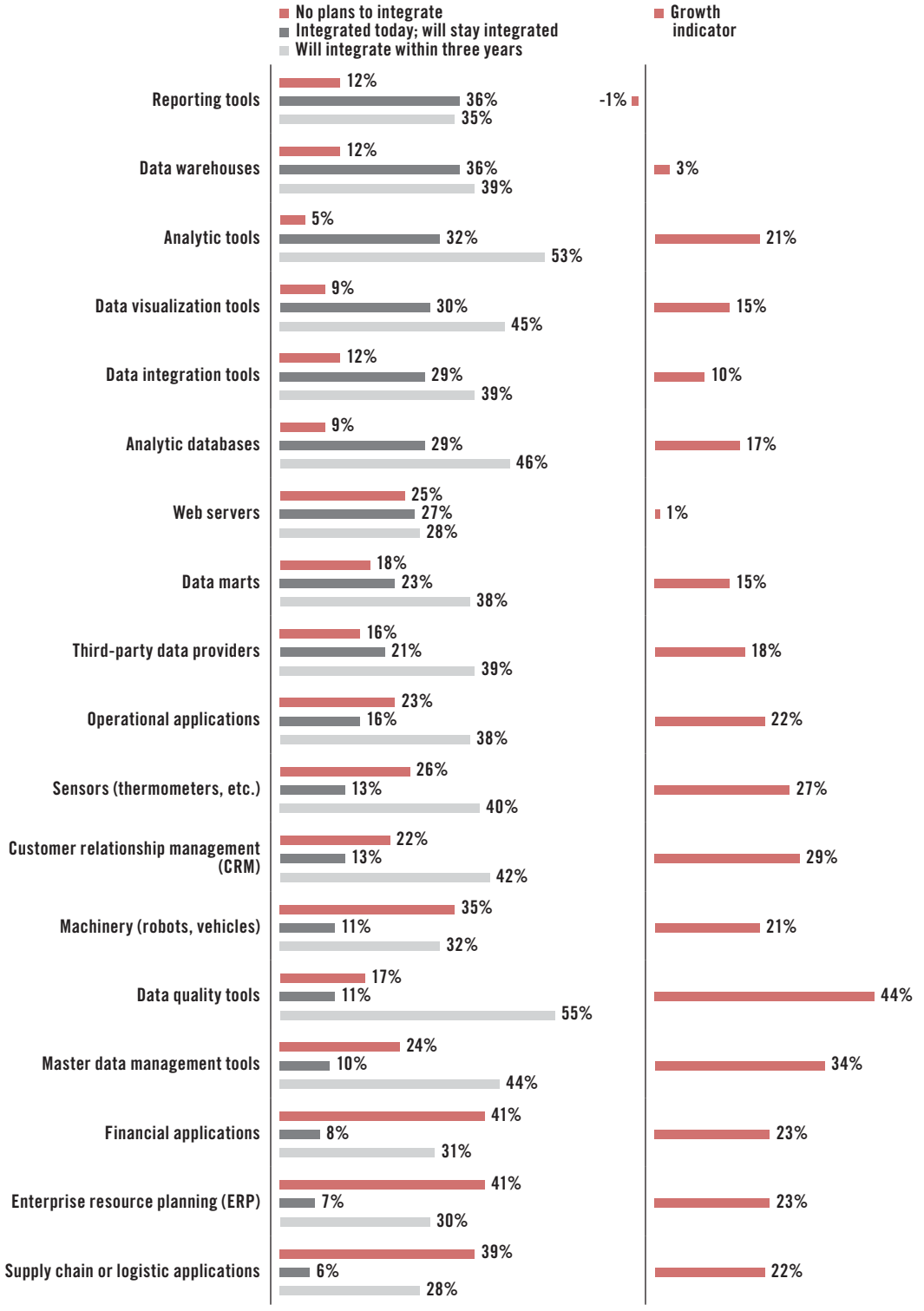


Figure 18. *Based on 99 respondents who have Hadoop experience. The charts are sorted by "Integrated today," in descending order.*

OSS Hadoop Tools in Use Today and Tomorrow

To sort out which Hadoop products are in use today (and will be in the near future), this report's survey asked: "Which of the following Hadoop technologies are in production in your organization today? Which will go into production within three years? For which do you have no plans?" (See Figure 19.)

It's no surprise that most respondents are already using the Hadoop "big five." These include the foundational platforms and frameworks, namely HDFS (70%) and MapReduce (68%), plus the MapReduce development environment called Pig (51%). Also included are the data management frameworks Hive (70%) and HBase (39%).

Popular utilities and shells are also commonly used today. At the top of this category are Zookeeper (37%) and Hue (35%). Both will see a rise in adoption within three years (22% and 19%, respectively).

Though relatively new, YARN is already entrenched among Hadoop users. Roughly one-third of respondents have YARN in production today (32%), with another third set to deploy it within three years (33%). Very few respondents have "no plans" for YARN (13%).

Real-time and near-real-time technologies are rare but are set for much greater adoption. That's a good thing because this is one of the most prominent holes in Hadoop's technology and use. In this category are Spark (28% today; 36% in three years) and Storm (19% today; 31% in three years). Samza is very rare today (2%) and many users have no plans for it (42%).

Query technologies are progressing nicely. Impala has been around for a few years; 18% of respondents are using it now and another 30% are committed to future use. Drill is brand new, but already has respondents using it (5%), plus another 26% committed.

ETL functionality has a foot in the door. In particular, Tez is relatively new but already has respondents using it (16%), with another 23% expected within three years. Kettle is rare (7%), with few users committing to future use (14%).

> **EXPERT COMMENT THREE CS ARE COMING IN HADOOP'S FUTURE.**
>
> "Most of the upcoming changes I see in and around Hadoop involve what I call 'the three Cs,' namely consolidation, clarification, and convergence," said a data expert based in the UK. "Given the plague of new vendors and new product divisions at older vendors, there will be the inevitable market consolidation, where some small vendors will be acquired by larger ones and where some products or functions within products will quietly go away (whether commercial or open source or a mix of both).
>
> "Clarification is really needed because we're currently confused about what data and processing belongs inside Hadoop versus outside it. This is seen in arguments about SQL on Hadoop versus SQL off Hadoop; structured versus unstructured data; and reporting versus analytics. Beware that the solutions to these arguments may be 'everything both on and off Hadoop.'
>
> "Finally, I expect to see a Darwin-style convergence, where Hadoop evolves to fit the enterprise data environment, and therefore ends up looking more like a relational database management system. At the same time, databases are evolving to address the environmental traits that Hadoop was built for, namely cheap linear scalability with highly diverse and raw data. So, who knows? The two may meet in the middle!"

Which of the following Hadoop technologies are in production in your organization today? Which will go into production within three years? For which do you have no plans?

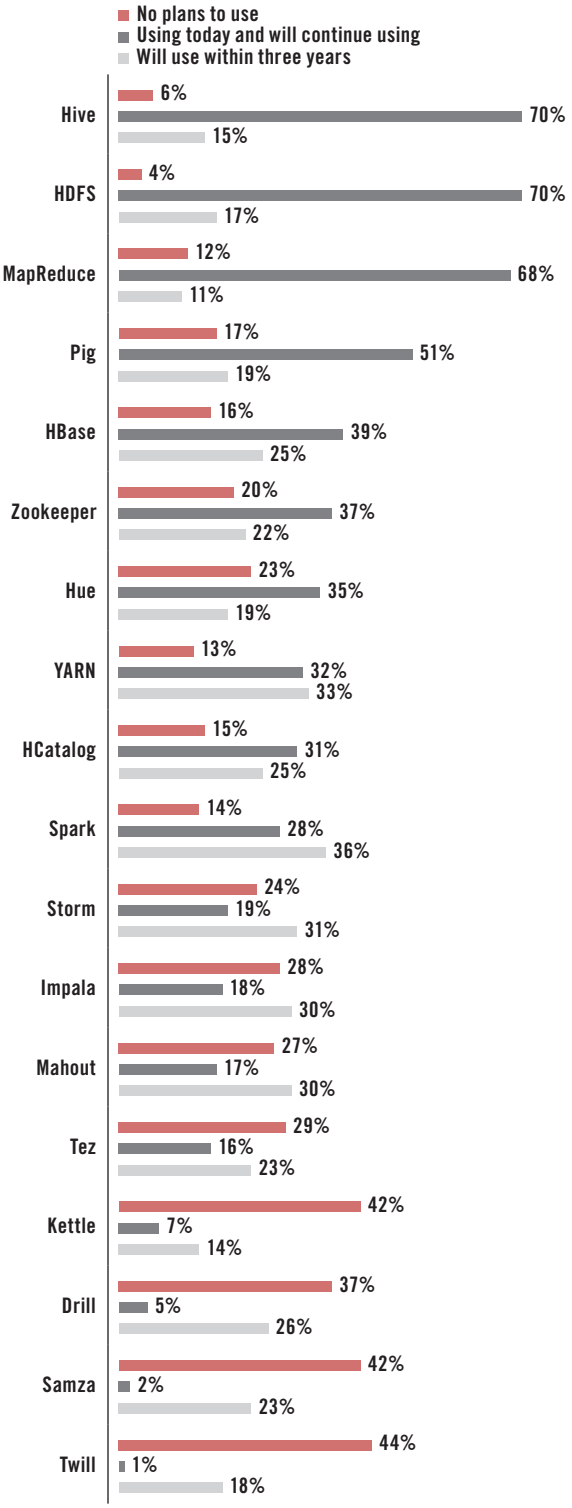


Figure 19. *Based on 99 respondents. Sorted by "Using today."*

Vendor Platforms and Tools in the Hadoop Ecosystem

The firms that sponsored this report are all good examples of software vendors that offer tools, platforms, and professional services that support Hadoop, and some offer distributions of Hadoop itself. So let's take a brief look at the product portfolio of each. The sponsors form a representative sample of the vendor community. Yet, their offerings illustrate different approaches for extending Hadoop functionality and manageability to make it more conducive to enterprise deployments.[7]

Actian Corporation

The Vortex Edition of the Actian Analytics Platform provides open, fast, and complete support for SQL in Hadoop environments. This is important because TDWI sees SQL-based analytics as the form of analytics most aggressively adopted by users at the moment. Obviously, SQL-based analytics requires low-latency, standard SQL on a highly scalable platform. Hadoop alone cannot satisfy those requirements. However, the Actian Vector X100 engine layered atop Hadoop can. This innovative combination of technologies enables users to run high-performance SQL analytics natively in Hadoop, which is equivalent to what some users do on relational platforms, but with greater speed and scalability and at less cost.

The Vortex Edition of the Actian Analytics Platform provides full ACID compliance and update capabilities, along with native DBMS security. It also includes a visual data flow framework for data blending, enrichment, and data science, plus YARN-certified administration. It includes predictive and graph analytics engines as well, addressing a wide range of analytic challenges from discovery analysis to churn prediction to network analysis and more.

Cloudera

Cloudera is a leading provider of Apache Hadoop-based software, services, and training. Cloudera offers both CDH (Cloudera's open source distribution, which includes Hadoop) and Cloudera Enterprise, which adds compliance-ready security and governance. These additions are enabled by Cloudera Navigator (auditing, data lineage, and metadata management), Apache Sentry (unified authorization), Cloudera Manager (administration) and Cloudera Support. With over 1,400 partners, Cloudera has the largest Hadoop partner network, to connect to the systems you're already using.

Cloudera's platform provides the speed, scale, and enterprise-grade features needed to build a complete enterprise data hub. On one platform, you can run batch (Apache Spark), interactive (Cloudera Impala), and real-time workloads (Apache Spark Streaming and Apache HBase), so more data is opened up to more users for more insights. The platform also features flexible deployment options for on-premises, cloud, and hybrid deployments. For cloud deployments, Cloudera has partnerships with Microsoft Azure and Amazon Web Services, and also offers Cloudera Director, the portable, self-service cloud deployment and management tool.

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

EXASOL

EXASOL AG provides EXASolution, an analytic database management system that achieves high performance with linear scalability by combining in-memory technology, columnar compression and storage, and massively parallel processing (MPP). Furthermore, EXASolution supports ANSI-standard SQL, but with extensions that allow users to apply SQL operations to an unlimited range of data structures. EXASolution is a standalone DBMS, but it can be configured to serve as a front end to Hadoop via the EXAPowerlytics Hadoop integration service. This configuration provides a SQL-off-Hadoop solution that enables users to leverage their SQL tools and experience, while working with Hadoop mixed data types at query speeds unobtainable from most other Hadoop solutions. This configuration also lays a foundation for popular Hadoop use cases, such as data warehouse extensions, queryable archives, and advanced analytics (especially SQL-based analytics). For example, most EXASOL customers use a reporting or analytic tool to connect to tables in EXASolution, tables that may be sourced from multiple systems (Hadoop and otherwise) using virtualization and NoSQL techniques.

IBM

IBM InfoSphere BigInsights Enterprise Edition integrates open source Apache Hadoop with enterprise functionality to deliver large-scale analysis with built-in resiliency and fault tolerance. The software supports structured, semi-structured, and unstructured data in its native format for maximum flexibility. It adds features for administration, discovery, development, provisioning, security, and support, along with best-in-class analytical capabilities. The result is a solution for complex, enterprise-scale projects based on Hadoop.

Moving on to related products, IBM Big Sheets is a Web-based spreadsheet application built for business users who demand visualization with all Hadoop data. Similarly, the new IBM Watson Explorer is for exploring large Hadoop datasets. IBM Big Text supports an annotation query language for natural language processing that's optimized for parsing massive record-based datasets (such as logs and streams). Finally, IBM Adaptive MapReduce is fully compatible with Apache MapReduce but has better performance, scheduling, and checkpointing. Related tools for Hadoop from the IBM InfoSphere BigInsights product family include BigSQL and BigR.

MapR Technologies

MapR provides a complete distribution for Apache Hadoop that is deployed at thousands of organizations globally for production, data-driven applications. MapR focuses on extending and advancing Hadoop, MapReduce, and NoSQL products and technologies to make them more feature rich, user friendly, dependable, and conducive to production IT environments. For example, MapR is leading the development of Apache Drill, which will bring ANSI SQL capabilities to Hadoop in the form of low-latency, interactive query capabilities for both structured and schema-free data. As other examples, MapR is the first Hadoop distribution to integrate enterprise-grade search; MapR enables flexible security via support for Kerberos and native authentication; and MapR provides a plug-and-play architecture for integrating real-time stream computational engines such as Storm with Hadoop. For greater availability, MapR provides snapshots for point-in-time data rollback, mirroring, and a No NameNode architecture that avoids single points of failure within the system and eradicates bottlenecks to cluster scalability. In addition, it's fast; MapR holds the Terasort, MinuteSort, and TPC-x-HS world records.

MarkLogic

MarkLogic Corporation offers a commercial, schema-agnostic NoSQL database management system that can sit atop Hadoop to support real-time transactional applications, leverage HDFS as a storage tier, and connect with MapReduce/YARN for ETL, analytics, or enrichment. MarkLogic runs on clustered hardware, like Hadoop, so the two can share nodes to get the most value out of the cluster. MarkLogic runs atop many file systems, including multiple versions and distributions of HDFS. At MarkLogic's core is an inverted index, as in a search engine. Data in MarkLogic is fully indexed for effective queries, multiple indexed views of data, and security at both file and index levels. This enables both search and query, plus the ability to build indices on the fly, such as range indices for columns or unique indices based on recent query activity. One of MarkLogic's differentiators is the triple index, which yields fast look ups and very efficient joins.

MarkLogic has mature DBMS functionality that can satisfy demanding IT organizations. These functions make Hadoop more like the DBMSs that enterprise IT has experience with, while still supporting Hadoop's massive amounts of multi-structured data.

Pentaho

Pentaho, Inc., is known for its unified, open, embeddable, and pluggable platform that tightly couples data integration and business analytics so organizations can easily consolidate, orchestrate, and analyze all data regardless of source. Pentaho Data Integration is the heart of the Pentaho platform, providing a visual and simplified experience for processing data at scale. The platform enables developers and analysts to ingest, blend, and process vast amounts of data from Hadoop, NoSQL, data warehouses, analytical databases, and data-driven applications to ensure the most valid and trusted analytics are delivered in a timely manner. Pentaho's Adaptive Big Data Layer insulates enterprises from the shifting sands of the big data market by allowing IT organizations to link and transform data based on business requirements rather than technical requirements of the underlying Hadoop distribution. For maximum portability, the Pentaho Adaptive Big Data Layer supports the latest Hadoop distributions from Cloudera, Hortonworks, and MapR, and it provides plug-ins for NoSQL databases (Cassandra and MongoDB) and connectors for specialized data stores (Amazon Redshift and Splunk).

SAS

SAS is the leader in business analytics and data management solutions that deliver value to customers by enabling them to make better decisions, faster. SAS looks at the big data challenge holistically; our support for Hadoop spans the entire life cycle—from data management to data discovery to analytics model development to deployment.

You can access and integrate data from Hadoop, push SAS processing and data quality functions into the Hadoop cluster, or lift data into memory and perform distributed data processing, data exploration, analytic computations, and more.

- SAS Data Loader for Hadoop enables business users to profile, transform, cleanse, and join data on Hadoop without requiring specialized skills or writing code.
- SAS Visual Analytics and SAS Visual Statistics enable users to load data into memory, quickly discover relationships, and build predictive models.
- SAS In-Memory Statistics offers an interactive programming environment with the latest machine-learning techniques to analyze diverse data in Hadoop.

SAS can help unlock the value stored in Hadoop, improve the productivity of your data scientists, and generate timely and precise insights.

Talend

Talend Inc. offers a portfolio of integration tools for data, application, and process integration, all in a single unified platform. Recent offerings from Talend have extended the portfolio into big data integration. Using Talend Studio, developers can create integration solutions for Hadoop without writing complex code. The Studio includes over 800 connectors and components, including native support for Hadoop, NoSQL, and many structured and unstructured data sources. Hence, developers can start working with Hadoop and NoSQL databases immediately, without a long learning curve, which in turn accelerates the time to use for users. The resulting solutions generate optimized, native code (e.g., MapReduce, Pig, HiveQL, Java) that leverages the massively parallel processing power of Hadoop distributions to provide extreme scale for users. Users can load, transform, enrich, and cleanse data inside Hadoop without additional storage or computing expense. Running data quality functions inside Hadoop increases data accuracy and standardization so users can make more informed decisions based on trusted data.

Trillium Software

Trillium Software has over 20 years of experience as a leading vendor of data quality tools and services. The Trillium Software System (TSS) and methodology are known for scalability and high performance in enterprise scenarios that involve very large datasets and real-time data, making TSS well positioned to operate with various types and uses of big data. Recently, Trillium extended its toolset to run natively in Hadoop via full integration with Hive and MapReduce. Trillium adds to the Hadoop ecosystem by bringing much-needed data quality functions to the Hadoop user base, which so far has scrimped on data quality and other essential data management disciplines as they struggled to deploy early-phase systems. Hadoop users can now apply TSS to improve data-driven insights, power BI and analytics with reliable data, extend complete views of customers, and maximize operational excellence. After all, data must be high quality and fit for purpose before a business can achieve a substantial ROI on big data initiatives, BI and analytics tools, and related investments. This has long been true of enterprise data, and now it's true of big data.

USER STORY THE ANALYSIS OF WEBSITE VISITOR BEHAVIOR IS A LOW-RISK BUT VALUE-ADDING FIRST USE CASE FOR HADOOP.

"We started experimenting with Hadoop about two years ago, leveraging our internal cloud," said Roni Schuling, an enterprise data architect at Principal Financial Group. "After six months, we partnered with one business unit to start a pilot on one use case. Then, about nine months ago, we deployed a vendor distribution of Hadoop on a physical implementation. This is a production system optimized for analyzing Web logs, so we can understand website visitor behavior. Since people have to log into our website to see most of their secure content, we know exactly who they are. Therefore, we can correlate their site behavior with more traditional enterprise data we have about them.

"Hadoop proved to be a good fit for this initial, well-defined use case. Although this solution is in production, we still consider it a proof of concept for the rest of the enterprise. We proved that Hadoop is useful for us. Even so, we're not satisfied with everything in our current distribution. As we expand our user base and use cases, additional security measures will be needed. Additionally, without a more intuitive consumption capability, we are forced to build Hadoop technical skills and rely on data placement to get the most value. We're hoping to build upon what we've learned with our initial Hadoop distribution to expand capabilities for doing more with data in its original raw form. We'll continue to evaluate our current distribution, to understand if we can close gaps with different distributions."

Top 10 Priorities for Enterprise Hadoop

In closing, we summarize the findings of this report by listing the top priorities for making Hadoop enterprise grade, including a few comments about why these priorities are important. Think of the priorities as recommendations, requirements, or rules that can guide user organizations into successful strategies for implementing enterprise-scope Hadoop.

1. **Be open to Hadoop and other new options.** This also means you should be open to the enterprise use of open source in general, advanced forms of analytics that are new to you, new structures of data, new sources for data, new paradigms for managing data, and new business methods that leverage big data for organizational advantage. In a way, Hadoop is emblematic, as just one example of how enterprise IT and business are changing. You can embrace and guide change so it leads to improvements, or you can maintain the status quo as opportunities pass by.
2. **Innovate with big data on enterprise Hadoop.** The vast majority of survey respondents (89%) consider Hadoop an opportunity for innovation. Some innovations create something out of nothing, as when you apply streaming big data to business monitoring. Some Hadoop-driven innovations are incremental, such as using big data to expand the data samples that mature data mining and statistical analysis applications depend on for accurate actuarial calculations and customer segments. Likewise, social data can take complete views of customers to a whole new level. Data aside, the low cost of Hadoop can lead to enterprise innovations in funding, sponsorship, infrastructure provision, and budgeting.
3. **Base Hadoop adoption on business and technology requirements.** If your organization is like those surveyed, requirements and benefits for advanced analytics, big data leverage, data exploration, extending older data management platforms, archiving, and cost containment will lead you to consider Hadoop. One of these alone is compelling enough. If your organization has all these requirements, they will lead you into the broad enterprise use of Hadoop described in this report.
4. **Know the hurdles so you can leap over them.** Common hurdles include the skills gap, weak business support, security issues, and excessive hand coding. Never let these stop you. As discussed in this report, survey respondents and interviewees alike have found workaround solutions for all these, and the feature functionality of the Hadoop ecosystem improves continuously to lessen the hurdles.
5. **Get training (and maybe new staff) for Hadoop and big data management.** Your focus should be on training and hiring data analysts, data scientists, and data architects who can develop the applications for data exploration, discovery analytics, archiving, and content management that organizations need if they're to get full value from big data. When in doubt, hire and train data specialists, not application specialists, to manage big data. Most BI/DW professionals are already cross trained in many data disciplines; cross train them more. Finally, close the skills gap further by leveraging self-service data access and preparation tools.

6. **Co-opt Hadoop to rethink the economics of data and content architectures.** Users interviewed for this report described how they've deployed multiple data platform types in their data warehouse environments, where each platform is best of breed for specific workloads and user requirements. Besides being a system architecture, this is also a new economic model in that each platform has different costs, and users steer data and processing to the cheapest platform that will do the job right. This system and economic model is well established in data warehousing and is proliferating to other enterprise content and data ecosystems. The low cost of Hadoop is the leading driver behind this enterprisewide change in IT portfolios and architectures, which is why Hadoop is a critical success factor for the model.

7. **Prepare for hybrid data ecosystems by defining places for Hadoop in your architecture.** Though Hadoop is going enterprise in scope, many users haven't yet deployed Hadoop in any context. Obvious benefit-driven starter use cases include data staging in a data warehouse environment, a collocation point for massive datasets to enable broad data exploration, processing data for advanced analytics, a replacement for archaic archives, and an extension to content management. Put these together and Hadoop's enterprise-scope value becomes apparent.

8. **Consider Hadoop use cases outside the usual BI/DW and analytic applications.** Archival and backup systems are outdated and ineffective in most firms. Business continuity and disaster recovery systems are almost as bad. Hadoop's scalability and low cost are compelling for these use cases but modernized so they are online and queryable, so it's no surprise that survey respondents and interviewees alike highlighted archiving as a use case they are adopting aggressively. Other non-BI/DW use cases include content management, document management, and records management. These show that Hadoop is not just for BI, DW, and analytics, but is broadening out to other enterprise use cases.

9. **Look for capabilities that make Hadoop data look relational.** Relational functions are key to the enterprise acceptance of Hadoop because high-profile use cases require them, such as SQL-based analytics, the management of tabular data, ELT transformation processing, and databases for operational applications. It's true that Hive and HBase can enable these cases to a certain degree, but data people want to leverage their existing ANSI SQL skills and portfolio of SQL-based tools. Better support for SQL on and off Hadoop is coming from numerous vendors and open source contributors, but remember: improving SQL support in no way detracts from Hadoop's unique capabilities as an inherently NoSQL platform. Part of the power of Hadoop is its ability to support many approaches to many types of data. In that regard, Hadoop gets more diverse almost daily.

10. **Develop and apply a strategy for enterprise Hadoop.** If your organization is totally new to Hadoop, a POC project is in order. The POC should test the business value of multiple use cases. To keep that manageable, prioritize the use cases and start with one that is both easy from a technology viewpoint and potentially useful for the business. Common starting points include amassing big data for exploration, discovery, and a specific form of analytics. Force the POC team to test other use cases, but with diversity. Besides the obvious use cases in analytics, also test data warehouse extensions, archiving, content management, and storage provisioning. Likewise, the POC team should be as diverse as the use cases so that more of the enterprise gains Hadoop experience. That puts everyone in a better position for the true goal of the POC project: creating a prioritized list of Hadoop-based applications that will eventually stretch across the whole enterprise.

www.pentaho.com

Hadoop for the Enterprise: Enhanced with Pentaho

Big Data Integration and Analytics

Regardless of the data source, analytic requirement, or deployment environment, Pentaho allows you to turn big data into big insights. The **Pentaho platform** provides Hadoop users with visual development tools and big data analytics to easily prepare, model, visualize, and explore data sets. From **data preparation** to **predictive analytics**, Pentaho covers the data life cycle to remove complexity and reduce the time to realize value from big data.

Pentaho's extensible and scalable platform delivers secure, blended data on demand to ensure accuracy, compliance, and governance. Analytics-ready data is delivered to the people and processes that need it most, seamlessly meshing with existing IT architectures and enterprise security frameworks.

- Supports and orchestrates the complete analytic data flow from data ingestion, access, integration, and blending of big data with traditional data sources
- Supports high performance data enrichment, blending, and processing of data at scale inside Hadoop, enabling the **delivery of governed data sets** for interactive and predictive analytics
- Data visualizations, reports, and dashboards can easily be **embedded** in other applications or made available through self-service analytics
- Leverage the broadest spectrum of big data sources with the **Pentaho Adaptive Big Data Layer** to take advantage of and insulate from the fast moving and rapidly changing big data ecosystem
- Promotes **open, standards-based architectures**, easy to integrate with or extend to existing infrastructure
- Achieve precise control with a full breadth of administrative tools, including detailed performance monitoring, analytics content distribution, job management, and more

Get More from Hadoop

Visual Development for Hadoop Data Prep and Modeling: Pentaho's visual development tools drastically reduce the time to design, develop, and deploy Hadoop analytic solutions by as much as 15x, compared to traditional custom coding and ETL approaches.

Pentaho Visual MapReduce: Scalable in-Hadoop Execution: Pentaho's Java-based data integration engine works with the Hadoop cache for automatic deployment as a MapReduce task across every data node in a Hadoop cluster, making use of the massive parallel processing power of Hadoop. Pentaho includes native connections to HDFS, MapReduce, HBase, Imapala, and Hive.

YARN Integration for Easier Data Orchestration: **Pentaho and YARN** allow IT to design and execute high performance data orchestration without coding, allowing you to exploit the full computing power of Hadoop by leveraging existing skill sets and technology investments.

About Pentaho

Pentaho is building the future of business analytics. Pentaho's open source heritage drives our continued innovation in a modern, integrated, embeddable platform built for accessing all data sources. With support for all of the leading Hadoop distributions, NoSQL databases, and high performance analytic databases, Pentaho provides the broadest support for big data analytics, as well as integration and orchestration of big data and traditional sources.

TDWI RESEARCH

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on business intelligence, data warehousing, and analytics 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, data warehousing, and analytics 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.

tdwi
Advancing all things data.
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