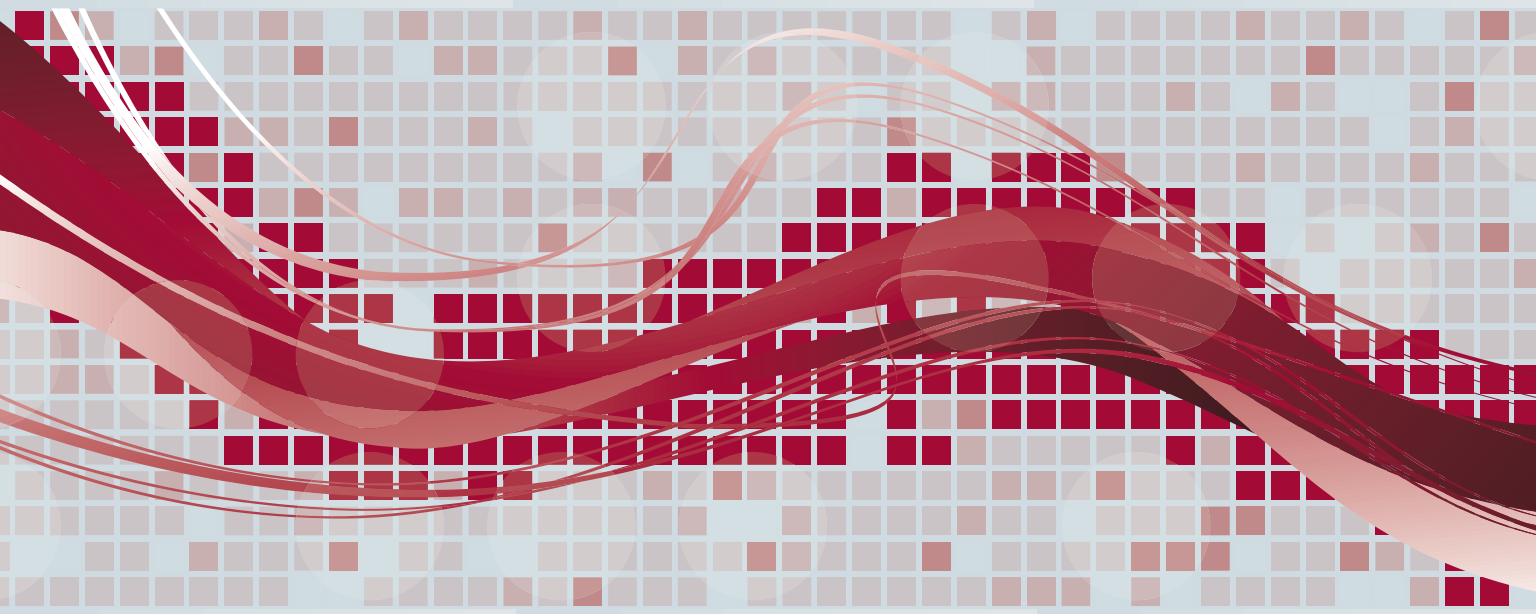
TDWI RESEARCH

TDWI BEST PRACTICES REPORT

SECOND QUARTER 2013

INTEGRATING HADOOP INTO BUSINESS INTELLIGENCE AND DATA WAREHOUSING

By Philip Russom



CO-SPONSORED BY

sas

tdwi.org

tdwi

INTEGRATING HADOOP INTO BI/DW

By Philip Russom

Table of Contents

About the Author

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

About TDWI

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

About the TDWI Best Practices Reports Series

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

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

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who 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: Jennifer Agee, Michael Boyda, and Denelle Hanlon.

Sponsors

Cloudera, EMC Greenplum, Hortonworks, ParAccel, SAP, SAS, Tableau Software, and Teradata sponsored the research for this report.

Research Methodology and Demographics

Report Scope. The purpose of this report is to accelerate users' understanding of the many new products and practices based on Hadoop technologies that have emerged in recent years. TDWI assumes that Hadoop usage will become mainstream in coming years. This report explains ways that Hadoop can be integrated with mature implementations for business intelligence, data warehousing, data management, and analytics.

Terminology. In this report, the term "Hadoop" refers to the Hadoop Distributed File System (HDFS) as well as the growing family of other open source software products and technologies available from the Apache Software Foundation and several software vendor firms.

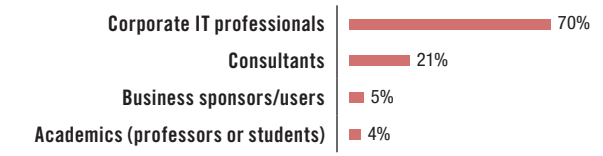
Survey Methodology. In early November 2012, TDWI sent an invitation via e-mail to the data management professionals in its database, asking them to complete an Internet-based survey. The invitation was also distributed via websites, newsletters, and publications from TDWI and other firms. The survey drew responses from 325 survey respondents. From these, we excluded incomplete responses and respondents who identified themselves as vendor employees. The resulting completed responses of 263 respondents form the core data sample for this report. Due to branching in the survey, some questions were answered only by respondents with hands-on experience with Hadoop.

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

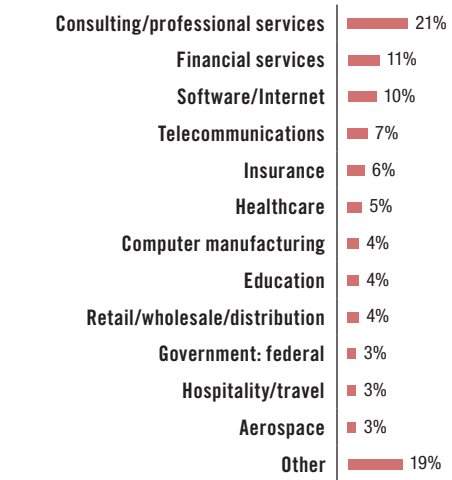
Survey Demographics. The majority of survey respondents are IT professionals (70%), whereas the others are consultants (21%), business sponsors or users (5%), and academics (4%). We asked consultants to fill out the survey with a recent client in mind.

The consulting industry (21%) dominates the respondent population, followed by financial services (11%), software/Internet (10%), telecommunications (7%), insurance (6%), and other industries. Most survey respondents reside in the U.S. (49%), Asia (20%), or Europe (17%). Respondents are fairly evenly distributed across all sizes of companies and other organizations.

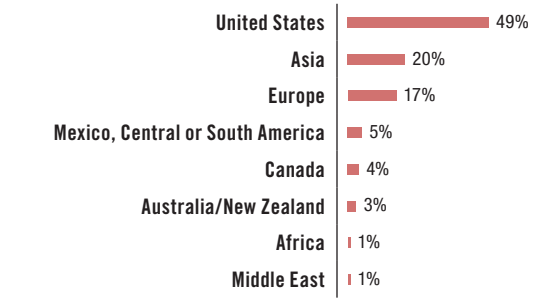
Position



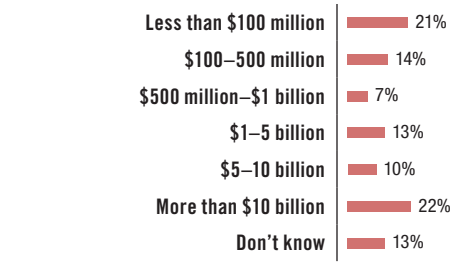
Industry



Geography



Company Size by Revenue



Based on 263 survey respondents.

Executive Summary

Hadoop is a family of multiple products and technologies, available via open source and from software vendors.

Apache Hadoop is an open source software project administered by the Apache Software Foundation (ASF). The Hadoop family of products includes the Hadoop Distributed File System (HDFS), MapReduce, Pig, Hive, HBase, and so on. These products are available as open source from ASF, as well as from several software vendors. The number of vendor products that integrate with Hadoop products increases almost daily. In this report, the term "Hadoop" usually means the entire Hadoop family of products, regardless of their open source or vendor origins. Some discussions focus specifically on HDFS.

Hadoop solves tough problems in BI/DW. Therefore, it will soon be integrated into far more BI/DW solutions.

Business intelligence (BI) professionals' interest in Hadoop has been driven up in recent years because Hadoop has proved its usefulness with the toughest challenges in BI today, namely big data, advanced analytics, and multi-structured data. For that reason, TDWI anticipates that Hadoop technologies will soon become a common complement to (but not a replacement for) established products and practices for business intelligence (BI), data warehousing (DW), data integration (DI), and analytics. Therefore, a wide range of user organizations need to prepare for Hadoop usage. Although it's true that Hadoop can be valuable as an analytic silo, most organizations will prefer to get the most business value out of Hadoop by integrating it with—or into—their BI, DW, DI, and analytics technology stacks.

Hadoop solutions are available today as a mix of open source, vendor products, and user designs.

According to this report's survey, users with hands-on Hadoop experience say it's still immature and needs serious improvements in security, administrative tools, high availability, and real-time operation. These and other problems are being addressed by the open source community of technical users, which continues to infuse innovation into existing Hadoop products as well as introduce new ones via ASF's incubation process. The pace of Hadoop innovation has accelerated because a number of software vendor firms now contribute to Hadoop's open source. The first wave of support for Hadoop technologies by vendor tools and platforms is already in place, with subsequent waves coming soon. The number of technical users conversant in Hadoop is increasing steadily.

Users are already using multiple Hadoop products with their systems for BI, DW, DI, and analytics.

According to this report's survey, the Hadoop products most commonly used today are (in priority order) MapReduce, HDFS, Java, Hive, HBase, and Pig. Those poised for greatest future adoption are Mahout, Zookeeper, and HCatalog. All of these have compelling use cases for BI, DW, DI, and analytics. In fact, survey respondents who have hands-on Hadoop experience say they've already integrated Hadoop with analytic tools, DWs, reporting tools, Web servers, analytic databases, and data visualization tools—showing that Hadoop is already established as a component within BI/DW technology stacks. Of these respondents, 78% feel Hadoop is a complement to a DW, not a replacement. Enabling big data analytics is the leading benefit of Hadoop, whereas a lack of Hadoop skills is the leading barrier. BI/DW aside, a few respondents also anticipate using Hadoop as a live archive (23%) or as a platform for content management (35%).

Brisk Hadoop adoption will change BI/DW for the better. Get ready!

Only 10% of organizations surveyed have a Hadoop implementation in production today, but a whopping 51% say they'll have one within three years. If this trend pans out, Hadoop will impact at least half of BI/DW environments soon. Hence, users need to prepare for Hadoop usage now.

The purpose of this report is to accelerate users' understanding of the many new Hadoop-based products that have emerged in recent years. The report also maps newly available Hadoop options to real-world use cases. This information can help user organizations successfully integrate Hadoop technologies into their BI portfolios and practices with maximum business value.

Introduction to Hadoop Products and Technologies

We've experienced years of both hype and enlightened discussions in the IT press and elsewhere, extolling Hadoop as an enabler for big data, analytics, and multi-structured data. Yet, despite the attention Hadoop has received, relatively few BI professionals and other IT personnel are familiar with it. For example, this report's survey asked: Do you feel you know what Hadoop is and does? A shocking 26% said no. Likewise, a mere 18% of survey respondents report having experience deploying and/or using a Hadoop cluster, and only two-thirds of the 18% have reached production deployment.

Ignorance of Hadoop is still common in the BI and IT communities.

Therefore, before discussing applications of Hadoop in BI/DW and strategies for integrating Hadoop with BI/DW technology stacks, this report must begin by defining what Hadoop is and does—as well as what it does not do so well. A fun and effective way to do this is to bust some of the most common myths and misconceptions harbored about Hadoop today.

Busting 10 Myths about Hadoop[1]

Although Hadoop and related technologies have been with us for over five years now, most BI professionals and their business counterparts still harbor a few misconceptions that need to be corrected about Hadoop and related technologies such as MapReduce. The following list of 10 facts will clarify what Hadoop is and does relative to BI/DW, as well as in which business and technology situations Hadoop-based BI, DW, DI, and analytics can be useful.

Fact #1. Hadoop consists of multiple products.

We talk about Hadoop as if it's one monolithic thing, but it's actually a family of open source products and technologies overseen by the Apache Software Foundation (ASF). (Some Hadoop products are also available via vendor distributions; more on that later.)

Hadoop comprises multiple products, available from multiple sources.

The Apache Hadoop library includes (in BI priority order): the Hadoop Distributed File System (HDFS), MapReduce, Pig, Hive, HBase, HCatalog, Ambari, Mahout, Flume, and so on. You can combine these in various ways, but HDFS and MapReduce (perhaps with Pig, Hive, and HBase) constitute a useful technology stack for applications in BI, DW, DI, and analytics. More Hadoop projects are coming that will apply to BI/DW, including Impala, which is a much-needed SQL engine for low-latency data access to HDFS and Hive data.

Fact #2. Hadoop is open source but available from vendors, too.

Apache Hadoop's open source software library is available from ASF at www.apache.org. For users desiring a more enterprise-ready package, a few vendors now offer Hadoop distributions that include additional administrative tools, maintenance, and technical support. A handful of vendors offer their own non-Hadoop-based implementations of MapReduce.

Fact #3. Hadoop is an ecosystem, not a single product.

In addition to products from Apache, the extended Hadoop ecosystem includes a growing list of vendor products (e.g., database management systems and tools for analytics, reporting, and DI) that integrate with or expand Hadoop technologies. One minute on your favorite search engine will reveal these.

[1] This section of the report was originally published as the expert column "Busting 10 Myths about Hadoop" in TDWI's *BI This Week* newsletter, March 20, 2012 (available at tdwi.org). The column has been updated slightly for use in this report.

Fact #4. HDFS is a file system, not a database management system (DBMS).

HDFS is not a DBMS. Oddly enough, that's an advantage for BI/DW.

Hadoop is primarily a distributed file system and therefore lacks capabilities we associate with a DBMS, such as indexing, random access to data, support for standard SQL, and query optimization. That's okay, because HDFS does things DBMSs do not do as well, such as managing and processing massive volumes of file-based, unstructured data. For minimal DBMS functionality, users can layer HBase over HDFS and layer a query framework such as Hive or SQL-based Impala over HDFS or HBase.

Fact #5. Hive resembles SQL but is not standard SQL.

Many of us are handcuffed to SQL because we know it well and our tools demand it. People who know SQL can quickly learn to hand code Hive, but that doesn't solve compatibility issues with SQL-based tools. TDWI believes that over time, Hadoop products will support standard SQL and SQL-based vendor tools will support Hadoop, so this issue will eventually be moot.

Fact #6. Hadoop and MapReduce are related but don't require each other.

Some variations of MapReduce work with a variety of storage technologies, including HDFS, other file systems, and some relational DBMSs. Some users deploy HDFS with Hive or HBase, but not MapReduce.

Fact #7. MapReduce provides control for analytics, not analytics per se.

MapReduce is a general-purpose execution engine that handles the complexities of network communication, parallel programming, and fault tolerance for a wide variety of hand-coded logic and other applications—not just analytics.

Fact #8. Hadoop is about data diversity, not just data volume.

Hadoop promises to extend DW architecture to better handle staging, archiving, sandboxes, and unstructured data.

Theoretically, HDFS can manage the storage and access of any data type as long as you can put the data in a file and copy that file into HDFS. As outrageously simplistic as that sounds, it's largely true, and it's exactly what brings many users to Apache HDFS and related Hadoop products. After all, many types of big data that require analysis are inherently file based, such as Web logs, XML files, and personal productivity documents.

Fact #9. Hadoop complements a DW; it's rarely a replacement.

Most organizations have designed their DWs for structured, relational data, which makes it difficult to wring BI value from unstructured and semistructured data. Hadoop promises to complement DWs by handling the multi-structured data types most DWs simply weren't designed for. Furthermore, Hadoop can enable certain pieces of a modern DW architecture, such as massive data staging areas, archives for detailed source data, and analytic sandboxes. Some early adoptors offload as many workloads as they can to HDFS and other Hadoop technologies because they are less expensive than the average DW platform. The result is that DW resources are freed for the workloads with which they excel.

Fact #10. Hadoop enables many types of analytics, not just Web analytics.

Hadoop gets a lot of press about how Internet companies use it for analyzing Web logs and other Web data, but other use cases exist. For example, consider the big data coming from sensory devices, such as robotics in manufacturing, RFID in retail, or grid monitoring in utilities. Older analytic applications that need large data samples—such as customer base segmentation, fraud detection, and risk analysis—can benefit from the additional big data managed by Hadoop. Likewise, Hadoop's additional data can expand 360-degree views to create a more complete and granular view of customers, financials, partners, and other business entities.

The Status of HDFS Implementations

As the 10 facts just listed indicate, HDFS and other Hadoop products show great promise for enabling and extending applications in BI, DW, DI, and analytics. But are user organizations actively adopting Hadoop?

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

To quantify this situation, this report's survey asked: When do you expect to have HDFS in production? (See Figure 1.) The question asks about HDFS because in most situations (excluding some uses of MapReduce) an HDFS cluster must be in place before other Hadoop products and hand-coded solutions are deployed atop it. Survey results reveal important facts about the status of HDFS implementations. (A slight majority of survey respondents are BI/DW professionals, so the survey results represent the broad IT community, but with a BI/DW bias.)

HDFS is used by a small minority of organizations today. Only 10% of survey respondents report having reached production deployment.

A whopping 73% of respondents expect to have HDFS in production. Of these, 10% are already in production, with another 63% upcoming. Only 27% of respondents say they will never put HDFS in production.

HDFS usage will go from scarce to ensconced in three years. If survey respondents' plans pan out, HDFS and other Hadoop products and technologies will be common in the near future, thereby having a large impact on BI, DW, DI, and analytics—plus IT and data management in general, as well as how businesses leverage them.

When do you expect to have HDFS in production?

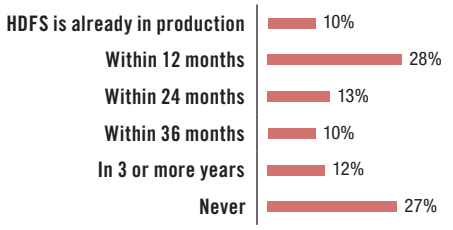


Figure 1. *Based on 263 respondents.*

Hadoop Technologies in Use Today and Tomorrow

As pointed out earlier, this report considers Hadoop an ecosystem of products and technologies. Note that some are more conducive to applications in BI, DW, DI, and analytics than others, and certain product combinations are more desirable than others for such applications.

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 and related technologies are in production in your organization today? Which will go into production within three years? (See Figure 2.) These questions were answered by a subset of 48 survey respondents who claim they've deployed or used HDFS. Hence, their responses are quite credible, being based on direct, hands-on experience.

HDFS and MapReduce are the most used Hadoop products today.

HDFS and a few add-ons are the most commonly used Hadoop products today. HDFS is near the top of the list (67% in Figure 2) because most Hadoop-based applications demand HDFS as the base platform. Certain add-on Hadoop tools are regularly layered atop HDFS today:

- **MapReduce (69%)** for the distributed processing of hand-coded logic, whether for analytics or for fast data loading and ingestion
- **Hive (60%)** for projecting structure onto Hadoop data so it can be queried using a SQL-like language called HiveQL
- **HBase (54%)** for simple, record-store database functions against HDFS's data

MapReduce is used even more than HDFS. The survey results (which rank MapReduce as slightly more common than HDFS) suggest that a few respondents in this survey population are using MapReduce today without HDFS, which is possible (as noted earlier). The high MapReduce usage also explains why Java and R ranked fairly high in the survey; these programming languages are not Hadoop technologies per se, but are regularly used for the hand-coded logic that MapReduce executes. Likewise, Pig ranked high in the survey, as a tool that enables developers to design logic (for MapReduce execution) without having to hand code it.

Some Hadoop products are rarely used today. For example, few respondents in this survey population have touched Chukwa (4%) or Ambari (6%), and most have no plans for using them (75% and 71%, respectively). Oozie, Hue, and Flume are likewise of little interest at the moment.

Expect aggressive adoption of Mahout, R, HCatalog, and Ambari.

Some Hadoop products are poised for aggressive adoption. Half of respondents (50%) say they'll adopt Mahout within three years,[2] with similar adoption projected for R (44%), Zookeeper (42%), HCatalog (40%), and Oozie (40%).

TDWI sees a few Hadoop products as especially up-and-coming. Usage of these will be driven up according to user demand. For example, users need analytics tailored to the Hadoop environment, as provided by Mahout (machine learning–based recommendations, classification, and clustering) and R (a programming language specifically for analytics). Furthermore, BI professionals are accustomed to DBMSs, and so they long for a Hadoop-wide metadata store and far better tools for HDFS administration and monitoring. These user needs are being addressed by HCatalog and Ambari, respectively, and therefore TDWI expects both to become more popular.

[2] Mahout does not require distributed computing, so the 50% may not all be used with HDFS.

Which of the following Hadoop and related technologies are in production in your organization today? Which will go into production within three years?

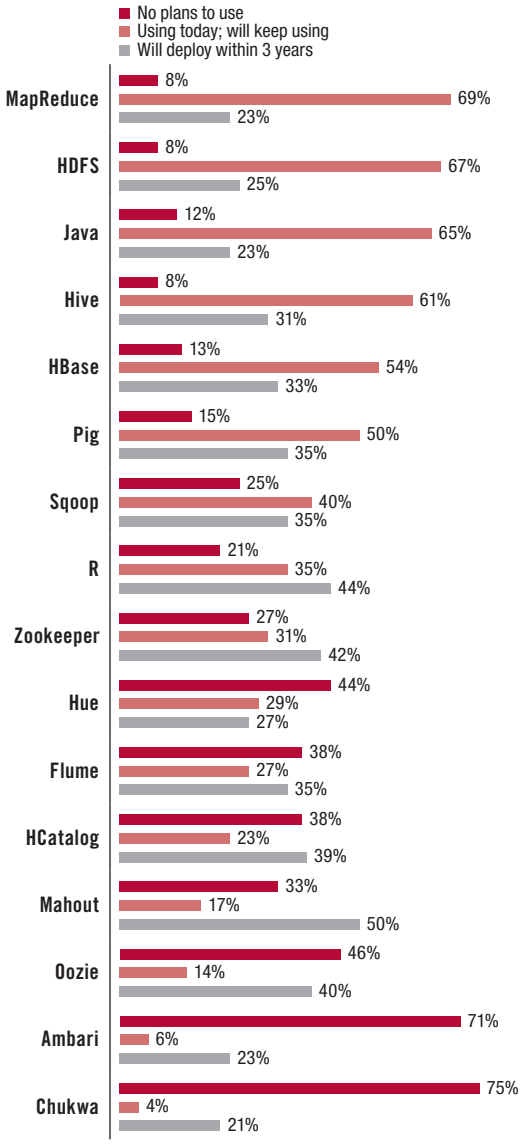


Figure 2. *Based on 48 respondents who have experience with Hadoop. The chart is sorted by the responses for "using today."*

USER STORY

HADOOP SCALES WELL, JUST AS ADVERTISED.

"About a year ago we started looking into ways of processing big data," said an IT manager at an insurance firm. "We saw that our traditional BI tools have mature functionality, but most of it won't scale to the extreme size of big data. So, we started evaluating tools, and concluded that Hadoop would be a good fit for us because of its scalability. Today, we have three different Hadoop solutions up and running.

> "The first two solutions analyze Web log data and display ad data, which is what Hadoop was designed for, and Hadoop scales well, just as advertised. The third solution is unique to us. It processes large quantities of sensor data coming from fleet vehicles to analyze performance characteristics of drivers and vehicles. Each record is simply a dozen numeric and text fields. But we've collected trillions of records over the years, so querying the data in a relational database took days. In Hadoop, rescoring analytic models about drivers and vehicles is easily done overnight."

Use Cases for Hadoop in BI, DW, DI, and Analytics

Hadoop: Problem or Opportunity for BI/DW?

Hadoop is an opportunity to embrace new analytics and their business benefits.

Hadoop is still rather new, and it's often deployed to enable other practices that are also new, such as big data management and advanced analytics. Hence, rationalizing an investment in Hadoop can be problematic. To test perceptions of whether Hadoop is worth the effort and risk, this report's survey asked: Is Hadoop a problem or an opportunity? (See Figure 3.)

The vast majority (88%) considers Hadoop an opportunity. The perception is that Hadoop products enable new application types, such as the sessionization of website visitors (based on Web logs), monitoring and surveillance (based on machine and sensor data), and sentiment analysis (based on unstructured data and social media data).

A small minority (12%) considers Hadoop a problem. Fully embracing multiple Hadoop products requires a fair amount of training in hand coding, analytic, and big data skills that most BI/DW and analytics teams lack at the moment. But few users (a mere 12%) surveyed consider Hadoop a problem.

Is Hadoop a problem or an opportunity?

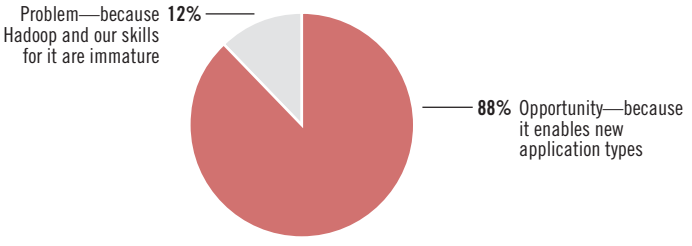


Figure 3. *Based on 263 respondents.*

Perceptions of Hadoop's Role in BI/DW

Most users perceive Hadoop as a complement to an EDW that enables new analytics with new big data types.

HDFS is an unlikely replacement for an EDW. (See Figure 4.) EDWs continue to be killer platforms for supplying clean, standardized, dimensional, aggregated, and SQL-addressable data with an audit trail for the most common BI deliverables, namely standard reports, newer styles of reports (dashboards and scorecards), performance management metrics, operational BI data, and cubes for online analytic processing (OLAP). Hence, an EDW supplies data for most of the deliverables of the average BI/DW program. Compared to an EDW, Hadoop products in their current state simply aren't as good at supplying data with such stringent requirements.

HDFS can augment and complement an EDW. (See Figure 5.) Although EDWs are killer platforms for supplying heavily processed data into most BI deliverables, there are other areas where the average EDW is not so well suited—but Hadoop products are. Hadoop excels with managing and processing file-based data, especially when the data is voluminous in the extreme and the data would not benefit from transformation and loading into a DBMS. In fact, for the kinds of discovery analytics involved with Hadoop, it's best to keep the data in its raw, source form. This is why Hadoop has such a well-deserved reputation with big data analytics.

Can the Hadoop Distributed File System (HDFS) replace your enterprise data warehouse (EDW)?

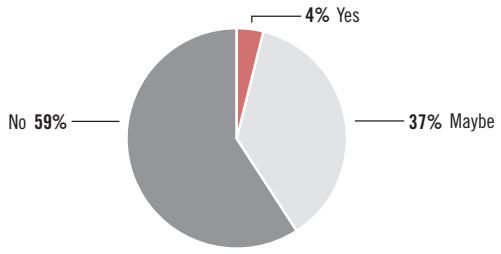


Figure 4. *Based on 263 respondents.*

Can HDFS augment your enterprise data warehouse (EDW) or other data infrastructure?

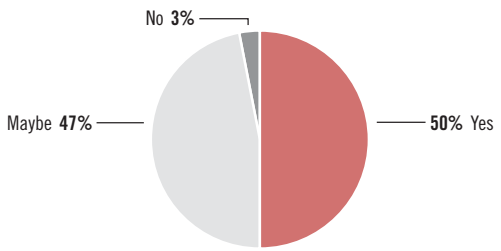


Figure 5. *Based on 263 respondents.*

Hadoop can expand areas of DW architecture that need scalability or multi-structured data.

You still need an EDW for its finesse with relational and dimensional data, plus its tough requirements for transformational processing and an audit trail. Hence, an EDW and HDFS are largely complementary. In addition, an impressive 78% of respondents (see Figure 7) feel that HDFS could be a useful complement to a DW, specifically for advanced analytics, as well as a DW's data staging area (41%) and a sandbox for ad hoc analytics (41%).

HDFS can expand an organization's analytic capabilities. (See Figure 6.) Many organizations have been like the proverbial deer in the headlights, frozen by the newness and enormity of big data. The right combination of Hadoop products can thaw "analysis paralysis" by enabling the management and processing of big data, for which traditional DWs and BI tools were not designed. For example, Hadoop cut its teeth on the analysis of petabytes of Web log data in large Internet firms, and now is being applied to similar analytic applications involving call detail records in telecommunications, XML documents in supply chain industries (retail and manufacturing), unstructured claims documents in insurance, sessionized spatial data in logistics, and a wide variety of log data from machines and sensors. Hadoop-enabled analytics are sometimes deployed in silos, but the trend is toward integrating Hadoop and EDW data at analysis time for maximal visibility into business performance and more complete 360-degree views of business entities.[3]

[3] For a more detailed discussion of advanced analytics with big data, see the 2011 TDWI Best Practices Report *Big Data Analytics*, available for free download at tdwi.org/bpreports.

Clearly, HDFS is not just for BI, DW, DI, and analytics. Survey respondents also see HDFS as a potential live archive for Web and other non-traditional data (52% in Figure 7), a live archive for traditional data (23%), or a repository for content, document, and records management (35%).

Can HDFS expand your analytic capabilities?

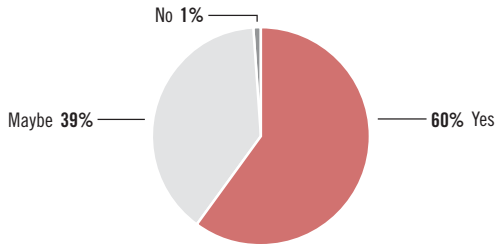


Figure 6. *Based on 263 respondents.*

In your perception, what would be useful applications of HDFS in your organization? Select all that apply.

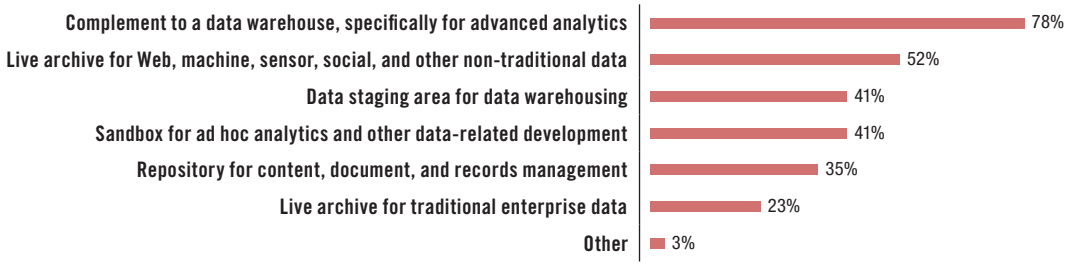


Figure 7. *Based on 712 responses from 263 respondents; 2.7 responses per respondent, on average.*

USER STORY

IT'S THE ADVERTISING, STUPID!

A user interviewed for this report recounted the following tale: "I know we look like an Internet firm, and that's how most people describe us. But, at the end of the day, we're really an advertising firm sitting atop content, and that's true of many other Internet firms. The content brings in website visitors, and our advertisers market to them through Web page ads. It's the same business model that the old paper magazines and newspapers had, but now it's online.

"So, we need answers to analytic questions, such as: What drives people to certain content? What's their profile? How do we draw more people to an area? Which advertiser is the best match for a content area or a visitor profile? Our advertising clients want answers to these questions, too, and they want answers frequently, despite the rising quantities of clickstream big data that the analyses are based on. Our Hadoop portfolio helps us scale up so we can deliver good answers frequently."

Benefits of Hadoop

In the perceptions of survey respondents, Hadoop has its benefits. (See Figure 8.)

Hadoop's leading benefits all concern advanced analytics for big data.

Enabling big data analytics is the leading benefit of integrating Hadoop into BI/DW. Most survey respondents recognize Hadoop's primary role as a big data source for analytics (71%). Similarly, HDFS' scalability also helps enable bigger and better practices in data archiving (20%) and schema-free data staging (19%), even with machine data from robots, sensors, meters, and so on (17%).

Hadoop supports exploratory analytic methods. Big data and analytics go together because analytic methods help user organizations get value from big data (which is otherwise a cost center) in the form of more numerous and accurate business insights (15%). However, these are not the same insights that a traditional BI/DW solution provides based on carefully prepared data about well-understood business entities. Since most big data comes from new sources seldom tapped by BI/DW, users can conduct information exploration and discovery (33%) and exploratory analytics with big data (48%) to discover new facts about a business, as well as its customers and operations.

Hadoop supports advanced forms of analytics. Online analytic processing (OLAP) is still the most common analytic technology in use today, but the exploratory analytics conducted with big data in a Hadoop environment typically involves so-called advanced analytics based on non-OLAP technologies such as data mining, statistical analysis, complex SQL, and so on (68%). This is often coupled with data visualization (25%).

Hadoop has benefits for DW environments, such as scalability and data diversity.

Hadoop can expand data warehouse environments. Many users surveyed feel that Hadoop is a good complement for a data warehouse (30%), probably because Hadoop's focus on advanced analytics complements the focus on reporting and OLAP typical of most DWs.

Hadoop scales cost effectively. HDFS is known to achieve extreme scalability (19%) on low-cost hardware and software (26%), so it can help users capture more data than before (24%).

Many types of analytic applications benefit from Hadoop. These include understanding consumer behavior via clickstreams (23%), sentiment analytics and trending (22%), recognition of sales and market opportunities (17%), detection of fraud (17%), definitions of churn and other customer behaviors (12%), and customer-base segmentation (11%).[4]

[4] For guidance in making a business case for Hadoop, see the 2011 TDWI Checklist Report *Hadoop: Revealing Its True Value for Business Intelligence*, available for free download at 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 8. *Based on 1,592 responses from 259 respondents; 6 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 9, page 16.)

Hadoop's leading barriers are adjustments to staffing/skills and DW architecture, plus the current immaturity of Hadoop tools.

Staffing and skills are the leading barriers to Hadoop implementations. As noted previously, the survey population for this report consists mostly of BI professionals. These people have strong data skills and their teams are staffed for data work. The challenge with HDFS and Hadoop tools is that, in their current state, they demand a fair amount of hand coding in languages that the average BI professional does not know well, namely Java, R, and Hive. However, this is not a showstopper; TDWI has seen a number of BI/DW teams successfully acquire the skills and staffing needed for Hadoop. As more and better development tools for Hadoop arrive from vendors and the open source community, the current excess of hand coding will give way to high-level automated approaches to BI, analytics, and data management development for Hadoop.

The current state of Hadoop development and admin tools are a problem. First of all, software tools for Hadoop are few and immature (28%), and they lack adequate metadata management (25%). Second, BI professionals are put off by Hadoop's poor reputation for handling data in real time (22%), which is required for common practices such as operational BI and ad hoc analytic queries. Users also have productivity concerns about the slow pace of hand-coded development (17%), and they feel that Hadoop's software tools need higher-level language support (16%), plus support for standard SQL (not Hive).

Despite the daunting list of current concerns, products in the broad Hadoop ecosystem (from both open source and vendors) are improving in all these areas and more. It's natural with open source software (OSS) that tools lag behind in the early days while a new platform such as HDFS is being developed and proved to be useful. Once the new platform is proved, the tools catch up. That's the lifecycle stage we're currently in with Hadoop.

Sell Hadoop internally based on the merits of the exploratory analytics it provides.

As with any new platform, selling Hadoop to your business is tricky. TDWI's experience is that organizations needing advanced forms of exploratory analytics for leveraging big data are automatically drawn to Hadoop. Organizations that don't fit this profile can suffer a lack of compelling business case (40%) or a lack of business sponsorship (41%). Hadoop's origins with Web data in large Internet firms is well known, but a variety of firms in other industries are now applying it to other types of big data, as discussed earlier in this report. As the definition of big data broadens, so does the profile of organizations that can make a solid case for Hadoop.

Architectural adjustments are needed when integrating Hadoop with BI/DW. As mentioned earlier, Hadoop is sometimes deployed in a silo, the same way some analytic applications are. TDWI's contention is that maximal business value comes from combining data and insights from both Hadoop and traditional BI/DW environments. So, even if Hadoop starts in a silo in your organization (and it has to start somewhere), integration with BI/DW should be a high-priority second step. Successful integration may require adjustments to an existing user-defined DW architecture (27%), which is part and parcel of architecting a new big data analytic system (32%).

Scalability is not a barrier to Hadoop usage. Apparently, Hadoop's strong reputation for extreme scalability is accepted credibly by survey respondents, because only 8% anticipate problems scaling up HDFS and related Hadoop products.

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

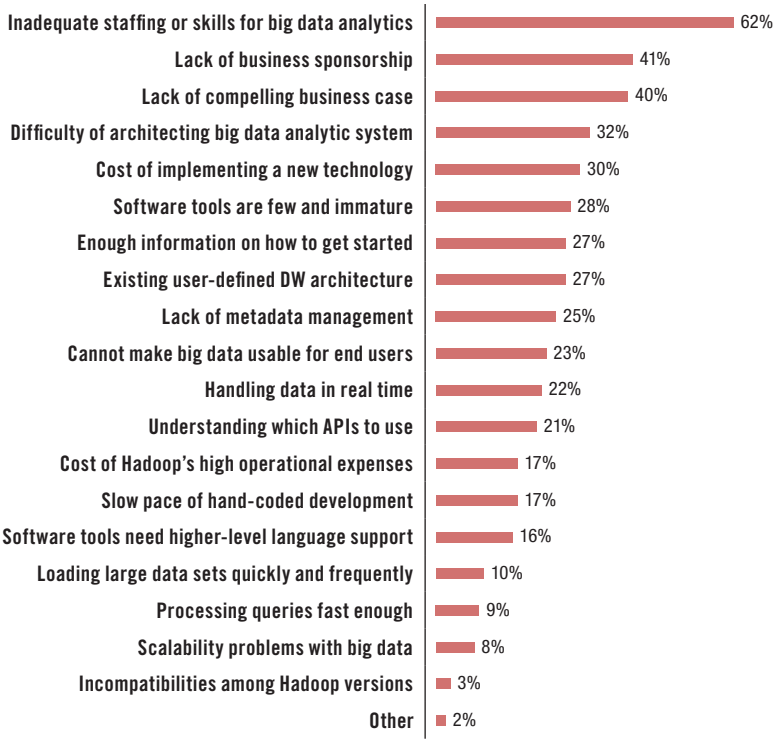


Figure 9. *Based on 1,186 responses from 256 respondents; 4.6 responses per respondent, on average.*

USER STORY

HDFS EASILY INTEGRATES INTO STANDARD DATA WAREHOUSE ARCHITECTURES.

"We have HDFS clusters and other Hadoop tools in our BI portfolio, and we've integrated them into our data warehouse architecture in two different ways," said a data warehouse architect at a large Internet firm. "First, we think of HDFS as a bigger and better data staging area, which doubles as a source data archive and a single platform for mixing structured and unstructured data. As you would do on any data staging area, we execute some of our ETL transformations, sorts, and joins on HDFS. Second, we use HDFS as if it were a non-dimensional warehouse for multi-terabyte, multi-structured, raw source data. In that scenario, we use Hadoop tools and newer analytic databases and data visualization tools that support Hadoop to query and analyze HDFS data directly. Now that an integrated architecture is in place, our next step is to get better at accessing and joining data from both the EDW and HDFS."

Emerging Best Practices for Hadoop

The survey responses discussed in this section of the report come from a subset of 48 survey respondents who report that they have deployed or used HDFS. 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 integrating Hadoop with BI/DW.

Reasons to Adopt Hadoop

To get a sense of why user organizations are deploying Hadoop, the survey asked a subset of respondents: Why did your organization adopt HDFS and related technologies? (See Figure 10.)

Hadoop users chose it for its scalability, exploratory analytics, and low cost.

Real-world organizations adopt Hadoop for its extreme scalability. Sixty-five percent of respondents with Hadoop experience chose Hadoop for scalability, whereas only 19% of the total survey population viewed scalability as an important benefit (refer back to Figure 8). In other words, the more users learn about HDFS, the more they respect its unique ability to scale.

Users experienced with HDFS consider it a complement to their DW. Roughly half of respondents believe this (52% in Figure 10), whereas only 15% think HDFS could replace their DW.

Half of Hadoop users deployed it as a platform for exploratory analytics (52%). However, one-third (35%) feel that HDFS and related technologies are also good for several applications (such as DW, archive, and content management).

Almost half of Hadoop users surveyed chose it because it's cost effective. Compared to other enterprise software, HDFS and its tools are low-cost software (42%), even when acquired from a vendor. They run on low-cost hardware (48%).

Why did your organization adopt HDFS and related technologies? Select all that apply.

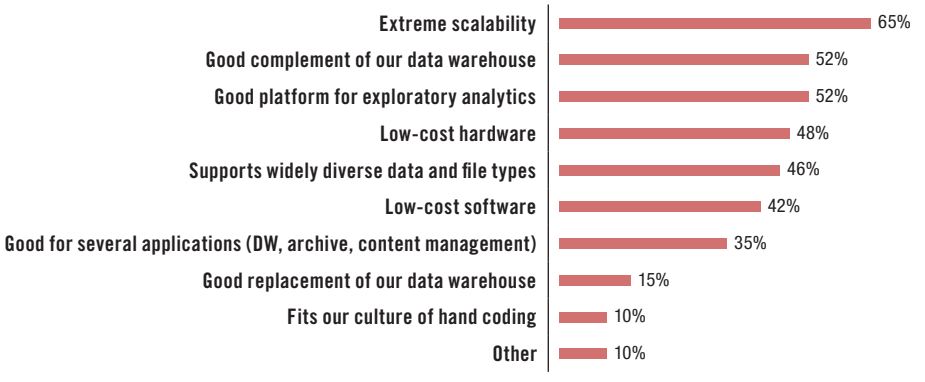


Figure 10. *Based on 180 responses from 48 respondents who have experience with Hadoop; 3.8 responses per respondent, on average.*

Hadoop is not free.

However, Hadoop is *not* free, as many people have mistakenly said about it. A number of Hadoop users speaking at recent TDWI conferences have explained that Hadoop incurs substantial payroll costs due to its intensive hand coding (normally done by high-payroll personnel such as data scientists) and its immature, non-productive tools for designing analytic solutions, optimizing those solutions, and maintaining an HDFS cluster. As more high-level tools arrive (tools that are more

productive and require far less coding), the payroll costs of Hadoop development should come down; furthermore, a wider range of developers (especially those for whom excessive hand coding is not appropriate) will be able to do a data scientist's job.

USER STORY

HADOOP'S COST-EFFECTIVE SCALING AND ARCHIVING FOR BIG DATA APPEALS TO MANY USERS.

"We run several HDFS clusters, ranging from small to very large," said Rick Miller, a senior director at Internet firm AOL. "We have a long history of using open source, and even contributing to it. We also look for cost-effective solutions. So, HDFS was an obvious choice.

"Furthermore, we do a lot of year-over-year analysis with giant collections of data of mixed schema types. As data volumes went through the roof, we knew we needed a platform for archiving extra-big data, but still be able to query and analyze it as needed. That also led us to Hadoop.

"By the way, we don't just use HDFS; we use all the Hadoop tools and others from Apache."

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 11.)

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

Data warehouse group (54%). In organizations that are intent on integrating Hadoop into the practices and infrastructure for BI, DW, DI, and analytics, it makes the most sense that the DW group or an equivalent BI team own and (perhaps) maintain a Hadoop environment.

Central IT (35%). Every organization is unique, but more and more, TDWI sees central IT evolving into a provider of IT infrastructure, especially networks, data storage, and server resources. This means that application-specific teams are organized elsewhere, instead of being under IT. In that spirit, it's possible that Hadoop in the future will be just another infrastructure provided by central IT, with a wide range of applications tapped into it, not just analytic ones.

BI/DW aside, Hadoop may be owned by application teams or provided by IT as common infrastructure.

Application group (29%). Many types of big data are associated with specific applications and the technical or business teams that use them. Just think about Web logs and other Web data being generated or captured by Web applications created by a Web applications development team. In those cases, it makes sense that an application group should have its own Hadoop environment.

Research or analysis group (25%). Many user organizations have dedicated teams of business analysts (recently joined by data scientists), who tackle tough new business questions and problems as they arise. Similar teams support product developers and other researchers who depend heavily on data. These relatively small teams of analysts regularly have their own tools and platforms, and it seems that Hadoop is joining such portfolios.

Business unit or department (15%). Most analytic applications have an obvious connection to a department, such as customer-base segmentation for the marketing department or supplier analytics for the procurement department. This explains why most analytic applications are deployed as departmentally owned silos. This precedence continues with big data analytic applications, as seen in the examples given in the previous two bullets.

Who owns and maintains your Hadoop environment? Select all that apply.

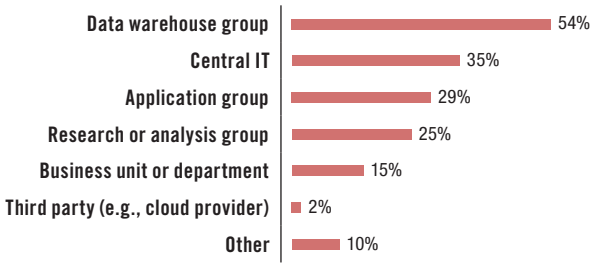


Figure 11. *Based on 82 responses from 48 respondents who have experience with Hadoop; 1.7 responses per respondent, on average.*

Job Titles for Hadoop Workers

One way to get a sense of what kinds of technical specialists are working with HDFS and other Hadoop tools is to look at their job titles, so this report's survey asked a subset of respondents to enter the job titles of Hadoop workers. (See Figure 12.) Many users are concerned about acquiring the right people with the right skills for Hadoop, and this list of job titles can assist in that area.

Hadoop workers are typically architects, developers, and data scientists.

Architect. It's interesting that the word *architect* appeared in more job titles than any other word, followed closely by the word *developer*. Among these, two titles stand out—data architect and application architect—plus miscellaneous titles such as system architect and IT architect. Most architects (regardless of type) guide designs, set standards, and manage developers. So architects are most likely providing a management and/or governance function for Hadoop, since Hadoop has an impact on data, application, and system architectures.

Developer. Similar to the word *architect*, many job titles contained the word *developer*. Again, there's a distinction between application developers and data (or BI) developers. Application developers may be there to satisfy Hadoop's need for hand-coded solutions, regardless of the type of solution. As noted, some application groups have their own Hadoop clusters. Of course, data and BI developers bring their analytic expertise to Hadoop-based solutions.

Data scientist. This job title has slowly gained popularity in recent years, and seems to be replacing the older position of business analyst. Another way to look at it is that some business analysts are proactively evolving into data scientists, because that's what their organizations need from them. When done right, the data scientist's job involves many skills, and most of those are quite challenging. For example, like a business analyst, the data scientist is also a hybrid worker who needs knowledge of both business and data (that is, data's meaning, as well as its management). But the data scientist must be more technical than the average business analyst, doing far more hands-on work writing code, designing analytic models, creating ETL logic, modeling databases, writing very complex SQL, and so on. Note that these skills are typically required for high-quality big data analytics in a Hadoop environment, and the position of the data scientist originated for precisely that. Even so, TDWI sees the number of data scientists increasing across a wide range of organizations and industries, because they're needed as analytic usage gets deeper and more sophisticated and as data sources and types diversify.

Analyst. Business analyst and data analyst job titles barely registered in the survey. Perhaps that's because most business analysts rely heavily on SQL, relational databases, and other technologies for structured data, which are currently not well represented in Hadoop functionality. As noted, some analysts are becoming data scientists, as they evolve to satisfy new business requirements.

Miscellaneous. The remaining job titles are a mixed bag, ranging from engineers to marketers. This reminds us that big data analytics—and therefore Hadoop, too—is undergoing a democratization that makes it accessible to an ever-broadening range of end users who depend on data to do their jobs well.

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

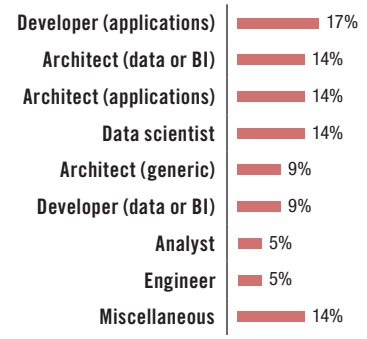


Figure 12. *Based on 58 responses from 48 respondents who have experience with Hadoop; 1.2 responses per respondent, on average.*

USER STORY

HADOOP CAN BE FAST AND CHEAP TO SET UP.

"An applications team at my firm just deployed a small HDFS cluster of about a dozen nodes," said a BI director at a TV cable and satellite company. "The firm just launched the beta of a new product that streams video content to several types of mobile devices. The application that streams the video generates a ton of data about streaming speeds and interruptions, device types, content consumed, capacity consumed, geographic locations, and the usual date and time stamps. All that data is critical to understanding how well—or not—the new system is working, so it can be improved before its first general release.

"The application team hadn't anticipated the need for analysis, so they hurriedly looked for a cheap, scalable platform they could set up and get results from quickly. HDFS, coupled with Hadoop analytic tools, fit the bill. Within hours of downloading the free open source software from Apache.org, they were getting their first counts of events on the streaming video system, which they charted using Google Maps and Google Docs."

HDFS Clusters and Nodes

It's still early for Hadoop adoption. Median organizations have only two HDFS clusters.

Number of HDFS clusters per enterprise. One way to measure the adoption of HDFS is to count the number of HDFS clusters per enterprise. Since many more people have downloaded HDFS and other Hadoop products than have actually put them to enterprise use, it's best to count only those clusters that are in production use. The vast majority of survey respondents (and, by extension, most user organizations) do not have HDFS clusters in production, so this report identified 32 respondents who do, and asked them about their clusters. (See Figure 13.)

When asked how many HDFS clusters are in production, 32 survey respondents replied in the range 1 to 100. Most responses were single-digit integers, which drove the average number of HDFS clusters down to 12 and the median down to 2. Parsing users' responses reveals that more than half of respondents have only one or two clusters in production enterprisewide at the moment, although one-fifth have 50 or more.

Note that ownership of Hadoop products can vary, as discussed earlier, which affects the number of HDFS clusters. Sometimes central IT provides a single, very large HDFS cluster for shared use by departments across an enterprise. Sometimes departments and development teams have their own.

Number of nodes per HDFS cluster. We can also measure HDFS cluster maturity by counting the number of nodes in the average cluster. Again, the most meaningful count comes from clusters that are in production. (See Figure 14.)

The few HDFS clusters in production today are fairly mature, with 45 nodes on average and a median at 12.

When asked how many nodes are in the HDFS cluster most often used by the survey respondent, respondents replied in the range 1 to 620, where one-third of responses were single digit. That comes to 45 nodes per production cluster on average, with the median at 12. Half of the HDFS clusters in production surveyed here have 12 or fewer nodes, although one-quarter have 50 or more.

To add a few more data points to this discussion, people who work in large Internet firms have presented at TDWI conferences, talking about HDFS clusters with approximately 1,000 nodes. However, speakers discussing fairly mature HDFS usage specifically in data warehousing usually have clusters in the 50- to 100-node range. Proof-of-concept clusters observed by TDWI typically have 4 to 8 nodes, whereas development clusters may have but 1 or 2.

Enter an integer representing the number of HDFS clusters in production across your enterprise:

Range	=	1 to 100
Average	=	12
Median	=	2

Figure 13. *Based on 32 respondents who have experience with Hadoop, as well as an HDFS cluster in production.*

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

Range	=	1 to 620
Average	=	45
Median	=	12

Figure 14. *Based on 32 respondents who have experience with Hadoop, as well as an HDFS cluster in production.*

Data Volume Managed by HDFS

Judging by the relatively small counts of HDFS clusters and nodes just discussed, most of the organizations in this survey population are in an early phase of Hadoop and HDFS usage. The same conclusion can be drawn from the rather modest data volumes they're managing in Hadoop today. (See Figure 15.)

Today. Sixty percent of users surveyed are managing less than 10 terabytes of data in Hadoop today. Even so, a respectable 21% are already managing data in the 10- to 100-terabyte range. A few brave souls (8%, who probably work for Internet firms) are managing roughly a half petabyte of data today, which is more indicative of what Hadoop can do at the high end.

Hadoop data volumes are modest today, but the "half-petabyte club" will quintuple in three years.

In three years. The number of user organizations in the "Hadoop half-petabyte club" will shoot up, from 8% today to 46% in three years, as organizations managing modest volumes today mature into larger data sets and integrate with more applications (including BI, DW, DI, and analytics).

What's the approximate total volume that your organization manages in Hadoop, both today and in three years? Select one per row.

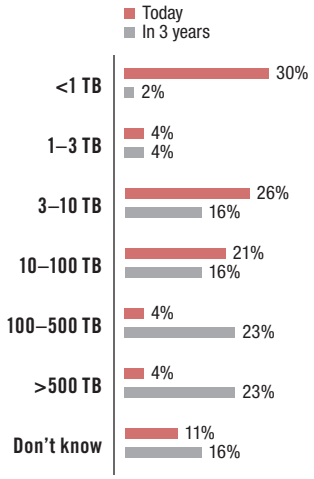


Figure 15. *Based on 48 respondents who have experience with Hadoop.*

USER STORY

INFORMATION LIFECYCLE MANAGEMENT FOR BIG DATA

"Our data warehouse environment includes an EDW on a popular relational database, two very large HDFS clusters (each managing petabyte-scale data), a data warehouse appliance, and miscellaneous analytic databases," said Rick Miller, a senior director at Internet firm AOL. "We know the things each platform does well, and we know approximately how much managing a terabyte of data costs on each. Plus, we understand how the value of data varies from one data set, source, or topic to the next. So we manage big data and other data on the cheapest platform that will get the job done. As we get better at this, we find ourselves managing and processing more data on HDFS and the appliance, and less on the EDW. We've even reduced the EDW license—much to the chagrin of the vendor!"

Data Latency Issues with HDFS

Hadoop is inherently latent, but there are work-arounds for time-sensitive data.

HDFS is inherently batch oriented, as are most Hadoop tools (MapReduce, in particular; HBase can be an exception). This imposes fundamental limitations on the speed and frequency with which data is loaded into or retrieved from HDFS via Hadoop tools. In other words, Hadoop suffers some data latency problems, in its current state. This is a bit ironic, considering that Hadoop is very much a big data platform, and some forms of big data stream in real time. This includes various types of machine data coming from sensors, robots, vehicles, mobile devices, and so on.

Hadoop workers deal with Hadoop's inherent data latency in a variety of ways. For data that needs on-demand, real-time access, they manage the data in a database management system whenever possible. As with any batch-oriented system, some Hadoop jobs can be run multiple times intraday—not just overnight—to load very recent Web log appends or blocks of machine data. For example, 25% of users surveyed load HDFS every 90 minutes or so. (See Figure 16.) Some users process streaming data in real time using systems for complex event processing or operational intelligence, then persist that big data in Hadoop for further study offline.

So, there are options today for bringing Hadoop closer to real time, and more are coming in the near future in the form of improvements to MapReduce, Hive, and HBase. Furthermore, SQL interfaces that run directly on top of HDFS are coming from a number of database, BI, and data integration software vendors. Especially promising is Impala, an open source project recently registered with the Apache Software Foundation. Among other things, Impala delivers support for standard SQL and low-latency data access. Both of these are critical if BI/DW professionals are to adopt Hadoop in appreciable numbers. Impala enables analysts to iteratively query data stored by HDFS or HBase, but from familiar SQL-based BI and analytic tools.

How frequently do you load new files and data into HDFS?

Figure 16. *Based on 48 respondents who have experience with Hadoop.*

Hadoop Functionality that Needs Improvement

Hadoop is still rather young, so it needs a number of upgrades to make it more palatable to BI professionals and mainstream organizations in general. Luckily, a number of substantial improvements are coming.

Hadoop users' greatest needs for advancement concern security, tools, and high availability.

Security. Hadoop today includes a number of security features, such as file-permission checks and access control for job queues, but the preferred function seems to be service-level authorization. This is the initial authorization mechanism that ensures clients connecting to a particular Hadoop service have the necessary, preconfigured permissions. Add-on products that provide encryption or other security measures are available for Hadoop from a few third-party vendors. Even so, there's a need for more granular security at the table level in HBase, Hive, and HCatalog.

Administration. As noted earlier, much of Hadoop's current evolution is at the tool level—not so much in the HDFS platform. After security, users' most pressing need is for better administrative tools (35%), especially for cluster deployment and maintenance (19%). The good news is that a few vendors offer tools for Hadoop administration, and a major upgrade of open source Ambari is coming soon.

High availability. HDFS has a good reputation for reliability due to the redundancy and failover mechanisms of the cluster on which it sits. However, HDFS is currently not a high availability system, because its architecture centers around NameNode. This is the directory tree of all files in the file system, and it tracks where file data is kept across the cluster. The problem is that NameNode is a single point of failure. The loss of any other node (whether intermittently or permanently) does not result in data loss, but the loss of NameNode brings the cluster down. The permanent loss of NameNode data renders the cluster's HDFS inoperable, even after restarting NameNode.

A BackupNameNode is planned to provide high availability for NameNode, but Apache needs more and better contributions from the open source community before it will be operational. There's also Hadoop SecondaryNameNode (which provides a partial, latent backup of NameNode) and third-party patches, but these fall short of true high availability. In the meantime, Hadoop users protect themselves by putting NameNode on especially robust hardware and by regularly backing up NameNode's directory tree and other metadata.

Latency issues. A number of respondents are hoping for improvements that will overcome the data latency of batch-oriented Hadoop. They want Hadoop to support real-time operation (31%), fast query execution (29%), and streaming data (25%). These will be addressed soon by improvements to Hadoop products such as MapReduce, Hive, and HBase, plus the new Impala query engine.

Development tools. Again, many users need better tools for Hadoop, including development tools for metadata management (25%), query design (23%), and ANSI-standard SQL (31%), plus a higher-level approach that results in less hand coding (31%).

Across the broad range of functionality available in the Hadoop family, which functions need improvement the most? Select five or fewer.

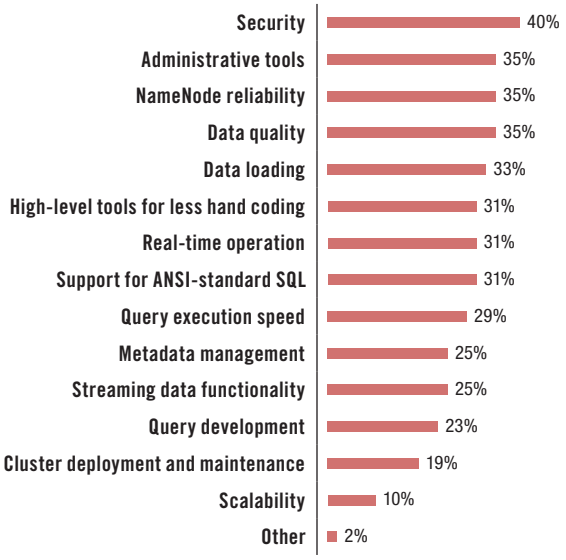


Figure 17. *Based on 195 responses from 48 respondents who have experience with Hadoop; 4 responses per respondent, on average.*

USER STORY

HADOOP ISN'T FOR EVERYONE.

A BI director interviewed for this report told TDWI: "Vendors with Hadoop distributions are courting us with demos and brochures. But their case studies are for firms that don't resemble us, so their stories seem irrelevant to my team and upper management. The hardware would be cheaper on Hadoop, compared to analytic systems we already have, but the total cost of Hadoop seems higher to me, thanks to the hefty payroll programmers pull down.

"Half of my team does statistical analysis with high-level tools, so they're turned off by the excessive hand coding of Hadoop. The other half consists of the usual BI pros, and they know enough to know that Hive won't do what SQL does for them. Their SQL-based tools don't yet support Hive. So far, we haven't identified a compelling use case for Hadoop that would fit our needs and placate our concerns."

Trends among Tools and Platforms Integrated with Hadoop

By this point in the report, we've seen a number of data, tool, and platform types that could be integrated with (or simply used around) the tools and platforms of Hadoop. These range from data warehouses to reporting tools to various forms of big data. Regardless of what project stage you're in relative to integrating Hadoop with BI, DW, DI, and analytics, knowing the available options is fundamental to making good decisions about approaches to take and software products to evaluate.

In your organization, with which platform and tool types is Hadoop integrated today? With which will Hadoop be integrated within three years? Make one selection per row.

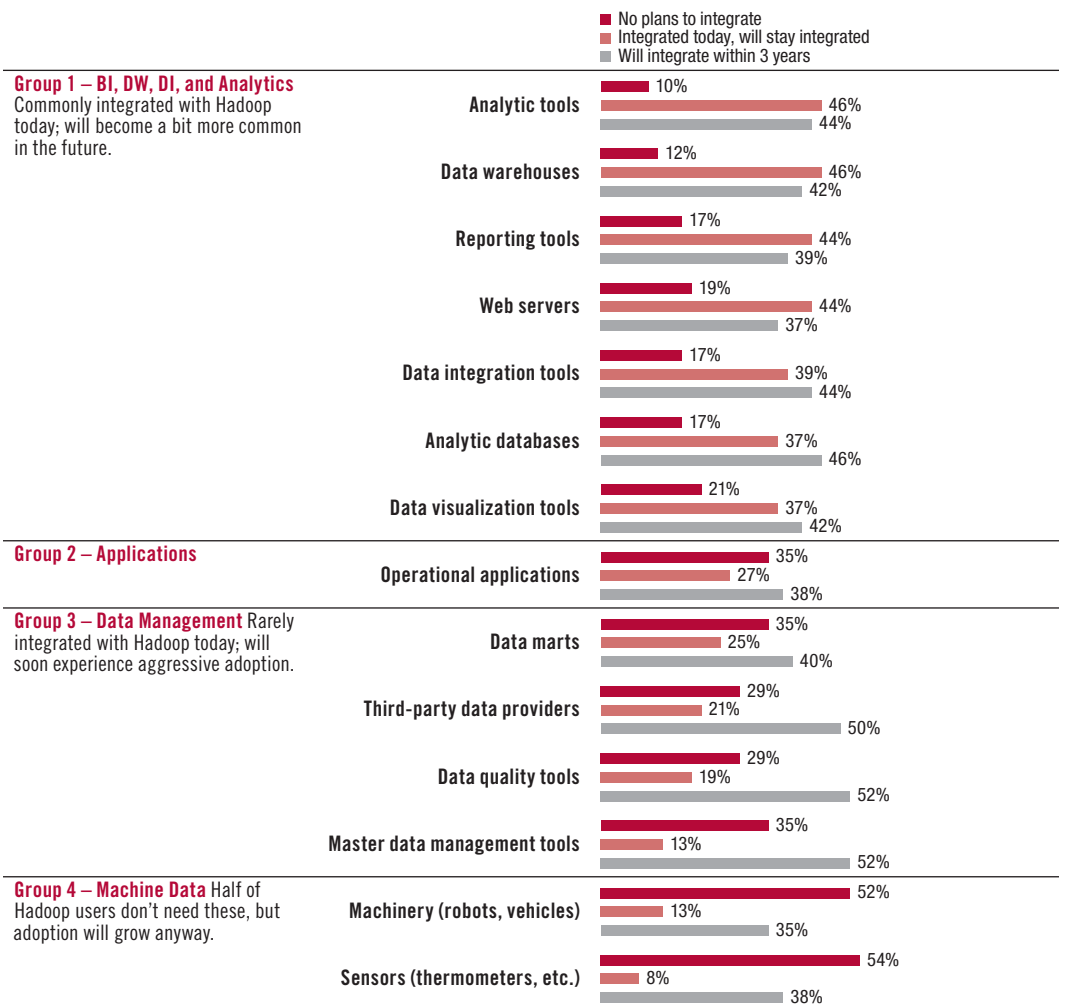


Figure 18. *Based on 48 respondents who have experience with Hadoop. The chart is sorted by "integrated today," in descending order.*

To quantify the level of integration of Hadoop and to draw the big picture of available options, TDWI presented survey respondents with a long list of data, tool, and platform types. (See Figure 18.) Concerning the list of Hadoop integration possibilities presented in the survey, TDWI asked

There are several tool, platform, and data types you might integrate with Hadoop. Choose carefully.

respondents: "In your organization, with which platform and tool types is Hadoop integrated today? With which will Hadoop be integrated within three years? Make one selection per row." Each row represents a data, tool, or platform type with which Hadoop could be integrated or simply used. For each row, respondents selected one of three multiple-choice answers:

- No plans to integrate
- Integrated today; will stay integrated
- Will integrate within three years

Survey responses show which platforms are in use with Hadoop today.

Note that the question in Figure 18 was presented only to survey respondents who have direct experience with Hadoop; therefore, their responses are an accurate portrayal of Hadoop's integration with BI/DW today, as well as committed users' intentions for the future.

Survey responses also reveal which platforms will see brisk adoption.

The list is a catalog of the most likely options for Hadoop integration. The responses indicate which options successful users are employing today, as well as which they anticipate employing in a few years. From this information, we can quantify trends and project future directions for Hadoop integration with BI, DW, DI, and analytics. We can also deduce priorities that can guide users in planning their future efforts in Hadoop integration. It's useful to know in what order successful users adopt the options, because that enables the effective planning of project lifecycle stages.

Groups of Tool and Platform Types Integrated with Hadoop

Figure 18 paints the "big picture" of the integration and use of certain tools and platforms with Hadoop. For example, four groups of options cluster together based on similar usage levels, as well as similar functionality. (See the four groups labeled in Figure 18.)

Group 1 – BI, DW, DI, and Analytic Tools and Platforms

Figure 18 is sorted, in descending order, by the survey result values for "integrated today," followed by values for "will integrate." In this sort order, the tools and platforms most often integrated with Hadoop today have bubbled up to the top of the chart. Group 1 tools and platforms—already the most popular among Hadoop options today—will get even more popular soon, because sizable percentages of respondents selected these as something they will integrate in the near future.

Group 1 is the epitome of integration between Hadoop and BI/DW.

Group 1 brings together most of the options related to BI, DW, DI, and analytics. This corroborates TDWI's assertion that integrating BI/DW platforms with Hadoop can be done and has already been done successfully by a number of diverse user organizations. It also shows that BI/DW tools and techniques are critical success factors for the leveraging of Hadoop and the big data it manages, as well as the business value that follows from such activities. Therefore, users should assume that all implementations of Hadoop—and, by extension, most big data platforms—should integrate BI, DW, DI, and analytics for maximum business value.

Analytic tools are the most common add-on integrated with Hadoop. Analytic tools appear at the top of the chart in Figure 18 because today they are used with Hadoop more than any other tool or platform type (at least, in the user population surveyed). This makes sense, because the primary application of Hadoop is to analyze big data. HDFS has no analytic functions built in, so these must come from analytic tools integrated with it.

It's rare to find HDFS without analytic tools.

Analytic tools are popular among Hadoop users today (46%), and they will become even more popular. An additional 44% of survey respondents say they will integrate analytic tools with Hadoop within three years.

However, not all organizations surveyed need analytics for Hadoop; as mentioned earlier, non-analytic applications for Hadoop include archiving and content management. Note that non-analytic applications for Hadoop constitute a minority practice, as seen by the mere 10% of respondents who say they have no plans to integrate analytic tools with Hadoop.

Data warehouses are commonly integrated with Hadoop. At 46%, DWs are tied for first place with analytic tools. This corroborates TDWI's position that DWs and Hadoop products are not mutually exclusive, which means they can coexist productively in real-world user organizations.

Reporting tools are in use with Hadoop today. These include Apache Hive, but also mature tools from established BI vendors. Although they haven't gotten much press, most of the leading BI vendors have built gateways, ODBC/JDBC drivers, and APIs that connect their reporting and analysis tools with HDFS or layers above it, such as Hive, HBase, and MapReduce. Support for HCatalog and Impala is coming.

Analytic databases are coming on strong with Hadoop. Thirty-eight percent of respondents use them with Hadoop today, and 46% more will within three years. These are relational database management systems (RDBMSs) that were designed and built from the bottom up for data warehousing and analytics. Some are known for their columnar data stores, in-memory databases, or appliance packaging. Most of them come from fairly young vendors that have sprung up in the last 10 years, although a few are offered by more mature RDBMS vendors.[5]

Analytic databases are most often deployed as edge systems that complement a reporting-oriented DW. Analytic databases are known for high-performance SQL, even with multi-terabyte volumes of data in poor condition (which describes most big data). Those two facts account for an increasingly popular configuration, which is to move terabytes of non-traditional data from HDFS to an analytic database, where the SQL support and data latency are far better than in HDFS and the scalability is almost as good. Some analytic databases now support HDFS and other Hadoop products so well that the analytic database can be layered over HDFS for direct, near-time data access. In this configuration, the analytic database becomes a SQL-compliant "front end" for HDFS.

Data visualization tools are becoming de rigueur with Hadoop and big data in general. Over one-third of respondents report using data visualization tools with Hadoop today (38%), and another 42% anticipate doing so within three years. Modern tools for data visualization support a number of reporting and analysis methods—not just advanced forms of visualization. Yet, visualization is a great way to explore data and discover unknown facts, which is why it's a great fit for the discovery analytics typically done with big data. In addition, leading data visualization tools work directly with Hadoop data, so that large volumes of big data need not be processed and transferred to another platform.

Group 2 – Applications

Integrating Hadoop with a variety of operational applications seems to be a mid-level priority at the moment, taking a backseat to BI/DW integration. But progress is being made—usually in the form of data aggregation, not application integration. For example, many operational applications generate big data as a by-product of a business process, as in the Web logs generated by Web servers and Web applications or the logs that are constantly appended to by sensors, robots, and other machinery. This differs from data products that are considered high value, such as financial transaction data or the golden record of a customer.

Similar to a DW, HDFS commonly aggregates massive amounts of application data.

[5] For a more detailed discussion of analytic databases and their requirements, see the 2012 TDWI Checklist Report *Analytic Databases for Big Data*, available for free download at tdwi.org/checklists.

High-value data products continue to be managed and nourished in relational databases, whereas file-based, big data by-products are finding a useful home aggregated in Hadoop (without true integration back to originating applications), where their value increases as users learn how better to recycle, study, and leverage them via BI/DW. This style of Hadoop usage follows a recognizable data warehousing strategy based on data aggregation, but with data types, sources, and volumes rarely seen in traditional data warehousing.

Group 3 – Data Management Tools and Platforms

Group 3 consists of a number of data management practices and their attendant platforms, namely data marts, third-party data, data quality, and master data management. Although popular in the traditional world of structured data, these are still rare in the new Hadoop world of mixed data types and data structures. In fact, a fair number of respondents (29% to 35%) say they have no plans for these data management practices. Yet, all these are poised for aggressive growth, with 40% to 52% of respondents anticipating use within three years. For that level of adoption to happen, both Hadoop and data management need to evolve.

Both Hadoop and data management practices need to evolve before the two can come together.

The data management practices and platforms of Group 3 (and others) were designed for structured data and relational databases, with heavy reliance on metadata, standard SQL, and random access to atomic data. HDFS and other Hadoop products were designed for the opposite—mostly schema-free, multi-structured data in a non-indexed file system. Despite this glaring mismatch today, Hadoop and data management in general are coming closer. This report has already noted the numerous improvements and new functionalities coming to Hadoop, many of which make the Hadoop ecosystem more compatible with how relational databases, SQL, and their semantics work. From another direction, vendors offering products for Group 3 practices (and especially for data integration, which is in Group 1) are steadily rolling out support for HDFS, MapReduce, Hive, HBase, HCatalog, and so on.

Other issues exist, but are evolving appropriately. For example, the methods of advanced analytics executed against big data (such as data mining, clustering, statistics, and sentiment) typically require lightly processed data or raw source data. Transforming that kind of data via standardization, cleansing, and modeling can omit or obscure the data nuggets that advanced analytics needs to produce insights. Vendor and open source tools aside, users are currently evolving their data management practices to accommodate the peculiarities of data preparation for advanced analytics with Hadoop and other big data.[6]

Group 4 – Machine Data

Not much machine data is managed by Hadoop today. But this is coming.

The types of machine data encompassed by Group 4 are the least addressed Hadoop options today. This is seen in the fairly high percentages of respondents (52% and 54%) who say they have no plans to integrate data from machines or sensors, respectively, into Hadoop. This could be because most machine data streams 24/7 in real time. As discussed in previous sections of this report, the Hadoop environment in its current state is batch-oriented and therefore prone to excessive data latency. Some users collect machine data in logs and latently load the log files into HDFS, similar to what they do with Web logs. This way, at least they can study machine data offline, after the fact.

Luckily, new Hadoop tools and improvements to existing ones are coming soon to address the need for real-time data processing. As Hadoop's support for real time increases, so will users' integration of machine data into Hadoop. This is suggested by the 35% and 38% of survey respondents who say they will be integrating machine data into Hadoop within three years.

[6] For more information on this topic, see the 2009 TDWI Checklist Report *Data Requirements for Advanced Analytics*, available at tdwi.org/checklists.

USER STORY

HDFS ON A CLOUD

"Some time ago, my employer committed us to moving as many applications as possible off premises onto a certain provider's cloud," said a BI professional at an insurance firm. "So far, most of our BI and DW platforms have ported just fine to the cloud, including the EDW. And the vendor distribution of HDFS we use and other Hadoop tools ported great, too. The catch is that our statistical analysis tool won't run on the chosen cloud provider, so now we're extracting and moving big data off the cloud daily to get it to the on-premises statistical analysis tool. So far, the daily data migrations work just fine, despite the large data sets.

"Our BI environment is now a hybrid mix of HDFS and an EDW, with traditional BI and Hadoop analytic tools joining structured and unstructured data, both on cloud and on premises."

Vendor Platforms and Tools that Support Hadoop

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 to integrating Hadoop with BI, DW, DI, and analytics.[7]

Cloudera

Cloudera is a leading provider of Apache Hadoop–based software, services, and training, enabling data-driven organizations to derive business value from all their data while simultaneously reducing the costs of data management. CDH (Cloudera's distribution including Apache Hadoop) is a comprehensive, tested, and stable distribution of Hadoop that is widely deployed in commercial and non-commercial environments. Organizations can subscribe to Cloudera Enterprise, comprising CDH, Cloudera Support, and the Cloudera Manager, to simplify and reduce the cost of Hadoop configuration, rollout, upgrades, and administration. Cloudera also provides Cloudera Enterprise Real-Time Query (RTQ), powered by Cloudera Impala—the first low-latency SQL query engine that runs directly over data in HDFS and HBase—to enable interactive BI on Hadoop. As a major contributor to the Apache open source community, with tens of thousands of nodes under management across customers in every industry, and a partner program that includes more than 490 tools, applications, systems and SI partners, Cloudera's depth of big data experience and expertise is profound.

EMC Greenplum

EMC Corporation is the world's leading provider of data storage platforms and other information infrastructure solutions. EMC acquired Greenplum in 2010 and has since built it up as the EMC Data Computing Division. Greenplum customers are some of the largest firms in the world, and they regularly deploy Greenplum products on grids or clouds to scale up to very big data and advanced forms of analytics. Greenplum Database is known for its shared-nothing massively parallel processing (MPP) architecture, high-performance parallel dataflow engine, and gNet software interconnect technology. Greenplum Database supports high-speed, parallel connectivity to Hadoop and offers a SQL front end that works directly on top of HDFS. Greenplum HD is an open source, enterprise-ready Hadoop distribution, and Greenplum Chorus is a productivity platform for collaborative research and analysis that supports data exploration and analysis on HDFS. The three are integrated together in the Greenplum Unified Analytics Platform (UAP), which may be deployed on Greenplum Data Computing Appliances.

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

Hortonworks Hortonworks focuses on innovating the core of open source Apache Hadoop in ways that make Hadoop enterprise grade and therefore more applicable to more user organizations. Hortonworks' strategy is to distribute 100% open source Apache Hadoop, with additional operational, data, and platform services from the open source community, all packaged as the Hortonworks Data Platform (HDP). Multi-tenancy is built into HDP, so it can be a shared enterprise infrastructure instead of a silo, and HDP 1.2 beefs up security, which is the leading concern of Hadoop users. Hortonworks is a major contributor to open source Hadoop technologies, and it has recently shown leadership in the design of Apache HCatalog (metadata services for the Hadoop ecosystem), Apache Ambari (cluster management and monitoring for HDFS), and high availability for NameNode in Hadoop 2.0.

ParAccel ParAccel Analytic Platform is built from the ground up to execute high-performance analytics on big data. ParAccel establishes a bidirectional, node-to-node connection with HDFS to become an extension of the Hadoop cluster. This allows organizations to do out-of-the-box analytics on Hadoop data without additional programming that would require a data scientist or other expert. ParAccel's integration with HDFS enables users to utilize the lingua franca of analytics, namely standard SQL. ParAccel also optimizes the SQL queries, whereas Hadoop today has no query optimization. Furthermore, ParAccel can share analytic processing with Hadoop. Hence, ParAccel can serve as a SQL-compliant front end for Hadoop. ParAccel offers a full set of On Demand Integration modules so the analyst can share data and processes with unlimited numbers of other data sources while eliminating latency issues and supporting real-time analytics. ParAccel also comes with 600+ advanced functions ready to be used on Hadoop data.

SAP SAP provides a comprehensive set of solutions for big data, including analytic applications, rapid deployment solutions, BI and advanced analytic tools, analytic databases, data warehousing solutions, and information management tools. Furthermore, SAP enables its customers to integrate Hadoop into their existing BI, advanced analytic, and data warehousing environments in multiple ways, giving customers the ability to tailor Hadoop to their needs. Many customers are deploying Hadoop alongside SAP HANA, an in-memory database used for real-time analytics and other applications. Customers can use SAP Data Services to search and load data from HDFS or Hive into SAP HANA or SAP Sybase IQ. Furthermore, SAP BusinessObjects BI, SAP Visual Intelligence, and SAP Predictive Analysis users can query Hive environments, giving business analysts the ability to explore Hadoop data directly. Finally, customers can federate queries across SAP Sybase IQ and Hadoop environments, or alternatively run MapReduce jobs across an SAP Sybase IQ cluster.

SAS In recent years, SAS has focused on SAS High-Performance Analytics and is now leading a broader strategy to complement and apply Hadoop throughout the big data analytics lifecycle—from data preparation and data exploration to model development and deployment. For example, SAS Visual Analytics allows users to access the shared memory of the Hadoop cluster to offer an in-memory analytics platform, enabling users to explore and visualize massive amounts of data in the Hadoop ecosystem. SAS High-Performance Analytics Server is an in-memory solution that empowers analysts to develop predictive models and deliver insights in minutes, without data movement outside the Hadoop cluster. From a data management perspective, SAS provides access to Hadoop (via HiveQL) using SAS/ACCESS software. Users can also apply existing MapReduce, Pig, or Hive code from within the SAS environment. SAS Data Integration Server provides an intuitive, graphic interface to integrate and transform data to and from Hadoop. With SAS metadata, data lineage, and security, customers can continue to integrate their data management and analytic investments with Hadoop.

Tableau Software

Tableau is known for its strong visualization features, which allow analysts and other users to explore big data and discover previously unknown facts about the enterprise using easy-to-use, drag-and-drop methods. Analytics aside, Tableau is also used as an all-purpose BI platform that can be applied to either enterprise or departmental needs. For users with big data managed by Hadoop, Tableau easily connects at the Hive level to Apache Hadoop distributions from Cloudera and MapR. High latency is one of the barriers to Hadoop usage. Tableau jumps this hurdle by connecting live to a Hadoop cluster and extracting the data into Tableau's fast in-memory data engine. With a single click between the options of live connect or in-memory analytics, users can quickly analyze samples of data in memory, then reconnect to run a live query—without waiting for MapReduce queries to complete.

Teradata

In recent years, Teradata has broadened its product portfolio considerably to keep pace with new market demands for DW tools and discovery platforms that support big data, advanced analytics, and Hadoop. Teradata acquired Aster Data and its patented SQL-MapReduce; it brings much-needed SQL support to MapReduce, and has been extended recently to support SQL-H, which (primarily through HCatalog) provides access to data in Hadoop environments. Furthermore, Teradata struck an agreement to resell Hortonworks' distribution of Apache Hadoop. All these and more have come together in a best-of-breed solution called the Teradata Unified Data Architecture (UDA). This includes workload-specific platforms for every user in the enterprise, so they get seamless insight without worrying about the underlying technologies. UDA platforms ease system administration through value-added software integration, unify access and monitoring, and enable proactive customer support. An example of a UDA platform is the recently announced Teradata Aster Big Analytics Appliance, a ready-to-run platform that is preconfigured and optimized specifically for big data storage and analysis. The appliance runs the Aster SQL-MapReduce and SQL-H technologies on a time-tested, fully supported Teradata hardware platform. It also runs the Hortonworks Data Platform (HDP), and users can flexibly configure appliance nodes for Aster, HDP, backup, or combinations.

Top 10 Priorities for Integrating Hadoop into BI/DW

In closing, let's summarize the findings of this report by listing the top 10 priorities for using Hadoop with business intelligence and data warehousing, 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 integrating Hadoop with BI, DW, DI, and analytics.

1. **Embrace the new tool and platform ecosystem of Hadoop.** Do it for the benefits: 48 survey respondents with hands-on Hadoop experience said they adopted Hadoop for its extreme scalability, exploratory analytics, low cost, and support for multi-structured data. Make a business case for Hadoop based on these and the analytic applications they enable.
2. **Know the 10 myths of Hadoop and bust them daily.** Misconceptions abound. Expect to spend time educating peers and management about what Hadoop can and cannot do and why the organization needs it. Set proper expectations by stressing that Hadoop is a complement to existing systems, not a replacement.
3. **Don't be fooled: Hadoop isn't free.** This is the 11th myth of Hadoop. Hadoop's intense hand coding means many hours for high-payroll programmers. Budget, staff, and train accordingly. Large clusters demand electricity and administration, despite cheap servers.

4. **Get training (and maybe new staff) for new Hadoop.** TDWI's take is that it's easier to train a BI professional in Hadoop than it is to train an applications developer in BI and Hadoop. Inexplicably, some organizations have taken the latter route. Take the former.
5. **Look for capabilities that make Hadoop data look relational.** Let's be honest: This is what true data people such as BI professionals want. Luckily, Hadoop's upcoming upgrades and new products supporting SQL, metadata, real-time query, and more tabular functions are exactly what data people hope for from Hadoop.
6. **Expect to wait a while for certain Hadoop functionality to mature.** You may wish to put off using Hadoop's weaker functionality, namely SQL support, real-time query, admin tools, and machine data. Until those functions are updated, learn and use Hadoop's strengths, namely extreme scalability, multi-structured data management, and hand-coded analytics.
7. **Beware siloed analytics, including Hadoop implementations.** After all, the goal is to integrate Hadoop with your well-integrated BI/DW environment, not proliferate twenty-first-century spreadmarts. Someone (probably not you) should decide whether the HDFS cluster will be departmentally owned (like a lot of analytic applications) or shared enterprise infrastructure supplied by central IT. Even if Hadoop begins in a silo (and you do have to start somewhere), make integration with BI, DW, DI, and analytics a second-phase priority. To make the integration happen, look for products (both open source and vendor built) that enable the integration points discussed elsewhere in this report.
8. **Adjust your data warehouse architecture to make a place (or places) for Hadoop.** There are many areas within standard DW architectures where HDFS and other Hadoop products can make a contribution, namely in data staging, archiving detailed source data, managing non-structured data, managing file-based data, data sandboxes, more processing power for an ETL hub or ELT push down, and anywhere you might use a non-dimensional operational data store. This richness of possibilities alone sells many people on Hadoop for BI/DW.
9. **Set up a proof of concept (POC), if you haven't already.** It's open source. Download HDFS from Apache.org or look for an introductory or developer's distribution from a vendor. HDFS is an obvious choice, but think carefully about what's next. Layer Hive over HDFS (and later HBase), if you have a POC application in mind that just needs straightforward queries. Layer MapReduce over HDFS (later Pig to simplify MapReduce development) if the application you have in mind requires analytics.
10. **Develop and apply a strategy for Hadoop integration into BI/DW.** It can be as simple as the plan for a POC described above, or you can start there and develop a more detailed plan for when to adopt which Hadoop tools and technologies, as well as when to turn on Hadoop support in your traditional BI/DW tools. Likewise, the plan should incorporate other priorities explained here, such as DW architecture adjustments, ownership, training, staffing, and budgeting. But don't work out those details until your team has decided which analytic applications would make good test cases and POCs and, therefore, should be developed in early phases. Based on survey results, candidate applications for early phases might study consumer behavior via clickstreams, sentiment analytics and trending, sales and market opportunities, fraud, churn, or customer-base segmentation.

sas.com/Hadoop

SAS® and Hadoop—Combine the Big Data Crunching Capabilities of Hadoop with SAS® Advanced Analytics

SAS is pleased to sponsor the TDWI Best Practices Report *Integrating Hadoop into Business Intelligence and Data Warehousing.* The leader in business analytics software and services, SAS has extended its advanced analytics capabilities to a variety of database and storage vendors, including Hadoop. SAS provides a richer visual and interactive Hadoop experience, offering seamless access to Hadoop capabilities such as the Pig and Hive languages and the MapReduce framework. It is all part of a larger SAS strategy to manage the entire analytics life cycle, from data to decision.

How SAS® Can Help

- **Easily access and use big data stored in Hadoop.** SAS/ACCESS® provides data access to Hadoop (via HiveQL). Users can access Hive tables as if they were native SAS data sets, and then apply text mining and predictive analytics to data stored in Hadoop.
- **Maximize Hadoop's distributed processing capability.** SAS helps execute Hadoop functionality with Base SAS by enabling MapReduce programming, scripting support, and the execution of HDFS commands from within the SAS environment. This extends support for Pig, MapReduce, and HDFS commands.
- **Effectively manage Hadoop using SAS Data Management**. SAS Data Management technologies help users get more value from data residing in Hadoop with existing familiar SAS technology. It provides intuitive graphic interfaces to access and transform data in Hadoop file systems. With shared SAS metadata and security components, customers can continue to integrate their data management and analytic investments.
- **Visualize your data stored in Hadoop, discover new patterns, and publish reports.** SAS Visual Analytics is an in-memory solution for exploring data very quickly. It enables you to identify opportunities for further analysis and convey visual results via Web reports or mobile devices.
- **Use Hadoop for big data analytics**. SAS High-Performance Analytics brings a set of in-memory products that allows you to develop analytical models using all data, not just a subset, and deliver rapid insights. You can run frequent model iterations and use sophisticated analytics to get answers to questions you never thought of or had time to ask.

How SAS® Is Different

- **Comprehensive analytic support for Hadoop.** SAS/ACCESS not only retrieves big data stored in Hadoop's distributed file system; it also allows you to incorporate and use other Hadoop capabilities such as the Pig and Hive languages and the MapReduce framework.
- **Flexible architecture.** SAS offers a flexible approach to hardware or database vendors by working with users to deploy the correct mix of technologies, including the ability to deploy Hadoop with other data warehouse technologies.
- **Complete lifecycle support.** SAS supports the entire analytics life cycle, from data preparation and exploration, model development, and production deployment and monitoring.

Learn more at sas.com/Hadoop.

TDWI RESEARCH

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

1201 Monster Road SW
Suite 250
Renton, WA 98057-2996

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

106436_S108845.0413