WHITE PAPER

Trends in Enterprise Hadoop Deployments

Sponsored by: Red Hat

Ashish Nadkarni Laura DuBois

October 2013

IDC OPINION

Businesses are in the midst of a transformation — a journey that will ultimately make them analytics driven to remain competitive in the 21st century. Examples of analytics-driven activities include service innovation, customer experience improvements, detection and remediation of anomalies, and reduction of time to market for goods and services. To meet the need for analytics-driven activities, businesses are collecting, analyzing, and storing more data, often from sources that did not exist a few years ago. They are also implementing newer workflows that allow them to quickly and continuously assess the results of these analytics platforms — and change their business functions accordingly. This means that newer analytics platforms like Hadoop have to be more agile than their predecessors and need to support continuous value derivation loops implemented via these Big Data workflows. As this white paper illustrates, the key findings of IDC's *Red Hat Hadoop Usage Survey* indicate that the implementation of analytics platforms like Hadoop impacts the choices businesses have to make with respect to their infrastructure and data management. Such infrastructure needs to support:

- A scale-out compute platform that leverages commodity components but also doubles as an economically feasible persistent storage platform
- Structured and unstructured data sets that are associated with NoSQL, MPP, and traditional SQL databases
- Data management systems like file systems that augment or replace the native capabilities of the platform itself
- The capture and collation of data from multiple sources, including high-speed streaming

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

METHODOLOGY

This IDC white paper presents results from an IDC survey of IT and storage professionals that covers investment plans for, current adoption of, attitudes toward, and requirements for Hadoop solutions. This white paper is based on a survey (conducted in August 2013) of 202 IT professionals responsible for management, administration, and/or decision making for their firm. The surveyed firms are currently using Hadoop or considering (i.e., an evaluation or in planning stages) a Hadoop implementation.

To maintain consistency in the results, IDC provided each of the survey respondents with definitions of Big Data and analytics terms and concepts. These definitions can be found in the Appendix.

SITUATION OVERVIEW

Big Data and analytics will continue to be a key constituent of infrastructure spending as businesses undergo a transformation to being data driven. To that effect, businesses are seeking newer data sources with the goal of seeking correlation and causality patterns. They are also storing this data longer, even data that has been analyzed before, as they move from search-based analytics to discovery-based analytics. This data-driven transformation means that the infrastructure spending is not just to accommodate this data deluge but also to support newer open source–based analytics platforms like Hadoop that present their own set of challenges in the datacenter.

While the survey focused on the deployment of Hadoop in the enterprise, IDC notes that many of the challenges faced by businesses are not unique to Hadoop. Rather, they extend to other analytics platforms — both commercial and open source based.

As an example, platforms like Hadoop lack enterprise-grade resiliency and data management capabilities. Therefore, businesses that need the functionality of Hadoop but cannot afford to compromise on service-level objectives are forced to augment their Hadoop deployments with commercial add-on solutions.

 ©2013 IDC

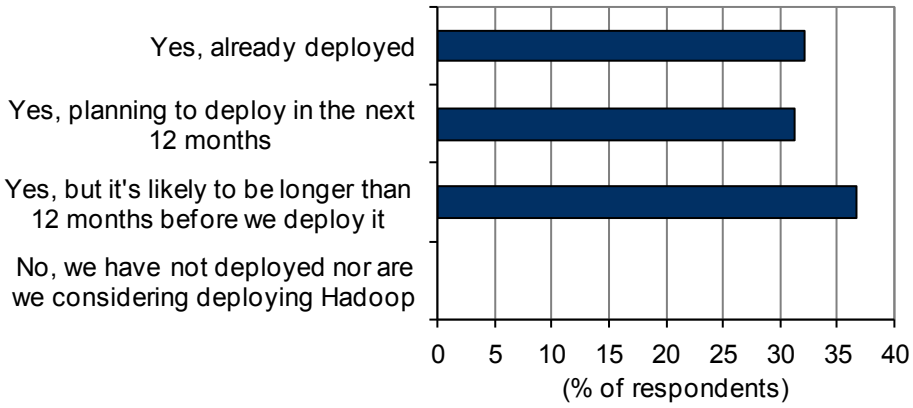
Hadoop in the Enterprise

To participate in the survey, firms had to be using Hadoop or considering a Hadoop implementation. Figure 1 illustrates the number of respondents who either have deployed or are considering deploying Hadoop. 32.2% of respondents indicated that their firms have existing Hadoop deployments. An additional 31.2% indicated that they had plans to deploy Hadoop within 12 months. And finally, 36.6% said that their Hadoop deployment schedule could go beyond 12 months.

FIGURE 1

Hadoop Deployment Status

Q. Has your organization deployed or considered deploying Hadoop?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

Primary Use Cases for Hadoop

Figure 2 illustrates some of the use cases for Hadoop. This question was posed as a multiple-select question, and the results show that several businesses use Hadoop in more than one way:

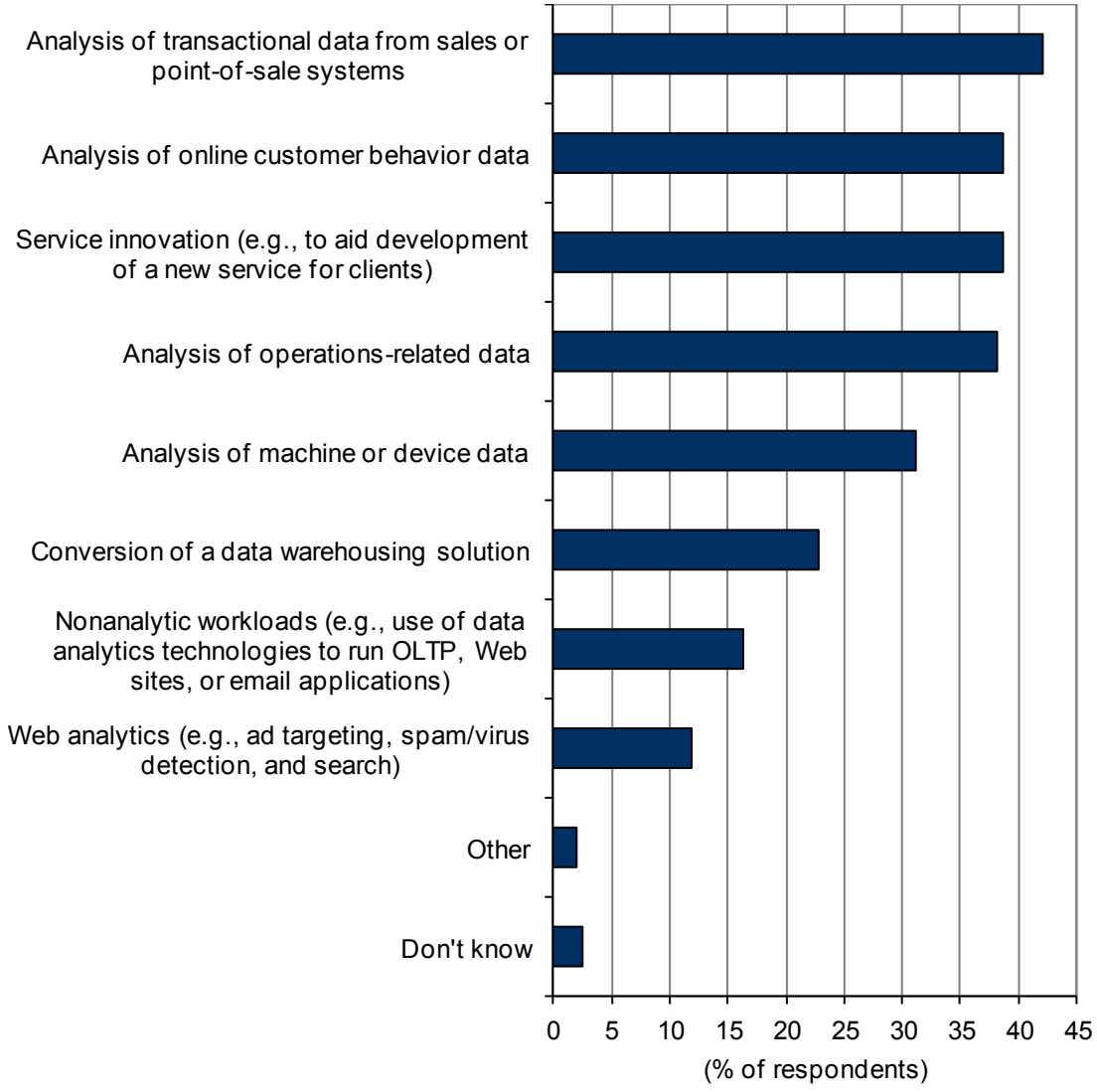
- Analysis of raw data — whether it is operations data, data from machines or devices, data from point-of-sale systems, or customer behavioral data gathered from ecommerce or retail systems — is one of the dominant use cases for Hadoop.
- Nearly 39% of respondents indicated that they use Hadoop for service innovation, which includes the analysis of secondary data sets for modeling of "if-then" scenarios for products and services.
- The less popular use cases for Hadoop include its deployment as a platform for nonanalytic workloads (e.g., in conjunction with an SQL overlay for OLTP). Many businesses also use Hadoop as a replacement for their older data warehouse databases.

- Hadoop is gaining traction as a Web analytics and content-sharing platform. Several businesses use it in Web analytics (e.g., ad targeting, spam/virus detection, and search) or as a persistent storage platform for content sharing or to serve meta-messaging applications.

FIGURE 2

Primary Use Cases for Hadoop

Q. What are the primary use cases for Hadoop in your organization?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

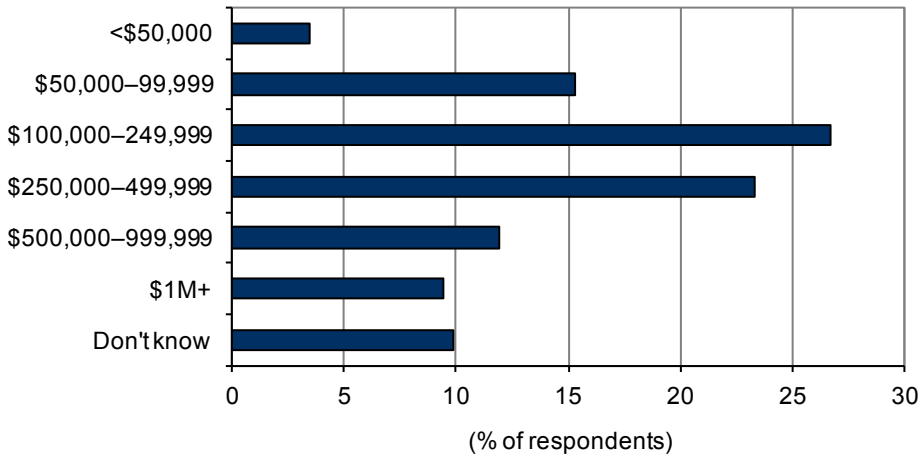
Hadoop Data Migration Costs

Figure 3 illustrates the direct and indirect costs associated with migrating data into Hadoop. Since most of the data sets are not native to Hadoop — and in many cases reside inside multiple applications in multiple locations — getting data into Hadoop can be quite a taxing and expensive proposition. Nearly 50% of respondents indicated that their firms pay anywhere from $100,000 to $500,000 in data migration costs. Some of these costs are one-time costs associated with developing data conversion modules, while some of these costs are recurring costs associated with ongoing conversions. Figure 3 highlights the fact that open source variants of analytics platforms are not cheap to deploy. It is therefore imperative that businesses make every attempt to measure the ROI of such platforms. This data is illustrated in Figure 5.

FIGURE 3

Hadoop Data Migration Costs

Q. Approximately what was the cost in dollars to migrate data into the Hadoop environment(s)?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

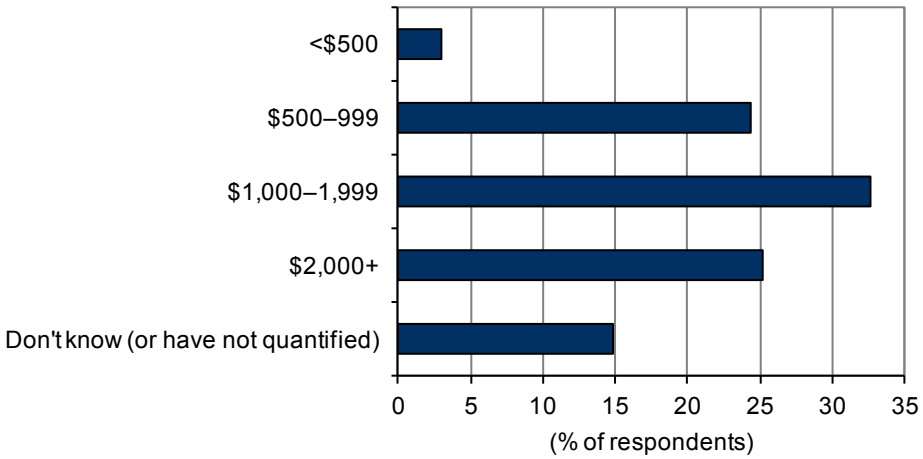
Ongoing Support Costs

Figure 4 indicates the annual support costs per server in the Hadoop cluster. Since Hadoop is a node-based shared nothing architecture that can be built using commodity components, most of these costs are management costs. IDC expects these costs to be in line with the support costs for other application servers in the infrastructure. The unweighted median annual cost per server is $1,600.

FIGURE 4

Ongoing Support Costs for Managing the Hadoop Cluster

Q. Please state the ongoing support costs, per server per year, of managing the Hadoop cluster.



n = 202

Base = all respondents

Notes:

Cost is dollars/server/year.

This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

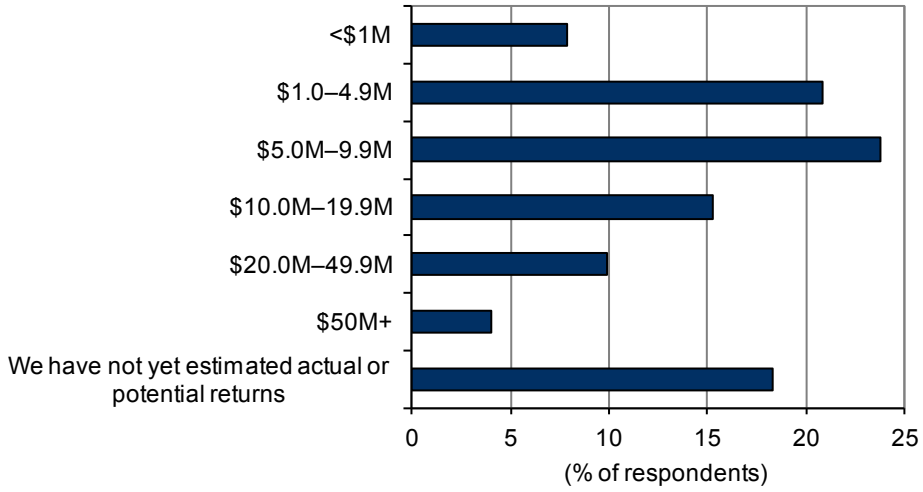
Quantified Business Benefits

Figure 5 illustrates the quantified business benefits derived from deploying Hadoop. This is an important data point — something that should be mandatory for every business that has deployed or is considering deploying Hadoop. This is especially true when the use cases do not directly support a profit center (e.g., analysis of operations data). While the vast majority (nearly 82%) of respondents were able to provide ballpark ranges for quantified benefits, the fact that about 18% of respondents could not provide such information is troubling. IDC expects the situation to change as the pressure builds on the business sponsors and stakeholders to justify their investments in Hadoop.

FIGURE 5

Quantified Business Benefits of Deploying Hadoop

Q. What is the quantified business benefit in dollars that your organization has realized, or expects to realize, in the first three years by implementing Hadoop?



n = 202

Base = all respondents

Notes:

Data represents a three-year dollar impact.

This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

Hadoop Enterprise Deployment Patterns — Infrastructure

This section deals with some of the Hadoop infrastructure patterns observed by IDC via the survey. The biggest impact of deploying Hadoop is felt on the overall datacenter infrastructure (note that IDC did not survey businesses that run Hadoop in public clouds such as those from Amazon).

Polyglot Persistence

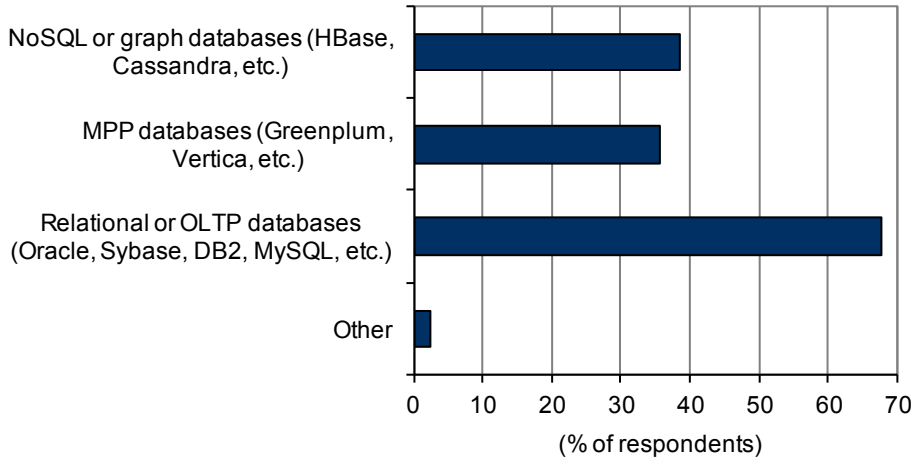
As businesses morph themselves into data-driven entities, they have no choice but to embrace polyglot persistence for their data — a condition in which they move away from a monolithic, one-size-fits-all approach to storing data. The fact that structured data sets (that can be organized to fit a predetermined schema) form only a small component of the overall data pie means that other types of databases like NoSQL and MPP databases are increasingly finding a place in the analytics infrastructure.

Figure 6 illustrates how this landscape looks in the enterprise. Nearly 39% of respondents indicated that they use NoSQL databases like HBase, Cassandra, and MongoDB, while nearly 36% indicated that they use MPP databases like Greenplum and Vertica in conjunction with Hadoop in addition to traditional relational or OLTP databases. This situation also underscores the importance of causality and correlation — in which traditional structured data sets are analyzed in conjunction with unstructured data from newer sources. Examples of such use cases are social media, customer behavior, call logs, and fraud detection.

FIGURE 6

Database Usage with Hadoop

Q. Which of the following do you use in conjunction with Hadoop?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

Adoption of Commercial Hadoop Variants

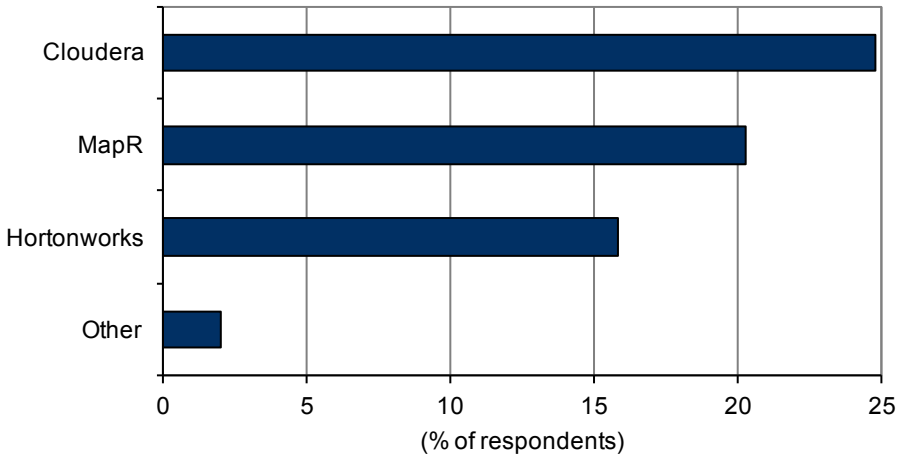
The key to wide-scale adoption of open source platforms like Linux in the enterprise has been the availability of commercial variants. Suppliers that package such variants ensure that businesses can avail themselves of the same level of service and support that they are used to with commercial offerings. The Hadoop platform is undergoing a similar transformation — which is one of the main reasons why there has been a surge in enterprise adoption. Suppliers like Cloudera, Hortonworks, MapR, and Intel have created Hadoop distributions, each with its unique value-added benefits.

Figures 7 and 8 essentially validate the success of Hadoop to commercial variants and attribute this success to the support offered by the aforementioned suppliers. The three leading suppliers — Cloudera, MapR, and Hortonworks — dominate the enterprise Hadoop scene. When asked about the reasons for selecting the said distributions, the vast majority of respondents cited support, management, and storage costs. In other words, the "DIY" model for Hadoop appeals to only a few businesses; the rest mostly go for commercial variants, much like the route they opted for with Linux.

FIGURE 7

Hadoop Distribution Choice

Q. What Hadoop distribution do you use?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

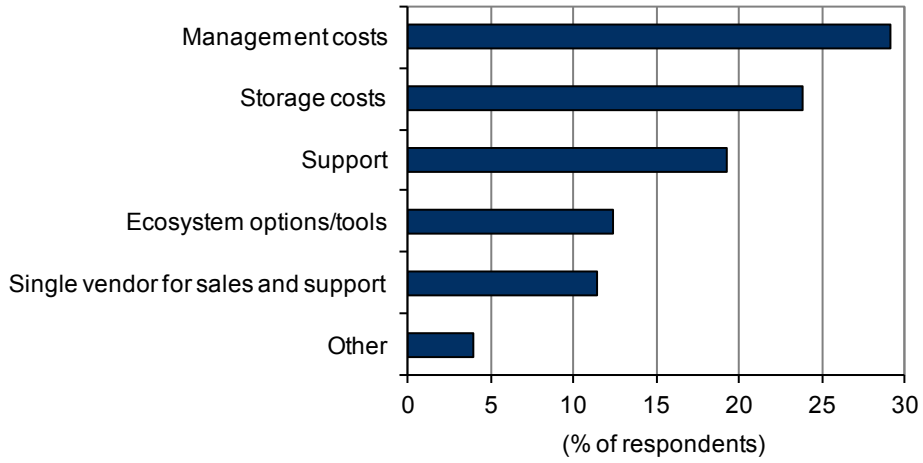
Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

FIGURE 8

Primary Reason for Choosing Hadoop Distribution

Q. What was the primary reason for selecting this distribution?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

Alternative to the Hadoop Distributed File System

One of the challenges in deploying stock Hadoop distributions involves dealing with the limitations of the underlying distributed file system that is used for persistent data storage. Known as the Hadoop Distributed File System (HDFS), this distributed file system has many advantages over monolithic file systems; however, it also lacks some of the resiliency and robust data management capabilities offered by other commercial and open source file systems. Examples of such capabilities are lack of tiering and the much advertised single point of failure with the name node.

Figure 9 offers a slightly mixed view of how businesses perceive the limitations of HDFS. While the majority of respondents said they use a different file system, a significant portion of businesses still use HDFS as the persistent data store. IDC believes such businesses usually fall into two camps:

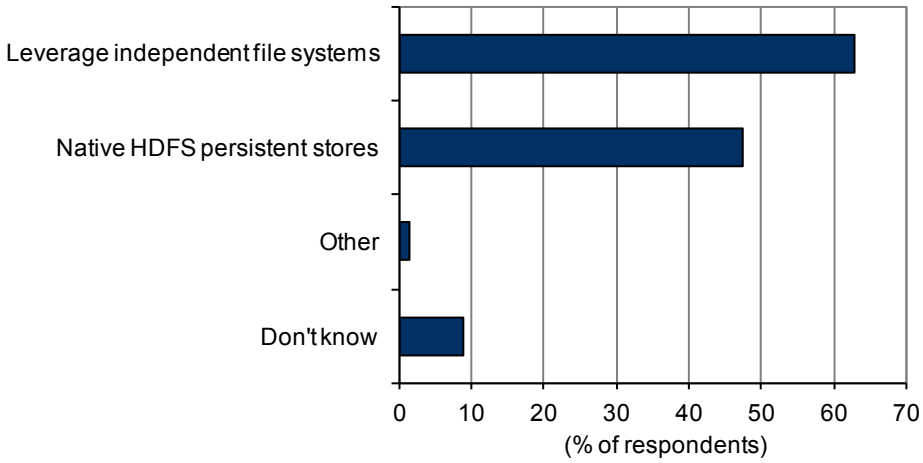
- The first camp is made up of businesses that do not use Hadoop in mission-critical environments or do not deal with critical data sets (which are purged after analysis). These businesses therefore are more risk tolerant in terms of failures or can deal with lack of data management capabilities since they treat Hadoop as a transient platform.
- The second camp is made up of businesses that leverage hardware and/or virtualization capabilities to manage resiliency and/or tiering. Such businesses factor in the use of external resiliency or data management capabilities in the TCO model for Hadoop.

 ©2013 IDC

FIGURE 9

Alternative to HDFS

Q. Has your organization deployed or do you plan to deploy the following for your Hadoop infrastructure?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

Figure 10 illustrates the types of alternative file systems that are used by businesses in enterprise Hadoop deployment. File systems like IBM's General Parallel File System (GPFS), Red Hat's GlusterFS, and EMC's Isilon OneFS that have earned a reputation for their robust scale-out capabilities are clearly preferred as alternatives to HDFS. The "other" category is made up of several distributed file systems and object storage platforms, mostly from start-ups that are gaining in popularity because of the robust enterprise-grade features of their file systems/object platforms. All of these file systems or object platforms offer a common set of benefits to the enterprise, as illustrated in Figure 11:

- They support HDFS connectors, allowing the Hadoop compute operations to seamlessly run on data that resides on non-HDFS data stores (regardless of whether this data is hierarchically organized or stored in flat namespaces).
- In most cases, these file systems are offered as a part of scale-out file-based or object-based storage platforms. The use of such platforms allows businesses to gain economies of scale in the datacenter and also perform in-place analytics on data that would have existed anyway on such platforms (e.g., archived user data or unstructured data gathered via the Internet).

- Tiering and built-in data management capabilities of such file systems mean that businesses can reduce their overall TCO and use these storage platforms for long-term data retention. An example of built-in data management applies to businesses that need to ensure that the data in Hadoop is stored in a compliant manner.

- The resiliency capabilities of such file systems mean that businesses can treat Hadoop as the sole enterprise analytics platform, use Hadoop for storing mission-critical data, and, in many cases, do away with traditional data mining and warehousing platforms. One of the most commonly used examples here is the fact that HDFS creates multiple copies of data. The use of smart protection techniques such as erasure codes does away with multiple copies of data.

- The built-in limitations of HDFS contribute to the mediocre performance of Map/Reduce operations in Hadoop. HDFS is really designed to do just one thing and one thing only: form a general-purpose file system for unstructured data. It lacks some of the robust features demanded by enterprises in file systems. An example that supports this argument is how HDFS lacks the ability to make use of server-side flash. Another example is how replication in HDFS has to be turned off to limit network chattiness and the resulting latency. File systems that are purpose built for the enterprise readily address these limitations with HDFS. Furthermore, suppliers like Red Hat have gone to great lengths to demonstrate how the use of their file system in place of HDFS improves the overall experience of Hadoop.

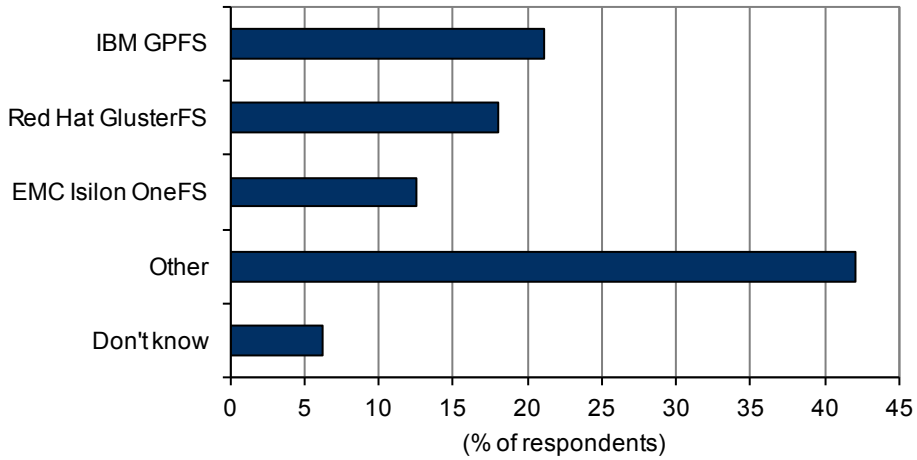
It needs to be noted that most of the storage suppliers that offer Hadoop connectors for their software-based or hardware-based platforms have partnerships with the well-known Hadoop distributors. This arrangement is similar to arrangements that storage vendors have with application vendors — this model goes a long way to convincing businesses that when there are issues, the storage vendor and the Hadoop distributor are engaged to jointly resolve the issue.

The "people element" in the respondent answers indicates that businesses often perform extensive testing and/or evaluation before settling on a configuration that they believe will work for them — this, as IDC has discovered in other research, can be anywhere from a few weeks to a few months (and, in rare cases, spans more than two years).

FIGURE 10

Most Used HDFS Alternative

Q. Which independent file systems or database platforms do you use or plan to use with your Hadoop infrastructure as a replacement for or to augment HDFS (Hadoop data store)?



n = 128

Base = respondents who deployed or plan to deploy independent Hadoop file systems

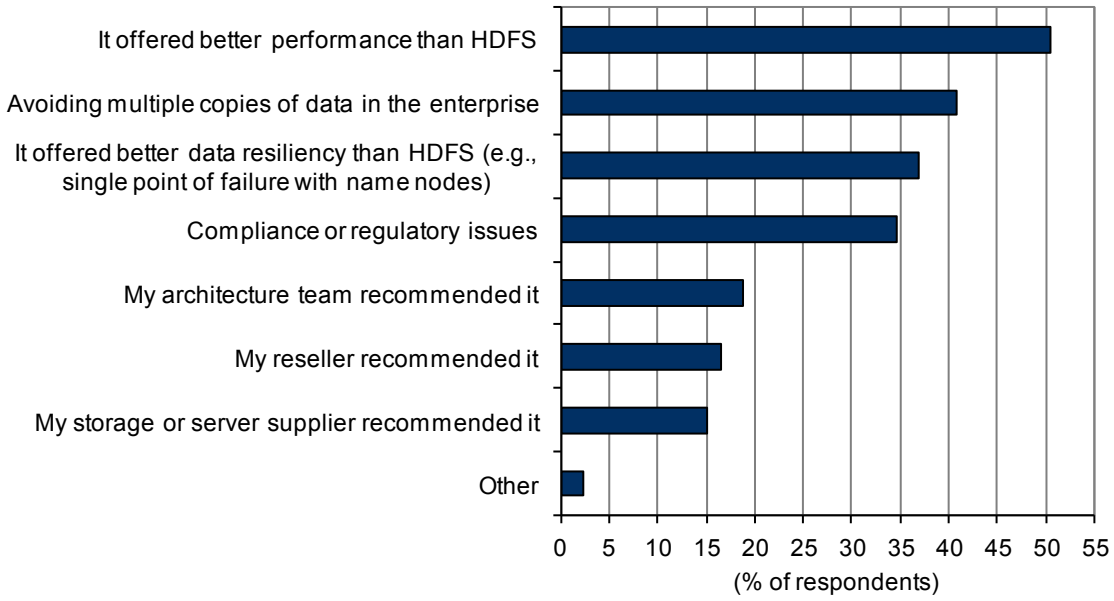
Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

FIGURE 11

Reasons for Choosing HDFS Alternative

Q. What were your reasons for selecting an alternative data store to augment or replace HDFS?



n = 127

Base = respondents who deployed or plan to deploy independent Hadoop file systems

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

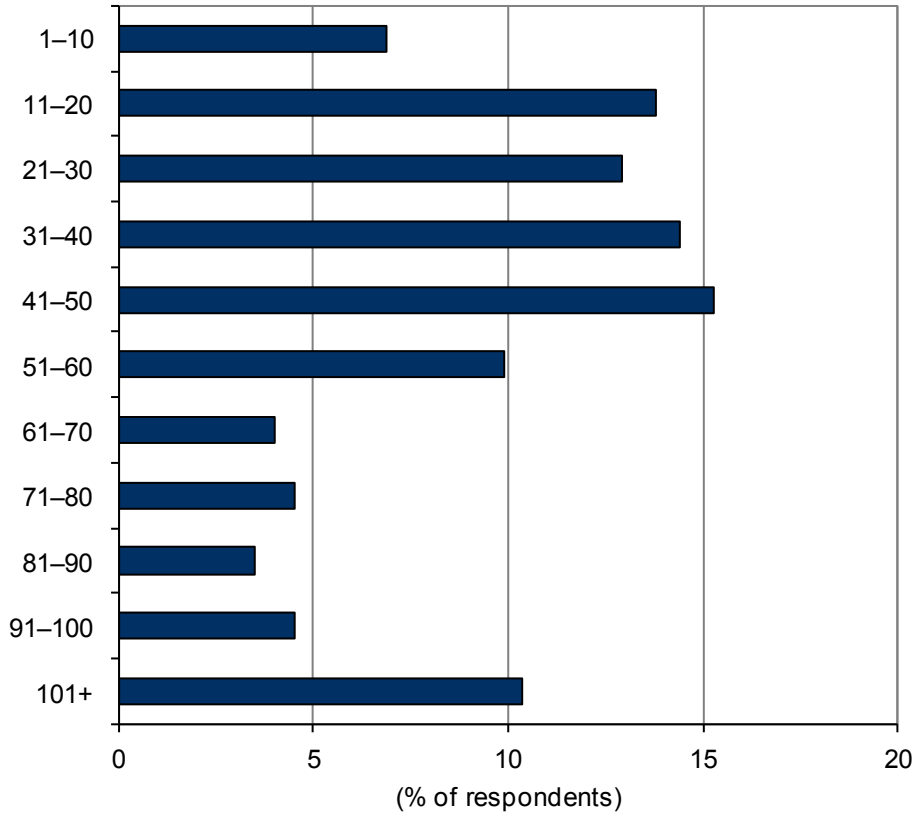
Common Hadoop Configurations in the Enterprise

Figure 12 illustrates the size of Hadoop clusters in the enterprise. Most respondents indicated that their Hadoop clusters are somewhere between 20 nodes and 60 nodes in size (with a growing number indicating that they could go beyond 100 nodes). Many of the respondents indicated that their businesses had more than 1 Hadoop cluster in their environment.

FIGURE 12

Number of Hadoop Nodes Available

Q. How many Hadoop nodes do you have in all clusters in place today (or are you planning to deploy)?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

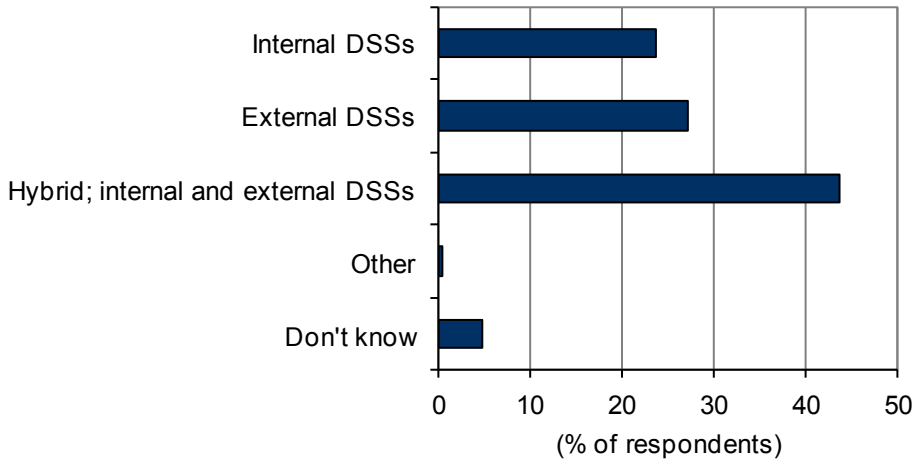
Persistent Storage for Hadoop

Figure 13 indicates the storage configuration that most businesses select for their Hadoop clusters. The majority of respondents indicated that they still use either external disk systems or a combination of internal and external disk systems as persistent storage targets. Businesses strictly using internal disk systems are still in the minority.

FIGURE 13

Hadoop Storage Infrastructure

Q. Based on the definition of centralized versus distributed storage architectures, what type of architecture are you considering or have you already deployed (for Hadoop)?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Data is not weighted.

Use caution when interpreting small sample sizes.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

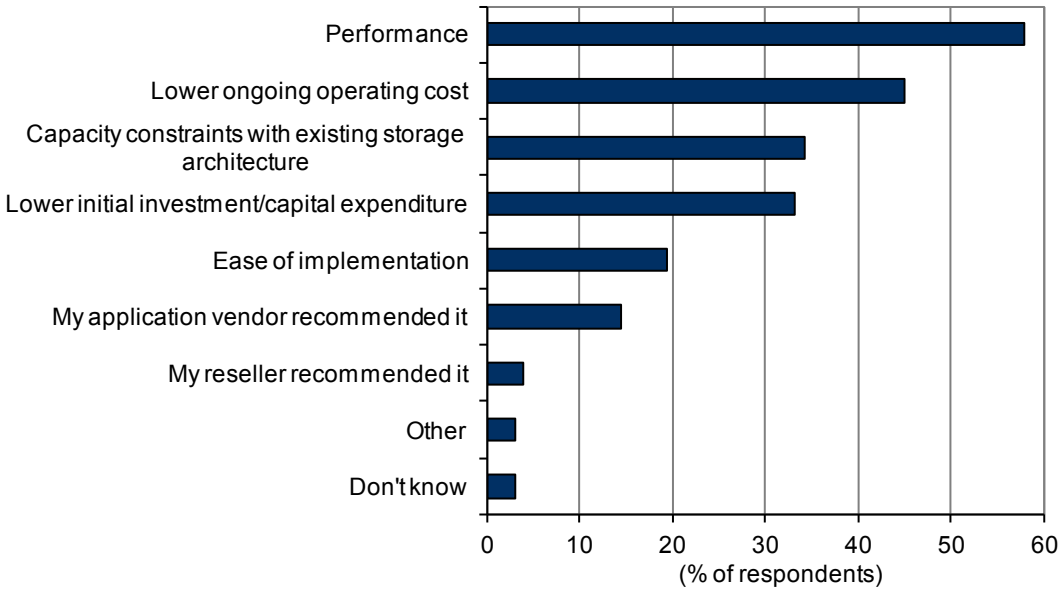
Figure 14 illustrates some of the reasons for selecting the said configuration. Performance remains a key requirement, followed by capex and opex costs. Businesses that opted for internal or hybrid platforms did so because of capacity constraints with existing storage architecture but may not have necessarily bought into the robustness of the solution they chose.

 ©2013 IDC

FIGURE 14

Reasons for Hadoop Storage Choice

Q. What were the primary drivers of selecting this architecture?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

The Hadoop platform is made up of two components — a distributed compute engine and a persistent data store that leverages a shared nothing node-based architecture. Both components have been architected to use commodity "off the shelf" components. Specifically for storage, this means that all of the persistence requirements can be met by stuffing the node (which is a server with computing power) with internal disks. This is the same design philosophy that is being adopted by suppliers when they design scale-out file-based and/or object-based platforms.

IDC believes that in almost all scenarios and for most workloads, this commodity-centric node-based design performs as well as or even better than traditional shared everything architectures that most enterprises are used to. However, for many enterprises that still deploy all storage via Fibre Channel SANs, this is a radical shift — something that cannot meet their requirements for an enterprise-grade application. Such businesses still use external disk systems that are attached via Fibre Channel or InfiniBand or use any of the Ethernet/IP-based file and block protocols.

IDC feels that businesses are in the midst of a transformation and that ultimately they will be forced to adopt node-based designs for distributed computing applications like Hadoop. Until then, they will leverage scale-up solutions for scale-out applications — more for peace of mind than for any tangible benefits.

Hadoop Enterprise Deployment Patterns — Data Management

This section deals with some of the Hadoop data management patterns observed by IDC via the survey.

Data management inside the Hadoop environment is a key challenge that many businesses have to deal with. The data management challenge includes issues such as data migration, data growth, managing data sources, compliance, and data protection. Collectively, the issues illustrate that deploying Hadoop in the enterprise is no different from deploying a complex, commercial multitiered application stack.

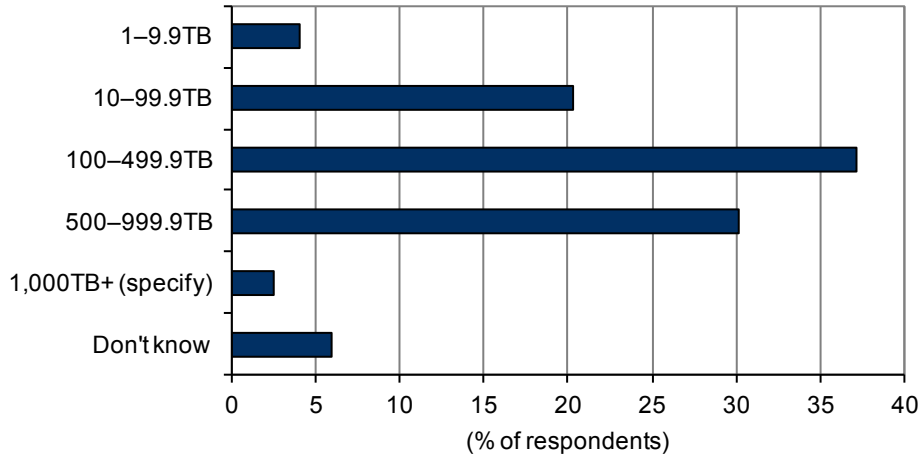
Terabytes Under Management

Figure 15 illustrates the distribution of total managed terabytes in enterprise Hadoop environments. The majority of respondents indicated that their Hadoop configurations exceeded 100TB, which would qualify these configurations as Big Data environments, according to IDC's definition. When the data in Figure 15 is combined with the data in Figure 16, which shows storage infrastructure growth distribution to be around 30% on average, it is conceivable that most Hadoop enterprise deployments will qualify as Big Data configurations from a "terabytes managed" perspective. Because of the nature of data stored in (and, more importantly, purged from) Hadoop, it is rare to find petabyte-scale Hadoop deployments, although arguably, those will eventually be commonplace as more businesses use Hadoop as a persistent storage platform for postprocessed (or postanalyzed) data.

FIGURE 15

Size of Storage Infrastructure Used for Hadoop

Q. How many terabytes are being/will be managed in the disk and SSD (i.e., primary) storage component of your Hadoop infrastructure?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

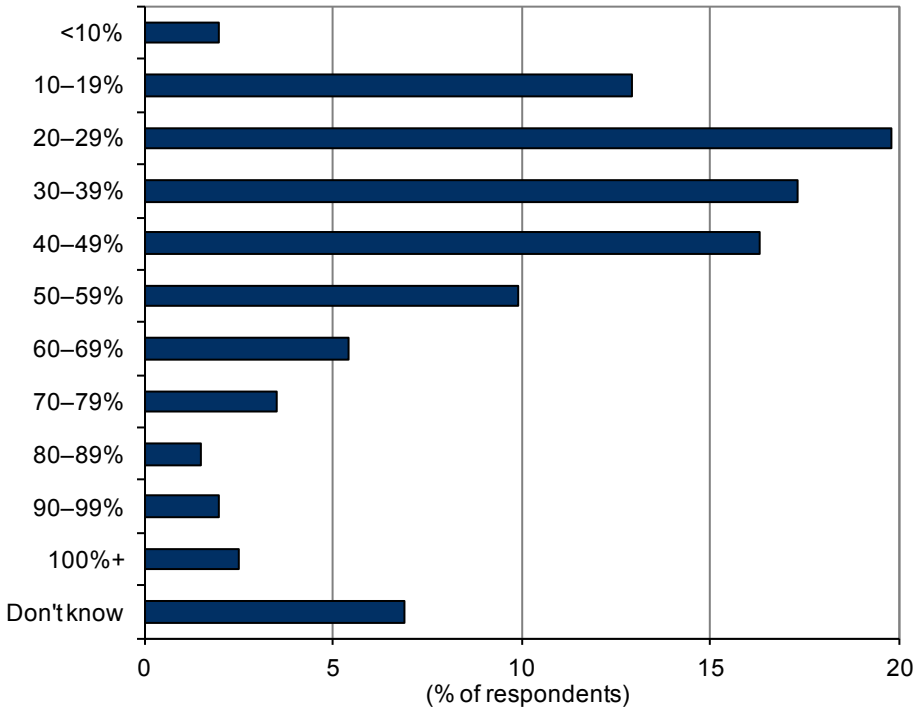
Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

FIGURE 16

Hadoop-Related Storage Infrastructure Growth

Q. What kind of storage growth are you seeing in your Hadoop environment?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

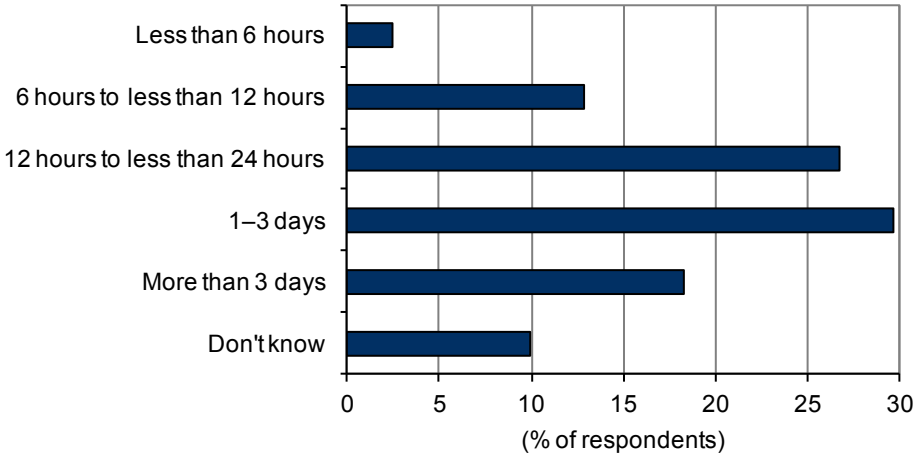
Data Ingestion into Hadoop

Figure 17 lays the groundwork for this section. Specifically, it illustrates that for most businesses, the initial data migration/conversion process is quite a chore — the majority of respondents stated that it took their businesses anywhere from one to three days or even more in many cases to perform the initial data ingestion. That time frame is staggering and lays bare the challenges with data formats, data currency, and maintaining coherency and consistency with external data sources.

FIGURE 17

Initial Data Migration Time

Q. How long did it take you to do the initial migration of data into the Hadoop cluster?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

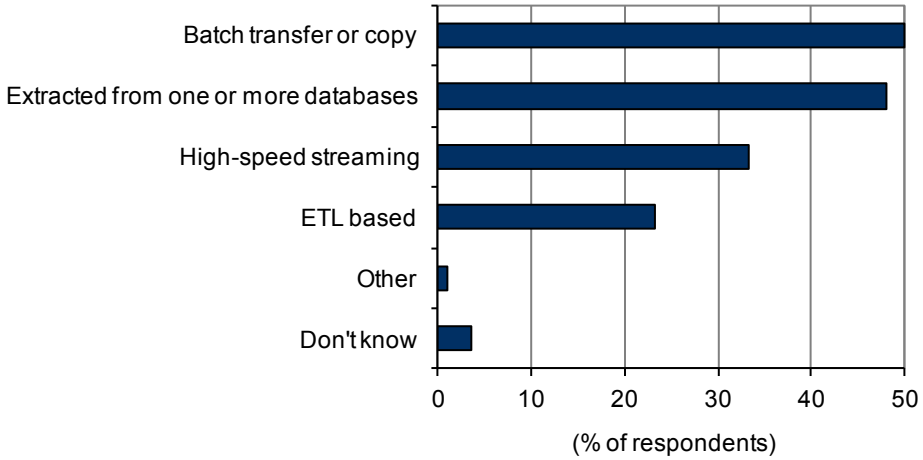
 ©2013 IDC

Figure 18 expands upon the challenges by illustrating the diverse data ingestion mechanisms used to ingest data into Hadoop. These range from high-speed streaming to ETL loads. Traditional database extraction methods, including batch transfers and copies, are also widely used. Whenever the data ingestion process is a one-time process or uses protocols that cannot maintain currency by default, the business has to rely on an additional layer of scripting or currency application to overcome that limitation.

FIGURE 18

Hadoop Data Capture Approach

Q. How is the data moved into or captured by the Hadoop cluster in an ongoing manner?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

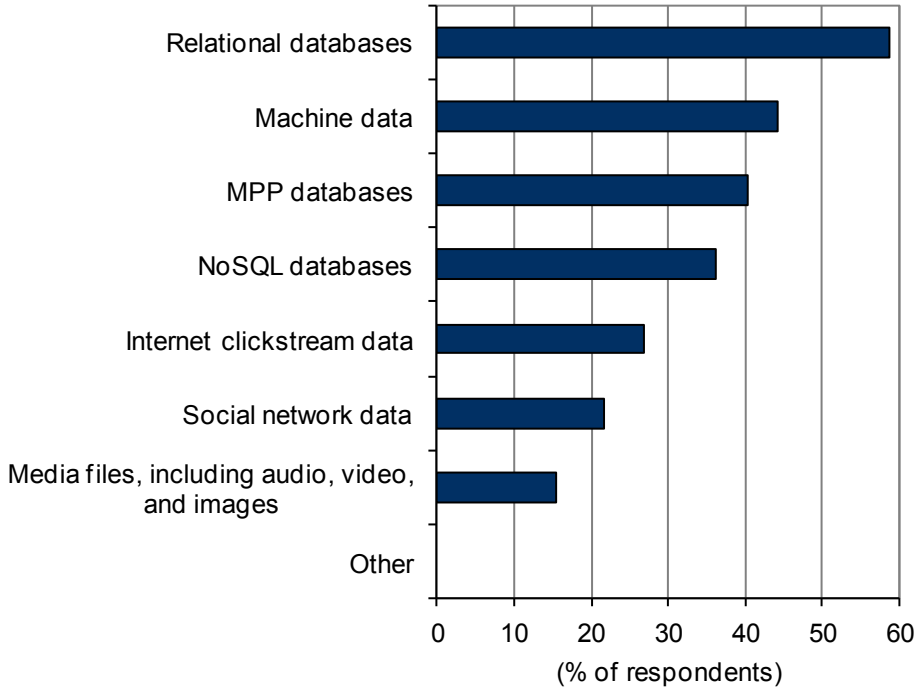
Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

Figure 19 illustrates the diverse sources used to ingest data into Hadoop. Relational databases still constitute the bulk of the data sources, followed by machine data — which includes point-of-sale devices, sensors, and datacenter appliances as well as any equipment that is capable of logging telemetry data.

FIGURE 19

Data Sources Feeding into Hadoop

Q. Which data source(s) are the data sets extracted from feeding into Hadoop?



n = 97

Base = respondents who extracted data from one or more databases

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

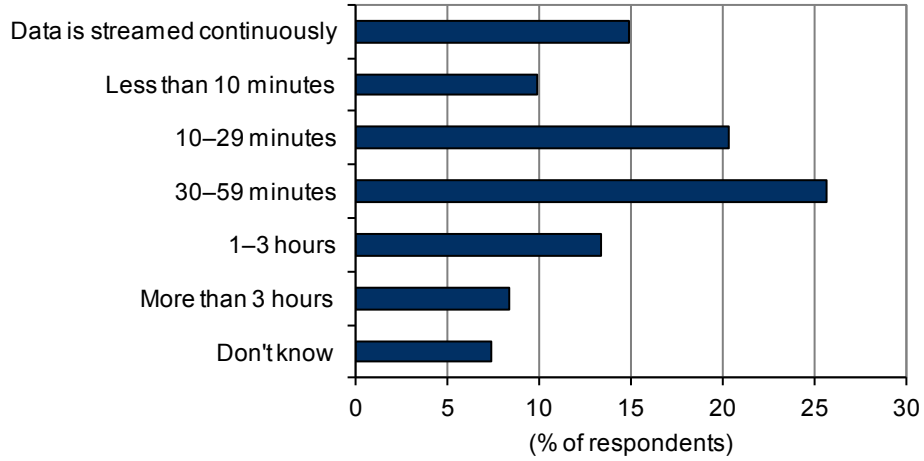
 ©2013 IDC

Figure 20 illustrates typical times for data imports into Hadoop. The majority of respondents said that it takes anywhere from 10 minutes to 3 hours. An increasing number of respondents said that data is streamed continuously. IDC believes that as the "shelf life" of data decreases, businesses will have no choice but to stream data continuously into Hadoop so that the data can be analyzed and correlated or its causal values derived before the data sets expire.

FIGURE 20

Time Taken to Import Data into Hadoop

Q. How long does it typically take to import data into the Hadoop cluster?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

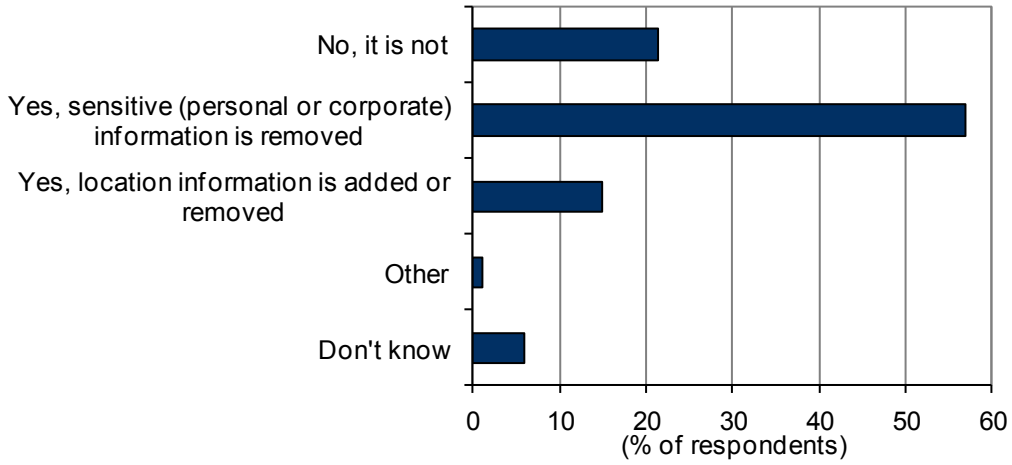
Data Security and Protection

Figures 21 and 22 illustrate how businesses handle data security during ingest and export. An overwhelming majority of respondents said that their data is modified in some fashion before it is imported into Hadoop. The same holds for data when it is exported from Hadoop. Examples of such modifications include scrubbing of data to remove personal or sensitive information, addition or removal of location information, and addition or removal of any traceable tags. Businesses that are required to certify compliance or fall under regulatory requirements are especially sensitive to security preprocessing. Hadoop is not yet a certified compliant application in many industries like healthcare (HIPAA) in its default form, and therefore, businesses in such industries have to mandatorily treat it like an "external" application.

FIGURE 21

Data Scrubbing/Processing Prior to Import into Hadoop

Q. Is data scrubbed or processed before it is imported into Hadoop?



n = 202

Base = all respondents

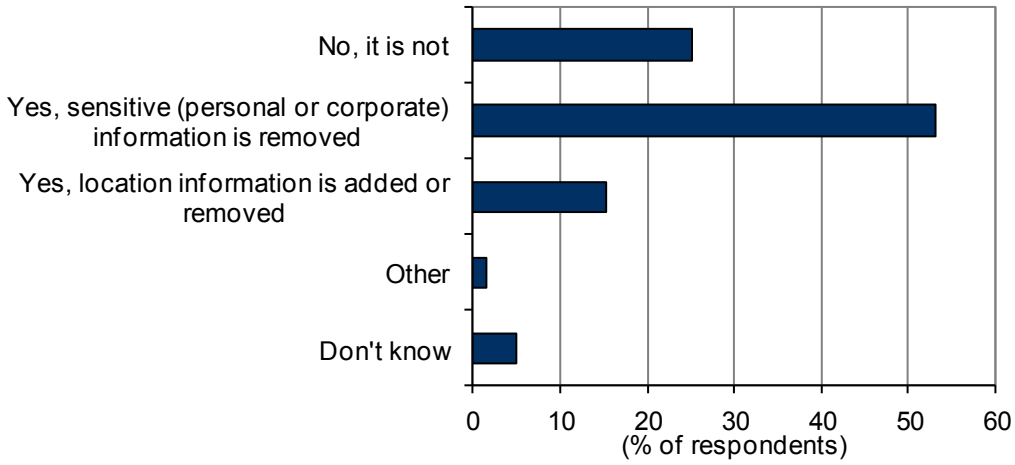
Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

FIGURE 22

Data Scrubbing/Processing After Export from Hadoop

Q. Is data scrubbed or processed after it is exported from Hadoop?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

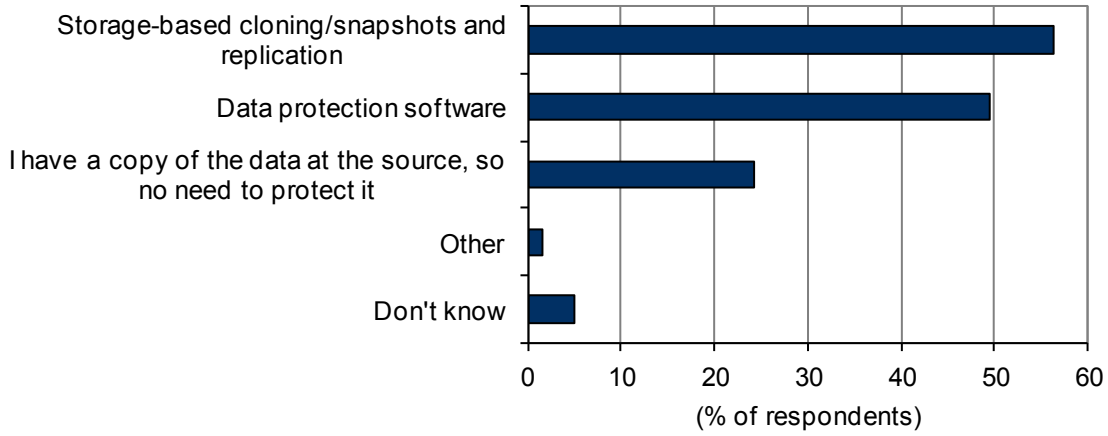
 ©2013 IDC

Figure 23 illustrates how businesses protect data inside Hadoop. Such protection mechanisms are in addition to the replication capabilities of HDFS when it is used as the persistent storage layer. The majority of businesses that use hybrid or external disk systems treat Hadoop like any other application and therefore use array-based snapshots and cloning techniques. A good number of businesses also rely on data protection software in situations where the data inside Hadoop does not change that often and therefore the Hadoop cluster can be periodically backed up, like any other application. Needless to say, a significant number of businesses do not back up their Hadoop cluster at all because they have a copy of the data elsewhere.

FIGURE 23

Data Protection Policies Before Processing

Q. What kinds of data protection policies are in place in your Hadoop infrastructure before data is processed or analyzed?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

Storage Media for Pre- and Postprocessed Data

Figures 24 and 25, respectively, illustrate the storage/application medium in which data is stored before and after it is processed or analyzed in Hadoop. Hadoop is an analytics platform that can store and analyze structured and unstructured data. In fact, several businesses leverage Hadoop to correlate or derive causality between disparate structured and unstructured data sets. Depending on the type of persistent storage mechanism used, it is easy to estimate the type of data sets used:

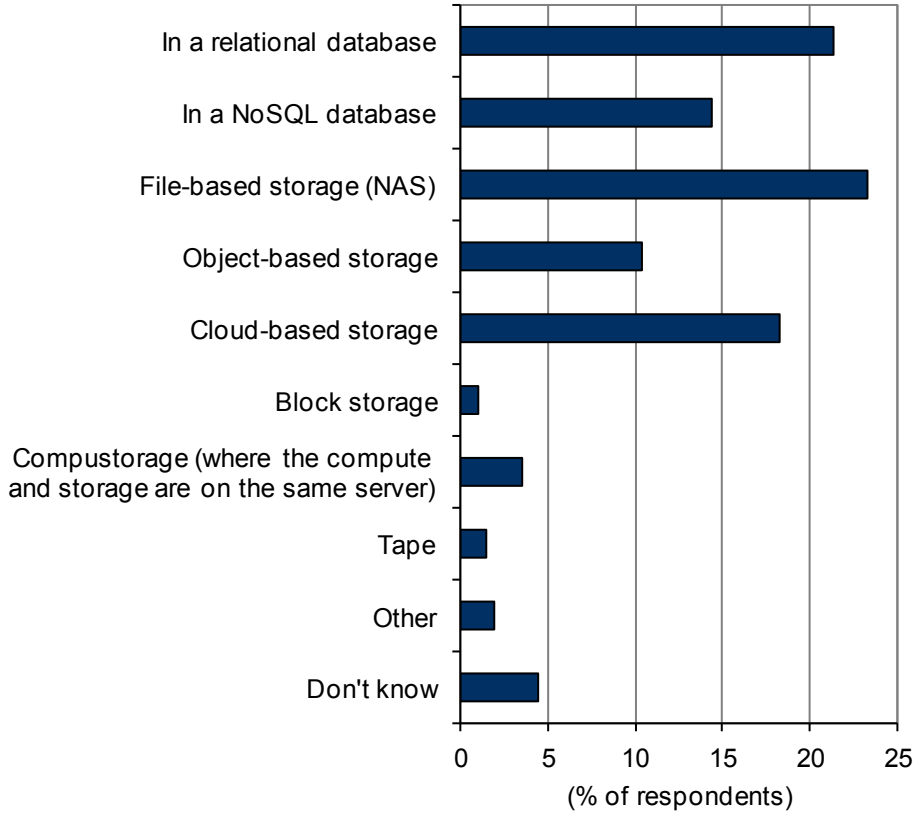
- If a relational database is used, it means that the data set is preformatted to match the database schema(s).

- In the case of NoSQL databases, the data sets are more than likely sourced from Web 2.0 applications since NoSQL databases lack fixed schemas and are application programmable.
- In the case of file-based (NAS), object-based, or cloud-based storage, the data set is more than likely unstructured data gathered from human interactions, sensors, or machines.
- Many businesses choose to store pre- and postprocessed data on tape. This is primarily for cost reasons and in situations where the data import or export is not time critical.

FIGURE 24

Data Storage Medium Before Processing

Q. On what storage medium is data stored before it is processed or analyzed?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

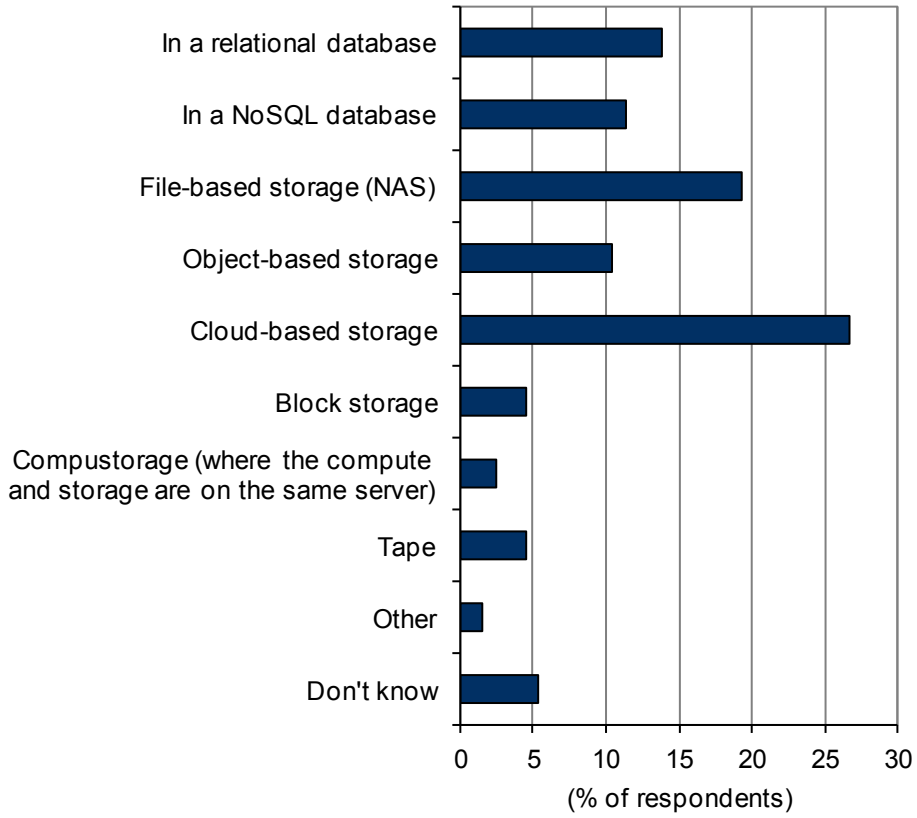
Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

FIGURE 25

Data Storage Medium After Processing

Q. On what medium is data stored after it is processed?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

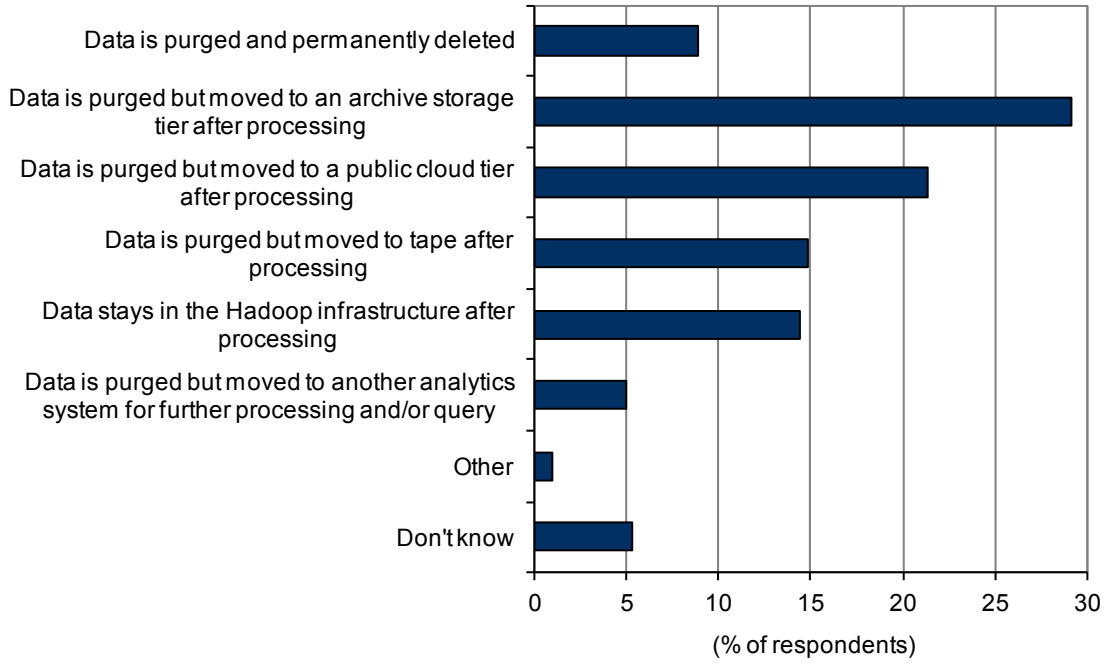
Data Retention Policies

Figures 26–29 illustrate the diversity in data retention policies for enterprise Hadoop deployments. Most respondents indicated that there are some data retention policies for data before and after it is processed. In other words, very few businesses permanently delete pre- and postprocessed data — most of them store data from 6 months to 3 years. As far as the storage media are concerned, in most cases, this data is moved to a lower-cost tier, which in some cases also includes tape.

FIGURE 26

Data Retention Policies for Raw Data

Q. What kinds of data retention policies are in place in your Hadoop infrastructure for the raw data after it is processed or analyzed?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

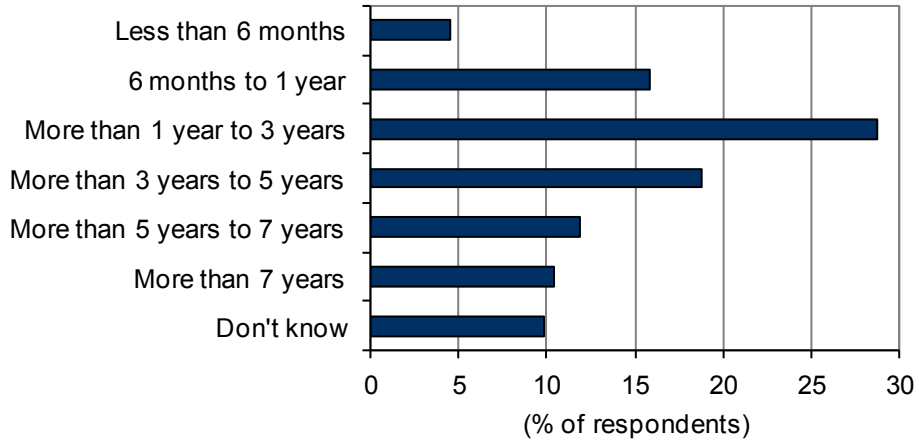
Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

FIGURE 27

Raw Data Retention Time

Q. How long is the raw data retained after it has been processed/analyzed?



n = 202

Base = all respondents

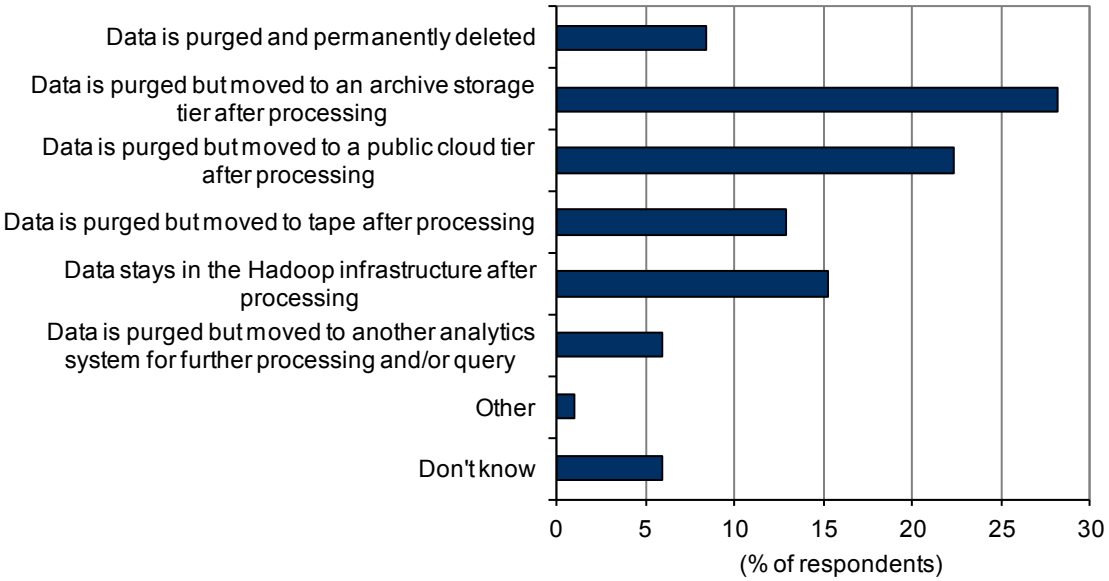
Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

FIGURE 28

Data Retention/Archival Policies After Processing

Q. What kinds of data retention or archival policies are in place in your Hadoop infrastructure for the resulting data set after the raw data is processed or analyzed?



n = 202

Base = all respondents

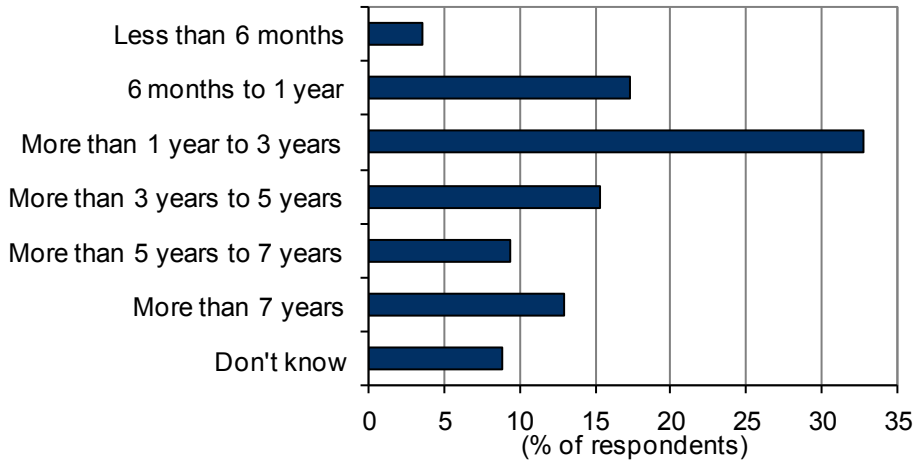
Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

FIGURE 29

Processed Data Retention Time

Q. How long is the resulting data retained after the raw data set has been processed/analyzed?



n = 202

Base = all respondents

Note: This survey is managed by IDC's Quantitative Research Group.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

CHALLENGES AND OPPORTUNITIES

The results of this survey indicate that in spite of the overall success of Hadoop, businesses still face some real challenges when it comes to deploying and managing Hadoop.

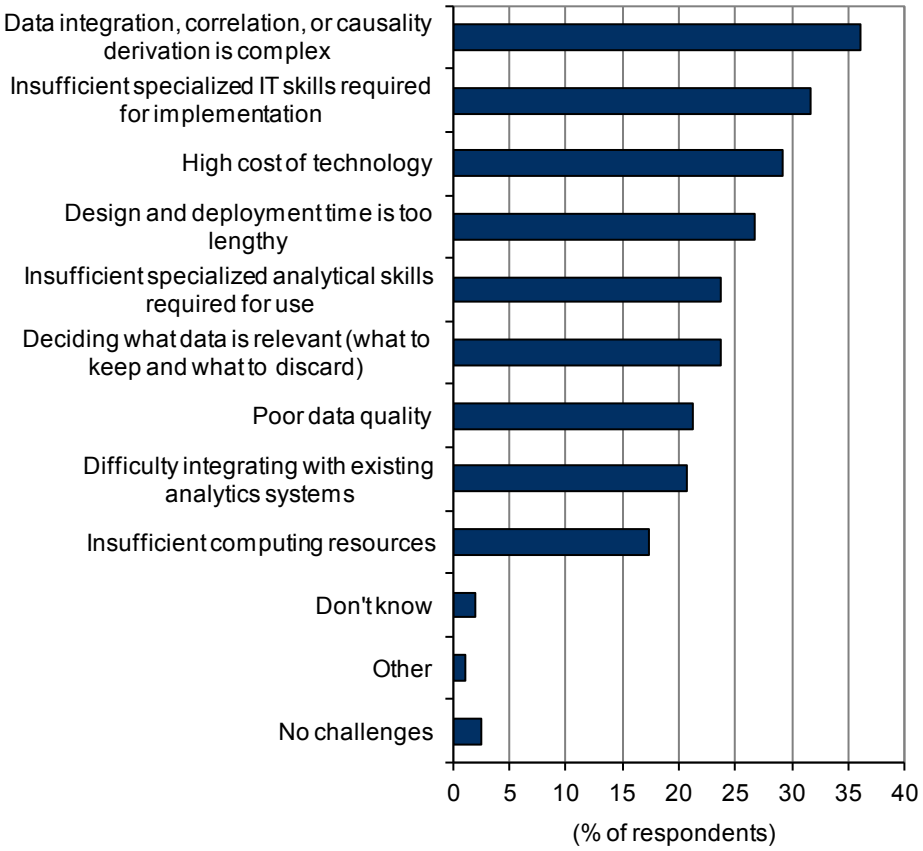
Figure 30 illustrates that from an application/software perspective, the biggest challenge has to do with data integration, correlation, or causality derivation. In second place and third place, respectively, are lack of IT skills and the high cost of technology. From a skills shortage perspective, there are two emerging groups of specialists: individual data scientists and firms that specialize in data science, Big Data infrastructure, and applications as well as their integration.

 ©2013 IDC

FIGURE 30

Hadoop Challenges: Application/Software Component

Q. Thinking of the application/software component of your Hadoop and analytics environment, which of the following best describes the challenges you faced or are facing with your infrastructure?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

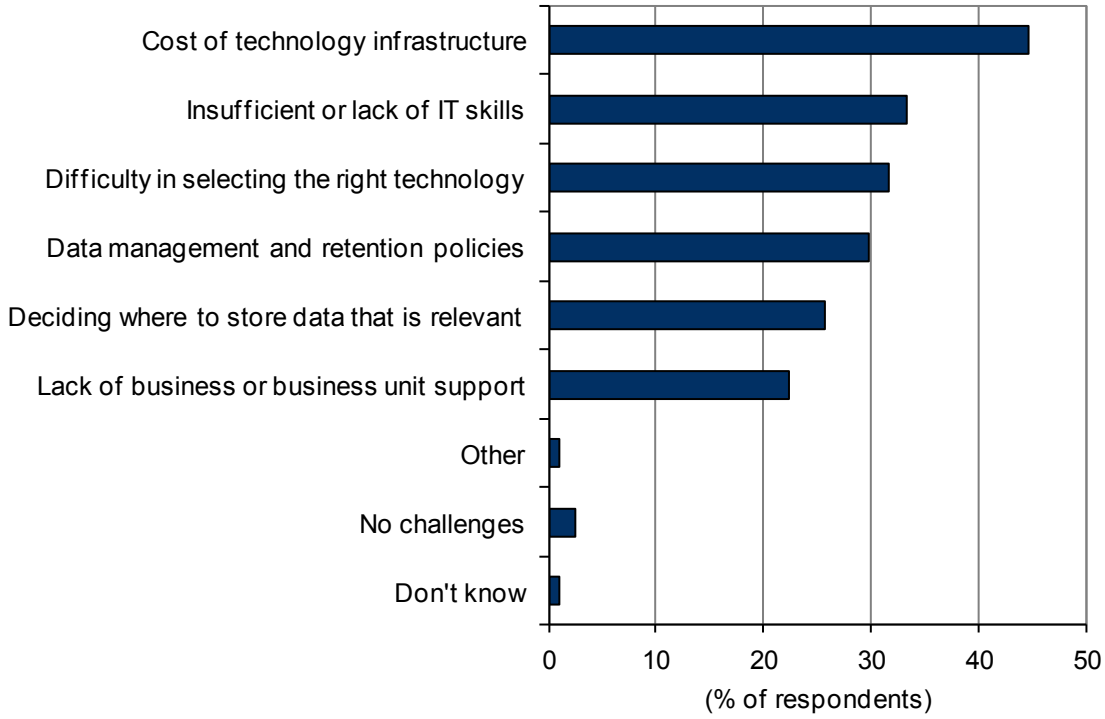
Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

Figure 31 illustrates that from an infrastructure perspective, the picture is not that different. Businesses complain about the high cost of technology and the lack of IT skills when it comes to Big Data (Hadoop) infrastructure. The lack of insight into what to store and what to purge means that by default businesses store data regardless of its value — thereby adding to the overall cost of the infrastructure.

FIGURE 31

Hadoop Challenges: Infrastructure/Hardware Component

Q. Thinking of the infrastructure/hardware component of your Hadoop implementation environment, which of the following best describes the challenges you faced or continue to face with your infrastructure?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

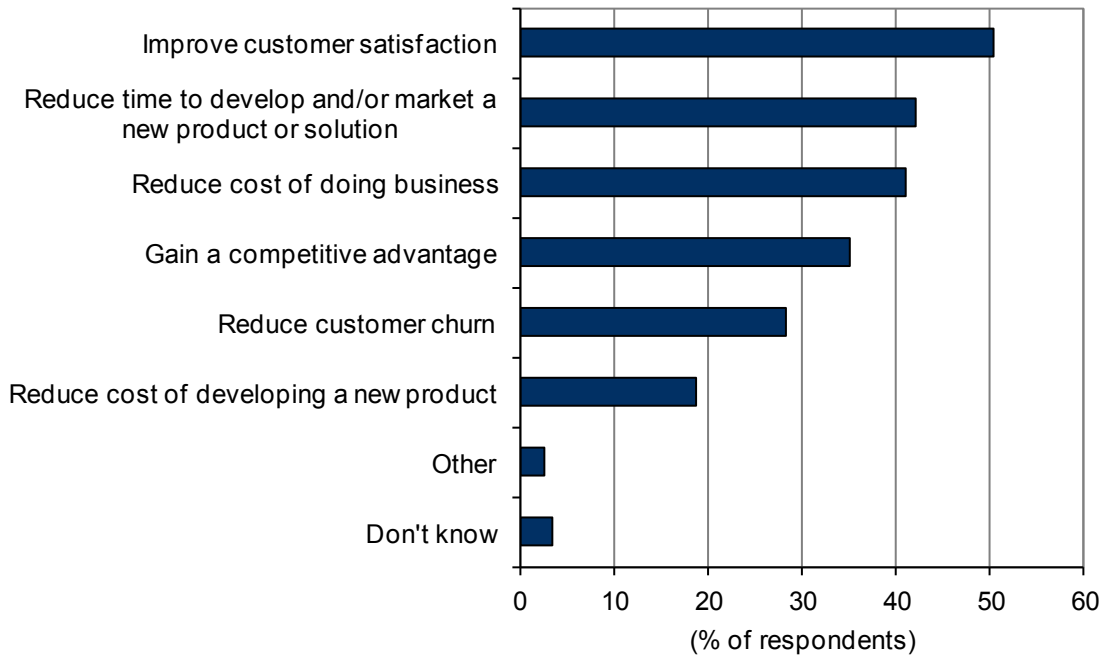
FUTURE OUTLOOK

Figure 32 rounds out the key findings of the survey by illustrating the main business challenges that businesses have solved or intend to solve by deploying Hadoop. Almost 51% of respondents cited the ability to improve customer satisfaction as the reason for deploying Hadoop. This was followed by 42.1% of respondents who cited reduced time to develop and/or market a new product or solution and 41.1% who cited reduced cost of doing business. Slightly more than 35% of respondents listed the ability to gain a competitive advantage as the reason for deploying Hadoop, followed by 28.2% who listed reduced customer churn. Finally, nearly 19% of respondents cited reduced cost of developing a new product.

FIGURE 32

Business Challenges Solved by Hadoop

Q. What business challenges have you solved or do you intend to solve with this Hadoop deployment?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Multiple responses were accepted, so total will not sum to 100%.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

THE RED HAT HADOOP SOLUTION

A key finding of this survey is that many businesses, especially in the enterprise, heavily leverage external storage and data management solutions to enhance the functionality of Hadoop. Most of the functional enhancements are in the areas of data management, resiliency, and performance. Businesses tend to put their faith in storage solutions from platform suppliers like IBM, Red Hat, and EMC as persistent storage media for Hadoop.

The use of a general-purpose platform also supports a broad set of use cases beyond just structured data. Mixed workflows are best served by a general-purpose platform (like Red Hat Storage Server and IBM GPFS) rather than specialized platforms (like HDFS) that lack robust support for mixed workflows.

Since Hadoop is a Linux-based software solution that is designed to leverage commodity hardware, it tends to have a natural affinity toward software-based storage offerings that make use of general-purpose distributed file systems. It is not surprising, therefore, to see IBM GPFS, Red Hat GlusterFS, and EMC Isilon OneFS listed as the top 3 file system choices in Figure 10.

Of these three systems, only Red Hat offers an integrated open source–based enterprise Linux platform that combines a distributed file system with a Hadoop connector, enterprise middleware, and the ability to run Hadoop computational workloads natively. Except for solutions from a few start-ups, in almost all cases, the storage platform remains separate from the computational platform, which goes against the core principle of the Hadoop platform.

Some of the key benefits of the Red Hat Hadoop solution are:

- Reduced management costs by leveraging the Red Hat Enterprise Linux and Red Hat Storage Server frameworks
- A high-performance file system to augment HDFS but maintain out-of-box HDFS API compatibility:
  - Built-in data protection and resiliency
  - Elimination of name node bottlenecks and a single point of failure
  - Built-in disaster recovery with georeplication (Red Hat Storage Server)
  - POSIX compliance for simple data ingest/export
  - The ability to maintain data locality as the cluster scales, thereby reducing network chatter and improving cluster efficiency
  - Red Hat Enterprise Linux that is qualified to run add-on analytics tools like Pig, Hive, Mahout, Avro, and Lucene
  - Red Hat JBoss Middleware, which provides the necessary middleware technologies to combine structured and unstructured data, high-velocity and historical data, and federated data access to reduce the information gap by

cost effectively making all data available for analytics (The key foundational technologies to enable this objective are JBoss Data Virtualization to handle data at rest, including Hadoop, and JBoss A-MQ messaging to handle data in motion in a reliable, scalable, and secure way.)

- The ability to run Hadoop on OpenStack using Red Hat Storage and Red Hat Enterprise Linux OpenStack Platform with Savanna

CONCLUSION

Big Data and analytics will continue to influence and morph businesses everywhere. Many of these changes will continue to disrupt how business is conducted. Only a few of these changes will be evolutionary in nature. Regardless of the type of change, the one thing that all these changes will have in common is that they will leave an indelible mark on how businesses leverage data — businesses will become data driven to the point where they adopt a hawkish, discovery-oriented mindset. This will have a profound impact on how infrastructure is deployed and consumed:

- Traditional environments will eventually make way for next-gen analytics environments that are more suited to a service-based infrastructure with mixed workflows.
- Infrastructure will move from a capex-heavy model to an opex-heavy model. Businesses will leverage the cloud not just for preprocessed and postprocessed data but also for on-demand computing.
- Big Data and analytics will have to leverage the shared IT infrastructure to maximize ROI. IT will rely on general-purpose platforms to build this infrastructure to gain economies of scale.

APPENDIX

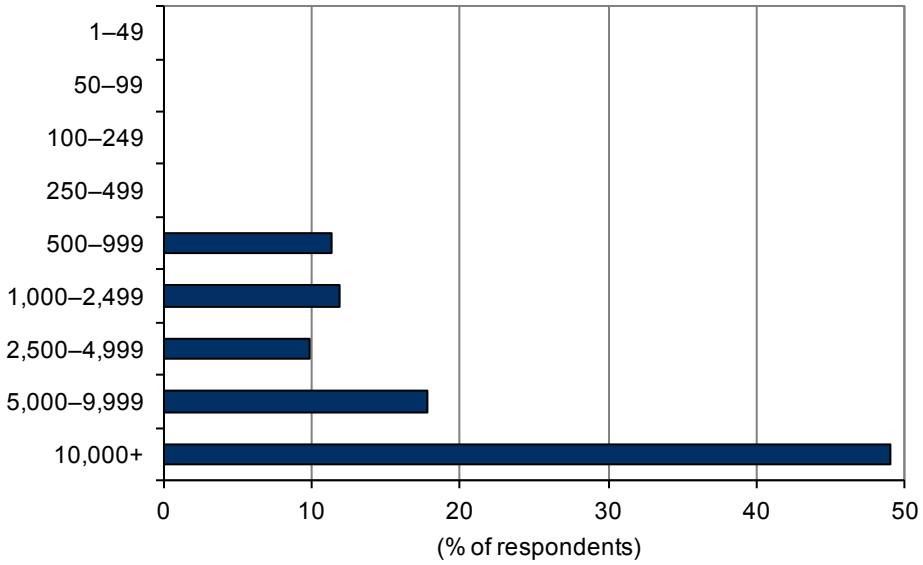
Screening Questions

Figures 33–36 list the screening questions that were posed to respondents before they took the survey. The goal of these questions was to ensure that the results were accurate and consistent.

FIGURE 33

Respondents by Number of Full-Time Employees

Q. Approximately, how many full-time employees are employed by your organization (including all branches, divisions, and subsidiaries)?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Data is not weighted.

Use caution when interpreting small sample sizes.

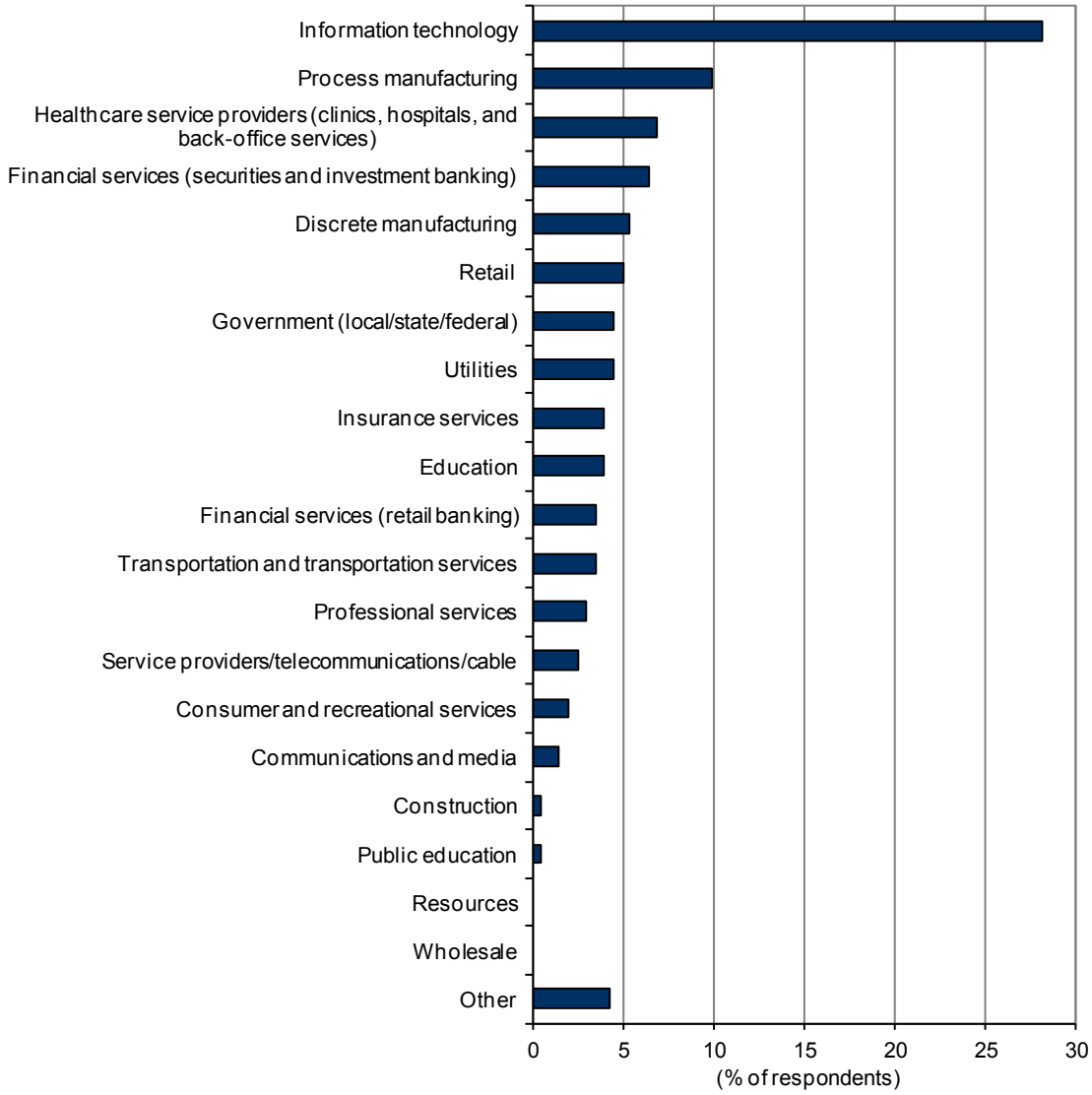
Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

FIGURE 34

Respondents by Primary Business Activity

Q. Which of the following best describes your organization's primary business activity?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Data is not weighted.

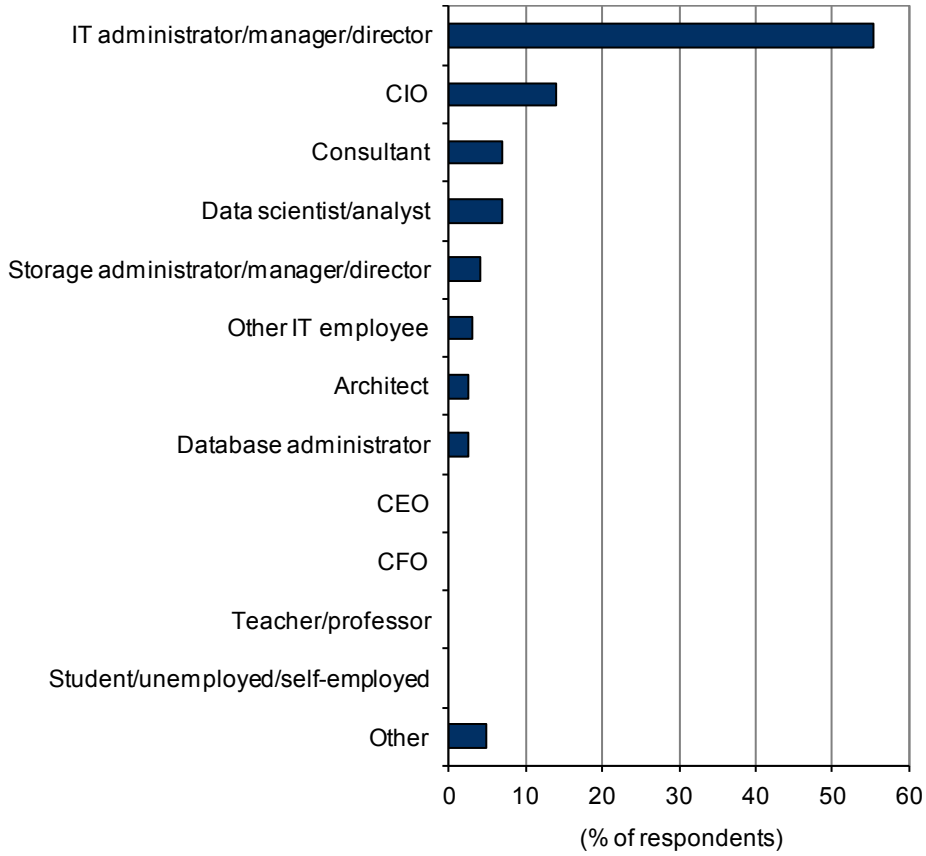
Use caution when interpreting small sample sizes.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

FIGURE 35

Respondents by Job Title

Q. Which of the following best describes your title?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Data is not weighted.

Use caution when interpreting small sample sizes.

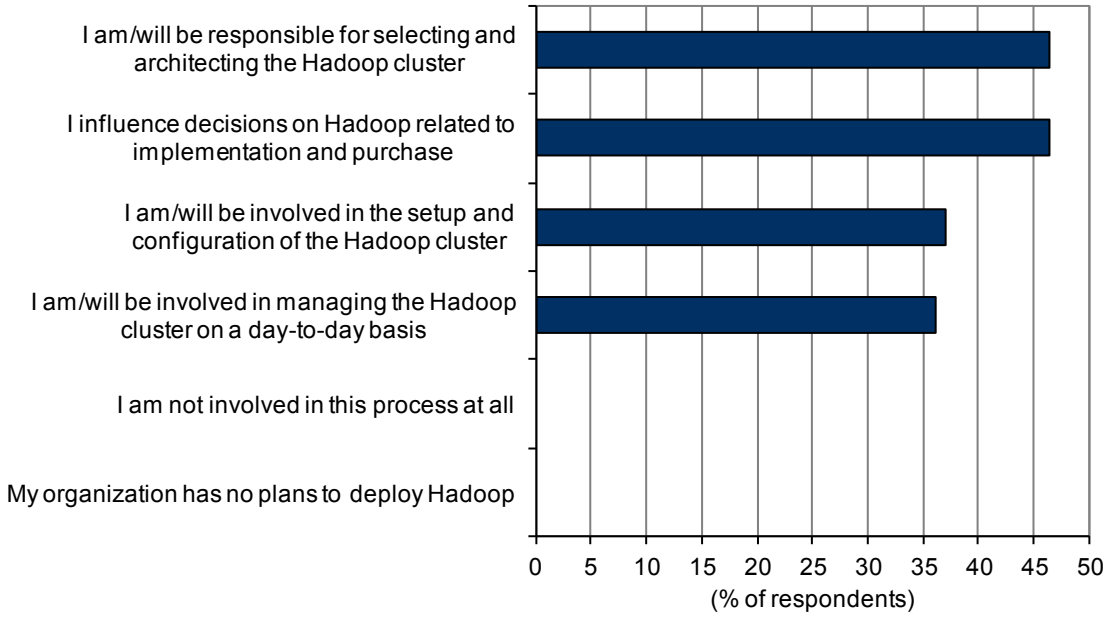
Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

 ©2013 IDC

FIGURE 36

Respondents by Role in Hadoop

Q. Which of the following best describes your role in your organization's current or planned implementation of Hadoop?



n = 202

Base = all respondents

Notes:

This survey is managed by IDC's Quantitative Research Group.

Data is not weighted.

Multiple responses were accepted, so total will not sum to 100%.

Use caution when interpreting small sample sizes.

Source: IDC's *Red Hat Hadoop Usage Survey,* August 2013

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2013 IDC. Reproduction without written permission is completely forbidden.