White Paper

Data Analytics Infrastructure and the Essential Data Lake: A Global Study

Sponsored by: Dell EMC

Ashish Nadkarni
February 2017

IDC OPINION

Digital transformation (DX) – a technology-based business strategy – is essential for businesses to embrace today to be prepared to thrive in a digital economy. Organizations that transform themselves digitally are the ones that efficiently harness information – derived from analyzing large and diverse sources of data – to maintain or increase their competitive differentiation. Efficient and timely harnessing of information depends on how they combine technology, people, and intellectual property. From a technology perspective, next-generation applications (NGAs) and infrastructure solutions – collectively referred to as IDC's 3rd Platform technologies (see Appendix for definition) – support organizational goals for information-backed agility. Specifically, by embracing cloud-native applications, big data applications, and business intelligence applications (collectively referred to as data analytics applications), businesses can:

- Create greater operational efficiencies
- Improve the customer experience
- Build deeper relationships with their partners
- Accelerate time to market for new products and services

IDC believes that by 2018, one-third of industry company leaders will be disrupted by 3rd Platform-based competitors. The time to act is now. A data analytics strategy is therefore not only critical to an organization but also essential and urgent. For businesses to fully benefit from implementing a capable analytics ecosystem, they need an equally capable infrastructure to support it. This infrastructure needs to be agile, scalable, and secure.

Most 3rd Platform applications are designed with a core "analytics first" design principle. An "analytics first" – implemented as a consolidated *data analytics environment* approach – makes it easier to provide a unique user experience as a part of the application workflow. Businesses are beginning to realize the benefits of designing their data analytics environments with *data lakes* as an infrastructure layer for consolidating, aggregating, and analyzing multiple and disparate data sets. In recently conducted research on how businesses are transforming themselves, IDC found what infrastructure attributes are crucial to the functioning of a data analytics ecosystem.

METHODOLOGY

In 2016, Dell EMC commissioned IDC's Infrastructure Research Team to conduct a web-based, global survey of 1,105 IT professionals at businesses that have evaluated and deployed or are in the process

of deploying data analytics infrastructure in their environment. The goal of the survey was to better understand the profile of infrastructure on which data analytics environments are deployed and specifically examine interest in adopting data lakes for analytics. Survey data figures can be found in the Appendix at the end of this white paper.

SITUATION OVERVIEW

Digital transformation – a technology-based business strategy – is essential for businesses to embrace today to be prepared to thrive in a digital economy. Organizations that transform themselves digitally are the ones that efficiently harness information – derived from analyzing large and diverse sources of data – to maintain or increase their competitive differentiation.

Businesses are engaging in analytics-based digital transformation initiatives (see Appendix for a detailed definition of DX) to create greater operational efficiencies, improve the customer experience, build deeper relationships with their partners, and accelerate development and time to market for new products and services (to name a few). Any business that isn't actively planning on a digital transformation of its business and operations model risks being left behind (refer to Figure 1; see Appendix for all figures in this white paper). Many of these initiatives impact multiple business units, making such initiatives strategic to the firm's future (refer to Figure 2).

Next-Generation Applications Power Digital Transformation

From a technology perspective, NGAs – which include mobile, cloud, data analytics, and social applications and platforms (collectively referred to as IDC's 3rd Platform technologies) – support organizational goals for information-backed agility, which is crucial for DX.

Much of this effort is unique to the business, and therefore, the implementation is highly customized. Many NGAs are commercially procured; however, a growing number of these are open source based – commercial variants or freely downloaded and customized variants (refer to Figure 3). The selection criteria for commercially procured and open source–based applications are largely similar. These include security, ease of management, backup and disaster recovery, ease of operations, and cost (refer to Figures 4 and 5).

NGAs are designed to harness and act on information that is derived from analyzing large and diverse sources of data to maintain or increase competitive differentiation. This data is "pooled" into a single repository for search and discovery as well as causality and correlation analytics using algorithms designed specifically by data scientists.

NGAs therefore have big data and business intelligence (collectively referred to as data analytics applications) components at the core. Data analytics environments – that power NGAs – pull data from various databases, high-speed streaming, batch transfers, and ETL (extract, transform, and load – three database functions that are combined into one tool to pull data out of one database and place it into another database; refer to Figure 6) tools. Data analytics as a part of the NGA stack makes it easier for businesses to address challenges such as improving customer satisfaction, reducing the cost of doing business, gaining a competitive edge, and mitigating risks to the business (refer to Figure 7). Many of these challenges are top priorities for lines of businesses (LOBs) – and therefore LOB managers are the primary sponsors of the corresponding NGA/data analytics deployments.

Data Analytics Infrastructure: Crucial for NGAs

Businesses define the successful outcome of a business analytics deployment in terms of the tangible and intangible value they derive from it. Infrastructure is a critical cog in the data analytics engine that powers NGAs and therefore successful DX initiatives.

Businesses generally prefer to start with standardized infrastructure components and add custom components for their use case (refer to Figure 8). In some cases, they are compelled to build infrastructure from the ground up. However, the more customized the infrastructure, the more an operations burden on the IT organization — hence the preference to keep it as standard as possible. In many cases, the data analytics infrastructure has higher and more stringent SLAs than the rest of the infrastructure, arguably because of the importance of the platform to the business (refer to Figure 9). In such cases, it is all the more important to keep the infrastructure components as standard as possible so that servicing them is easier. Many businesses prefer a multitenant/shared tenancy (business units and workloads, respectively) model because of cost and economies of scale (refer to Figures 10 and 11), though such models carry the risk of "noisy neighbor" issues, which can manifest themselves in terms of performance and latency issues at the individual-application level.

In the same vein, businesses consider metering and chargeback capabilities as essential capabilities for the infrastructure — primarily because of two reasons (refer to Figure 12). First, they need to measure the value of the deployment, and second, they need to ensure a mechanism or on-ramp to a more opex-heavy model. People resources also count toward this value (refer to Figure 13).

When asked, respondents identified in-place analytics, scale out, security and compliance, and business continuity as the top 4 crucial attributes of the data analytics infrastructure.

A crucial requirement for data analytics deployments is the ability to gain insight from combing through various data sources, many of which have limited shelf life. In-place analytics is an efficient mechanism to gain insights from rapidly perishable data. Naturally, the infrastructure — and specifically the storage infrastructure — needs to support multiple storage and nonstorage access protocols. Figure 14 illustrates that more and more businesses prefer that the storage layer support database (ODBC and JDBC) and object (S3, Swift, and CDMI) protocols in addition to traditional file (including HDFS) protocols. Figure 15 shows the kinds of data protection policies in place in the data analytics infrastructure before it is processed/analyzed.

Cost, performance, and capacity drive the preference for on-premises file storage, off-premises (public) cloud storage, and on-premises object storage media for storing data before it is analyzed. Such data sets could contain raw data sets downloaded from the internet or extracted from external databases and applications (refer to Figures 16 and 17).

After the data is analyzed, many businesses prefer to simply purge data from the analytics platform onto an archive or a lower-cost tier (refer to Figure 18). Such tiers include an archive tier (potentially object storage), public cloud storage, or even tape. Some businesses choose to leave the data in place, potentially because of regulatory and compliance requirements. For most businesses, storing the data for one to five years is the norm, while for some, it is longer — even as long as seven-plus years, again given regulatory or compliance requirements (refer to Figure 19).

Businesses cite availability, real-time analytics capabilities, and high data ingest capabilities (speed of ingest) as the top 3 requirements for the storage tier (refer to Figure 20). These are followed by

seamless and on-demand change of capacity and performance capabilities as needed. Metering came in as the least important.

Data Lakes: Building Blocks for Data Analytics Infrastructure

Data lakes can be considered to be a foundational component of a data analytics environment. Data lake platforms incorporate key infrastructure requirements highlighted previously, chief among them being:

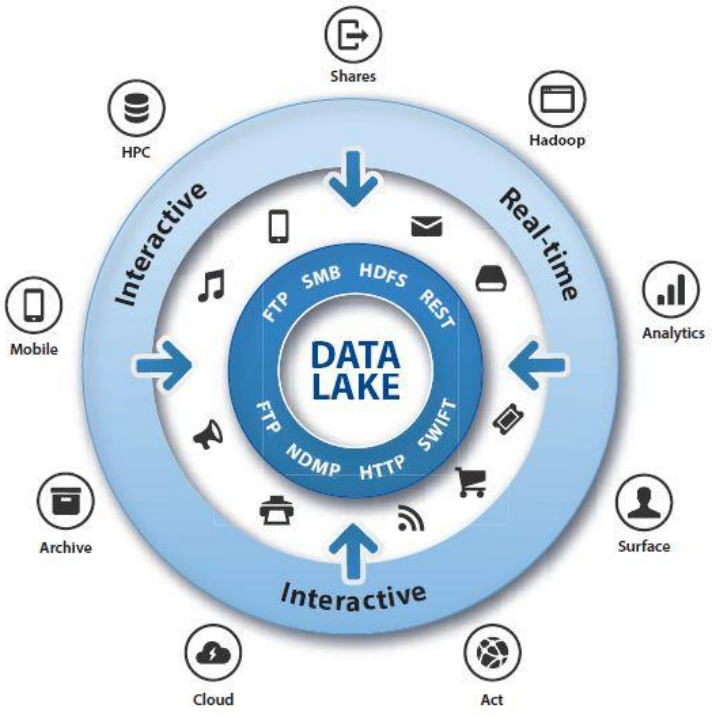
- In-place analytics (multiprotocol access with reduced data movement)
- Scale-out (independent scaling of performance and capacity in a cost-efficient manner) security and compliance
- Business continuity

IDC defines data lakes as big data repositories that are designed to securely store and analyze data with reduced movement from one system to another. A data lake enables disparate data types to be consolidated onto a single, scalable, extensible, and agile repository. (See Appendix for a detailed definition of data lakes.)

Data lakes are procured and operate as storage systems. Data lakes support robust enterprise SLAs for availability, resiliency, protection, and security; are optimized for analytics workloads; and offer out-of-the-box integration with most analytics platforms like Hadoop. A logical representation of a data lake is shown in Figure 21.

FIGURE 21

Illustration of a Data Lake



Source: IDC, 2017

Businesses prefer four main deployment models for data lakes (refer to Figure 22). These include appliance based, cloud based, direct-attached storage based, and the separation of compute and storage. This indicates that data lake platforms have to be software defined and support flexible deployments.

Most businesses have data lakes listed on their immediate deployment road map (refer to Figure 23). For such businesses, the requirements for data lakes are:

- **Technical requirements** include multiprotocol data ingest (i.e., the ability to move data into the data lake via different protocols), built-in data resiliency and protection, and the ability to analyze very large data sets. These are followed by high-speed data ingest, in-place analytics (the ability to analyze data without moving or copying data to another platform), and enterprise-grade security. Storage-specific requirements — which are not unique to data lakes — are built-in fully automated tiering, on-demand scalability, and copy data management (refer to Figure 24).
- **Business requirements** include risk mitigation, the ability to reduce operational costs (including people) and capital expenditure, and tangible infrastructure ROI. These are followed by multi-BU collaboration, workflow management, infrastructure and process efficiency, the ability to improve existing product and service offerings, and new product development (refer to Figure 25).

Data Lake with Dell EMC

The Dell EMC Isilon is ideal for meeting the technical and business requirements listed above. First the multiprotocol ingest of data (a crucial function in data analytics environments) quickly and reliably ingests data into the data lake using protocols closest to the workload generating the data. This makes the platform very friendly for complex big data workflows while providing risk mitigation and a tangible ROI.

A scale-out data lake provides key capabilities to eliminate silos of data; secure and protect information assets; and support existing and next-generation workloads while speeding time to insights. Starting with a scale-out data lake, organizations can:

- Invest in the infrastructure today to get started
- Realize the value of data, store, process, and analyze it in the most cost-effective manner
- Grow capabilities as needs grow in the future

This enables organizations to store everything, analyze anything, and build a solution with the best ROI. By decoupling storage from analysis and application, organizations gain flexibility to choose between a larger number of strategies to deploy solutions without risking data loss, cost overruns, and data set leaks. The data lake offers organizations the capability to simplify the IT infrastructure tier, secure and protect data efficiently, and get to insights faster.

FUTURE OUTLOOK

IDC believes that data lakes will become an integral part of a data analytics infrastructure in the coming years. As businesses master the science of identifying and collating data from various sources and converting it into consumable nuggets of information for their various organizational units, they will no doubt be compelled to establish data lakes — upon which various analytics workloads can

concurrently operate. Such data lakes will enable existing workloads as well as be future-proof to seamlessly support new applications and workloads.

CHALLENGES/OPPORTUNITIES

Data Analytics Infrastructure Challenges

Data lakes hold out a promise to consolidate storage for multiple workloads and applications onto a single shared storage platform, to reduce costs and complexity in their environment, and to make analytics infrastructure efficient, agile, and scalable. However, challenges remain in their adoption — many of which are directly associated with the rollout of a bigger companywide analytics environment. Vendors like Dell EMC, with a reputation in the enterprise, should seek to address them with their data lake solutions.

Businesses can and still do face challenges with their data analytics infrastructure. Key challenges include:

- **Software/application challenges:** High cost of technology, application compatibility, data relevancy and integration challenges, and managing scale and complexity (refer to Figure 26)
- **Hardware/infrastructure challenge:** Cost and scalability of technology infrastructure, data integration, challenges with technology selection, and the lack of or insufficient IT skills (refer to Figure 27)

Many of these challenges lead businesses to shy away from deploying a greenfield data analytics infrastructure. This may be a short-term decision, as the forces of DX are too strong to be ignored (refer to Figure 28).

Furthermore, not all businesses are sold on data lakes yet. Budget is cited as a main reason, followed by a perceived increase in complexity and the need for additional resources to support them. Businesses also (to a lesser extent) fear the overhead of changing workflows to accommodate data lakes (refer to Figure 29).

Opportunity for Dell EMC

The reasons cited previously collectively present solid opportunities for vendors like Dell EMC to double down on their efforts to convince businesses of the simplicity of their solution. IDC conducted in-depth interviews to confirm if businesses were in fact able to realize tangible benefits by deploying a data lake for analytics. IDC recommends that Dell EMC highlight the key capabilities that address many of the challenges listed previously:

- **Application compatibility and integration:** An IT executive surveyed put it this way: "An immediate tangible benefit was speed — speed of data ingest and speed for processing the data in the lake. Connecting the Hadoop environment with the data lake using the HDFS connector was seamless." (Note: The HDFS Connector is a component of Hadoop — an open source distribution managed by the Apache Foundation. It allows the export of data from an ingestion application like Kafka into Hadoop.)
- **Data relevancy:** Another IT executive surveyed said: "The main benefit was the assurance that we had *all* our available data collected and stored in one place. With data in a single place, reporting goes from being a chore to a breeze, specifically, month-end reporting. All we need to do is make sure that all the systems present the data to the data lake, regardless of the

source and at the exact same time. Sorting and sifting through the data is much easier this way."

Furthermore, Dell EMC can showcase how its solutions scale in a capex-friendly fashion and how a vast array of ecosystem partners can assist in mitigating challenges with technology selection, integration, and operations.

CONCLUSION AND ESSENTIAL GUIDANCE

As businesses propel into the IDC 3rd Platform era, they will want to tap into new data sources and adopt new data storage formats, all geared toward gaining competitive intelligence in a quick, secure, and cost-effective manner. It is imperative that businesses that are consolidating their big data and business analytics environments take a serious look at creating purpose-built data repositories for their analytics environments. Such repositories form an essential cog in the development of a unified data fabric. All data lakes are not designed to serve the same mission, so careful analysis is a must to ensure that there is a match with other software layers of the data analytics stack.

IDC believes that data lakes should be a foundational component of enterprise analytics infrastructure strategy that spans current-gen and next-gen applications. As businesses learn to collate data from various sources and convert it into consumable nuggets of information for their various organizational units, they will no doubt be compelled to establish data lakes — upon which various workloads can concurrently operate. Such data lakes will enable existing workloads, as well as be future-proof, to seamlessly support new applications and workloads.

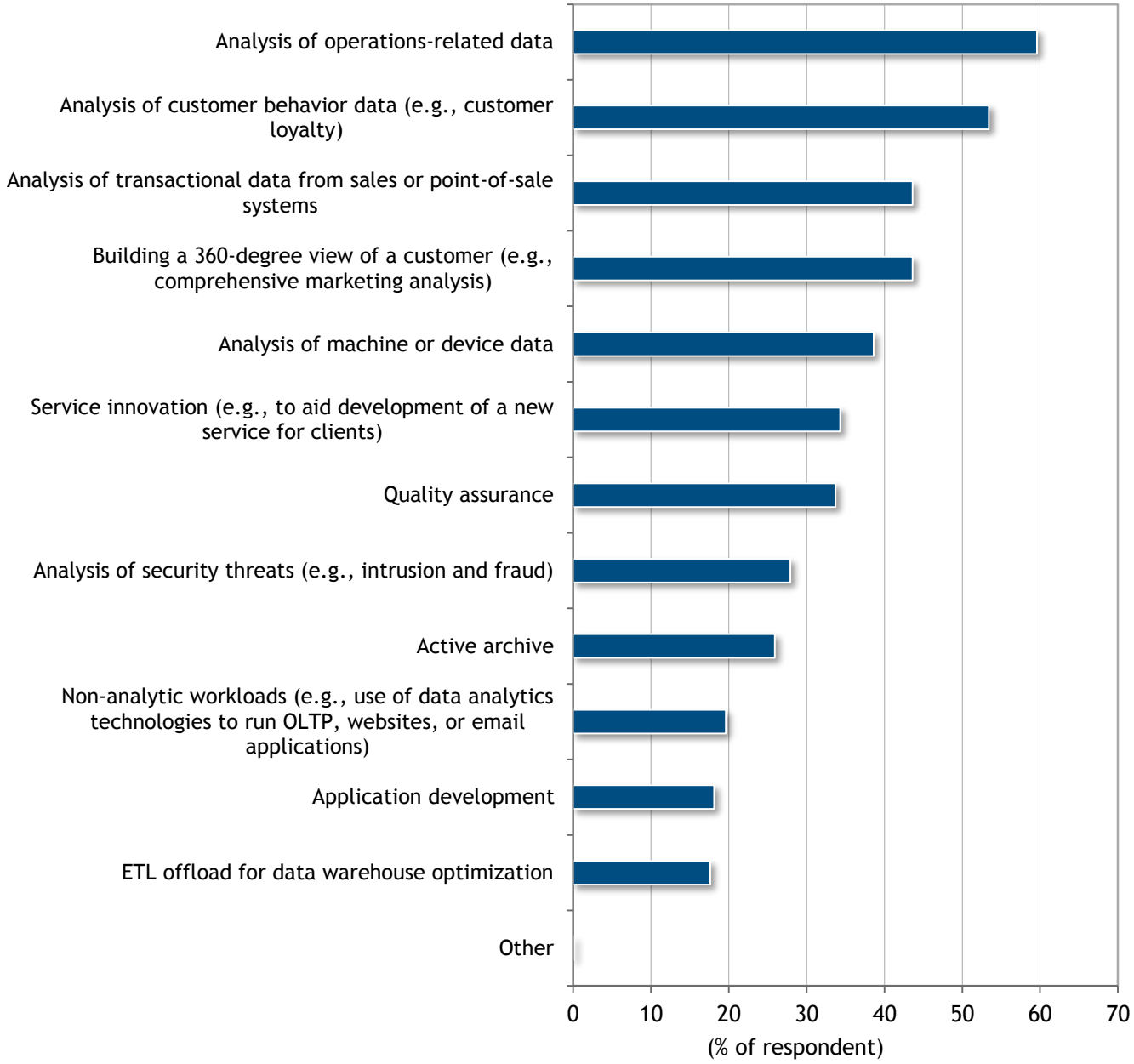
Products like Dell EMC Isilon possess the necessary attributes such as multiprotocol access, availability, and security to provide the foundations to build a data lake for many analytics workloads.

APPENDIX

FIGURE 1

Primary Use Cases

Q. What are the primary use cases for deploying or planning to deploy data analytics?



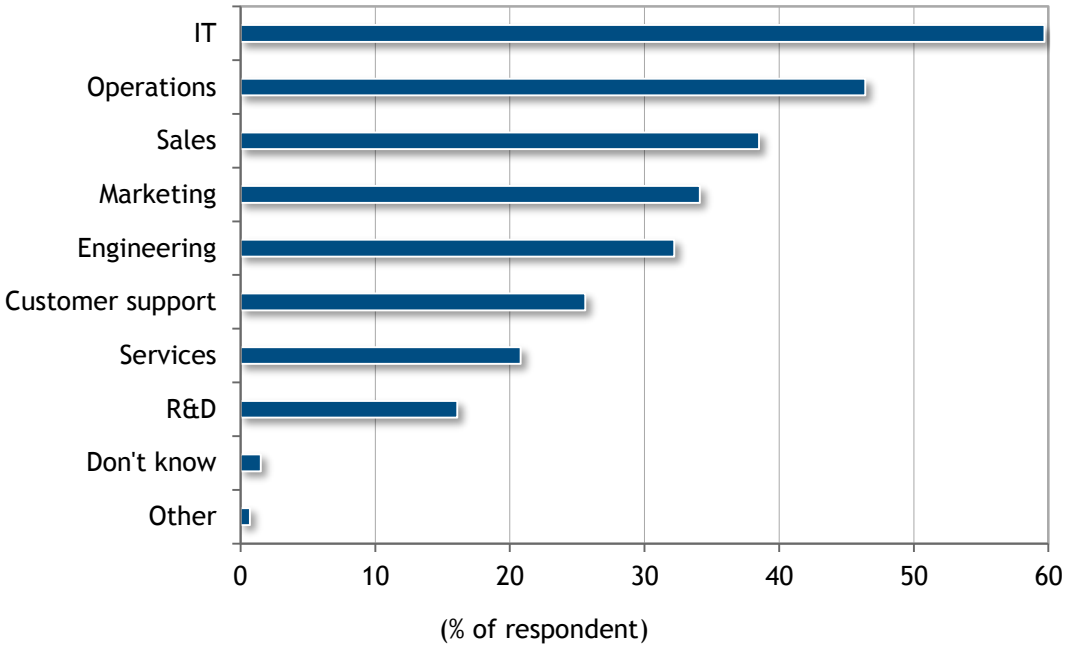
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 2

Primary Sponsors of Infrastructure

Q. *Who in the organization are/were THE PRIMARY sponsors, drivers, and supporters for this infrastructure?*



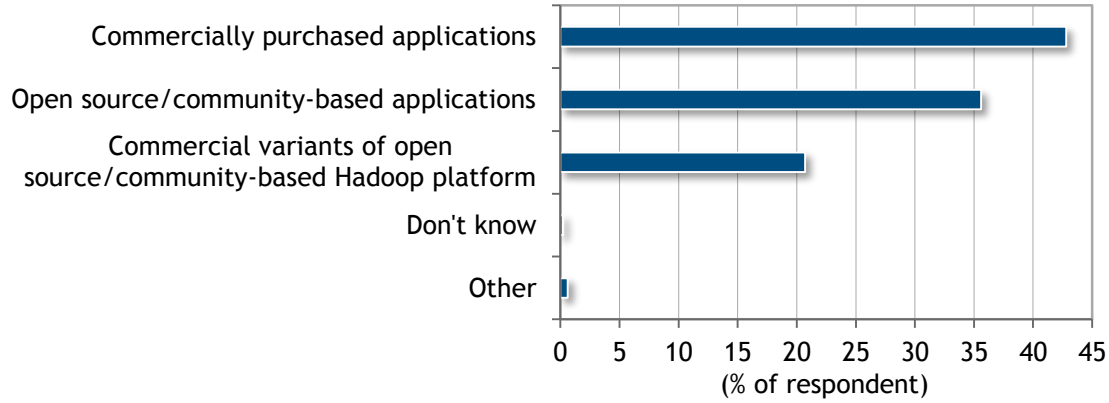
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 3

Application Type

Q. *What type of data analytics applications does/will your organization primarily run?*



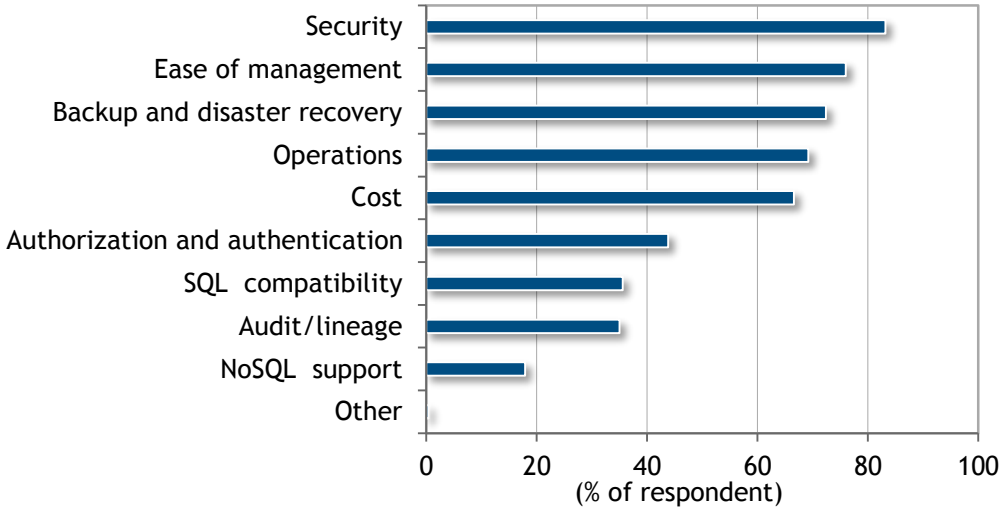
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 4

App Selection Criteria

Q. *What were the criteria your organization used to decide to use commercially purchased applications?*



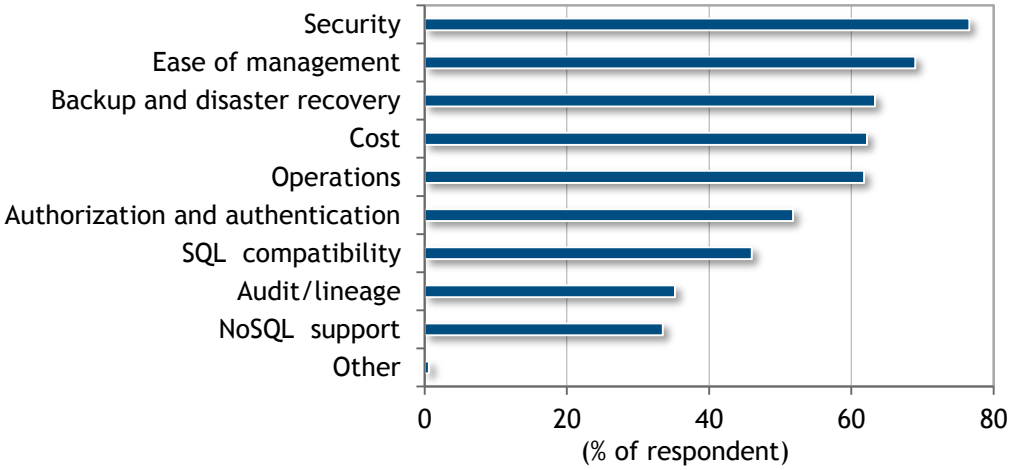
n = 480, Base = respondents indicated organization primarily run commercially purchased applications

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 5

Open Source App Selection Criteria

Q. *What were the criteria your organization used to decide to use open source/community-based applications?*



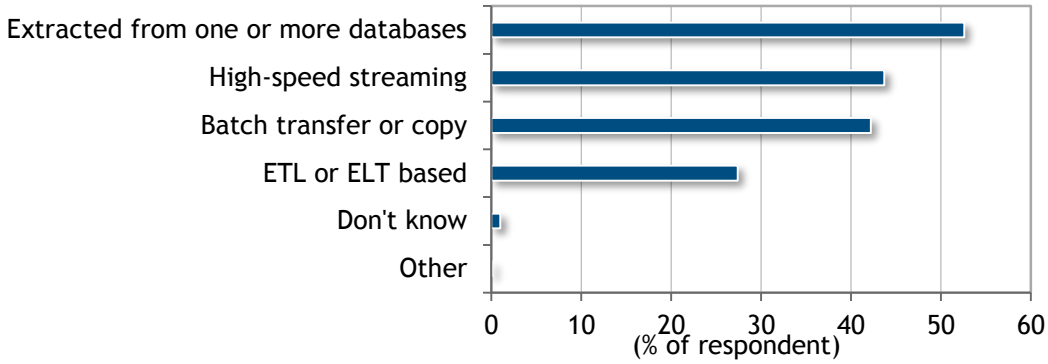
n = 386, Base = respondents indicated organization primarily run open source/community based applications

Source: IDC's *Data Lake Survey*, September 2016

FIGURE 6

How Data Is Harvested

Q. *How does your organization harvest data?*



n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 7

Business Challenges Addressed

Q. *What business challenges are you addressing with this data analytics deployment?*

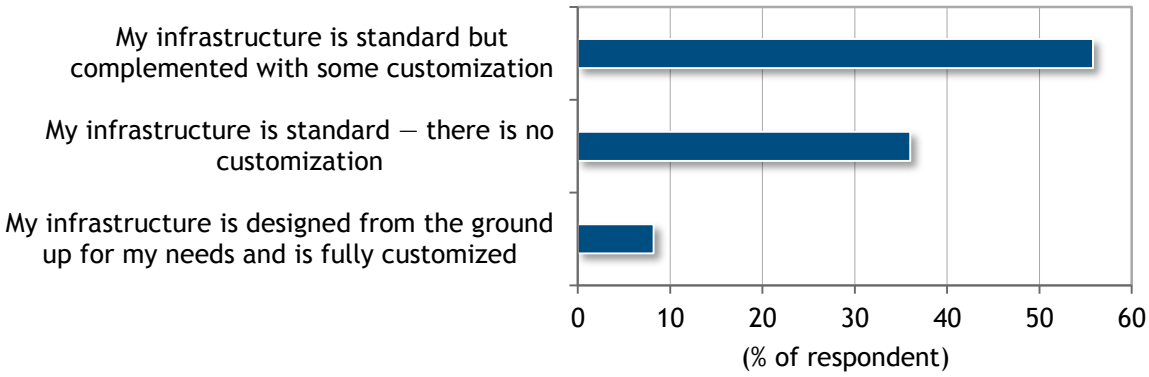
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 8

Level of Customization

Q. *What is the level of customization in your data analytics infrastructure?*



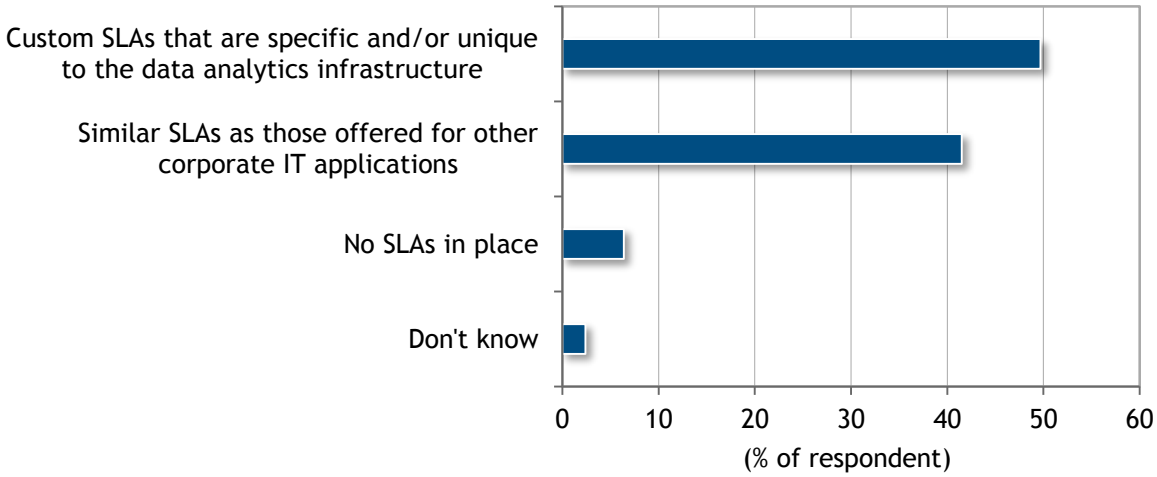
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 9

Service Levels Offered

Q. *What kind of SLAs do you offer for the storage component of your data analytics infrastructure?*



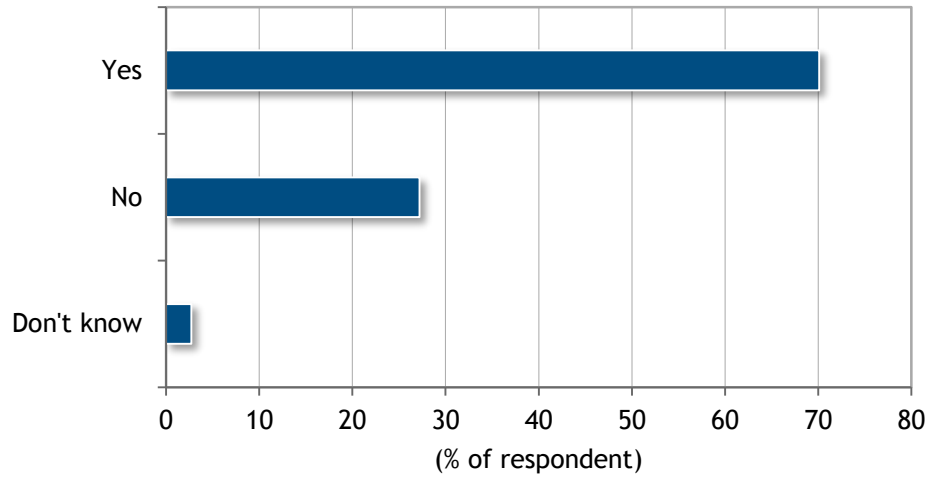
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 10

Shared Versus Dedicated Tenancy

Q. *Do you have a chargeback or metering system in place for this data analytics infrastructure?*



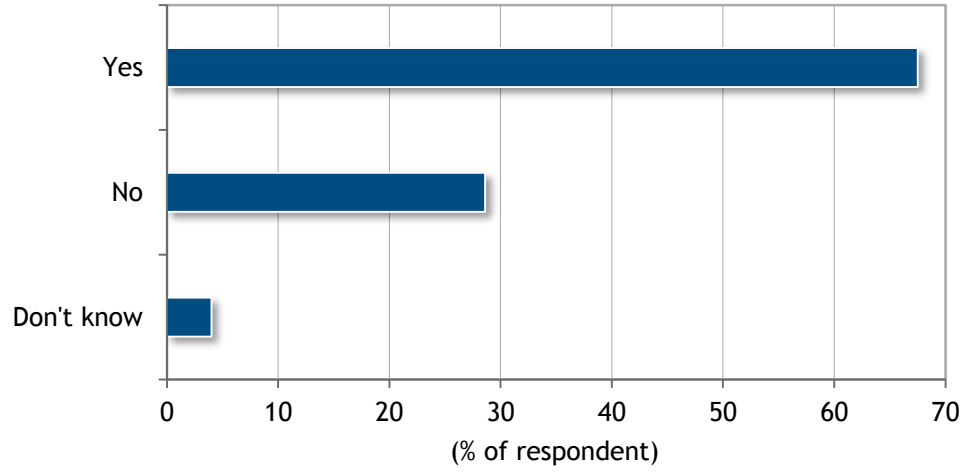
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 11

Collocation of the Infrastructure

Q. *Are your analytics workloads collocated along with non-analytics workloads on your data analytics infrastructure?*



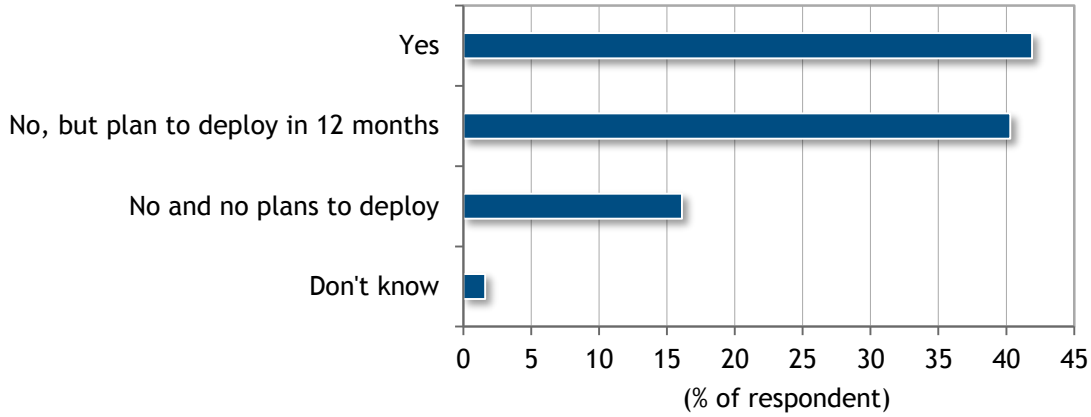
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 12

Metering/Chargeback Policies

Q. *Do you have a chargeback or metering system in place for this data analytics infrastructure?*



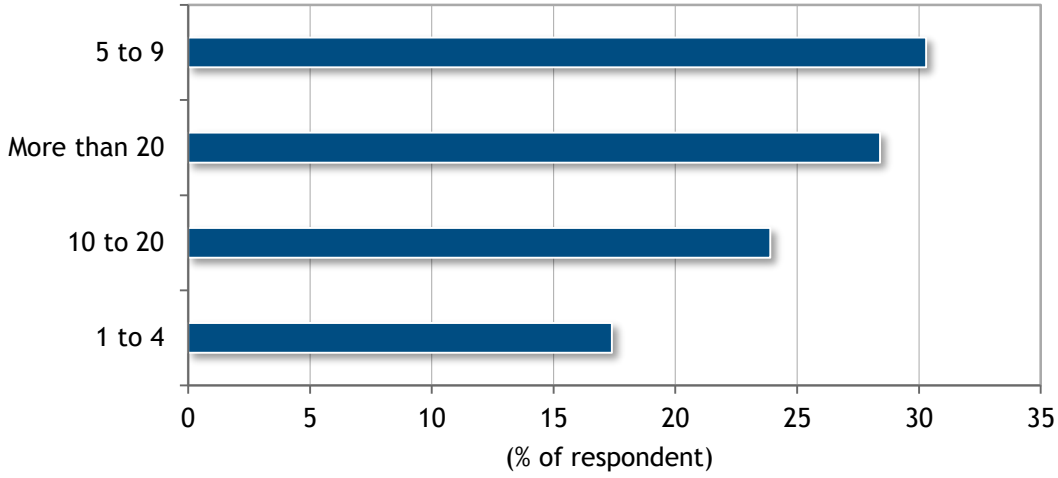
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 13

People Resources

Q. *How many people resources are used to manage this infrastructure?*



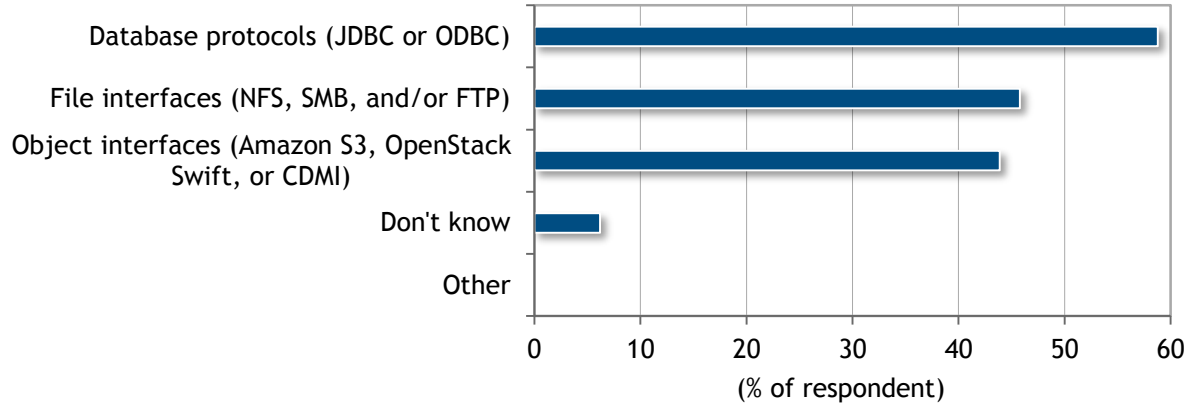
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 14

Preference for Data Access Protocols

Q. What kind of data access interfaces/protocols are used to capture and/or ingest data into your existing enterprise storage?



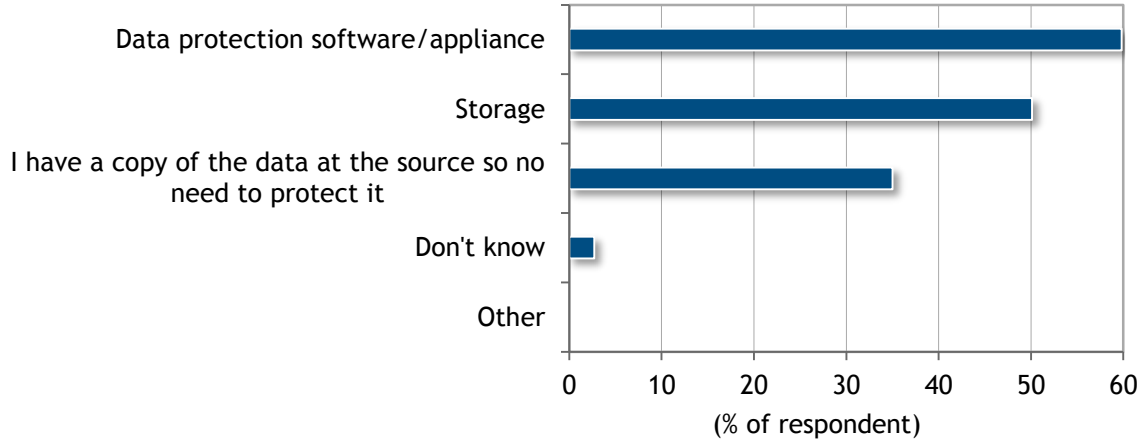
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 15

Data Protection Policies

Q. What kinds of data protection policies are in place in your data analytics infrastructure before it is processed/analyzed?



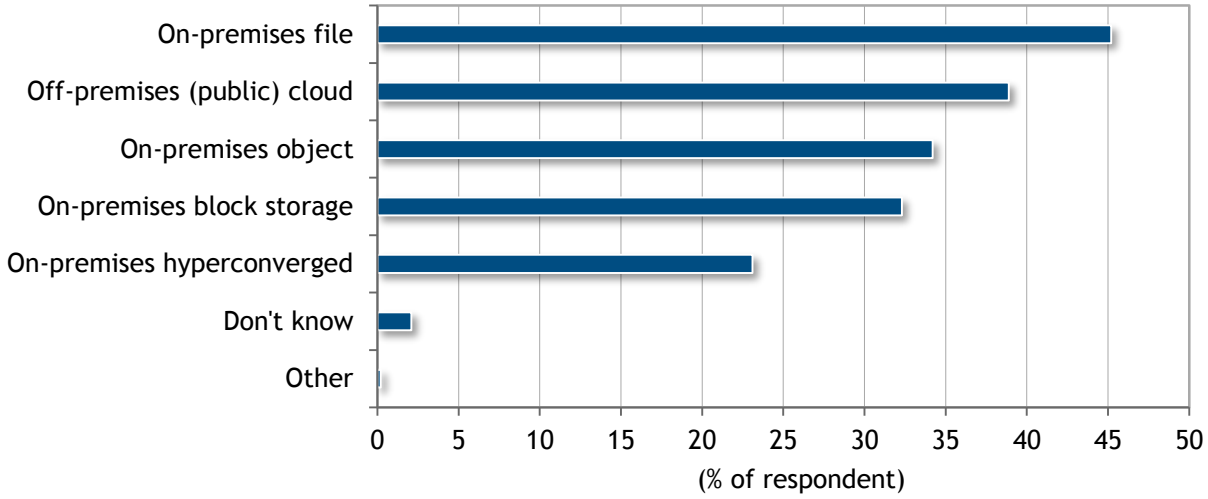
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 16

Storage Media for Pre-Analysis Data

Q. On what storage medium is data stored in preparation for analysis (i.e., before it is analyzed)?



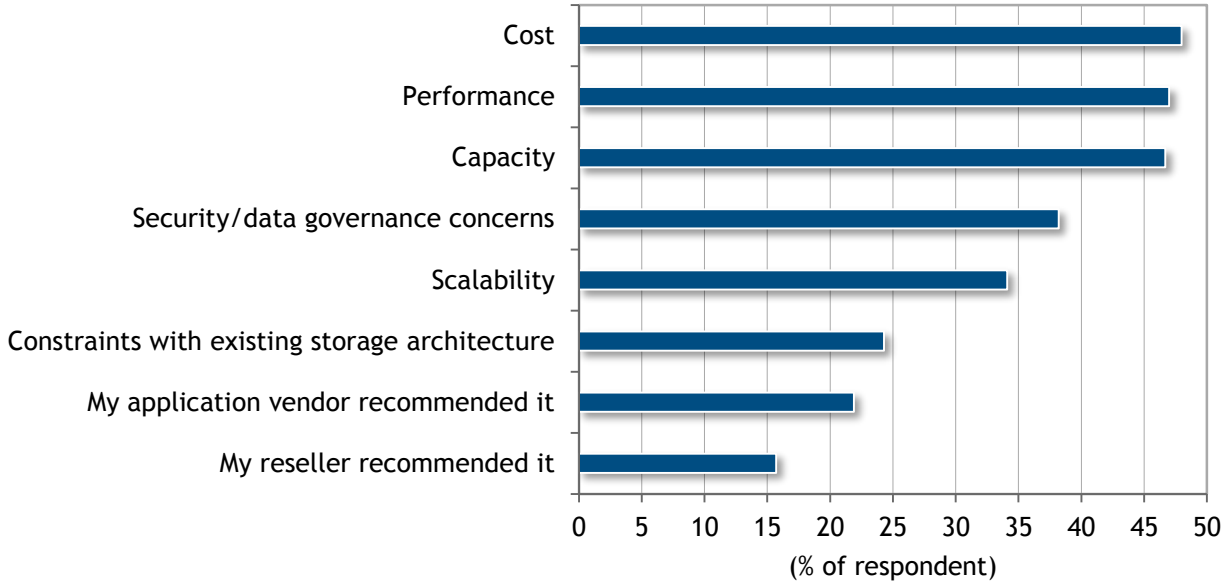
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 17

Storage Media for Pre-Analysis Data Drivers

Q. What were the primary drivers of selecting these media?



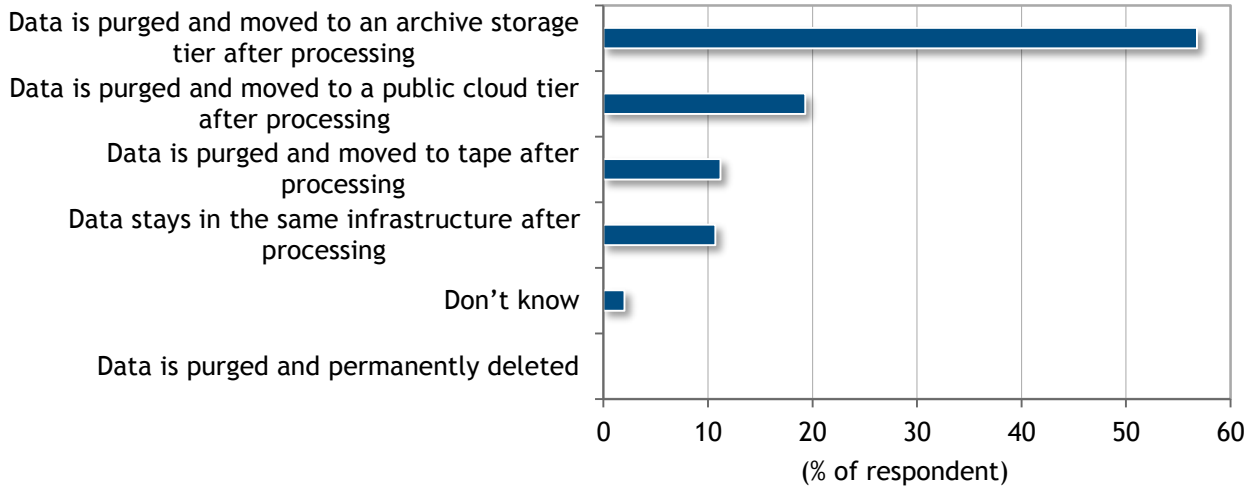
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 18

Storage Media for Post-Analysis Data

Q. *What kinds of data retention and/or archival policies are in place in your data analytics infrastructure for the raw data after it is processed/analyzed?*



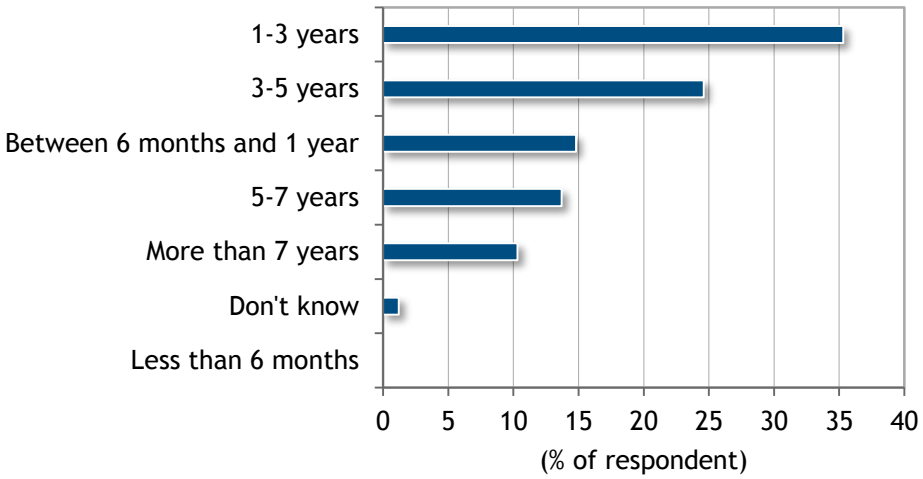
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 19

Data Retention Policies

Q. *How long is this data retained?*



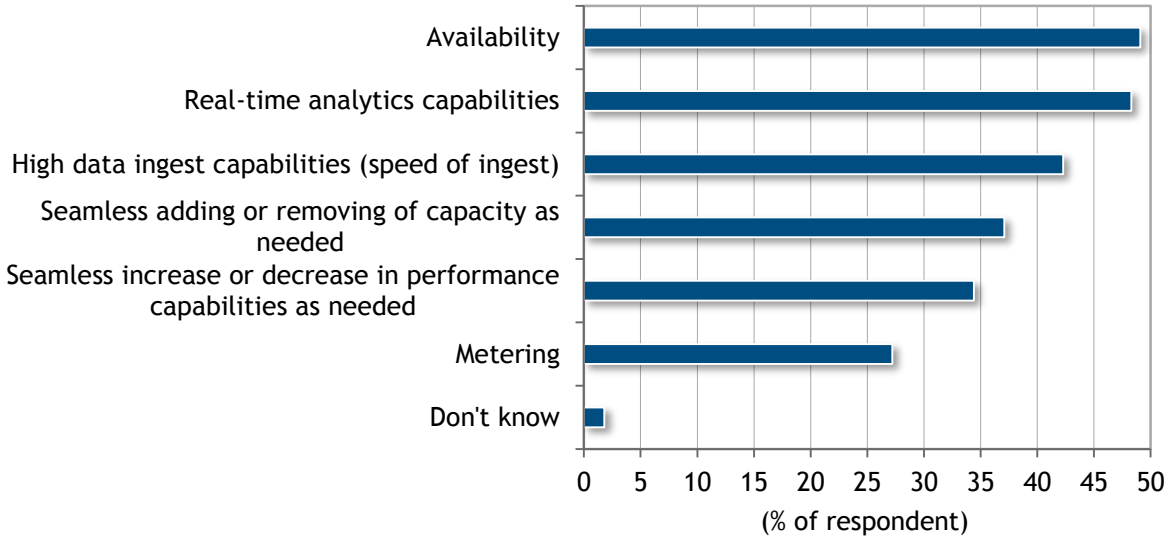
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 20

Storage Attributes

Q. What kinds of storage attributes are critical to the functioning of your data analytics infrastructure?

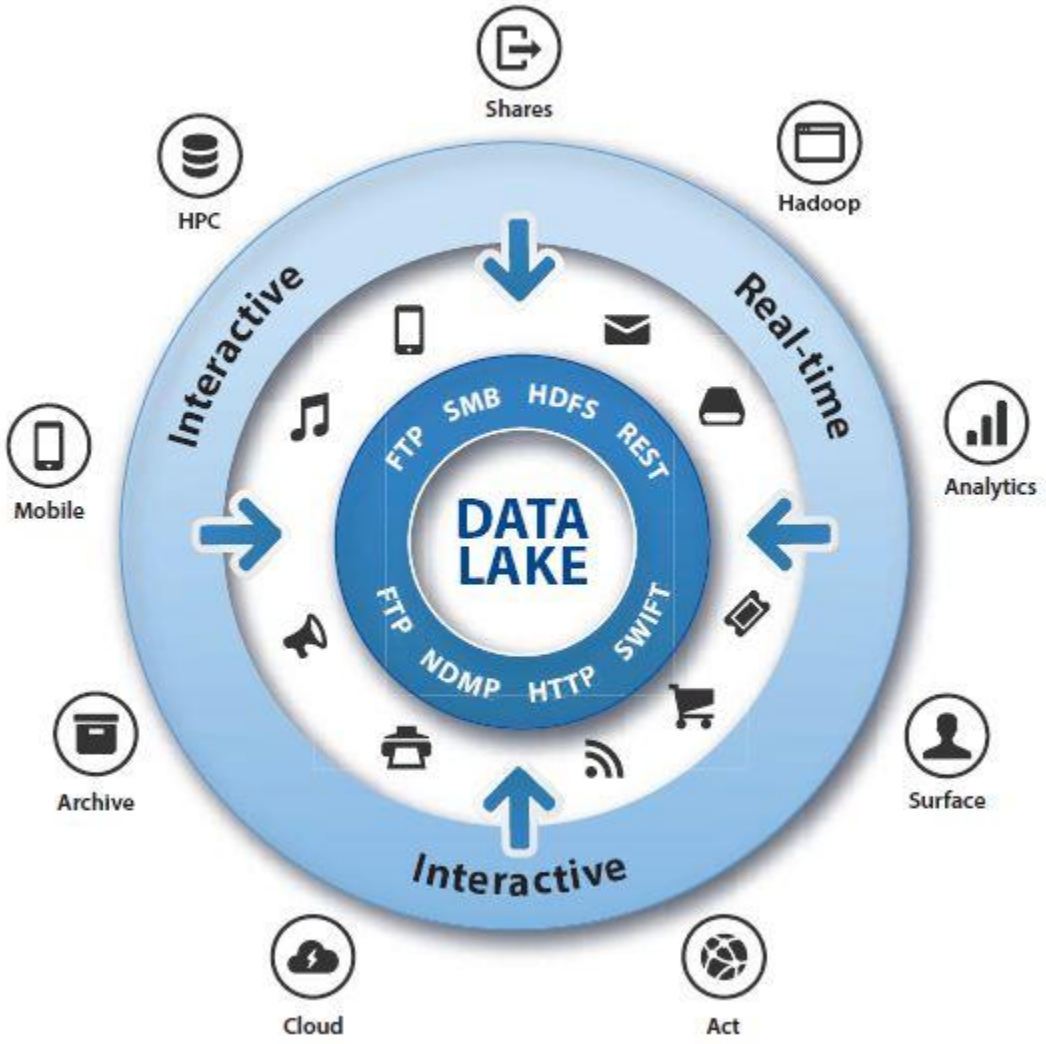


n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 21

Illustration of a Data Lake

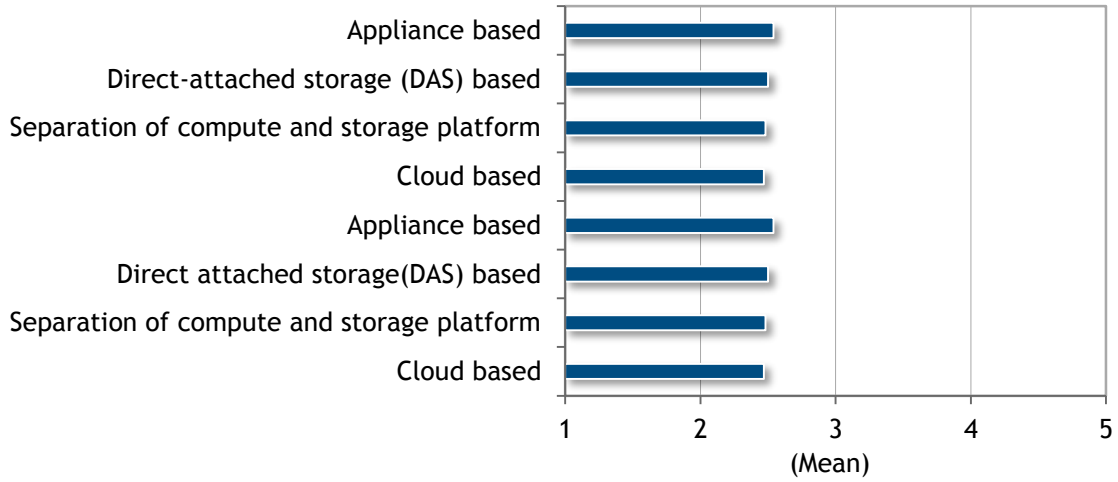


Source: IDC, 2017

FIGURE 22

Preference for Delivery Models

Q. Please rank the following in terms of their viability as an analytics infrastructure platform.



n = 1,105

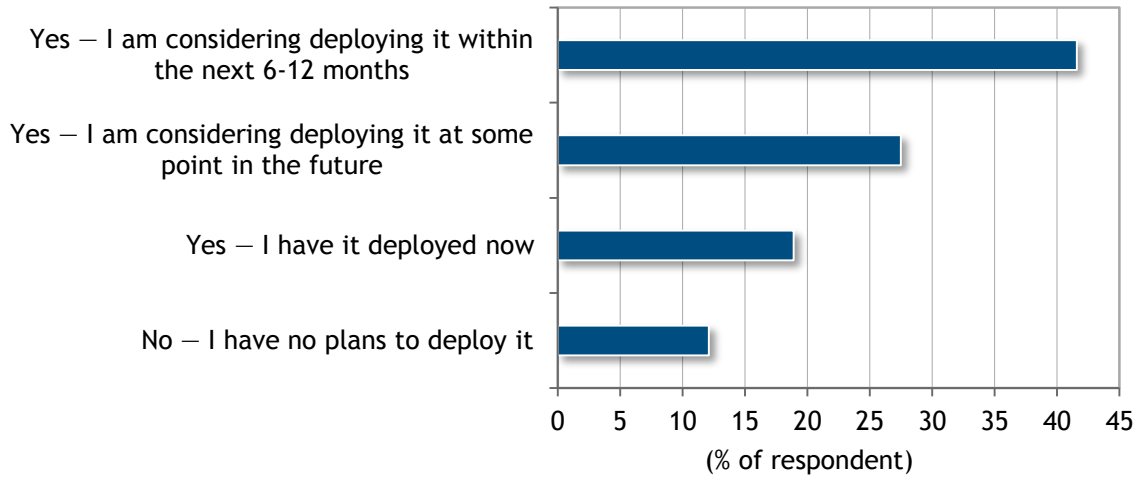
Note: The data is measured on a scale of 1 to 5, where 1 = the least viable and 5 = the most viable.

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 23

How Prevalent Are Data Lakes?

Q. Given the definition of data lakes, do you believe you have deployed or are considering deploying a data lake in your environment?



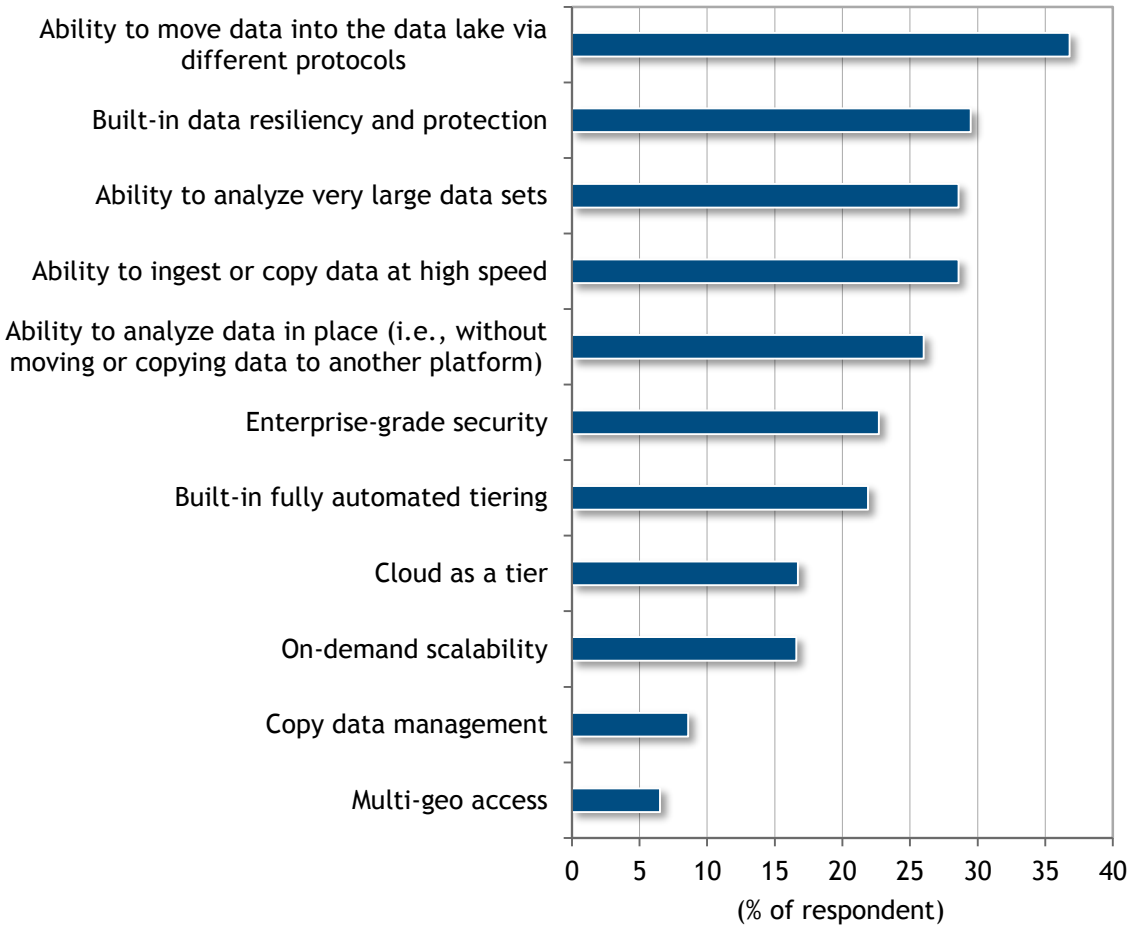
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 24

Technical Capabilities

Q. Given that you have deployed data lakes, what technical capabilities are you benefitting from the most?



n = 210

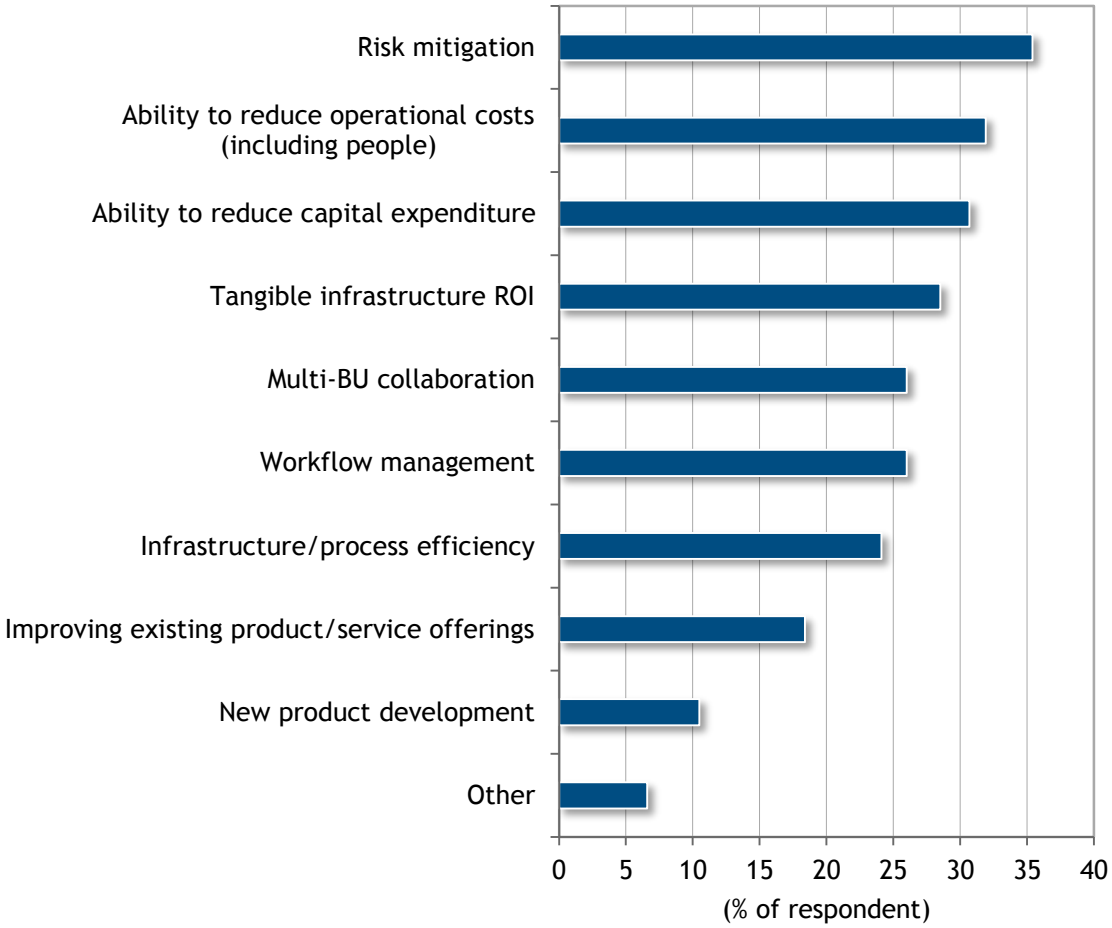
Base = respondents who have deployed data lake in their environment

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 25

Business Capabilities

Q. What business capabilities are you achieving with your data lake?



n = 210

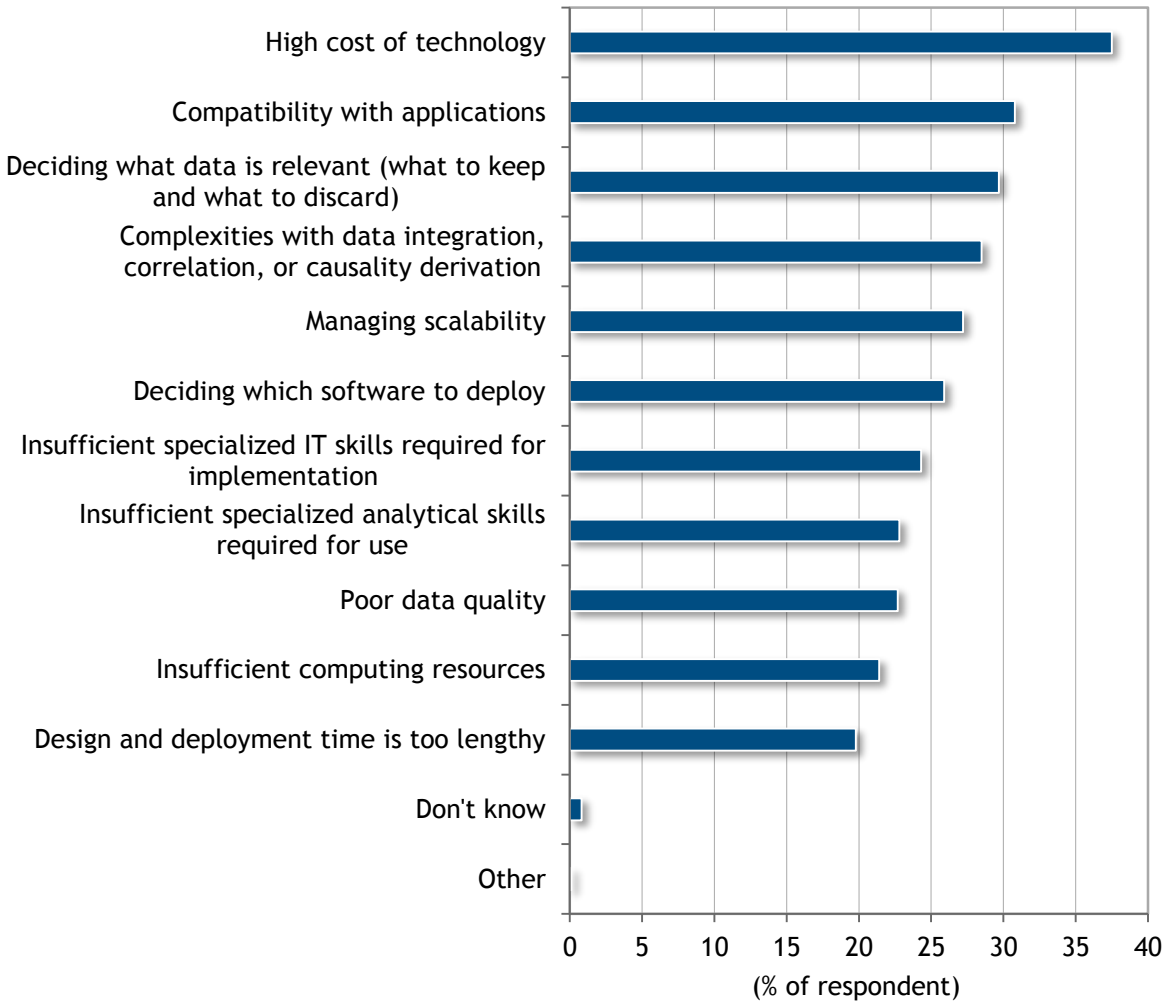
Base = respondents who have deployed data lake in their environment

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 26

Software/App Challenges

Q. *What are the software/application challenges that you face/expect to face with your data analytics infrastructure?*



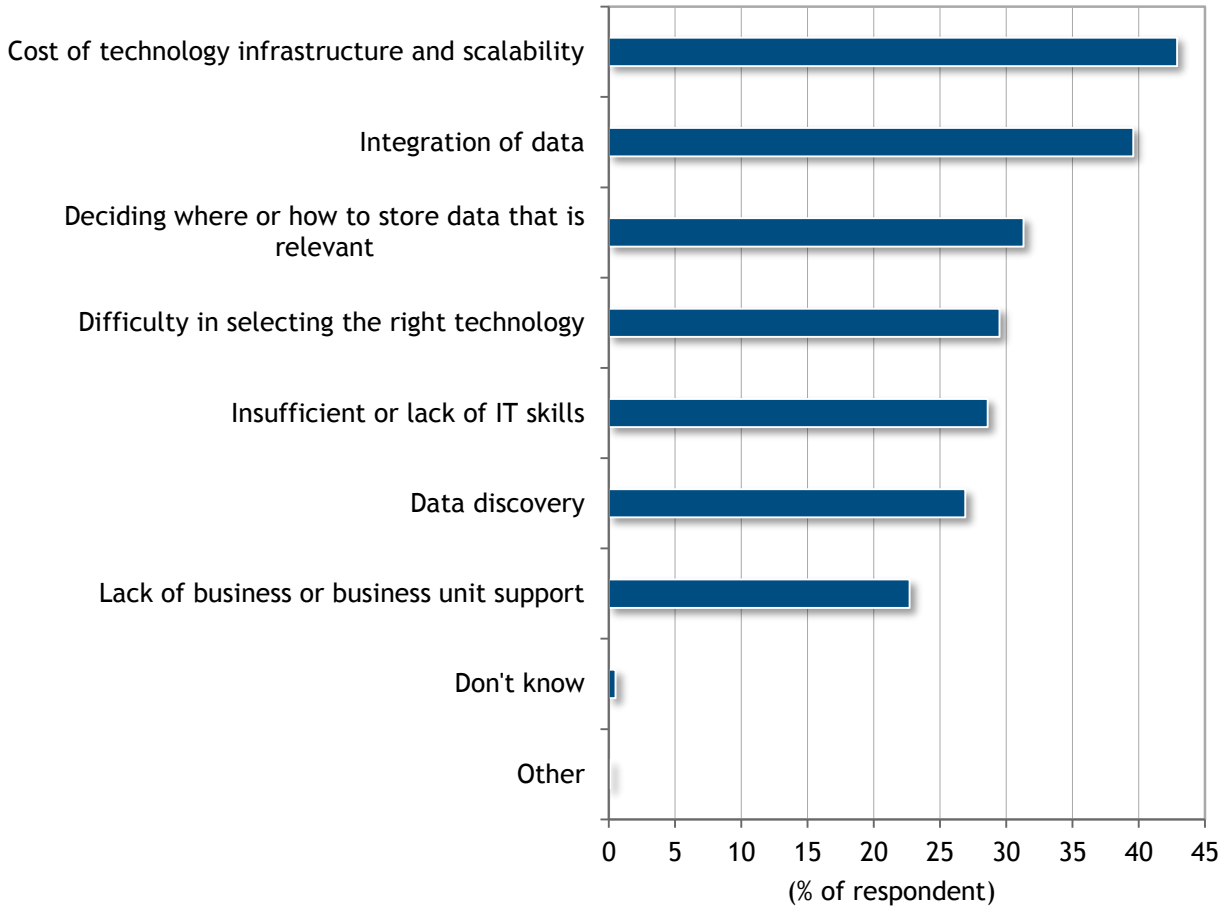
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 27

Hardware Challenges

Q. *What are the hardware challenges you face/expect to face with your data analytics infrastructure?*



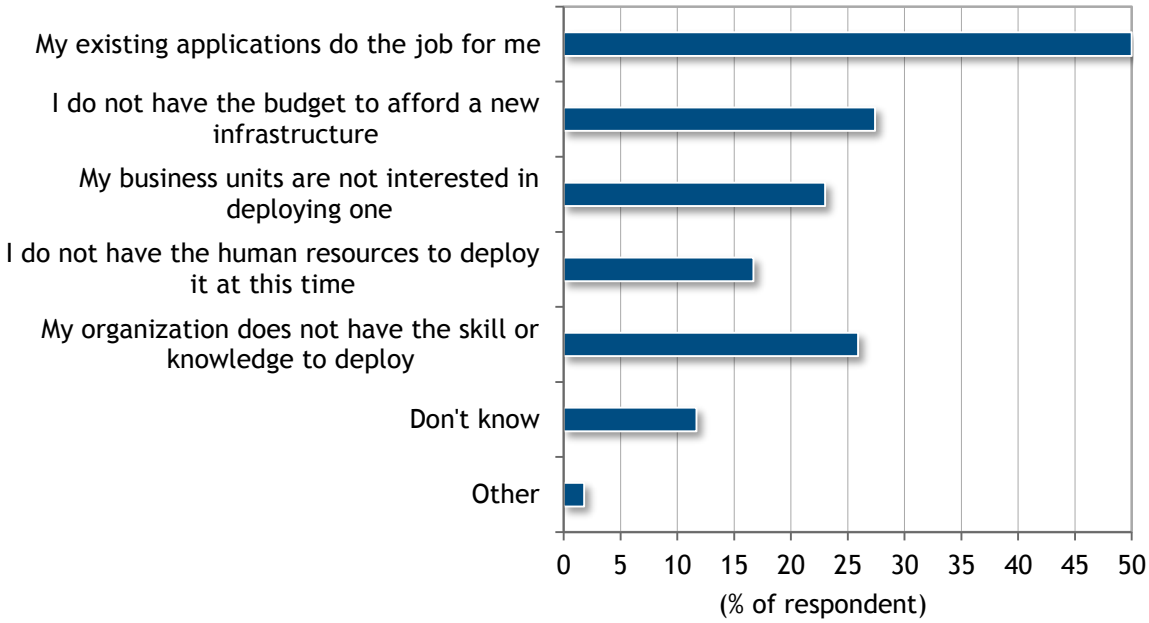
n = 1,105

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 28

Decision Against Deploying

Q. *What were the reasons your organization decided against deploying a data analytics infrastructure?*



n = 115

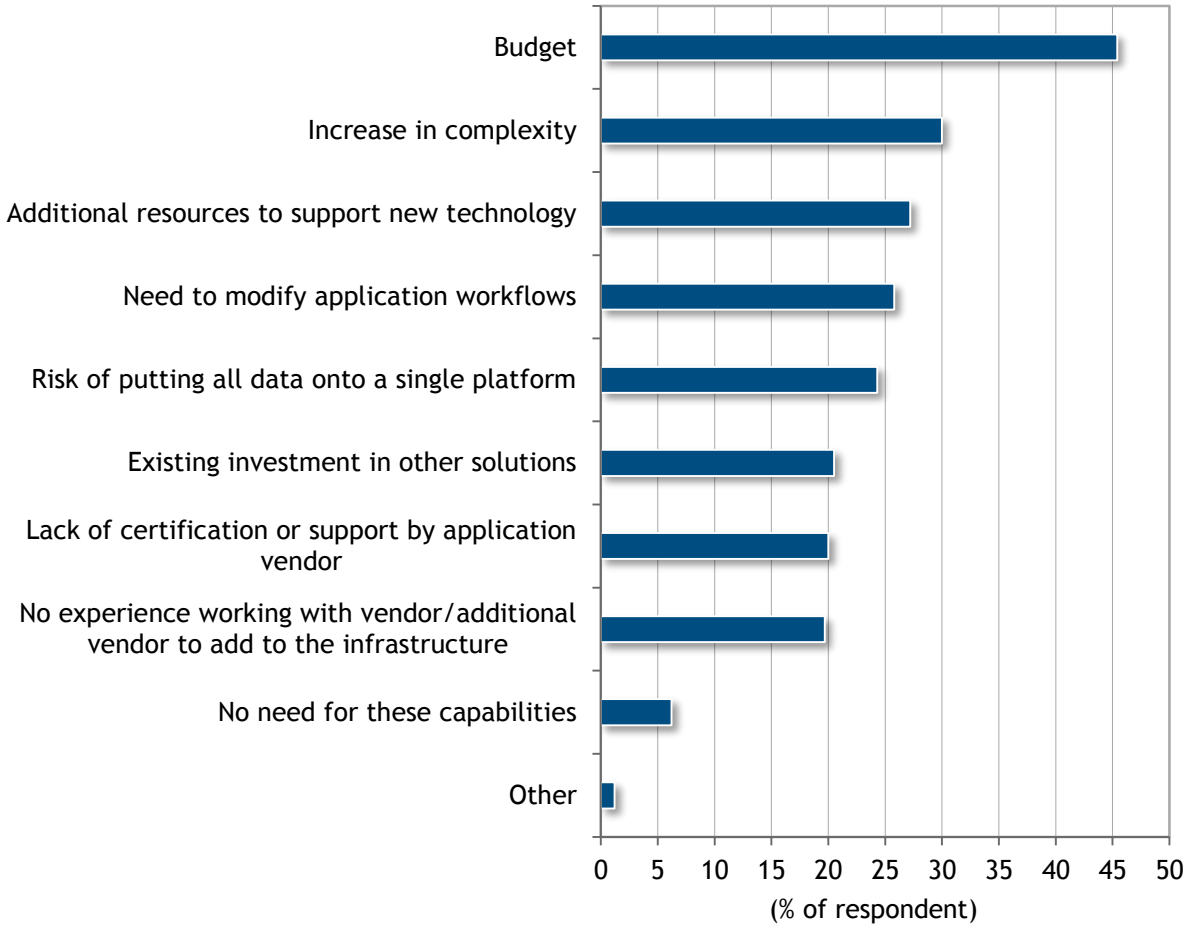
Base = respondents who have no plans to implement a data analytics infrastructure

Source: IDC's *Data Lake Survey,* September 2016

FIGURE 29

Reluctance, if Any, in Deploying a Data Lake

Q. *What hesitance, if any, do you have in deploying a data lake?*



n = 895

Base = respondents who have not deployed data lake

Source: IDC's *Data Lake Survey,* September 2016

DEFINITIONS

The following definitions were provided to survey respondents and are used throughout this white paper.

Digital Transformation

Digital transformation (DX) is the use of 3rd Platform technologies by enterprises to create value and competitive advantage through new offerings, new business models, and new relationships. The IT

market will become dominated by 3rd Platform tools that need to be connected to digital transformation initiatives in order to be relevant. For organizations, transforming themselves digitally involves:

- Automating business processes
- Interconnecting systems of record, engagement, analytics, and insight
- Managing, interpreting, and acting on "exascale" data
- Embracing IDC's 3rd Platform technologies, including next-gen apps

Organizations that thrive in the digital economy will leverage technology, people, and intellectual property to own the customer experience.

Data Lakes

Like big data repositories, data lakes can be thought of as a corpus of unstructured and semi-structured data collected and collated from different sources (like streams) into a single unified data pool (hence the term *data lake*). A data lake offers multiple access points for data "on-ramping," meaning support for standard network access protocols (NFS, CIFS, pNFS) as well as RESTful object interfaces by which applications can write data into the repository. However, more crucially, a data lake supports the storing of the data in a manner agnostic to how it is moved into the repository and in a manner that makes it easier for adjacent data analytics workloads to analyze the data. Specifically, the data is stored using open standards rather than in proprietary formats. Data lakes can be considered a central "deep storage" repository for consolidating different types of unstructured, semistructured and, to some extent, structured data.

Data lakes are seen as a solution to the data deluge and access conundrum faced by businesses that cannot be solved using big data repositories built on a single platform like Hadoop. Akin to enterprise data warehouses, data lakes allow disparate and incoherent data types to be consolidated onto a single, scalable, extensible, and agile storage platform. IDC expects that most data lakes will need to support:

- **Multiformat multiprotocol data ingest and access.** Data lakes need to support data to be ingested (i.e., placed on them) via a variety of file, object, and even block interfaces that include, but are not limited to, NFS, pNFS SMB, NDMP, HDFS, or RESTful object interfaces (such as OpenStack Swift, Amazon S3, and CDMI) by which applications can write data into the repository. Access mechanisms can be open, standards based or, where required, application specific. The expectation then is that this same data should be consumable (read: accessible) via different mechanisms or interfaces without the need to copy, replicate, or export it (into a different format).
- **Access-agnostic storage.** Data lakes should not make any presumptions on the manner in which the data is ingested or accessed — the two mechanisms could be completely different (e.g., data ingested via NFS could be accessed via HDFS or via an API). This also means that unlike typical file-based storage platforms, data lakes should make their metadata extensible and programmatically accessible, beyond the normal expectations of a specific access interface or API. Data lakes can therefore be suitable for both traditional workloads, such as home directories, file shares, sync-and-share applications, as well as Hadoop, and next-generation business and social analytics and cloud and mobile applications.
- **Deep storage with "infinite" scalability and efficiency.** Data lakes should support unprecedented nondisruptive scalability and agility. Data lakes should also support efficiency, for both upstream and downstream tiering. The platform should make use of upstream (flash) tiering that support efficient analytics workloads of hot data sets and downstream (cloud, cold

storage) tiering for storing inactive data sets but, crucially, should support data movement between tiers depending on the I/O activity and decay. In other words, the highly active data is placed on a tier optimized for performance (dollar per IOPS), while inactive data is placed on a tier optimized for capacity (dollar per gigabyte). Data lakes also need to have built-in data optimization, protection, and availability mechanisms that exceed the service levels established for the workloads operating on them. As a consolidated repository, it is essential that the platform also has a built-in robust data loss prevention (DLP) mechanism.

- **Support the three AAAs of security.** Given the fact that most data lakes will need to store sensitive data sets, it is essential that they support robust authorization, audit, and authentication mechanisms for users and applications. They also need to support inline and at-rest data encryption.

With a platform that acts like a data lake, businesses can minimize fragmentation and gain better and consistent insight into their entire data.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com