DATA LAKE

ADOPTION AND MATURITY SURVEY FINDINGS REPORT

UNISPHERE RESEARCH

database TRENDS AND APPLICATIONS

Radiant Advisors

DATA LAKE

ADOPTION AND MATURITY SURVEY FINDINGS REPORT

TABLE OF CONTENTS

Data Lake Adoption and Maturity Survey Findings Report

By Radiant Advisors

IINTRODUCTION

In Q2 2015, Radiant Advisors and Unisphere Research, a division of Information Today, Inc., released "The Definitive Guide to the Data Lake," a joint research project that sought to clarify the industry's emerging data lake concept through combining fundamental data and information management principles with the experiences of early implementations to provide deeper insight and a practicable perspective into how current data architectures are transforming into modern data platforms.

Following the success of that report, Radiant Advisors and Unisphere Research reconvened with the support of sponsoring vendors Teradata, Hortonworks, Attunity, and HP Data Security to launch a supplementary research study on the current state of data lake adoption maturity. By surveying both current and potential adopters in the industry, this study was designed to document key perceptions, challenges, and successes by focusing on data organization, integration, security, and definitional clarification to address key areas of concern and interest in ongoing data lake adoption. The intent of the survey and this corresponding report is to understand and share the current and planned adoption of technologies in the Hadoop ecosystem, intended specifically for a data lake strategy, and to learn how adopting companies are addressing critical data lake success factors, including rethinking data for the long-term, establishing governance first, and tackling security needs upfront. The survey and report also identify emergent areas of concern and new areas of clarification needed for data lake maturity.

We surveyed 385 IT practitioners and stakeholders at organizations within a variety of industries who subscribe to *Database Trends and Applications* and *Big Data Quarterly* magazines; contributing respondents were, therefore, from a highly technical audience. Approximately 60% were IT and database administrators and staff or were otherwise identified with IT operations, while the remainder held CXO positions or other executive leader business roles. There was also a small sample of responders from academia, including roles as both student and professor (less than 5%).

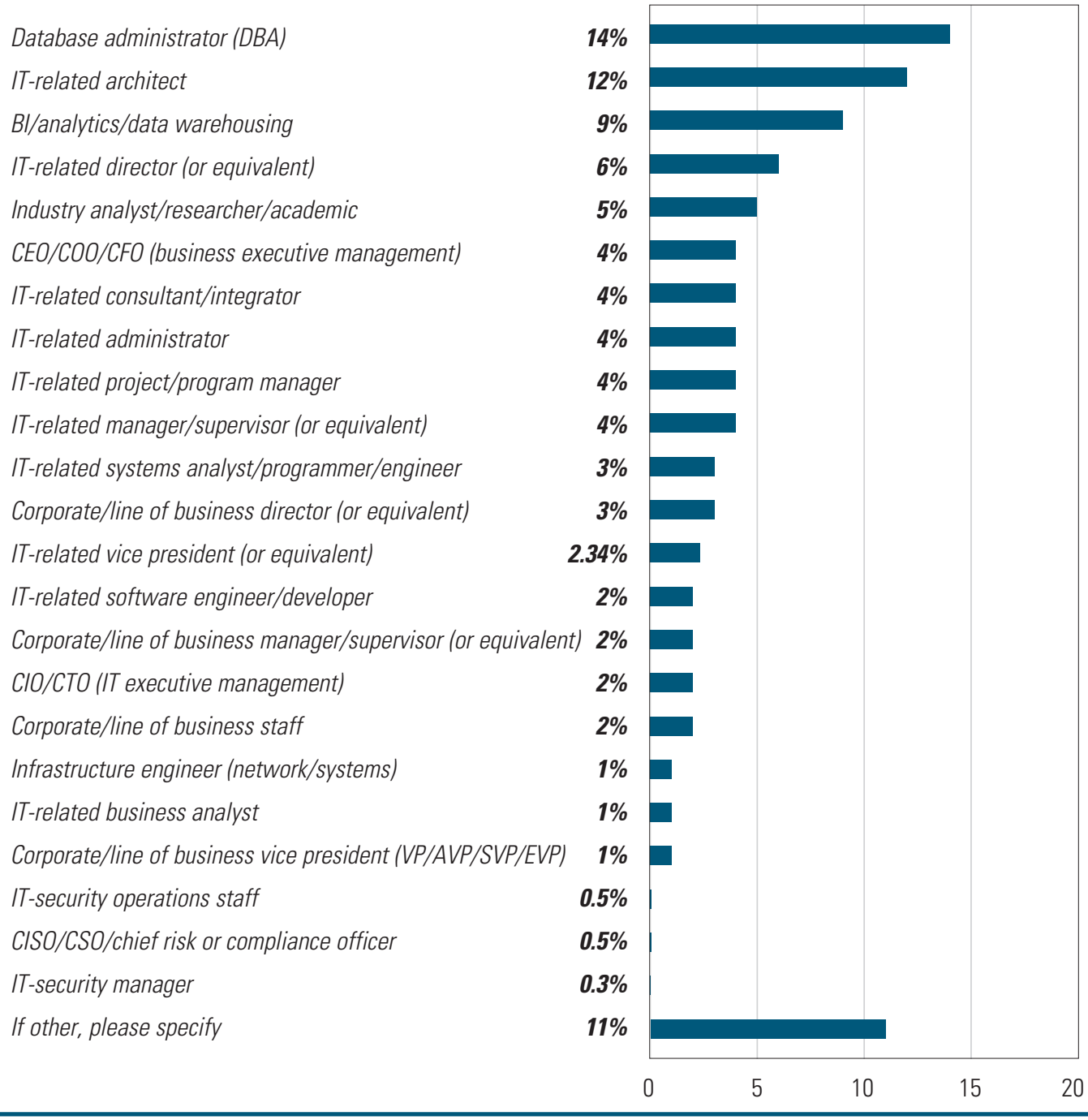
At the organizational level, the most dominant organization sizes in the survey were those with more than 20,000 employees worldwide (30%) and those of less than 250 employees (21.81%). Of these organizations, a broad spectrum of industry verticals were reached by the survey, with finance and software being the most highly represented at 13% and 11%, respectively. Other well-represented sectors included education, government, and manufacturing. Finally, regarding geographic engagement, North America represented the largest region represented in the survey, with over 80% of responders.

High-Level Key Findings

- The data lake is increasingly recognized as both a viable and compelling component within a data strategy, with companies large and small continuing to move towards adoption.
- Clear early use cases exist for the data lake.
- Governance and security are still top-of-mind as key challenges and success factors for the data lake.

Each of these findings will be explored in the subsequent sections below. As a disclaimer, please note that some totals in the study do not equal 100% due to rounding or the acceptance of multiple responses per question.

Which of the following best describes your role at your organization?



UNDERSTANDING THE "DATA LAKE"

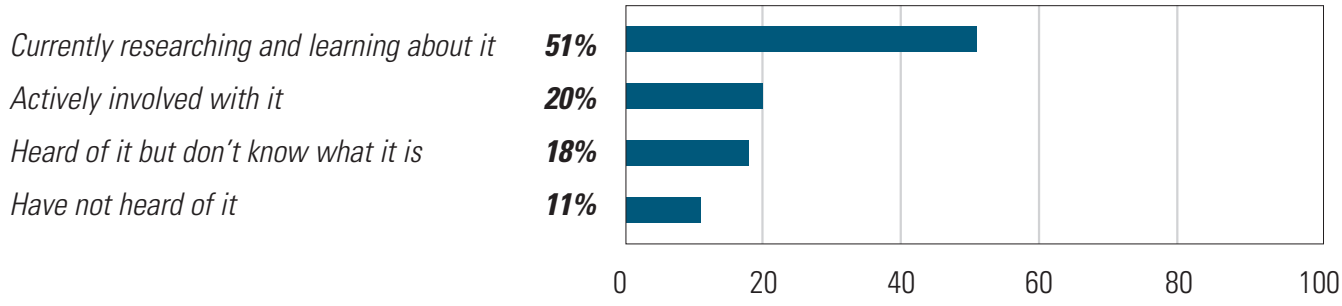
One of the most convoluted conversations to date regarding the data lake has been on providing an exact definition. "In the Definitive Guide to the Data Lake," we defined the data lake as an architectural strategy and an architectural destination, thus addressing both the end state architecture and establishing an adoption and transformation strategy for data architecture-related decisions on the journey to the data lake. While adopters are not ready to accept a single definition (only 1.12% felt that the data lake is well-defined and consistent at a detailed level), 17% of responders do feel that the data lake is defined sufficiently at a conceptual level to be able to move forward, and many (another 33%) feel that this clarification is improving. However, an equally significant segment (nearly 30% of responders) feel that the concept is only somewhat becoming clearer, and 19.66% feel that the industry is flat when it comes to adding clarity to the term. Currently, respondents indicate that the definition of the data lake is becoming more solidified, but it is nowhere near completely crystallized yet.

However, while lack of definition consensus was evidenced in our survey results, among our responders there were clear trends on the data lake definition that fall in line with the definitions established in "The Definitive Guide to the Data Lake" report. Of these, the most common selections for the data lake descriptors were either as data architecture for IT (59%) or as data strategy for IT (67%). Designations as a business strategy or concerns for cost savings for big data storage were among the least common selections for the appropriateness of the data lake.

Though hype and backlash over definitions of the data lake remain top-of-mind, adoption of this strategy and architecture is nevertheless moving forward. Over 32% of responders have an approved budget to launch an initiative within the next fiscal year with another 15% having already submitted budget for adoption. As would be expected, there are another 35% still researching and very interested in more information to solidify a strategy decision. Today, the quest for information has begun to translate into knowledge and understanding as a large percentage—roughly 50%—of survey respondents is actively learning about how to capitalize on the benefits of a data lake in a modern data architecture strategy. Understanding begets action, and an additional 20% of respondents say that they are moving beyond education and awareness and are currently involved in data lake initiatives. Thus, it would seem that the murkiness of the data lake is improving through companies driven by business value and expecting to learn and clarify as they go prior to best practices being established.

While only a little over 1% of respondents were ready to dismiss the data lake concept as purely

What is your familiarity with the term "Data Lake?"



marketing hype, many are unsure of exactly what they think of the data lake, with 15% of respondents worried that the data lake is too risky and carries too many data governance and security concerns to be seriously considered as an architectural strategy.

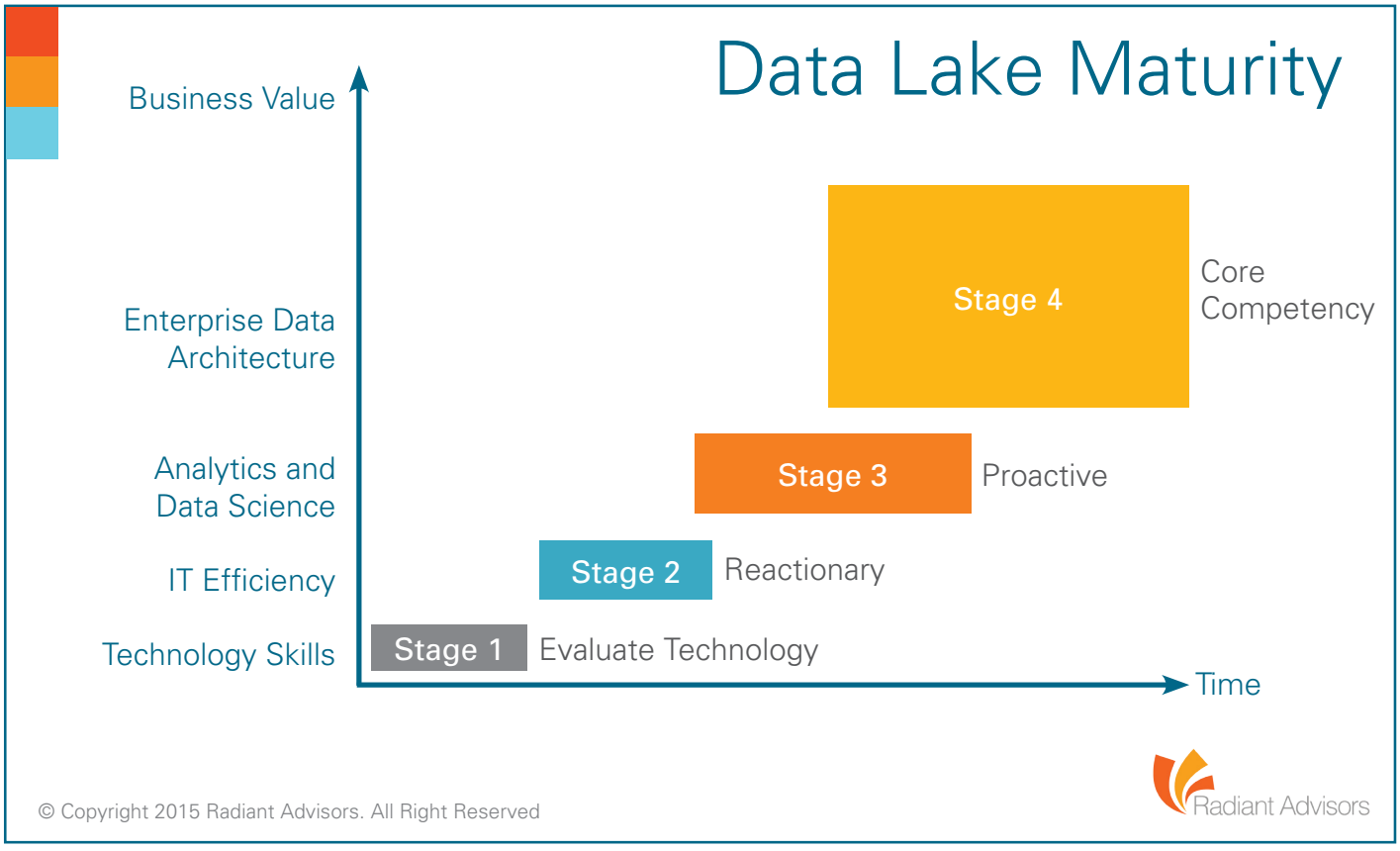
Relatively even percentages of respondents feel that the data lake is neutral and/or positive (good for IT data management), 27% and 24%, respectively. The largest majority (32%) of responders took the next step by responding that the data lake is more than positive, but that is valuable, defined in the survey as good for business analytics and discovery. Ultimately, over half (56.4%) of respondents reported a favorable impression of the data lake concept for either data management or data discovery reasons. While this is the impression of the majority, 20% of respondents still worry that the data lake is too risky and/or a marketing gimmick to be of use and another 27% are on the fence; continued clarification and prescriptive guidance on the definition and data lake use cases is needed.

ASSESSING DATA LAKE MATURITY AND COMMITMENT

Alongside our earlier data lake definition(s), Radiant Advisors also proposed a four-step data lake maturity model that provided a framework for adopters to move along their maturity journey, from Stage 1 of evaluating technology—those "kick the tires" big data pilot projects that focus on specific business project-based outcomes—to Stage 4, continuous optimization with the data lake as a core competency. In this final stage of maturity, we see Hadoop fulfilling a foundational component of the enterprise data architecture strategy, and supporting more of the operational, analytic, and big data workloads with both persistence and data engine layers.

As many companies continue their journey in building a modern data architecture, Hadoop adoption is prolific in the market and beginning to edge above the 50th percentile as quantified in this survey. For example, in our survey, 55% respondents were currently using Hadoop. Of these, 42% are running four or more clusters, though the number of nodes is less certain with responses ranging roughly equally in the 1–4, 5–8, 17–32, and 33–100 ranges (it's fair to note, too, that at least one-third of responders were uncertain how many nodes were in use by their organization). This is consistent with research that shows big data projects in early stages are siloed and not aware of each other causing the need for a single consolidated data lake strategy.

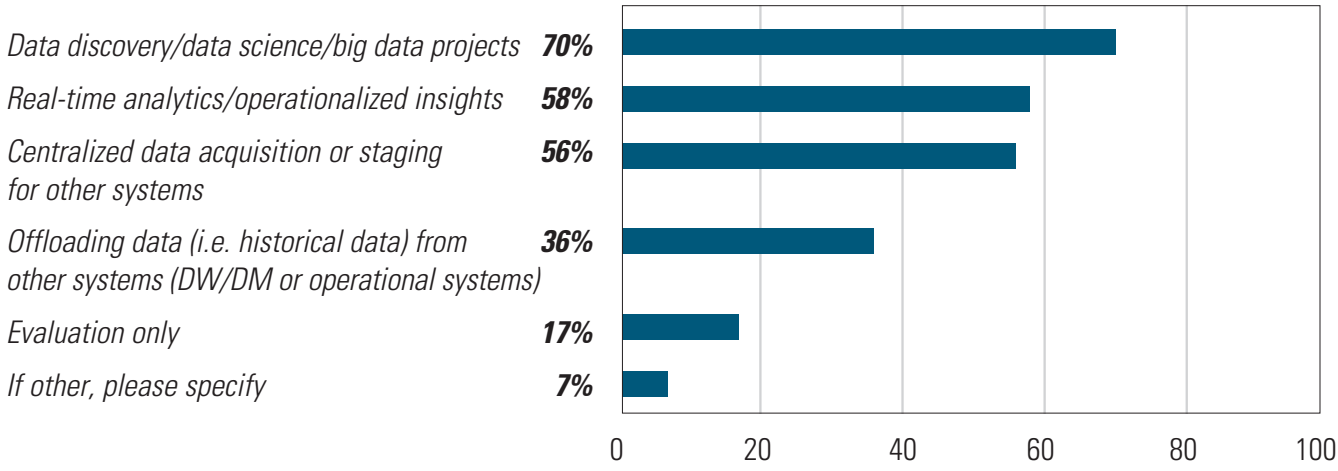
Aligning survey responses with our maturity model, we see many companies advancing as suggested in our earlier report, with most companies assessing their progress in Stages 1 through 3 of the maturity model. The largest majority (70%) are using Hadoop for data discovery, data science, and big data projects, which aligns with Stage 3 of our maturity model as the data lake begins to reach the tipping point that will move



the concept into its place as a core part of IT strategy. Additionally, of those currently using the data lake for other initiatives and use cases, data discovery, data science, and big data projects continue to be the largest trend for use cases in the next 12 months (80%) with real-time and operationalized analytics emerging as the second highest priority use case (77%). Again, these trends and predictions align to our earlier statements that the data lake will unify data discovery, data science, and enterprise BI.

Like Hadoop, which continues in adoption, commitment to the data lake strategy is growing. We saw this evidenced in the survey findings in terms of monies being allocated to current or future data lake initiatives, as reported earlier. However, we also see this in market perceptions of how organizations are committing to data lake strategies.

What use cases are Hadoop cluster(s) primarily being used for currently?



DATA LAKE KEY CHALLENGES AND CRITICAL SUCCESS FACTORS

A DATA LAKE has the potential to produce compounding value as it grows and becomes a fundamental part of an enterprise's overall data strategy. However, as adoption continues, critical data lake success factors exist to address challenges faced by companies. Three of these success factors are: rethinking data for the long term (data needs – known and unknown—now must be balanced for requirements of reusability); establishing governance first, and tackling security needs up front. These success factors align with survey results, with respondents reporting the most worrisome challenges as metadata management issues (71%), security (67%), and governance (71%).

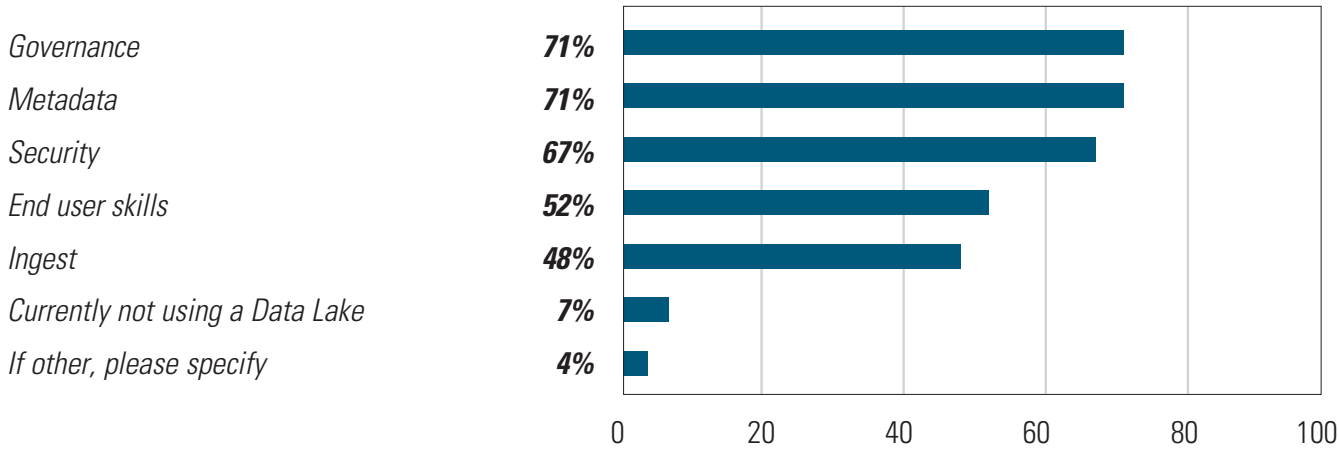
In the subsections we will explore governance and security in more detail, but alongside these key challenges for adoption success there are a few additional obstacles reported in the survey to which we'd like to give attention. First, the availability and competencies of Hadoop skillsets was a noticeable area of concern (roughly 55%). This is a reflection of the continued gap in knowledge and skills available, which is due in part to the rapid influx of new and improved technologies. Second, tying back to discussions above, almost 50% of responders still require an agreed-upon definition and strategy for the data lake. Budget-oriented and data integration challenges also command a large percentage (42% and 41%, respectively) of obstacles and concerns. Where and how to integrate the data lake with other components of the ecosystem still seem ambiguous, with over 47% of responders reporting that there is loose integration between the data lake and other components while, similarly, nearly 37% reported that there is very tight integration.

Governance

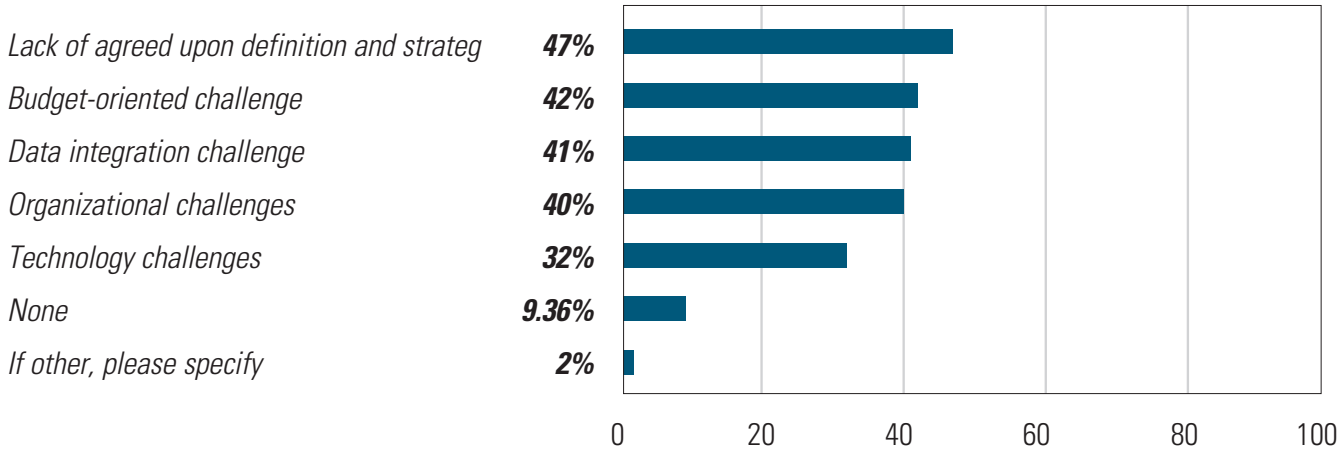
The need for data governance has always reinforced a framework for business drivers and risk management related to data and information polices. As validation to our earlier assessments, 62% of responders in the survey noted that governance is a "must have" from the beginning. While separately 31% made the distinction that governance should not obstruct its function and can be added incrementally.

Ultimately, governance establishes a common understanding for how everyone will work with data in the data lake. While in the past this framework has provided information policies to the enterprise

What are the key challenges you have experienced in making the Data Lake concept a reality?



What are the obstacles for your company in achieving its Data Lake goals?



through the definition and assignment of data owners, stewards, and specialists, these should now be extended to the data lake. Moreover, new data governance policies will need to focus on data ingestion, the evaluation of data sources (internal, external, and acquired third party), and specialized user data sets. Enabling governed data discovery will also be paramount to address data accessibility and how discovered insights, context, and analytic models will be institutionalized within the enterprise. It should also define key technology requirements, such as access controls, mobility, and security.

Security

Along with governance, security needs should likewise be tackled at the beginning of the data lake adoption process. A data-centric security approach provides a broad perspective by which to think about data from creation to consumption.

In our survey, approximately 42% of survey responders selected that a data lake strategy cannot be started without a security framework in place. Additionally, another 45% of responders felt that this security framework must be consistent or even more robust than all other currently available IT database security policies. Concerns about data encryption continue to be an obstacle for data lake adoption for data throughout its entire lifecycle.

Authentication and authorization will be cornerstones of data security in the data lake—authentication to verify that the users are who they claim to be, and authorization to permit access to a secure, centralized repository of data available through revocable credentials. Encryption—and likewise, decryption—will also have a role across all forms of data consumption from the data lake. Ultimately, security needs to be prioritized and "coherent, pervasive, and dynamic" by default.

CONCLUSION

The journey to the data lake has officially begun. Today, the data lake is increasingly recognized as both a viable and compelling component within a data strategy, with companies large and small moving towards adoption.

The absence of a clear, consistent definition is perceived as a hindrance for adoption, yet conceptual agreement about the purpose for a data lake as it relates to both data architecture and data strategy for IT is as an adequate substitute to allow early adopters to begin actively adopting the data lake as a key component of their modern data architectures. As adoption increases and companies progress along the maturity curve, moving beyond awareness to active initiatives, top concerns, and challenges continue to be governance and security, factors that must be addressed at upfront. Additional critical success factors—such as Hadoop skillsets, budget, and data integration—will persist. However, as the value of a data lake for data management, analytics, and discovery become more widely publicized with feasible use cases, adoption and maturity are likely to increase.

As organizations uncover keys to success, best practices, and long-term data strategies, we believe that the data lake will become a more common and fundamental component of the data environment.

Sponsors

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities.

The Hortonworks Data Platform provides an open platform that deeply integrates with existing IT investments and upon which enterprises can build and deploy Hadoop-based applications.

Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop.

www.hortonworks.com

Teradata helps companies get more value from data than any other company. Our big data analytic solutions, integrated marketing applications, and team of experts can help your company gain a sustainable competitive advantage with data. Teradata helps organizations leverage all their data so they can know more about their customers and business and do more of what's really important.

www.teradata.com

Think Big, a Teradata company, offers big data roadmap, architecture, engineering and ongoing support services for data lake and analytic solutions.

A TERADATA COMPANY

http://thinkbig.teradata.com

Attunity:
Right Data. Right Place. Right Time.

Attunity is a leading provider of data management software solutions that enable moving, preparing and analyzing data efficiently to streamline operations, increase productivity and improve decision-making. Attunity enables data access, management, sharing and distribution of data, including Big Data, across heterogeneous enterprise platforms, organizations, and the cloud. The company's software solutions include data replication, data flow management, test data management, change data capture, data connectivity, file transfer and replication, data warehouse automation, data usage analytics, and cloud data delivery. For more information, visit www.attunity.com and join our communities: Twitter, Facebook, LinkedIn, and YouTube.

www.attunity.com

HP Security-Data Security drives leadership in data-centric security and encryption solutions. With over 80 patents and 51 years of expertise we protect the world's largest brands and neutralize breach impact by securing sensitive data-at-rest, in-use and in-motion. Our solutions provide advanced encryption, tokenization and key management that protect sensitive data across enterprise applications, data processing IT, cloud, payments ecosystems, mission critical transactions, storage, and big data platforms. HP Security - Data Security solves one of the industry's biggest challenges: simplifying the protection of sensitive data in even the most complex use cases.

www.voltage.com