research

BEST PRACTICES REPORT

Q1 2017

Data Lakes

Purposes, Practices, Patterns, and Platforms

By Philip Russom

tdwi

Transforming Data
With Intelligence™

Research Sponsors

Diyotta

HPE Security–Data Security

IBM

SAS

Talend

BEST PRACTICES REPORT

Q1 2017

Data Lakes

Purposes, Practices, Patterns, and Platforms

By Philip Russom

Table of Contents

About the Author

PHILIP RUSSOM, Ph.D., is senior director of TDWI Research for data management and is a well-known figure in data warehousing, integration, and quality, having published over 550 research reports, magazine articles, opinion columns, and speeches over a 20-year period. Before joining TDWI in 2005, Russom was an industry analyst covering data management at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and consultant, was a contributing editor with leading IT magazines, and was a product manager at database vendors. His Ph.D. is from Yale. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

About TDWI

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

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies and is supplemented by surveys of business intelligence professionals. To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving business intelligence problems or an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical BI issues. To suggest a topic that meets these requirements, please contact TDWI senior research directors Fern Halper (fhalper@tdwi.org), Philip Russom (prussom@tdwi.org), and David Stodder (dstodder@tdwi.org).

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who responded to our requests for phone interviews. Second, our report sponsors, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: James Powell, Lindsay Stares, James Haley, Michael Boyda, and Denelle Hanlon.

Sponsors

Diyotta, HPE, IBM, SAS, and Talend sponsored this report.

Research Methodology and Demographics

Report Scope. Many organizations are under pressure to capture, manage, and leverage both new big data and exploding volumes of traditional enterprise data. At the same time, many analytics applications demand that old and new data be consolidated at scale to enable broad data exploration and analytics correlations. The *data lake* has come forward as a new data-driven design pattern for persisting massive data volumes characterized by diverse data types, structures, sources, containers, and frequencies of generation.

Audience. This report is geared for business and technical managers responsible for implementing and modernizing data environments that consolidate both traditional enterprise data and new big data, a use case for which the data lake was created.

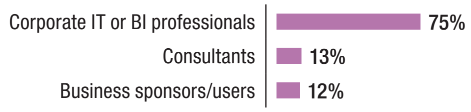
Survey Methodology. In November 2016, TDWI sent an invitation via email to the data management professionals in its database, asking them to complete an Internet-based survey. The invitation was also distributed via websites, newsletters, and publications from TDWI and other firms. The survey drew responses from 273 survey respondents. From these, we excluded respondents who identified themselves as academics or vendor employees. The resulting complete responses of 252 respondents form the core data sample for this report.

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

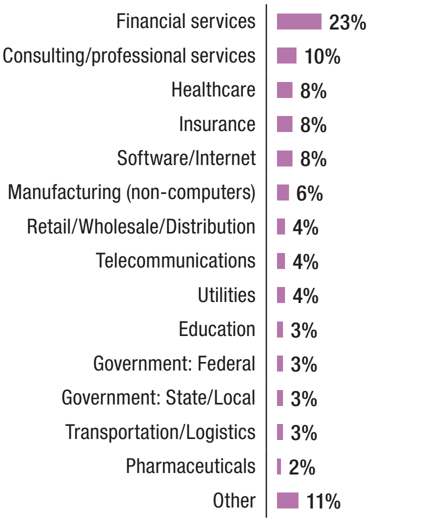
Survey Demographics. The majority of survey respondents are IT or BI/DW professionals (75%). Others are consultants (13%) and business sponsors or users (12%). We asked consultants to fill out the survey with a recent client in mind.

The financial services industry (23%) dominates the respondent population, followed by consulting (10%), healthcare (8%), insurance (8%), software/Internet (8%), non-computer manufacturing (6%), and other industries. Most survey respondents reside in the U.S. (49%), Europe (22%), or Canada (10%). Respondents are fairly evenly distributed across all sizes of companies and other organizations.

Position

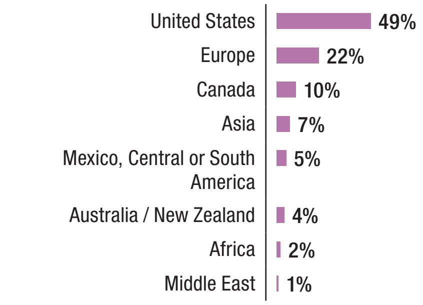


Industry

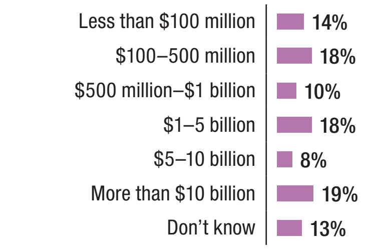


("Other" consists of multiple industries, each represented by less than 2% of respondents.)

Geography



Company Size by Revenue



Based on 216 to 252 survey respondents.

Executive Summary

A data lake is a collection of data organized by user-designed patterns.

When designed well, a data lake is an effective data-driven design pattern for capturing a wide range of data types, both old and new, at large scale. By definition, a data lake is optimized for the quick ingestion of raw, detailed source data plus on-the-fly processing of such data for exploration, analytics, and operations. Even so, traditional, latent data practices are possible, too.

Data lakes are already in production in several compelling use cases.

Organizations are adopting the data lake design pattern (whether on Hadoop or a relational database) because lakes provision the kind of raw data that users need for data exploration and discovery-oriented forms of advanced analytics. A data lake can also be a consolidation point for both new and traditional data, thereby enabling analytics correlations across all data. With the right end-user tools, a data lake can enable the self-service data practices that both technical and business users need. These practices wring business value from big data, other new data sources, and burgeoning enterprise data; these assets are not mere cost centers. Furthermore, a data lake can modernize and extend programs for data warehousing, analytics, data integration, and other data-driven solutions.

The business need for more analytics is the lake's leading driver.

The chief beneficiaries of data lakes as identified by this report's survey are analytics, new self-service data practices, value from big data, and warehouse modernization. However, lakes also face barriers, namely immature governance, integration, user skills, and security for Hadoop.

The data lake is top of mind for half of data management professionals, but not a pressing requirement for the rest. A quarter of organizations surveyed already have at least one data lake in production, typically as a data warehouse extension. Another quarter will enter production in a year. At this rate, the data lake is already established, and it will be common soon.

An explosion of non-relational data is driving users toward the Hadoop-based data lake.

Most users (82%) are beset by evolving data types, structures, sources, and volumes; they are considering a data lake to cope with data's exploding diversity and scale. Most of them (68%) find it increasingly difficult to cope via relational databases, so they are considering Hadoop as their data lake platform. Seventy-nine percent of users that already have a lake say that most of its data is raw source with some areas for structured data, and those areas will grow as they understand the lake better.

Data lakes are owned by data warehouse teams, central IT, and lines of business, in that order. Data lake workers include an array of data engineers, data architects, data analysts, data developers, and data scientists. One-third of those are consultants. Most full-time employees are mature data management professionals cross-trained in big data, Hadoop, and advanced analytics.

Most data lakes enable analytics and so are owned by data warehouse teams.

Most data lakes focus on analytics, but others fall into categories based on their owners or use cases, such as data lakes for marketing, sales, healthcare, and fraud detection. Most use cases for data lakes demand business metadata, self-service functions, SQL, multiple data ingestion methods, and multilayered security. Hadoop is weak in these areas, so users are filling Hadoop's gaps with multiple tools from vendor and open-source communities.

Hadoop outpaces relational databases as a platform for lakes, but some users use both.

There are two broad types of data lakes based on which data platform is used: Hadoop-based data lakes and relational data lakes. Today, Hadoop is far more common than relational databases as a lake platform. However, a quarter of survey respondents who have data lake experience say that their lake spans both. Those platforms may be on premises, on clouds, or both. Hence, some data lakes are multiplatform and hybrid, as are most data warehouses today.

To help users prepare, this report defines data lake types, then discusses their emerging best practices, enabling technologies, and real-world use cases. The report's survey quantifies users' trends and readiness for data lakes, and the report's user stories document real-world activities.

Introduction to Data Lakes

We're experiencing a time of great change as data evolves into greater diversity (more data types, sources, schema, and latencies) and as user organizations diversify the ways they use data for business value (via advanced analytics and data integrated across multiple analytics and operational applications). To capture new big data, to scale up burgeoning traditional data, and to leverage both fully, users are modernizing their portfolios of tools, platforms, best practices, and skills.

The data lake is a response to evolving big data and the need for more analytics.

One of the hotter areas in data modernization is the addition of data lakes to both greenfield and preexisting data ecosystems. Data lakes are already in production in many multiplatform data warehouse environments, advanced analytics applications, and the hybrid data ecosystems surrounding customer relationship management and sales force automation. TDWI feels that the data lake is here to stay, and many more organizations will adopt it for a growing list of use cases in enterprise analytics and operations.

The purpose of this report is to accelerate users' understanding of data lakes and their new best practices, use cases, strategies for organizing data lake projects, and data platforms and tools that are associated with lakes (whether from vendors or open source).

The Primary Characteristics of a Data Lake

Do you feel you know what a data lake is and does?

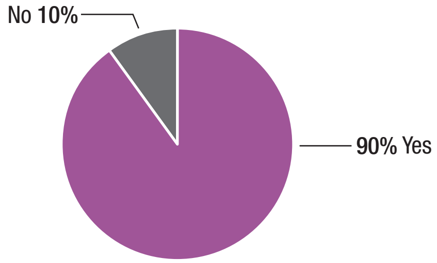


Figure 1. *Based on 225 respondents.*

Ninety percent of the people responding to this report's survey feel that they know what a data lake is and does (see Figure 1). This is impressive considering that the data lake is still very new. It's even more impressive when we consider how complex a lake can be and how closely it resembles other designs and practices in data management. To clear the confusion, let's define the data lake by examining its primary characteristics.

Users understand data lakes fairly well, even though they are new.

A data lake is a collection of data, not a platform for data. In the way that a database (defined as a collection of data and related elements) is managed with enterprise software called a relational database management system (RDBMS), a data lake is a collection of data (or multiple collections) that is usually managed on Hadoop, less often with an RDBMS.

A data lake is a design pattern populated with data, not a platform.

A data lake is a data-driven design pattern. A *design pattern* is a generalized, repeatable approach to commonly occurring situations in information technology solutions. Developers must flesh out a design pattern (based on current requirements) to create a finished solution. *Data-driven design patterns* tend to be localized data constructs such as data schema, models, tables, and record structures. They can also be data architectures, which tend to be large-scale combinations of multiple design patterns and other components. For example, consider data warehouse architectures, which include data marts, time series, dimensions, and operational data stores (ODSs). The data lake is an emerging data-driven design pattern, as are data vaults, enterprise data hubs, and logical data warehouses.

Most data lakes are on Hadoop, but some are on relational databases.

Hadoop-based data lakes versus relational data lakes. A common myth says that data lakes require open-source Apache Hadoop or a vendor distribution of Hadoop. The vast majority of data lakes deployed by users today are on Hadoop, and these are called Hadoop-based data lakes. However, a few data lakes are deployed atop RDBMSs, and these are called relational data lakes.

As we'll see in detail later in this report, some users need mature relational functionality in their data lake due to inherently relational analytics practices such as SQL-based analytics, SQL-based data exploration, SQL-based ELT pushdown, and online analytics processing (OLAP). Some of this functionality can come from tools layered onto Hadoop. However, for stringent relational requirements, some users prefer an RDBMS, plus the support for RDBMSs built into mature tools that users already have. In such situations, an organization may opt for the relational data lake, with no Hadoop involved.

Hadoop is the preferred platform for data lakes (but not the only one). Due to the great size and diversity of data in a lake, plus the many ways data must be processed and repurposed, Hadoop has become an important enabling platform for data lakes (and for other purposes, too, of course). Hadoop scales linearly, supports a wide range of analytics processing techniques, and costs a fraction of similar relational configurations. For these reasons, Hadoop is now the preferred data platform for data lakes—but not the only one.

On any platform, a data lake is usually large and ingests data quickly.

A data lake handles large volumes of diverse data. A data lake tends to manage highly diverse data types and can scale to handle tens or hundreds of terabytes—sometimes petabytes. This enables broad data exploration, the use of unstructured data, and analytics correlations across data points from many sources.

It ingests data quickly. A data lake ingests data in its raw, original state, straight from data sources, with little or no cleansing, standardization, remodeling, or transformation. These and other data management best practices can then be applied to the raw data very flexibly as future use cases demand. The early ingestion of data means that operational data is captured and made available for exploration, discovery, and reporting as soon as possible. Plus, data is landed and ready for transformations prior to loading elsewhere, such as into a data warehouse. Modern data is generated and pushed to the lake in multiple time frames and frequencies, so the lake's data integration infrastructure must support a long list of interfaces operating from nightly batch to intra-day microbatch to real-time and streaming.

The lake assumes that data will be repurposed often and unpredictably.

It allows for more data prep on the fly, not just ETL beforehand. A single data lake can support multiple functions. For example, the trend in data warehousing is to use the data lake as a new and improved zone for data landing and staging, while also using the same lake as a "sandbox" for broad data exploration and discovery-oriented advanced analytics. With data staging, all the old complex practices for ETL-ing data still apply (to get data ready for loading into warehouse structures, reporting, and OLAP). However, the new practice called data prep—an agile subset of data integration and quality functions—is required to construct new data sets on the fly while exploring and analyzing data. Data prep does not replace the established best practices of data management; both are required in a multipurpose, multitenant data lake.

It persists data in its original raw detailed state. A data lake focuses on detailed source data so that the source can be repurposed many ways as new requirements in advanced analytics evolve and emerge. This is important because the rapid pace of change within organizations and across marketplaces makes it difficult to foresee all the ways that data will need to be provisioned for analytics in the future. Even so, a data lake may also have areas within it for aggregated, transformed, and remodeled data. Similarly, so-called analytics sandboxes (populated by the lake's data) may or may not be persisted in the same lake.

It integrates into multiple enterprise data ecosystems and architectures. At the moment, TDWI regularly sees data lakes deployed as components within multiplatform data warehouse environments and hybrid data ecosystems for multichannel marketing. To a lesser degree, both Hadoop and the data lake are slowly entering other enterprise data ecosystems, including those for data archiving and content management.

Data-Driven Design Patterns that Resemble Data Lakes

Lakes, vaults, and hubs are similar, yet different.

A data vault is similar to a data lake. Like a data lake, a data vault is typically a large archive of detailed source data. Unlike a lake, the vault's data is standardized as it enters, to make data fit for purposes in data exploration and analytics. Furthermore, vaults usually have well-developed semantic and federated layers that provide additional structure in a virtual fashion. Data lakes tend toward this kind of light structuring after three or four project phases, whereas data vaults typically have such structuring designed into them from the first phase.

As noted earlier, lakes are usually on Hadoop, whereas vaults are almost always on large MPP configurations of RDBMSs. Hence, distinguishing a data vault from a relational data lake can be a hair-splitting semantic argument, at least at the platform level. Similarly, the data vault and the logical data warehouse have considerable overlap, in that both depend heavily on federated and virtual methods.

An enterprise data hub (EDH) can resemble a data lake. In a recent TDWI survey, among all the practices users could implement atop Hadoop, the one with the greatest anticipated growth over the next three years was the EDH.[1] As Hadoop users mature in their use of large repositories and other data-driven design patterns in a Hadoop environment, they need the governance, orchestration, security, and light structuring that a data hub can provide if they are to bring multiple use cases together in a controlled and governable multitenant environment.

Lakes and warehouses are increasingly used together.

The relational data warehouse and the Hadoop-based data lake coexist and complement each other. The two have substantial similarities and overlap, yet their strengths are mostly complementary, which is why an increasing number of data warehouse teams deploy both and integrate them tightly. In fact, a recent TDWI report revealed that 17% of surveyed data warehouse programs already have Hadoop in production alongside a relational data warehouse environment.[2] That's because the strengths of one compensate for the weaknesses of the other. They simply provide different sets of functions, thereby giving users twice the options. The table in Figure 2 compares and contrasts relational data warehouses and Hadoop-based data lakes.

[1] See the discussion around Figure 17 in *TDWI Best Practices Report: Hadoop for the Enterprise* (2015), online at tdwi.org/bpreports.

[2] See the discussion around Figure 16 in *TDWI Best Practices Report: Data Warehouse Modernization* (2016), online at tdwi.org/bpreports.

Relational Data Warehouse	Hadoop-Based Data Lake
RDBMS for relational requirements	Hadoop for diverse data, scalability, low cost
Data of recognized high value	Candidate data of potential value
Mostly refined calculated data	Mostly detailed source data
Known entities, tracked over time	Raw material for discovering entities and facts
Data conforms to enterprise standards	Fidelity to original format and condition
Data integration up front	Data prep on demand
Data transformed a priori	Data repurposed later, as needs arise
Typically schema on write	Typically schema on read
A priori metadata improvement	Metadata developed on read in many cases

Figure 2. *A generalized comparison of data warehouse and data lake characteristics.*

Compelling, Real-World Use Cases for Data Lakes

As we just saw, data lakes contribute heartily to a long list of technology scenarios, but what's in it for the business? The following real-world use cases answer that question:

Lakes enable exploration, discovery, and self-service.

Discovery of new insights and opportunities. Because big data usually comes from new sources, TDWI often refers to it as new data or new big data. The great promise and relevance of new big data is that it can be leveraged in new ways to develop new insights, which in turn can help organizations adapt to change in evolving business environments.

Self-service data exploration, data prep, and analytics. When a data lake (whether on Hadoop or RDBMS) is complemented with agile query tools and enhanced with business metadata, it can empower a broad range of users (even some business users) to explore new big data, build simple data sets, and create basic analyses. This is a high priority for many organizations.

Lakes enable new analytics and expand old ones.

Competing on analytics. New data-driven design patterns and data platforms integrate a broad range of data sources to create unique views into your customer base and marketplace.

Multichannel marketing. A mix of old and new data from websites, call center applications, smartphone apps, social media, third-party data providers, and internal touch points can reveal how your customers behave in diverse situations. The result is now being called the marketing data lake. It enables broad customer exploration and analytics, which improve the cross-selling, up-selling, account growth, acquisition, and retention that are goals for modern multichannel marketing.

Old and new data draw a more complete view of the customer. When an organization pursues the complete customer view—sometimes called the single view or 360-degree view—it amasses substantial data stores that are lightly structured. The view typically involves a wide record per customer, where each field quantifies a customer attribute, and each record is a row in a simple table. The "just enough" structure of this data (which needs some relational functionality, but not much) makes it ideal for a data lake, whether on Hadoop or an RDBMS.

Analytics with all the data. Given the right design patterns and data platforms, new big data can provide larger, broader data sets, thereby avoiding sampling errors and expanding existing analytics for risk, fraud, customer-base segmentation, and the complete view of the customer.

Real-time operations. This is so mainstream that there are now television commercials about how innovative firms capture and operationalize real-time data to approve insurance policies and residential mortgages in hours instead of weeks. A rep from one of these firms spoke at a TDWI conference, explaining the role of a relational data lake in real-time operations.

Data lakes can be extended to handle data in real time.

Sensor data and the Internet of Things (IoT). One of the biggest explosions of sensors is in businesses that rely on logistics. For example, a trucking firm used sensor data to prove how safely their drivers drive, which resulted in a million-dollar discount from their insurance company. Another trucking firm correlated sensor data with spatial coordinates to shorten delivery routes and delivery times, which boosted customer retention.

Streaming data. At a TDWI conference, a representative from a leading telco explained how data streaming from sensors—ingested into a data lake atop Hadoop—has brought unprecedented accuracy and timeliness to analytics for capacity planning, grid performance, and high availability. The company captures streaming data to spot performance trends that need immediate attention and persists streams for in-depth analysis later.

Decision-making value from unstructured text. The "killer app" for human language and other unstructured text is sentiment analysis, which has become almost commonplace as a new insight into customers and marketplaces. Larger, in-house solutions for sentiment analysis typically deploy a Hadoop-based data lake or something similar.

A data lake is a good choice for multistructured data.

Fraud and risk analytics. For example, a New York–based financial investment house loaded a few petabytes of emails into a data lake atop Hadoop. Using a variety of search and analytics technologies, they immediately discovered several cases of fraud and insider trading. The firm moved quickly to resolve these, thereby avoiding substantial losses and liabilities.

Analytics with miscellaneous server logs. Hadoop was originally designed to manage and process massive numbers of Web server logs. A data lake built atop Hadoop is ideal for scalable analytics with other logs as well—say, those from enterprise packaged applications.

Active data archiving. Archiving data on an enterprise level remains rather primitive, depending on ancient technologies such as magnetic tapes, optical disks, and offline processes. A Hadoop-based data lake can modernize archiving, so that it is online and accessible (with appropriate security) for searches and queries. This transforms an archive from an unused cost center to a valuable business tool.

USER STORY RETAIL MERCHANDIZING ANALYTICS IS A REAL-WORLD USE CASE FOR A DATA LAKE

"Our global business has multiple Hadoop clusters, each with a data lake. I work with a cluster located in the United Kingdom that supports analytics for the UK and our international retail businesses," said Zog Gibbens, an architect for data and analytics at global health and well-being enterprise Walgreens Boots Alliance. "Most of my work as an architect involves Hadoop because we are integrating it into some of our enterprise data ecosystems.

"The data lake's first production use case was studying the macro and micro spaces within each store, down to individual shelves and locations on them, to determine product sales performance and to optimize merchandizing. This application of the data lake improved analytics performance and delivered a business lift, especially from studying the propensity to purchase, resulting in a rise by a handful of percentage points on average.

"Here in the UK, our initial project with Hadoop and the lake was analytics for retail merchandizing for UK and international geographies. However, our colleagues in the U.S. started with Hadoop as an offload for their data warehouses."

Benefits and Barriers

Data Lake: Problem or Opportunity?

85% consider a data lake to be an opportunity, not a problem.

Most new best practices and technologies present previously unencountered challenges and problems. Users are right to ponder the balance of risk and reward before committing to new tech and practices. Such is the case with data lakes today. To gauge whether a data lake is worth the effort, this report's survey asked (see Figure 3): Is a data lake a problem or an opportunity?

The vast majority (85%) consider a data lake an opportunity. Responses to other survey questions reveal that users are very hopeful that data lakes will help them expand analytics programs, draw business value from new data assets, and extend data warehouses.

A small minority (15%) consider a data lake to be a problem. As we'll see in the next section of this report, many users are rightfully concerned about the difficulties of governing the content and use of a data lake. Similarly, they are aware of Hadoop's limited security functionality and their own nascent skills for Hadoop, big data, and data lakes. However, a growing number of organizations have worked through these concerns and barriers to embrace big data and similar new assets for business advantage.

Is a data lake a problem or an opportunity?

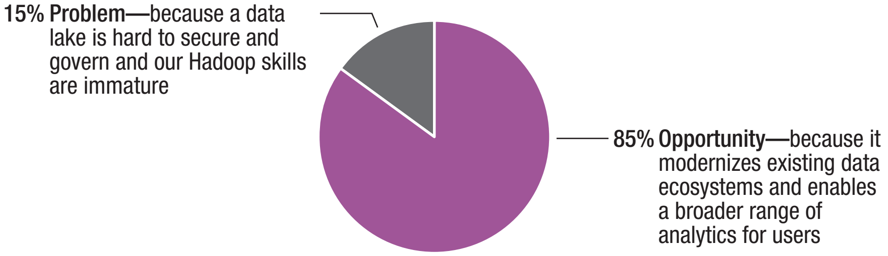


Figure 3. *Based on 237 respondents.*

Benefits of Data Lakes

The chief beneficiaries are advanced analytics, new data-driven practices, value from big data, and warehouse modernization.

In the perceptions of survey respondents, data lakes offer several potential benefits (see Figure 4). A few areas stand out in their responses:

Advanced analytics. The real driver behind most trends in IT and data management today is the growing number of firms, government agencies, and other organizations that need a broader range of analytics to compete, grow, retain customers, and achieve other organizational goals. Even when OLAP and older forms of analytics are in place, organizations need predictive and discovery-oriented analytics, based on advanced technologies for mining, clustering, graph, artificial intelligence, and machine learning. This report's survey is consistent with other surveys and studies from TDWI in that the most anticipated benefit of the data lake is advanced analytics (selected by 49% of respondents).

New data-driven practices. Tied for first place among benefits is the relatively new practice of data exploration (49%), sometimes called data discovery. The data lake can provide a scalable sandbox for exploring data integrated from multiple sources to discover new facts about the business and its customers, partners, and products. So they can study both old and new data, business and technical users alike are demanding data exploration along with other emerging practices that benefit from a data lake, such as self-service data access (24%) and data visualization (18%).

Business value from big data. Successful enterprises are not content to capture and manage big data and other new data assets as a cost center. Instead, they gain business value from new data, largely via analytics and reporting. A data lake can be a big data source for analytics (45% expect this benefit). Hadoop has become the preferred (but not exclusive) platform for big data and data lakes because adopters anticipate low-cost hardware and software (19%) and extreme scalability (19%).

Data warehouse modernization. Modernization continues to be a strong trend in data warehousing. Data lakes (whether on Hadoop or RDBMS) are regularly added to multiplatform data warehouse environments (DWEs) as part of the modernization process. Survey respondents agreed that a data lake can act as an extension of data warehouse storage (39%), as data landing and staging (36%), and as a strategy for data warehouse offload and cost reduction (34%). Likewise, the data lake can also be an extension of data integration (14%), often through pushdown processing.[3]

Diverse data structures. Another advantage respondents expect from a Hadoop-based data lake is the ability to capture and handle widely diverse data structures and file types (20%), including machine data from IoT, robots, sensors, meters, etc. (21%).

Other. One survey respondent selected "Other" and mentioned "quick access to data, [which] the business does not have today." Another respondent said, "We are likely to benefit [although] I have chosen to use a Web storage [provider] instead of Hadoop as data lake storage."

[3] For a more detailed discussion of the data lake's role in data warehouse modernization, see *TDWI Checklist Report: Evolving toward the Modern Data Warehouse* (2016), online at tdwi.org/checklists.

If your organization were to implement Hadoop-based data lakes, which of the following use cases would most likely benefit? Select six or fewer.

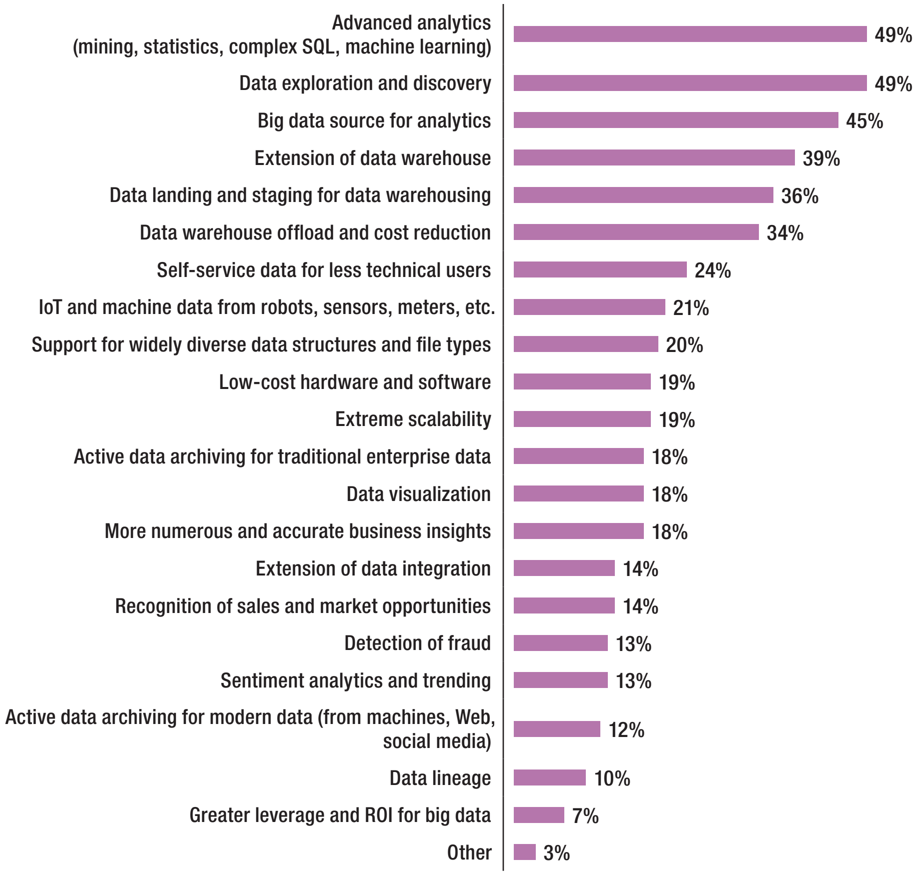


Figure 4. *Based on 1173 responses from 237 respondents. 4.9 responses per respondent on average.*

Barriers to Data Lakes

A data lake has its benefits, as we just saw. Unfortunately, it also has many potential barriers according to survey results (see Figure 5). The issues span across multiple areas:

Data governance. As noted earlier, the ungoverned dumping of data into a data lake can result in a so-called data swamp. Survey respondents are fully aware of this potential problem, and their responses rank a lack of data governance (41%) as their primary concern.

The leading barriers are governance, integration, lack of experience, privacy issues, and immature tech and practices.

Data integration. Data ingestion and its governance are critical success factors for a data lake, as discussed earlier. Survey respondents are accordingly concerned about their lack of data integration tools and skills for Hadoop (32%). The good news is that software vendors and the open source community have updated their data integration tools to support the interface, storage, and processing methods unique to Hadoop. In a related issue, data lakes need to be democratized by providing access for businesspeople and less technical users. Many users hope to address this issue with the self-service functionality and practices discussed elsewhere in this report.

Big data experience. The arrival of big data is what spurs most organizations to take an interest in data lakes and Hadoop. In these cases, users are new to big data, lakes, and Hadoop, and so they are naturally concerned about their inadequate skills for big data (32%), inadequate skills for Hadoop (32%), inadequate skills for designing big data analytics systems (24%), and inadequate skills for data lake design (23%). TDWI sees organizations rising to these challenges and succeeding by training existing data management employees, engaging consultants with big data experience, and, less often, hiring new employees with big data skills.

Business case. Data lakes are quite new and both business and technology personnel are still learning about them. This can lead to difficulty establishing a compelling business case (31%) or business sponsorship (28%). Obviously, a convincing business case is unlikely when the organization does not need a data lake (12%). TDWI sees successful business cases built on the business need for advanced analytics, broad data exploration, and value from big data.

Data privacy and compliance. A few survey respondents are concerned about a data lake's lack of data privacy compliance (17%) and the risk of exposing sensitive data such as personally identifiable information (28%). TDWI sees organizations overcoming such potential problems by extending their enterprise programs for data governance and/or stewardship to encompass the data lake, its data ingestion policies, and the usage of data in the lake.

Immature technology. Some organizations are aware of the immaturity of the data lake concept (27%), and so they are taking a "wait and see" position. As stated by one survey respondent, "[We are] still waiting to see what becomes best of breed in tools, i.e., Presto versus Spark and Hive versus Impala." A quarter or less of respondents are concerned specifically about Hadoop's immaturity with data security (26%), metadata management (24%), and ANSI-standard SQL (14%). Again, the vendor and open source communities regularly roll out advances in these areas, which should give users confidence.

Other. A number of survey respondents cited barriers based on "reluctance [from business users] to learn new tools" and "people's reluctance to change and allow something new."

In your organization, what are the most likely barriers to implementing Hadoop-based data lakes? Select six or fewer.

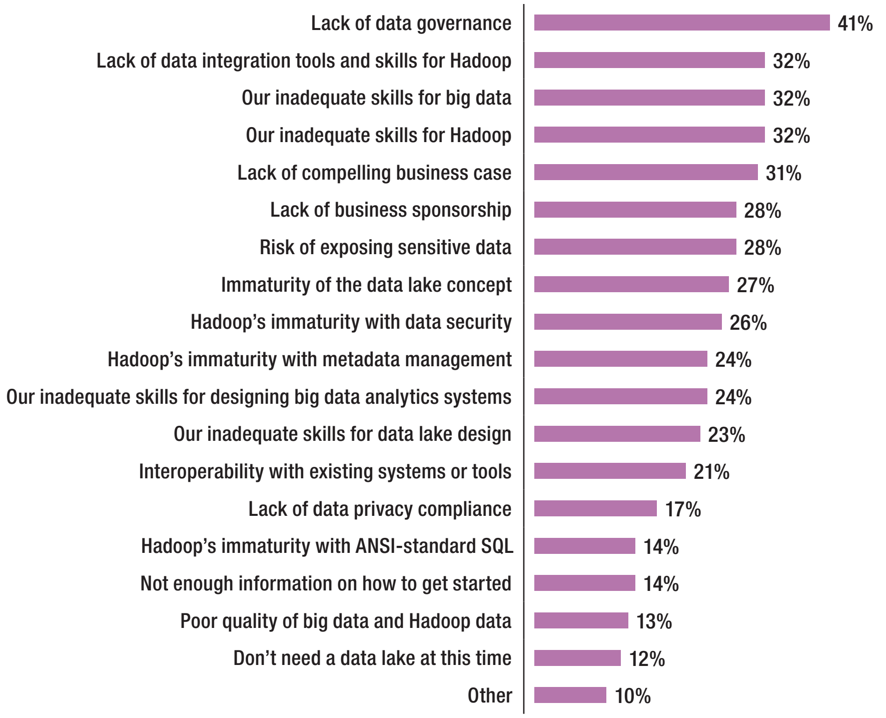


Figure 5. *Based on 1066 responses from 237 respondents. 4.5 responses per respondent, on average.*

USER STORY A DATA LAKE CAN BE STRATEGIC WHEN ALIGNED TO DELIVER SHORT-TERM AND LONG-TERM BUSINESS GOALS

"We've created an enterprise data strategy, supported fully by our CEO and executive vice presidents," said Shaun Rankin, the SVP of data management at Citizens Bank. "The data lake plays an important role in that strategy, and our current lake is sponsored personally by our chief data officer.

"Three drivers led us to a Hadoop-based data lake. First, our executives and their data strategy prefer a 'sourced once, used by everyone' approach, and the scalability of Hadoop, augmented with the right tools and governance, can enable that. Second, our technical users need to comingle diverse data from many sources for broad self-service exploration and analytics, and that's where the Hadoop-based data lake excels. Third, the low cost of Hadoop software and commodity hardware met our project's financial requirements.

"Based on those drivers, we selected a Hadoop distribution from a large vendor, and with their consulting help, we now have Hadoop and the data lake in production. Our plan will eventually integrate data from 75 account servicing systems into the data lake; we integrated 26 in 2016. Technical achievements in addition to our data lake so far include a new customer master with a handful of real-time interfaces. Business achievements based on these include improvements to commercial profitability, consumer applications, and risk mitigation. Upcoming goals include enterprise householding and self-service capabilities."

The State of Data Lakes

Why are Data Lakes important?

To gauge the urgency of data lake adoption, this report's survey asked respondents to rank the importance of the data lake relative to their organization's data strategy (see Figure 6).

The data lake is top of mind for half of data professionals.

How important is a Hadoop-based data lake for the success of your organization's data strategy?

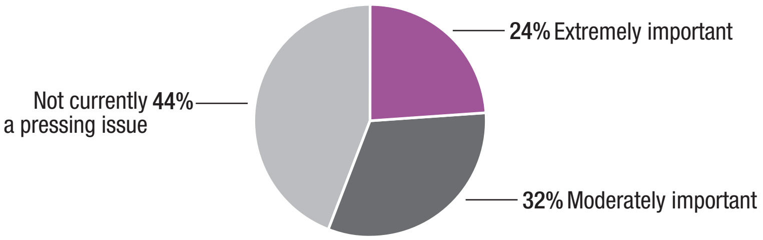


Figure 6. *Based on 225 respondents.*

For nearly half (44%) of survey respondents, the data lake is not a pressing issue. As one respondent put it, the data lake is "not important due to the lack of a business use case." Another said, "We don't have big data. Data quality is much more pressing." Yet another observed, "[Our] enterprise data is still fairly structured, and not high in the three Vs."

Over half of respondents (56%) recognize the importance of data lakes. A quarter (24%) feel the lake is extremely important, and an additional third (32%) see it as moderately important.

Why is the data lake important? To get their unvarnished opinions, we asked respondents to explain their answer using an open-ended question. The respondents' comments reveal a number of use cases, needs, and trends, as seen in the representative excerpts reproduced in Figure 7. Note that the quoted users work in many different industries and geographic regions. The data lake is top of mind for half of data professionals and their business sponsors in many contexts worldwide.

Users have specific and diverse reasons why the data lake is important.

In your own words, why is implementing a data lake important or not important?

- "[It is] important [to] update disparate data management solutions and position the organization to use modern tools to reduce time to delivery of data products." – Enterprise data architect, government, Canada
- "Implementing a data lake is important to enable our business to gain access and analyze data they have not been able to get in the past." – Enterprise data architect, insurance, U.S.
- "It is important to try to know our customers better, increase our analytics capabilities, and get value from data." – Chief data officer, financial services, Mexico
- "[A data lake is] important as storage for a variety of raw data and data types from various internal and external sources, [and it] can be implemented at a low cost." – Head of BI, financial services, Europe
- "[We] need flexible and cost-efficient storage for IoT data with great possibilities for scalable computing power [that is probably] cloud based." – Big data owner, manufacturing, Europe

- “Important so data can be homogenized and governed. Instead of [having] isolated repositories, a data lake can be the golden source if managed well.” – Big data lead, telecommunications, U.S.
- “[A data lake] allows end users a free-form, lightly governed analytics environment.” – Senior architect, professional services, U.S.
- “It has great potential to improve self-service and advanced analytics.” – Big data architect, telecommunications consultant, U.S.
- “Allows the rapid ingestion of large data volumes to be exposed to analysts for advanced analytics, data discovery, proof of concept, and visualization.” – Principal, professional services, Australia
- “The volume and diversity of data is not well supported by existing RDBMS technologies. Schema on write impedes ingestion of new data sources.” – Healthcare partner, professional services, U.S.
- “A data lake with an interface for [user access] will save a lot of time [versus] shredding and loading the data into a relational format.” – Senior director of risk analytics, financial services, U.S.
- “We need a hybrid data architecture that leverages our existing data management environment, and a data lake can be a big player.” – Data architect, telecommunications, Middle East
- “Important because we need to extend our classic data warehouse architecture with big data.” – Information architect, financial services, Europe

Figure 7. *Drawn from the text responses of 161 respondents.*

The Adoption of Data Lakes

The data lake is established and will soon become more common.

The survey asked respondents when they expect to have a data lake in production.

Roughly a quarter of organizations surveyed (23%, see Figure 8) already have a data lake in production. This shows that the data lake is already established as a real-world use case that works on a technical level and delivers value on a business level.

Another quarter (24%) anticipate a production data lake within twelve months. If this pans out, the number of deployed lakes will double within a year, and at that time roughly half of meaningful data programs will include a data lake. At this rate, in three years, as many as three-quarters of programs may include one or more data lakes. Admittedly, that’s rather optimistic; even if actual adoption falls short of the projection, the adoption rate will still be aggressive.

When do you expect to have a data lake in production?

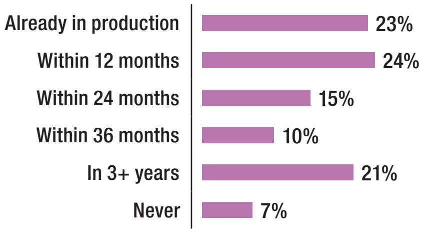


Figure 8. *Based on 252 respondents.*

USER STORY MANY ORGANIZATIONS ARE TODAY DEPLOYING THEIR FIRST DATA LAKE

"For our future data lake, we are currently working on a proof of concept project, with some hands-on work." said a data architect in a financial services firm in Germany. "We are planning on using Hadoop and the Apache ecosystem, and to this purpose we've selected one of the vendor distributions. Both during and after the proof of concept, we will have consultants with Hadoop and Apache experience to help us. Long-term, however, we plan to retrain existing employees with Hadoop and data lake skills, instead of hiring new ones. The data lake will be mostly built by us internally because we see this as a key differentiator for our business, and we are keen to develop and maintain these skills in-house.

"Our initial application for the data lake will be match-and-merge and quality checking of transaction records of various kinds, including the enrichment of data from disparate sources. We will also be experimenting with a number of technologies with regard to the reporting of data in a data warehouse environment. At a much later date, we hope to develop the data lake into a platform with which we can undertake further applications, such as analytics."

Characteristics of the Data in a Lake

Data is evolving. Survey respondents say that their data is evolving moderately (62%) or dramatically (20%, see Figure 9). Traditional data (mostly relational and structured, from standard enterprise applications) is being joined by new categories of big data, including data from sensors, handheld devices, machinery, Web applications, and social media. These bring with them new structures, interfaces, containers, and latencies.

The diversification of data is one of the issues a data lake addresses.

In turn, data management best practices are evolving to address the new data. Users are turning to new design patterns such as data lakes, vaults, and hubs on both old and new data platforms—namely RDBMSs and Hadoop—to accommodate the capture, storage, processing, analytics, and delivery of big data and other new data assets.

Is your data evolving, in terms of the diversity of its structures, data types, sources, management, and business uses?

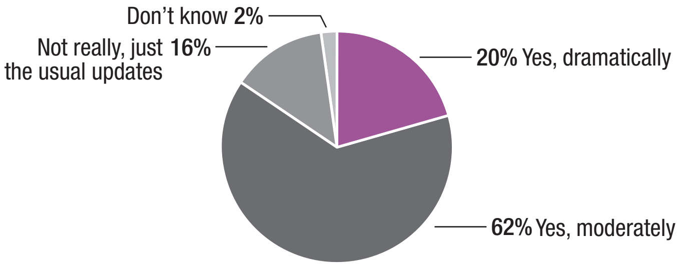


Figure 9. *Based on 225 respondents.*

Relational databases handle some types of big data well, other types not so well. Consequently, two-thirds of survey respondents say they are finding it increasingly difficult to manage new data in an RDBMS (see Figure 10). For the most part, it depends on the new data's structure or container.

Data's source structure can determine which data platform is best for a data lake.

For example, most machine data or streaming data is pushed out one message at a time, and each message contains a straightforward record structure. Even when messages are processed in real time, their records are also captured and appended to a log file or table for later processing and analytics. Some users choose to manage such data in an RDBMS-based data lake (or a traditional relational data warehouse), especially when their primary exploration and analytics methods demand relational processing, such as SQL and/or OLAP. To make this arrangement practical, however, the RDBMS platform usually grows into a substantial MPP configuration, which is very expensive in terms of software licenses, hardware, and maintenance payroll. For cost reasons, users may choose to take this lightly structured data to Hadoop and make do with Hadoop's nascent relational processing.

Finally, note that fully unstructured data—such as human language text from call center apps or social media—is not a good fit for RDBMSs, and over the years TDWI has seen many organizations try and fail to manage text in BLOBs (Binary Large OBjects) stored in RDBMSs. Similarly, container-based semistructured data—as in XML and JSON files—has had limited success in RDBMSs. Such data types tend to drive users toward Hadoop-based data lakes instead of RDBMSs.

As you adopt new big data's breadth of data types and structures, do you find it increasingly difficult to fit data into relational databases?

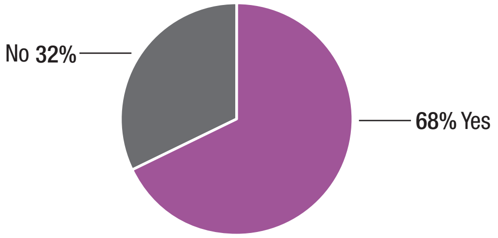


Figure 10. *Based on 225 respondents.*

Hype says that lakes are only for big data, but in fact they are for other data categories, too.

Surprise! There's more in the average data lake than just big data. Many data lakes manage a fairly even mix of modern and traditional data (39%, see Figure 11). More data lakes than not contain mostly traditional enterprise data (45%). TDWI believes that these two statistics hold true for both Hadoop-based and RDBMS-based data lakes. However, it seems likely that Hadoop is the platform when a lake contains mostly big data and other modern data (15%).

According to survey responses, the exclusive management of big data and other nontraditional data is a minority practice for data lakes (15%), whereas managing mostly traditional data is the majority practice (45%). TDWI suspects this will shift in the next few years, such that mixing modern and traditional data will become the leading practice. The change will be driven by more big data coming online and by more users understanding how to integrate and link old and new data accurately. Don't forget: as discussed earlier, the real driver is *analytics*. When you mix old and new data from diverse sources in diverse structures, the analytics correlations become highly detailed, resulting in richer facts and insights for the business to leverage.

Which of the following best describes the origins of data managed in your lake?

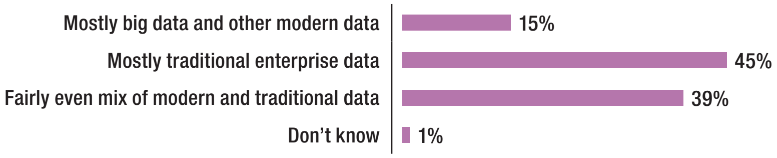


Figure 11. *Based on 72 respondents who have data lake experience.*

Data lakes manage mostly raw data, but ironically also include areas for structured data. According to this survey, the minority practice is to use the data lake to exclusively manage raw detailed source data (18% in Figure 12). The majority practice (a whopping 79%) is to have both raw source data and areas in the lake devoted to structured data. In an RDBMS-based data lake, the raw source would be in simple but large tables, containing millions of transactions—telco call detail records, customer events from CRM applications, etc. Other tables would contain the complete view of customers, prepped data for reports, etc.—data that was calculated, aggregated, refactored, or restructured from the raw source. In a Hadoop-based data lake, the same data may be there, but the raw source is in an eclectic mishmash of files and containers (at massive scale), and the restructured data is managed as Hive tables and/or HBase row stores (which are tiny by comparison).

It's a balancing act. A lake is faithful to data's raw state but also supports simple structures.

This combination isn't new. Many third normal form data warehouses and operational data stores fit this description. Plus, you get a similar effect when analytics sandboxes (which usually restructure data) are stored in the same lake from which their data came.

Furthermore, the structured data areas in an otherwise raw data lake can be a product of maturity or simply multiphased project planning. As users work with any data set, they come to understand which data needs to be reviewed or reported repeatedly, and they build target data structures accordingly.

Another way to think about it is to remember that reporting and analytics are two different practices. The lake's raw data is there for discovery-oriented exploration and analytics, whereas the structured areas are there for regularly refreshed reports and dashboards. (There are occasional exceptions, as when users perform analytics on structured data in a lake.) This division of data also supports a data integration scenario where the raw data is in a landing zone (which may double as an archive for analytics) and the structured areas are where data is staged prior to loading into a data warehouse, CRM application, etc.

Which of the following best describes the state of data in your lake?

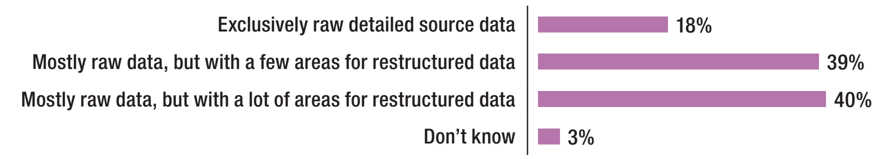


Figure 12. *Based on 72 respondents who have data lake experience.*

Data volumes in lakes are growing, as with any data-driven design pattern today. To quantify the growth rate, this report's survey asked: What's the approximate total volume that your organization manages in your Hadoop-based data lake(s), both today and in three years? According to survey results, in Hadoop-based data lakes today, most volumes range from 1TB to 100TB (66% of respondents, see Figure 13). However, in three years, these same data lakes will have expanded such that most volumes will range from 100TB to over 1PB (73% of respondents).

Data lakes are headed for the "petabyte club."

Note that the survey question specified Hadoop-based data lakes, and these aggressive growth rates are consistent with other TDWI studies involving Hadoop. For RDBMS-based data lakes, TDWI also anticipates impressive growth rates, but perhaps not as aggressive as in Hadoop environments.

What's the approximate total volume that your organization manages in your Hadoop-based data lake(s), both today and in three years?

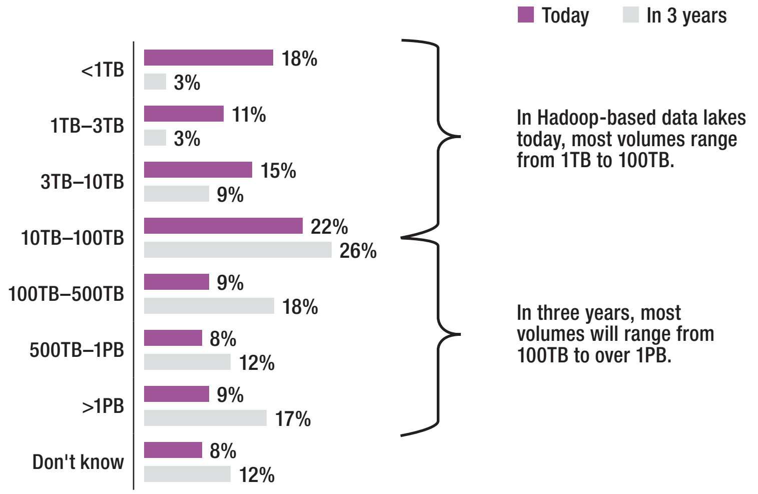


Figure 13. *Based on 71 respondents who have data lake experience.*

USER STORY A DATA LAKE CAN ABLY HANDLE DATA FROM MACHINES, IOT, AND STREAMS

"We've been planning our data lake for two years, and we look forward to going into production in another year," said a data architect at a firm that provides machines which manufacturers use for the injection molding of plastic products. "We collect data from our injection machines, so we can study it to understand how the machines perform, how our customers use them, and how we can improve the machines. It's not a lot of data, but the data is complex because it tracks many traits for each machine. All our machines are custom built, so each generates custom data structures.

"In the first phase, we will use our data lake to aggregate and improve our machine data, then develop new analytics that will improve our machines and customer service. Some of the machine attributes we'll study are the temperature of plastic resin, measure cycle times, oil pressure, and the speed of certain machine functions.

"Top-tier customers have configurations where data streams out of the injection machine. In a future phase of the lake project, we will capture streaming data in the lake so we can monitor the machine in real time and proactively alert the client of actions they should take. Years from now, we hope to have a suite of analytics applications our clients can use to study their machines' performance and compare them to similar ones based on data in the lake."

Organizational Matters

Data Lake Owners

Data warehouse teams and central IT are common data lake owners.

Because most data lakes are deployed on Hadoop, the ownership of the lake is similar to that of Hadoop. TDWI surveys have revealed aggressive adoption of Hadoop among data warehouse teams, which has outpaced adoption by other parts of the enterprise, including central IT organizations.[4] This trend with Hadoop ownership is reflected in Figure 14, which indicates that

[4] See the discussion around Figure 1 in *TDWI Best Practices Report: Hadoop for the Enterprise* (2015), online at tdwi.org/bpreports.

data warehouse groups (31%) are the leading owners of data lakes, with central IT (29%) in second place.

Several respondents selected "Other" (19%). Owners they identified include chief data officers, data governance organizations, data competency centers, and groups for data science and analytics. Just as some data warehouse groups report to business functions, some data lakes are owned by the finance or marketing departments.

Who owns and maintains your data lake?

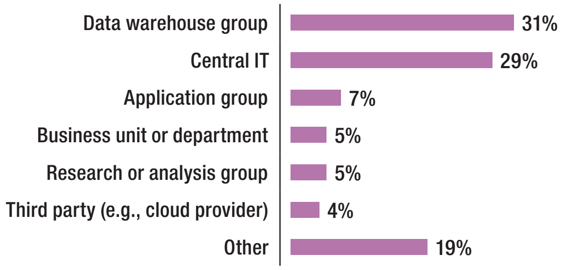


Figure 14. *Based on 75 respondents who have data lake experience.*

Data Lake Workers

According to our survey (see Figure 15), data lake and Hadoop workers include engineers (23%), architects (20%), analysts (18%), and developers (15%). As you might guess, most of these people have job titles that start with the word *data*. For example, most of the analysts are data analysts, but a few are systems analysts. Likewise, most of the architects are data architects, although some are solutions architects.

Of course, the data scientist is present (12%), as in other environments that involve advanced analytics and/or Hadoop. As with any data environment, there's a need for managers (6%), database administrators (4%), and data integration specialists (2%).

The job titles listed in Figure 15 show that data lakes are built and used by very technical people. Yet the users quoted in Figure 7 assume that a data lake should be democratized to enable a broad range of somewhat technical and nontechnical users. Hence, the data lakes in use today are, on average, not satisfying the requirements of business users.

Given that the data lake is still a new and emerging data-driven design pattern, it's possible that early project phases focus on enabling a relatively small constituency of data analysts and developers. If so, then it's possible that later phases will deploy self-service functions and practices that address the needs of larger constituencies of less technical business users. Organizations need to keep in mind that the business value and ROI of a data lake is not complete until the full range of users is enabled. Equally important, the data lake's role in innovation and change is not in full motion until all users are on board.

Enter the job titles of people who design and use data lakes using Hadoop technologies (140 characters maximum):

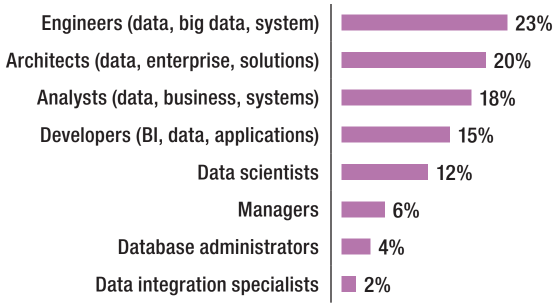


Figure 15. *Based on 100 responses from 61 respondents who have data lake experience. 1.6 responses per respondent, on average.*

Hiring and Training for Data Lake Skills

Fill the skills gap by training employees and hiring consultants.

Hiring new employees has proved ineffective (7% in Figure 16). For one thing, there aren't many data professionals available for hire who have hands-on experience with Hadoop and/or data lakes. For another thing, the few who exist (especially those who are data scientists) command high salaries.

Cross-training existing data management personnel is the most effective approach (38%). This avoids the perils of trying to find qualified new hires. Most data management professionals enjoy learning new skills. In addition, many managers of data teams find it easier to allocate workers when all of them are cross-trained.

Engaging consultants is a solid staffing approach (33%). As a matter of policy, many IT and data management teams bring in experienced consultants whenever they do anything that's new to them—and the data lake is very new. The consultants reduce risk, accelerate delivery, and teach employees new skills as they work through the project. Furthermore, there are many consulting firms available that have recently expanded their big data, analytics, data warehouse, and Hadoop practices to encompass data lakes.

Other (22%). Most of the comments entered by respondents described a scenario where employees are trained with new skills while the team is temporarily augmented by consultants who have data lake and/or Hadoop expertise. Sometimes the training is provided by a third party, and sometimes the training is a "knowledge transfer" process from consultants to employees.

How is your organization staffing Hadoop and data lake design? Select all that apply.

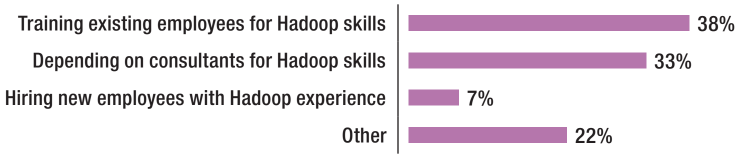


Figure 16. *Based on 73 respondents who have data lake experience.*

Governing a Data Lake to Keep It from Becoming a Data Swamp

Deteriorating into a data swamp is a real and likely peril for a data lake.

Note that it's easy to dump data in a data lake without ensuring an audit trail, data integrity, data quality, governance, stewardship, data life cycle management over time, and data protection. When a data lake is not managed and governed properly, it deteriorates into a data swamp, which is an undocumented and disorganized data store that is nearly impossible to navigate, trust, and leverage for organizational advantage. However, this risk is easily mitigated by data management best practices in data governance, curation, and stewardship. Furthermore, these people and process best practices can be automated by governance functions within data management software that help governance be more scalable, precise, and collaborative.

As with any important data asset, the data of a lake should at least be curated by a data steward who is responsible for driving improvements in the data. TDWI feels that the best stewards are businesspeople (not technical staff) because they can prioritize based on business need and keep data management work aligned with business goals. With data lakes, improvements should give priority to metadata, data quality, data lineage, just enough structure, and diversifying data and tools.

Data stewardship and governance are cures for the data swamp.

Stewardship aside, another viable remedy is the collaborative data governance program (which may subsume stewardship). The data lake is like any data platform or data store in that it needs data governance (DG) to keep its technical standards and business compliance high. DG usually takes the form of a board or committee, populated with a mix of data management professionals (who create enterprise standards for data) and business managers (who serve as data owners, stewards, and curators with a focus on compliance). All these people collaborate to establish and enforce policies that ensure data's compliant access, use, security, standards, and trust.

Most DG policies are applied and enforced via people and process. Others can be automated by the metadata management, business rules, and logic of data management solutions to detect, alert, and remediate when a situation does not conform to DG policies.

Implementers of a data lake should work with their enterprise DG board so that the lake and its data will comply with established DG policies. This should be done prior to designing the lake and loading it with data. Given that data lakes differ from older data store types, it is probable that yours will require new DG policies or revisions of older ones.[5]

Protecting a Data Lake via Data-Centric Security

It's not a matter of if—it's a matter of when your data lake will be hacked.

In the old days, a hacker breached your system, did little or nothing, and left. The hacker then bragged to peers about the accomplishment in public forums for everyone to hear. Those days are long gone.

Modern cyberattackers are far more ominous, especially when they operate as data thieves. They don't just hack their way in; their goal is to steal data so they can use it to plan larger and more lucrative crimes, such as identity theft. Cyberattackers are now organized and well-equipped with the tools, know-how, and technology to rapidly extract terabytes of high-value data assets from enterprises. Some even sell their expertise to nation-states. The enterprise is left with serious—even catastrophic—damages and liabilities that can easily take years and lots of money to remediate or litigate.

De-identify data so it's useless to cyberattackers and data thieves.

However, such risks and liabilities can be alleviated by de-identifying enterprise data so that internal users, applications, and business processes can work with data in its protected form, but external parties cannot read it at all. That way, when data is stolen, the thief has nothing to sell or commit a crime with.

[5] For an in-depth discussion of collaborative data governance, see *TDWI Checklist Report: Governing Big Data and Hadoop* (2016), online at tdwi.org/checklists.

With data-centric format-preserving encryption or masking technology, protection is applied at the data field and subfield levels. This preserves the characteristics of the original data, including numbers, symbols, letters, and numeric relationships (e.g., date and salary ranges), and it maintains referential integrity across distributed data sets. The de-identified protected form of the data can then be used in applications, analytics, data transfers, and data stores while being readily and securely re-identified when applications and users require it. Thus, when a data breach occurs, the enterprise has very little cleanup to do (except to tighten security), no public relations nightmares, and no costs for recovery or liabilities. The effects of a data breach are neutralized.

Hadoop is known to be an effective platform for data lakes. However, it's even better known for lacking enterprise-grade security functions. In its pure open source versions, Hadoop supports little more than Kerberos-based authentication, whereas it also needs data-centric security. Hence, new approaches to data-centric security are key to securing Hadoop-based data lakes and other Hadoop scenarios, such as data warehousing and marketing.

Given the escalating war on cybercrime, security upgrades in the form of data protection are strongly recommended for Hadoop installations. Furthermore, data should be protected before or during ingestion into Hadoop and its data lake. Luckily, security add-ons and upgrades for Hadoop are available from a number of vendors. When these are combined with open source ingestion tools, such as Apache NiFi, organizations can encrypt streaming IoT data at scale before it enters a Hadoop-based data lake for analytics.

Authentication is not enough. Hadoop needs other layers of security, too.

What should users look for in security tools? Consider at least three layers of security:

1. Hadoop (and by extension, data lakes built atop it) needs standard perimeter protection in the forms of authentication and authorization, as any IT system would.
2. It may also be useful to record an audit trail of access by users and tools that when studied can reveal security violations and enlighten capacity planning, data archiving procedures, and chargeback accounting. Operational metadata for the Hadoop environment, as described elsewhere in this report, can enable such audits.
3. Unlike the user-centric or application-centric security just mentioned, the data-centric security layer operates on or near the data to cleanse, block, and de-identify sensitive or high-value data. Again, the point is that the intruder gets nothing of value when stealing data from a secured Hadoop-based data lake.

Tools are available today that offer options for de-identification based on technologies for data encryption, masking, and tokenization. Look for tools that can de-identify data before it is landed in Hadoop. Also look for de-identification methods that allow users to read and analyze data in its protected form yet prevent cyberattackers and data thieves from reading and monetizing that same data. Finally, if your IT organization has standardized on LDAP or other directory services (as many have done), your Hadoop security approach should conform with IT standards when possible.[6]

[6] For a highly detailed explanation of data-centric security, see *TDWI Checklist Report: Data-Centric Security* (2016), online at tdwi.org/checklists.

> **USER STORY FOR DATA LAKES, DATA LANDING IS A COMMON BENEFIT AND DATA GOVERNANCE IS A COMMON CONCERN**
>
> "We provide IT consulting to a wide range of clients, but we also have pretty demanding needs for internal IT," said a senior IT engineer at a global consulting firm. "One of our internal teams is called 'The Landing Zone,' because their specialty is landing data and staging it for multiple applications. Their work increasingly includes Hadoop, big data, and data lakes. In fact, a recent project lands and integrates a few million records a day in a Hadoop-based data lake. The lake, in turn, supports data exploration and trending analytics, plus dashboards built with popular data visualization tools.
>
> "Now that we've seen how well the Hadoop-based data lake works, I personally would like to see a giant data lake stretching across the whole company. I really feel it could replace tons of data marts and provide a single repository that's easier to manage and control. Data governance issues are front and foremost for data lakes with me. I feel confident a data lake can be governed, but I'm having a bit of trouble convincing people. I think they'll come around."

Best Practices for Data Lakes

Categories of Data Lakes

TDWI has noticed that data lakes deployed in the real world tend to fall into recurring categories based on the larger application or data ecosystem the lake integrates with, the data domain managed by the lake, the department that commissioned the lake, or the industry the lake serves. Such distinctions are important to consider when designing a data lake because they affect the scope, tool functionality, and data content of a lake as well as its sponsor, funding, and user constituencies. Some examples follow.

Analytical data lakes. This can be as simple as a standalone, single-tenant data lake built for one analytics application ranging from sentiment analysis to money laundering detection. In other cases, the lake fulfills multiple analytical purposes, as in modern data warehousing, where a data lake supports data landing and staging (for data integration), data archiving and sandboxing (for data exploration and analytics), and warehouse extensions (which complement the central storage of the core warehouse).

Most data lakes have an analytical purpose and possibly an operational one, too.

Analytics aside, most data lakes have some departmental, data domain, or industry purpose in mind. Physically speaking, the data of these lakes may be located in one enterprisewide Hadoop cluster or RDBMS configuration (or on multiple platform instances). However, logically speaking, they are separate data lakes, as described below.

Marketing data lakes. This is a hot trend right now as marketers discover that a data lake is excellent for making analytical correlations and predictions across multiple customer channels, which in turn leads to higher conversion rates in cross-selling. The same lake can also enable new levels of granularity and insight for customer-base segmentation and complete views of customers. When extended with additional tools for orchestration, security, and governance, a marketing data lake may also serve as a customer data hub, which controls the compliant improvement and flow of customer data across an enterprise.

Similar to analytics apps and data marts before them, some data lakes focus on specific departments.

Sales performance data lakes. With the right data, a data lake can be a powerful prospect database for telemarketing, direct mail, and ad hoc prospecting. If the data lake also integrates historical data about long-term customers, a savvy account manager can come up to speed quickly when inheriting and servicing a mature account. A sales performance data lake is also useful for improving the accuracy of sales forecasts and optimizing sales territory assignments.

Some data lakes focus on the challenges and opportunities of a specific industry.

Healthcare data lakes. Industry-specific data lakes are emerging. For example, at a recent TDWI Conference, a representative from a U.S.-based healthcare provider talked about how a data lake enables researchers to study patient outcomes. The same lake also enables home caregivers to develop lists of high-risk patients (e.g., elderly, diabetic, or disabled) whom they target with specific services (e.g., nutrition, exercise, or physical therapy).

Financial fraud data lakes. Detecting fraud is challenging when it depends on correlations across numerous data points collected from multiple sources that represent a variety of entities and actions, strewn across extended time frames. After all, discovering that a single person was in close proximity to multiple accidents can indicate insurance fraud. Likewise, correlating people and companies can indicate collusion, insider trading, fraudulent transactions, or money laundering in financial services. A data lake helps by being the platform where all this diverse data comes together for broad data exploration and advanced analytics, usually with tools and technologies for mining, clustering, graph, and statistics.

Metadata Management for Data Lakes

To get full technical functionality and business value from a data lake, users need metadata management that supports multiple forms of metadata. This is a challenge for Hadoop-based data lakes because metadata management is still somewhat nascent in Hadoop. Some users prefer a relational data lake due to the mature metadata management of RDBMSs and related tools. Regardless of the lake's platform, TDWI finds most data lake users getting the rich metadata management they need via vendor products, especially data integration platforms. In Hadoop environments, many users and software vendors turn to open source Apache Atlas or solutions from Hadoop vendors such as Cloudera with Navigator.

Technical metadata is not enough. Many data lake practices require business metadata.

Technical metadata. This is the language that applications and their proprietary interfaces use to describe data, its structure, and lineage. Technical metadata usually consists of acronyms and codes, which is fine for technical developers or for software processes that access data without human intervention. Yet many user types find technical metadata indecipherable, including the business analysts, marketing specialists, and other business domain experts who are users of lake data.

Business metadata. Today, a growing number of nontechnical or mildly technical users want to work hands-on with data. Instead of raw technical metadata, this user constituency needs business metadata that employs human language descriptions of data that businesspeople can understand. Note that business metadata is created by technical users. For the technical user to create business metadata that's truly useful and accurate, mappings between metadata types should be based on a governed business glossary of terms that specifies the data owned by the business, expressed in industry and corporate standard language.

For many users, the point of implementing a data lake is to enable several self-service best practices including data access, exploration, discovery-oriented analytics, data prep, visualization, and analytics. Note that all these emerging self-service practices rely heavily on well-designed business metadata. Without it, nontechnical users cannot search and query a lake's data in a self-service fashion, which is key to getting full business value from the lake.

Advanced data environments need auditable operational metadata and tools for developing metadata on the fly.

Operational metadata. This form of metadata records details about data events such as source, access, load, alteration, copy, and movement. This is important in today's hybrid data ecosystems where data moves around a lot in the multiplatform environment, from the landing and staging to sandboxes to analytics tools, and then out to reports or to load target databases such as a warehouse. Hence, operational metadata can assist with data lake issues such as data lineage, redundancy, usage, collaboration, auditing, and security. Furthermore, operational metadata assists when governance requires a record of who (or what tool) accessed which data and when within the data lake (and other data stores).

Deduced and developed metadata. Many forms of big data lack metadata, for example, data from logs, streams, and social media. Users should look for tools (from vendor or open source communities) that can help data explorers and developers deduce and develop metadata as they search, query, and profile data in a lake. These "schema on read" functions must save deduced and developed metadata in a central repository for many users and applications to share. For Hadoop-based data lakes, look for tools that can "inject" metadata into files managed by Hadoop to make those files more readable and queryable.

Early Ingestion and Later Data Prep

Early ingestion gets the data in faster for earlier use and value. As more users want to conduct analytics with relatively fresh data, it's important to have a data-driven design pattern optimized for quick, continuous, and automated data ingestion—and that's why users are turning to data lakes. That way, data is available ASAP for exploration, reporting, analytics, and business monitoring. To speed ingestion, more data is being ingested in its original raw state or with minimal improvement up front; instead, raw data is processed and improved later, often on the fly in a new practice called data prep. Users don't spend as much time as in the past improving data on the off chance that it gets used. For new best practices in early ingestion and data prep to work properly, emerging design patterns (e.g., data lake, vault, or EDH) must handle a wide range of latencies for data ingestion, from traditional overnight batch to true real time.

Today's data ingestion is fast, simple, and focused on raw data collection.

Early ingestion has its benefits, yet you should not give in to the temptation to "dump" large amounts of arbitrary data into a data lake without proper data governance and metadata management. Instead, you should select data carefully, then document and improve it as it is ingested for data lineage and audit, metadata development, volume or node assignment, and compliance. Even within these strictures, a data lake can still be true to its primary mission, which is to provide detailed source material for exploration, analytical correlation, and future repurposing of data. Again, we see that finding the balance between raw data and restructured data is part of a data lake's design and the balance is a critical success factor.

Evolving best practice: How much do we transform data while ingesting it?

Many latencies of ingestion are possible with a data lake, but most ingestion cycles are batches or micro-batches executed a few times per day. Strictly speaking, this is not true real time, but it's a near-time best practice that suffices for business processes that benefit from data that's a few hours old. Relevant use cases include business monitoring, data exploration, and fuzzy metrics for performance management.

As users reach maturity with their Hadoop-based data lakes, many of them extend the lake into real-time operations so they can capture streaming data and react via real-time analytics and alerts. TDWI sees Hadoop users implementing Spark and Storm for these use cases. Both Hadoop and relational data lake users may get their real-time functionality from vendor products for complex event processing (CEP), event stream processing, or operational intelligence. In sophisticated implementations, the data lake can be a hub for real-time data in support of fast-paced business operations.

Data prep gets the data out easily and then quickly prepares it for the next purpose. A growing number of users need *data prep*, which is a select subset of the rich functionality that a mature, general-purpose data integration tool provides. The data prep subset is presented via a

Data prep is one of several new self-service practices seen with data lakes.

business-friendly user interface with business metadata (as a complement to technical metadata) designed for self-service data access and the creation of simple data sets.

Data exploration, prep, and analysis often go together in a multistep self-service process. As users access and explore data, they want to prepare a data set quickly based on what they discovered, then share the prepped data set with colleagues or seamlessly take it into analytics. These classes of users typically assume that metadata management (and especially business metadata) will be available in their data lake's toolset because data prep and other self-service practices require business metadata.

Data Integration Issues

Data prep is a practical expedient that complements sophisticated but daunting data integration.

Data lakes need both new data prep and mature data integration (DI). Note that the emerging tools and best practices for data prep (just discussed) do not replace the traditional best practices and mature tools of traditional DI, quality, complex aggregation, and advanced data structures (as required by most standard reports and data warehouses). The two are designed for very different classes of users in different development contexts. Although the old and new practices complement each other, both are required for comprehensive data management that meets the needs of today's diverse user constituencies, including data lake users who need self-service data exploration, data prep, and analytics.

DI's data flows stitch together the complex ecosystems where most data lakes live.

Successful data lakes depend on significant DI infrastructure. Given the great diversity of data types and sources seen in data lakes, DI must support a wide range of interfaces, platforms, data structures, and processing methods. Furthermore, DI helps define the key characteristics of a data lake such as early ingestion, metadata management, and repurposing data on the fly for queries, exploration, data prep, and other ad hoc practices.

Finally, most data lakes are integrated into a multiplatform data ecosystem similar to those for modern data warehousing and multichannel marketing. Data moves a lot in these ecosystems as it travels through ingestion, aggregation, analytical processing, target databases, and application synchronization. DI infrastructure is one of the golden threads that hold these complex data ecosystems together by enabling complex data flows. Likewise, DI drives data through the multiple zones that a mature data lake includes for landing, staging, data domains, analytics processing, and specific analytics applications.

Look for DI tools that are updated to support Hadoop's multiple interfaces.

For the success of Hadoop-based data lakes, DI tools must provide deep Hadoop support. A DI tool must make Hadoop easier to work with by supporting multiple interfaces to Hadoop, namely MapReduce, Pig, Hive, Sqoop, Spark, and so on. The tool should also enrich Hadoop with metadata (as discussed earlier) and enable pushdown processing into Hadoop without any programming required. To help future-proof the data lake, the DI tool should support all distributions and versions of Hadoop, plus provide updates for them.[7]

Adjust Data Management Best Practices to Fit the Data Lake

Older data management practices usually apply to data lakes but with some changes.

Perform more ad hoc data management. For example, instead of taking a week to design a data model, many users now create the model and populate it with data while they explore data in a data lake or similar data-driven design pattern. This is possible because today's advanced hardware and software have the speed and scale required to process and repurpose data on the fly at runtime, plus more agile development. Other examples of runtime data management include:

- Developing and capturing metadata as you ingest or explore the data in a lake[8]
- Applying data prep functions after exploration and before analysis

[7] For a more detailed discussion of data integration's role in successful data lakes, see *TDWI Checklist Report: Emerging Best Practices for Data Lakes* (2016), online at tdwi.org/checklists.

[8] For examples of how smart toolsets are enabling new directions for metadata management, see *TDWI Checklist Report: Governing Big Data and Hadoop* (2016), online at tdwi.org/checklists.

Data quality (DQ) efforts must be finessed with a data lake. Data standardization is the most commonly applied DQ task, yet it runs the risk of losing data values and structures that some forms of analytics need. For example, analytics algorithms—from fraud detection to customer-base segmentation—regularly look for outliers and nonstandard data. These "data nuggets" are too often stripped out during standardization, aggregation, and quality tasks before populating target databases, such as data warehouses. Data lakes persist data in its original state so that data nuggets are there for the kinds of exploration and analytics that need them.

Today, most data lake users limit DQ functions to light standardization for the sake of "just enough structure," as discussed earlier. This (along with metadata development) makes the lake's data easier to read for the query-driven processes of data exploration and SQL-based analytics. Data prep and query performance are likewise enhanced.

TDWI is confident that data lake users will eventually apply a broader range of DQ functions as data lakes and their uses reach maturity. In fact, users of the so-called marketing data lake are leading the way because the customer data domain is prone to DQ problems, and DQ functions are required to repurpose lake data for campaigns, segmentation, and syncing with enterprise applications.

Just Enough Structure

Users have now lived with data lakes long enough to know that the pure raw state of data seen in phase one of a project may give way to the structuring of some (but not all) data in later phases of the data lake.

Structure may increase over a data lake's life cycle.

For example, a marketing data lake may start by collecting data about customers drawn from numerous preexisting sources. Data is typically dumped from tables (in CRM, SFA, or marketing applications) or databases (containing complete views) or file-based behavioral data (containing Web logs or social media). Soon after, marketers and data analysts explore the big data and diverse data and "prove the concept" by discovering new facts, correlations, and insights. As a next step, they plan how to operationalize these discoveries into recurring tasks for reporting, analytics, operations, and data products. That's when they finally understand what kinds of structures need to be imposed on what data sets to get the full business impact and ROI from the data lake.

That process sounds similar to what we do in data warehousing, but there are significant differences. First, relatively few data subsets within the lake are restructured and remodeled. In other words, the vast majority of the data lake remains true to its original mission: to collect raw material for discovery and unforeseeable repurposing. Second, the resulting structures are simple, usually a few tables and keys or a record format for flat files. They get just enough structure, so they are fit for purpose with queries for exploration, recurring reports, set-based analytics, and data sync with operational applications.

A little structure is good. A lot can be bad.

Moderation is the rule. Just enough structure goes a long way and adds significant value. However, excessive structure may impede the data lake's mission of enabling discovery-oriented analytics.

Organize lake data in multiple "data zones." Users can add just enough structure to the design of a data lake by organizing data in multiple zones. In order of priority would be data landing zones (each tuned to an ingestion method), discovery zones (similar to analytics sandboxes), sync zones (where data is staged for sync with operational applications), and consumption zones (where data is certified ready for outbound use by users with appropriate authorization). Similarly, a marketing data lake may evolve zones for customer behavior analytics, customer-

Structure may be expressed by platform mechanisms or virtualization.

based segmentation, and 360-degree views of customers. The designer of the lake may organize the zones into groups for analytical versus operational use cases. Many variations are possible.

In a relational data lake, volumes and partitions may be defined via the utility tools of the RDBMS, or users can simply agree that certain tables represent certain zones. In a Hadoop-based data lake, users may define similar volumes, assign data to certain nodes, or design Hive tables or HBase row stores as homes for specific zones.

Express structure via virtualization without altering the raw source. Today's definition of the logical data warehouse relies heavily on logical approaches based on virtualization, federation, views, replication, and indexing. These approaches are also good for imposing just enough structure on a data lake without altering the original data. For example, the latest generation of data visualization tools (when pointed at lakes and other data sources) regularly executes federated queries and manages the result sets as materialized views.

> **USER STORY THE RELATIONAL DATA LAKE ENABLES THE REUSE OF PRIOR RELATIONAL SOLUTIONS AND OPTIMIZES WITH RELATIONAL CLOUDS**
>
> "We provide software that captures, integrates, and analyzes data from hardware for real-time location services—or RTLS for short," said Howard Fulks, a business intelligence architect at U.S.-based Intelligent InSites. "Most RTLS devices are tags and sensors of some sort, which may be standalone or embedded in machines, on wired or wireless networks. These devices transmit various data about the date and time, perimeter conditions (e.g., temperature), and entity identification (person, equipment type, supplies). However, our specialty is to track the location of entities so that healthcare workers can quickly and accurately find the patients, healthcare staff, and medical equipment that are critical to efficient, quality healthcare operations.
>
> "In upcoming months, we will roll out our first data lake for internal use as a consolidation point for exploring, discovering, and analyzing the spatial data our systems collect from RTLS devices. We hope to eventually have a secure multitenant data lake (probably cloud based) that our customers and partners can access for both operational and analytical purposes.
>
> "We decided that a relational data lake is a better fit for us as compared to a Hadoop-based data lake. That way, we can quickly fold in many of the data schema we've already developed for relational applications, and we can tap our deep skills for relational technologies. Furthermore, the relational data lake is something we can build and use in-house on our favorite RDBMS, on a small scale, then easily migrate to a cloud-based version of that RDBMS when our scalability and external access requirements demand it."

Data Lake Platforms and Architectures

Hadoop-Based Data Lakes Versus Relational Data Lakes

Straightforward questions can help you select a data platform appropriate to the data lake you'll design.

Before selecting a data platform for a data lake, determine the lake's relational requirements, especially those involving RDBMSs. Be sure to ask your team: Do we need advanced RDBMS data management functions, such as OLAP, materialized views, and complex data models (dimensional or hierarchical)? Do we need mature RDBMS functions, such as metadata, indexing, security, volumes, and partitioning? For ELT pushdown, will we have processing that demands an RDBMS (say, for complex table joins)? If your team answers yes to some of these questions, then you should at least consider an RDBMS for your data lake platform.

However, there are just as many compelling reasons to consider Hadoop for your data lake platform. Is your team under tight cost restrictions? Will the lake push the extremes of scalability? Does your team's culture work well with open source? Will the lake manage lots of file-based data? Does your team need a repository that can execute "in situ" a broad range of algorithmic analytics? Are relational requirements minimal for the data lake? Answering yes to some of these questions suggests you should consider Hadoop.

To get a sense of the platform choices data lake users are selecting in the real world, this report's survey isolated a subset of respondents who have direct experience designing or using a data lake. We asked that subset what type of platform their data lakes are on (see Figure 17).

For the data lake you use most, what type of data platform is it deployed on? Select only one.

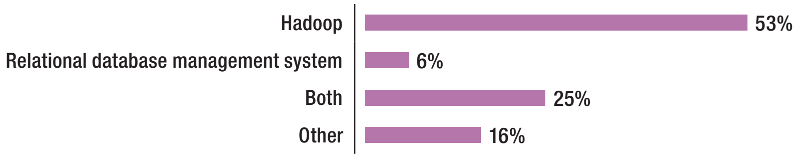


Figure 17. *Based on 75 respondents who have data lake experience.*

Hadoop is the preferred platform for data lakes (53%). This position was asserted earlier in the report, and here's the survey data to support that claim. In user interviews conducted for this report, lake users consistently cited linear scalability for storage, algorithmic analytical processing, and low cost as their primary decision points for choosing Hadoop.

A data lake can be deployed atop Hadoop, a relational platform, or both.

The relational data lake is real, but it's a minority practice compared to the Hadoop-based lake (6%). Relational lake users interviewed explained that their intense RDBMS and SQL requirements made the decision for them. However, TDWI has found other users that cite Hadoop's weaknesses as decision points, namely poor security, metadata, and SQL support.

Some data lakes span multiple platform types and instances. This explains why a quarter of survey respondents (25%) report using both Hadoop and an RDBMS. TDWI regularly encounters users with a modern data warehouse environment (DWE) that incorporates both (plus additional platforms based on columns, graph, clouds, etc.). In some DWEs, the relational warehouse manages multiple terabytes of raw detailed source data that has stringent relational requirements (along with the usual warehouse data for reports, OLAP, dashboards, etc.), while Hadoop manages most other data at scale. You can think of these as two data lakes or as one logical data lake that is physically distributed.[9]

Other (16%). Most of the respondents who selected "Other" entered a brand of cloud-based RDBMS or cloud-based Hadoop for their data lake's storage platform. This reminds us that cloud-based data platforms are here to stay and should be considered in any evaluation process.

SQL On Hadoop Versus SQL Off Hadoop

Let's look at even more questions to ask before selecting a data platform for a data lake. When looking into relational requirements, also look at your requirements for SQL. Is ANSI-standard SQL required? Do the lake users assume or demand SQL support? Do you anticipate using reporting or analytics tools that demand ANSI SQL? Will a substitute such as Hive on Hadoop suffice? How complex will the ad hoc queries of data exploration get? Do you need the mature query optimization of an RDBMS? What are the query performance requirements?

[9] For a detailed discussion of multiplatform data ecosystems, see *TDWI Best Practices Report: Evolving Data Warehouse Architectures* (2014), online at tdwi.org/bpreports.

Consider both SQL on Hadoop and SQL off Hadoop.

Also keep in mind that one of the hottest debates concerning Hadoop and SQL is the distinction between:

- **SQL on Hadoop:** SQL executes natively as a process inside Hadoop
- **SQL off Hadoop:** Tools outside Hadoop query Hadoop data[10]

TDWI puts the two together under the phrase *SQL for Hadoop*. Our survey asked Hadoop users to gauge their level of need relative to SQL for Hadoop (see Figure 18).

Which of the following best describes how you need SQL for Hadoop supported?

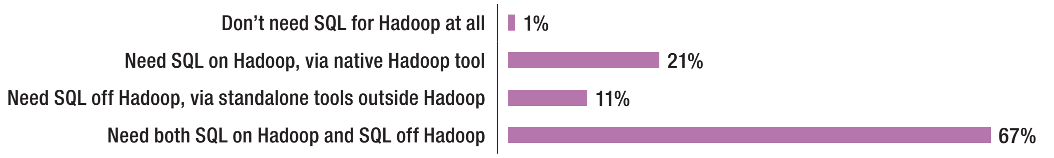


Figure 18. *Based on 72 respondents who have data lake experience.*

Hadoop users need SQL support, and most need multiple approaches.

Almost all Hadoop users need some kind of SQL support. A mere 1% report they have no need for SQL for Hadoop. This is not surprising given that most of Hadoop's use cases involve practices that tend to be wedded to SQL, namely decision making, exploration, and analytics.

The majority need both SQL on Hadoop and SQL off Hadoop (67%). There's a preference for SQL on Hadoop (21%) instead of SQL off Hadoop (11%). However, most Hadoop users surveyed need both. This makes sense for a variety of reasons. A data lake (especially a multitenant or multipurpose one) may need to satisfy the requirements of diverse user types, which leads to diverse, multiple approaches to data and tools. Furthermore, a diverse range of data integration and quality tools rely on SQL, too, and a data lake needs these for diverse requirements in data ingestion and data flows. Even traditional data warehouse and reporting environments include multiple SQL-based tools for the same reasons. Thus, the data lake is no exception.[11]

Why is SQL so critical to data lakes and similar design patterns?

SQL just had its 30th birthday, and it's as relevant now as ever.

SQL continues to be the language of data. Other query languages (namely, object query language and XQuery) have tried to displace SQL but have failed. Although Hive on Hadoop is being used a great deal by Hadoop users, most of those users are application developers. Data management professionals prefer SQL, no matter the platform.

SQL is familiar and it works. Many people (both business and technical) have SQL skills, which they use daily.

Tons of tools that could be used with a lake support ANSI SQL. These include tools for exploration, reporting, analytics, visualization, data integration, and data quality. Many users already have these in their software portfolios, and they wish to use them with a lake.

Data exploration is a top priority for most data lake users. Most of them want to explore the lake via ad hoc queries based on ANSI SQL. Without SQL, exploration is harder and slower, yielding less business value from a lake's data.

[10] For a related debate, replay the 2016 TDWI Webinar "SQL for Hadoop: When to Use Which Approach," online at tdwi.org.

[11] Note that the need for both SQL approaches to Hadoop is not new. See the discussion around Figure 15 in *TDWI Best Practices Report: Hadoop for the Enterprise* (2015), online at tdwi.org/bpreports.

Some emerging practices for self-service with lake data assume SQL. Besides data exploration, this also includes self-service analytics, visualization, and data prep. SQL aside, let's not forget that, at the other end of the spectrum, some users perform exploration and analysis via newer technologies for associative engines, semantic technologies, and natural language processing (NLP).

Data Lakes on Clouds

As noted earlier, cloud-based data platforms are here to stay and should be considered in any evaluation process, even when Hadoop is being considered. To gauge that progress, our survey asked Hadoop users among the respondents where their cluster is deployed (Figure 19).

Hadoop-based data lakes are typically on premises (52%) without a cloud. This survey result is consistent with most Hadoop users that TDWI has interviewed in recent years, who follow the traditional system integration route of assembling their own hardware and network into a cluster in-house. As an alternative, some hire a consulting or system integrator firm to do this work.

Hadoop is a bit behind in the trend toward cloud-based data platforms, but it will catch up.

Some Hadoop clusters are on a private cloud (16%). The term *private cloud* usually refers to an on-premises cloud set up by a user organization. Curiously, TDWI regularly finds private clouds for operational databases and data warehouses but not Hadoop. Perhaps as Hadoop proliferates it will appear on more private clouds.

Third-party clouds (12%) and managed service providers (7%) are emerging as Hadoop providers. TDWI has long noted the slow-moving trend toward third-party clouds for a wide range of IT systems and sometimes entire IT departments (as is common with IT outsourcing). As just noted, Hadoop is still somewhat new and is a bit behind that trend. Even so, TDWI is confident that within a few years Hadoop will be as common on various clouds as any IT system.

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

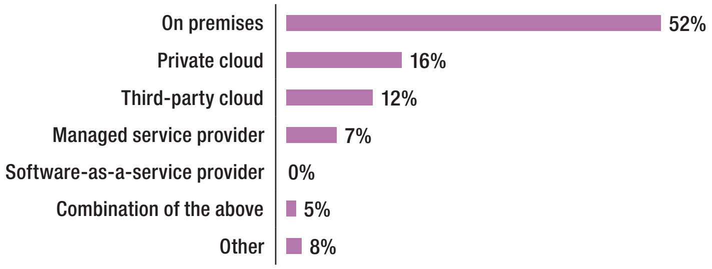


Figure 19. *Based on 75 respondents who have data lake experience.*

An elastic cloud has much to offer a data lake. Start-up costs are low and time-to-use is short with a cloud-based data lake solution. Plus, cloud-based data lake solutions tend to be highly available. However, the leading benefit is that a truly elastic cloud automatically provides and recovers server resources for optimal speed and scale as data processing and analytics workloads ramp up and subside. For the data professional, this greatly minimizes capacity planning and performance tweaks.

Data Lakes Integrated into Enterprise Data Architectures

Many data architectures today are multiplatform or will be soon.

The strongest trend in data warehouse architectures today is to diversify the portfolio of data platforms so that technical users can choose just the right platform for storing, processing, or delivering data sets and the products based on them, such as reports, analyses, and visualizations. In the modern multiplatform data warehouse environment (DWE), almost all core warehouses still run on RDBMSs, but they are integrated with other platforms—typically Hadoop and specialized RDBMSs (based on appliances, columns, clouds, or specific analytics such as graph).

This architecture is already established. A 2014 TDWI study shows that only 15% of data warehouses are on a single instance of one database brand. Most other data warehouses have diversified into a few platforms (37%) or many platforms (16%). At the other extreme, some "data warehouses" consist of many diverse purpose-built platforms without a true warehouse in the mix (15%).[12]

Monolithic warehouse architectures are being replaced by multiplatform ones.

In this hybrid environment, the core warehouse continues to be the preferred platform for reporting (from standard reports to dashboards), dimensional data (for OLAP, cubes, star schema, etc.), and data that requires extensive improvement or accuracy (e.g., financial reports).

However, the raw data for advanced forms of analytics is progressively being stored and processed on the other platforms of the DWE. This offloads the core warehouse so it can scale and focus on data that requires mature relational functionality (as reports and dimensional data do). This also takes raw detailed data to platforms that are well suited to advanced forms of analytics (based on mining, clustering, statistics, graph, etc.) at scale and at a reasonable cost.

The Hadoop-based data lake is emerging as a natural fit for the large volumes of data for advanced analytics that are being relocated as organizations modernize their DWEs. The trend is toward having the Hadoop-based data lake be the ingestion platform and analytics archive for the DWE, while sandboxing and set-based analytics may be done on specialty RDBMSs (but perhaps on Hadoop, too) and reporting and related functions are provisioned by the core warehouse.

The Hadoop-based data lake is becoming common in data warehouses.

The Hadoop-based data lake fits these and other trends quite well. The real driver is that enterprises need a broader range of analytics types so they can get better at making fact-based decisions, optimizing their organizational performance, and competing on analytics.

A Hadoop-based data lake is a standalone data platform, yet it coexists with and is tightly integrated with other data platforms in enterprise data architectures, as seen in the DWE example just discussed. Other examples include those found in marketing departments, multimodule ERP systems, and data-driven supply chains. Figure 20 pulls together and summarizes the many platforms and technical processes mentioned above into an architecture that TDWI sees regularly in DWEs and similar hybrid data ecosystems.

[12] See the discussion around Figure 10 in *TDWI Best Practices Report: Evolving Data Warehouse Architectures* (2014), online at tdwi.org/bpreports.

Architectural Summary of the Modern DWE, which includes a Hadoop-Based Data Lake

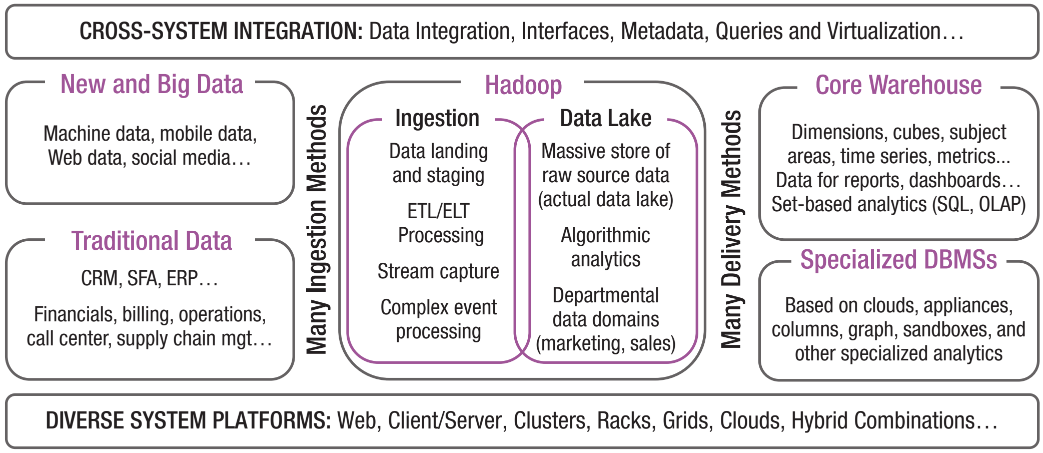

CROSS-SYSTEM INTEGRATION: Data Integration, Interfaces, Metadata, Queries and Virtualization…
New and Big Data
Machine data, mobile data, Web data, social media…
Traditional Data
CRM, SFA, ERP…
Financials, billing, operations, call center, supply chain mgt…
Many Ingestion Methods
Hadoop
Ingestion
Data landing and staging
ETL/ELT Processing
Stream capture
Complex event processing
Data Lake
Massive store of raw source data (actual data lake)
Algorithmic analytics
Departmental data domains (marketing, sales)
Many Delivery Methods
Core Warehouse
Dimensions, cubes, subject areas, time series, metrics...
Data for reports, dashboards…
Set-based analytics (SQL, OLAP)
Specialized DBMSs
Based on clouds, appliances, columns, graph, sandboxes, and other specialized analytics
DIVERSE SYSTEM PLATFORMS: Web, Client/Server, Clusters, Racks, Grids, Clouds, Hybrid Combinations…

Figure 20. *Source: TDWI.*

USER STORY THE JOURNEY TO COMMERCIAL SUCCESS VIA ANALYTICS ON DATA LAKES LEADS SOME FIRMS TO CLOUD-BASED SOLUTIONS

“One of the areas we support is epidemiology. It uses real-world data from medical records and insurance claims to support epidemiological and health economics studies, with a focus on the disease states our products treat,” said a senior director of big data and analytics for a large global pharmaceutical manufacturer. “It’s about enhancing our analytics capabilities broadly, not just about the data volume of big data. We figured that out early on, and the analytics must lead to some kind of commercial or operational success.

“To better support commercial success, we realized we needed to broaden our analytics approaches and scale up our stores of analytics data, so we started a pilot with a vendor distribution of Hadoop in the cloud. We continued to support other disparate platforms and repositories as well as proprietary tools for some applications. Yet we also wanted to test the feasibility of a data lake approach, leveraging the cloud and open source.

“Our analytics pilot drove business and technical benefits immediately, which convinced us that we made the right decisions concerning open source solutions (especially Hadoop), data lakes, and clouds. However, we suffered a series of problems with the vendor, so after more evaluations, we’re now migrating to a different cloud-based data platform with broader capabilities. The new platform has worked out so well with advanced analytics that the data warehouse team has decided to migrate to a cloud-based data warehouse. Thus far, we’re happy with cloud-based data solutions, and we anticipate expanding our cloud commitment.”

Vendor Platforms and Tools for Data Lakes

The firms that sponsored this report are all good examples of software vendors that offer tools, platforms, and professional services that are regularly involved in data lake design and use. The sponsors form a representative sample of the vendor community, and their offerings illustrate different approaches for growing and enhancing data lakes and related systems.[13]

Diyotta

Diyotta Inc. provides the Diyotta Modern Data Integration Suite, an integrated environment of multiple tools with a single console for all data management functions, from reusable designs to managing deployments. Diyotta's vision addresses the modern data, architectures, and practices that are discussed in this report. For example, the Modern Data Integration Suite is built for modern data ecosystems, which assume a mix of traditional and new data, structured and unstructured data, and very big data volumes and complexity. It's also built for today's multiplatform data environment, which assumes many databases and other data platforms of diverse types, vintages, and strengths from both vendor and open source communities on clouds and on premises. The Modern Data Integration Suite enables modern practices including data prep, data blending, and emerging data-driven design patterns (such as the data lake) whether on Hadoop, Spark, relational platforms, or a hybrid combination. To leverage the strengths of diverse data platforms, Diyotta takes the processing to where the data lives. This enables both hub and point-to-point architectures for design flexibility.

Hewlett Packard Enterprise (HPE)

HPE provides the wide range of software and hardware products and services that firms need as they transform into digital enterprises. All of HPE's offerings may be applied to data lake design and deployment, but a few stand out. For example, HPE's portfolio for big data analytics includes Vertica Advanced Analytics (for SQL-based analytics), IDOL (for unstructured data), and Haven OnDemand (for cloud-based big data analytics). For data-centric security, HPE provides advanced encryption, tokenization, and key management, which protect virtually unlimited sensitive data types across enterprise applications, including native support and integration for over 20 platforms. HPE SecureData protects the world's largest brands and neutralizes breach impact by securing data at rest, in use, and in motion. For data lakes on Hadoop or relational databases, HPE recommends that data-centric security be applied during ingestion. That way, authorized users can analyze the lake's data securely, but it's unusable by unauthorized parties, thereby alleviating business risk and liability. HPE Security simplifies the protection of sensitive data, even in the most complex use cases.

IBM

IBM sees the need for greater speed and scale as user organizations expand their data management solutions to embrace big data, leverage it through new analytics, and adopt hybrid architectures, which nowadays include a data lake and Hadoop. For these use cases, IBM offers IBM InfoSphere BigInsights, which enhances open source Apache Hadoop to make it enterprise-grade and to support the practices of data lakes, including data exploration, data prep, visualization, advanced text analytics, and data integration across both traditional sources and new big data sources. To enable deep analytics on lakes with big data, IBM offers PureData System for Analytics (an integrated and optimized appliance for analytics) and IBM dashDB (a managed data service on the cloud). Data streams are important sources for some data lakes and related data-driven design patterns, so IBM offers InfoSphere Streams, a complex event

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

processing system that supports the continuous capture and analysis of real-time data. Related tools for Hadoop and lakes from the IBM InfoSphere BigInsights product family include BigSQL and BigR.

SAS

SAS is a leader in big data management and advanced analytics solutions that help customers make better decisions, faster. The following capabilities—all supported by SAS—can enable data management and analytics for your data lake and other valuable repositories:

Data integration. Data movement, in-database processing, and native data access to traditional and emerging data sources such as Hadoop.

Data quality. Cleanse, standardize, and enrich data in real time and batch with prebuilt rules.

Self-service big data preparation. Business users profile, cleanse, and transform data on Hadoop without writing code.

Business glossary and metadata management. Track lineage, business rules, descriptive details, and workflow for improved governance of your data assets.

Event stream processing. Analyze real-time streams of data in motion for better decisions.

Data virtualization. Provide blended, secure views of your data without moving it.

Hadoop support. Access, deliver, and process data inside Hadoop across both the data management and analytics life cycle.

Visualization and advanced analytics. Deliver cutting-edge visualization and analysis capabilities without requiring analytical skills.

Talend

Talend delivers data lake agility for business and IT. Using Talend, organizations can collaborate more effectively on all their on-premises and cloud data lake projects using a future-proof, integrated data platform.

Talend was the first to deliver a real-time big data platform powered by Apache Spark and is continuing to innovate with its contributions to the Apache Beam project, the future of real-time processing. Talend takes the complexity out of integration efforts, so customers can ingest all their data at any speed and easily add metadata and trace lineage for their lakes. Companies can turn raw data into trusted insights through collaborative data governance and accelerate time to decision with smart data pipelines in real time.

Top 12 Priorities for Data Lakes

In closing, let’s summarize the findings of this report by listing the top 12 priorities for data lakes, including a few comments about why these priorities are important. Think of the priorities as recommendations, requirements, or rules that can guide user organizations through a successful implementation of a data lake.

1. **Consider a data lake for its business benefits.** For business users, a data lake is all about *analytics*. Even when a business has some forms of analytics (e.g., OLAP), they progressively need more advanced forms (e.g., predictive, mining, graph) to keep pace with evolving markets, customer bases, partners, and competitors. Similarly, a growing number of data-

savvy business users demand self-service data access, exploration, and visualization. Data lakes are known for early ingestion, which empowers a business to see and react to information sooner. A well-formed data lake with the right end-user tools can satisfy these business requirements.

2. **Consider a data lake for its technology benefits.** For technology users, the data lake is all about *free-form data aggregation*. That's because the discovery-oriented exploration and analytics that businesses are pining for today need large samples of data, lightly restructured (if at all) and aggregated from numerous sources. That's what the data lake is designed to do at scale.

3. **Know your data requirements and choose a data platform accordingly.** Even if you think you'll go with Hadoop, start by compiling your relational requirements in case those trump Hadoop. Also, that will help you plan your tooling for SQL on Hadoop and SQL off Hadoop. Expect to do both when your query requirements are diverse. Use the questions listed in this report to kick-start your requirements list for both relational and Hadoop requirements.

4. **Consider a hybrid architecture for your lake.** Never forget that the data lake is most valuable as an extension of an existing complex data environment (e.g., those for warehousing, marketing, supply chain, etc.), not so much as an independent data collection. That's why almost all the use cases in this report show a data lake integrated with these larger data ecosystems. The lake contributes to an already hybrid ecosystem, but a trend in data lakes is to make the lake a logical construct that is physically distributed over multiple platforms (as modern data warehouses are). In that sense, the lake itself becomes hybrid, which gives it a broader range of data types and analytics. As with data warehouses, the hybrid data lake combines Hadoop and an RDBMS (and maybe other platforms) to achieve that breadth.

5. **Expect to fill Hadoop's gaps with additional tools.** Some required tools complement Hadoop, such as those for data integration. Others fix the omissions of Hadoop, namely tools for metadata, security, and SQL support. The good news is that there are many tools available from vendors and open source. To simplify the portfolio of diverse tools needed, consider the vendors that provide multiple tools in a single unified environment, while supporting the full Hadoop ecosystem. Look for use cases that can be delivered successfully with mostly existing tools. After the value is proven, which should get you more budget, expand into additional tools.

6. **Select end-user tools that deliver business value.** Most businesspeople and some technical people will perceive the value of the data lake via the GUI of tools for exploration, data prep, visualization, and other analytics tasks. Know what your end-user constituencies need in that area, and help them find tools that will deliver the value they need. Be sure that end-user tools are complemented with tools for data-centric security so users can explore and analyze the lake's assets securely.

7. **Beware of data dumping.** Early hype around the data lake said that you could throw large data volumes into the lake arbitrarily, then let end users loose to fend for themselves. A number of publicized failures proved that assumption wrong. This kind of "data dumping" leads to redundant data (which skews analytics outcomes), nonauditable data (which no one will trust), and poor query performance (which kills the primary goal of the lake: exploration). In the worst cases, just accessing the data lake constitutes a compliance or privacy infraction.

   Don't be seduced by the magic of early ingestion and the linear scalability of Hadoop. Have a plan that determines exactly what data goes into the lake, based on the kinds of exploration and analytics required for priority users and applications, plus data landing and staging for data warehousing and related practices. Resist any data that is not specified in the plan. Note that your ingestion plan should include components for data-centric security.

8. **Design your data lake.** Once you have a plan for incoming data, think about how to organize volumes, partitions, and zones within the lake. Emerging best practices say that the typical zones are for data landing, data staging, data domains (e.g., customer data), departmental domains (e.g., data used by marketers), analytics archives, and analytics sandboxes. Once you know the zones, design data flows for moving data among them.

   Don't get carried away. A data lake is not a data warehouse, and a zone is not heavily structured like a subject area or dimension. Expect just a few zones, and they are for organization, not radical transformation or remodeling. Within each zone, the data is still in its raw state or slightly standardized, consistent with the data lake's focus on detailed source data for exploration and repeated repurposing.

9. **Focus on raw data, but expect more structure as your data lake matures.** There are now users with a couple of years (or more) of experience with data lakes, and they say it's like a lot of databases. Over time, you will understand which data subsets are accessed and restructured most, so you create data models and persist transformed data for users and applications that need it. That way, access performance and data consistency are improved.

   Again, don't get carried away or you may undermine the lake's reliance on raw data. In terms of data lake design, restructuring data usually results in rather simple record or tabular structures, usually achieved via light data standardization. Expect to revisit how data is organized in your data lake, similar how you would with any database but with much simpler actions and results. Also consider that restructuring data might mean that the data should leave the lake and go to a more structured environment, such as a data warehouse or mart. After all, one function of the lake is to feed other databases.

10. **Govern each data lake.** In an ideal world, you already have a data governance program that has created a library of policies for the compliant use of enterprise data, plus data standards to guide data quality and structure. As with any new data collection, the data governance board should vet a new data lake to specify which existing policies apply and to determine if old policies need revision to accommodate the data lake or if new policies are in order. Remember that Hadoop may also be new to your organization, so it may need separate vetting. Finally, governance is best when it's collaborative. New governance functions in vendor tools can capture and share data knowledge, as well as crowdsource citizen users to act as stewards for the lake's data quality, compliance, and usability.

11. **Cross-train data management specialists.** As mentioned previously, there are very few data management professionals available for hiring who have prior experience with data lakes and Hadoop. The people who are available tend to command rather high salaries. For these reasons, organizations prefer to cross-train existing employees in these skills instead of attempting new hires. This strategy pans out successfully because data management people love cross-training and learning new skills, and their value is raised in the process.

12. **Augment your staff with consultants who have data lake experience.** Although it's hard to find new employees with Hadoop and data lake skills, many consulting practices have upgraded to support such skills, and they have gained experience through multiple clients. When attempting something big that's new to you, turn to consultants and system integrators who have the appropriate experience. This will reduce project risks, shorten time to delivery, and provide a valuable transfer of knowledge from consultants to employees.

DIYOTTA
Modern Data Integration

Diyotta, Inc.

diyotta.com

Diyotta is a leading technology company that provides software to orchestrate and automate movement and integration of big data on Hadoop, MPP, and NoSQL platforms. The foundation of Diyotta's software is based on the five key principles of modern data integration, which were created out of the frustrations brought on by working with traditional data integration technologies and having to maintain custom coded scripts in big data environments.

The five key principles of modern data integration are:

1. Take the processing to where the data lives.
2. Fully leverage all platforms based on what they were designed to do well.
3. Move data point-to-point to avoid single server bottlenecks.
4. Manage all of the business rules and data logic centrally.
5. Make changes using existing rules and logic.

Organizations such as Sprint, Bank of Nova Scotia, Philip Morris International, and Health Lumen rely on Diyotta's technology for their big data movement and integration needs. To learn more about Diyotta and the principles of modern data integration, please visit diyotta.com.

Hewlett Packard
Enterprise

HPE Security–Data Security

voltage.com

HPE 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 ecosystems. HPE Security–Data Security solves one of the industry's biggest challenges: how to simplify the protection of sensitive data in even the most complex use cases.

IBM

ibm.biz/data_warehousing

IBM has the most complete set of capabilities to enable today's hybrid data warehouse, spanning from on-premises appliances (PureData System for Analytics) to Hadoop solutions (BigInsights) to data warehouse (dashDB) deployments that address virtually every information need: structured, semi-structured, or unstructured data, as well as hybrid deployments. To learn more visit ibm.com/datawarehouse and ibm.com/Hadoop.

SAS

sas.com/data

SAS is the leader in big data management and advanced analytics solutions. Through innovative analytics, business intelligence, and data management software and services, SAS helps customers at more than 80,000 sites make better decisions faster. SAS can help speed up the modernization of your data warehouse with the following capabilities:

- Data integration: Data movement, in-database processing, and native data access to traditional and emerging data sources like Hadoop
- Data quality: Cleanse, standardize, and enrich data in real time and batch with prebuilt rules
- Self-service big data preparation: Business users profile, cleanse, and transform data on Hadoop without writing code
- Business glossary and metadata management: Track lineage, business rules, descriptive details, and workflow for improved governance of your data assets
- Event stream processing: Analyze real-time streams of data in motion for better decisions
- Data virtualization: Provide blended, secure views of your data without moving it
- Hadoop support: Access, deliver, and process data inside Hadoop across both the data management and analytics life cycle
- Visualization and advanced analytics: Deliver cutting-edge visualization and analysis capabilities without requiring analytical skills

Data drives everything. Make sure it's right. Learn more at sas.com/data.

Talend

talend.com

Talend (NASDAQ: TLND) is a next generation leader in cloud and big data integration software that helps companies become data driven by making data more accessible, improving its quality, and quickly moving data where it's needed for real-time decision making. By simplifying big data through these steps, Talend enables companies to act with insight based on accurate, real-time information about their business, customers, and industry. Talend's innovative open source solutions quickly and efficiently collect, prepare, and combine data from a wide variety of sources allowing companies to optimize it for virtually any aspect of their business. Talend is headquartered in Redwood City, CA. For more information, please visit www.talend.com and follow us on Twitter: @Talend.

research

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

tdwi

Transforming Data
With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org