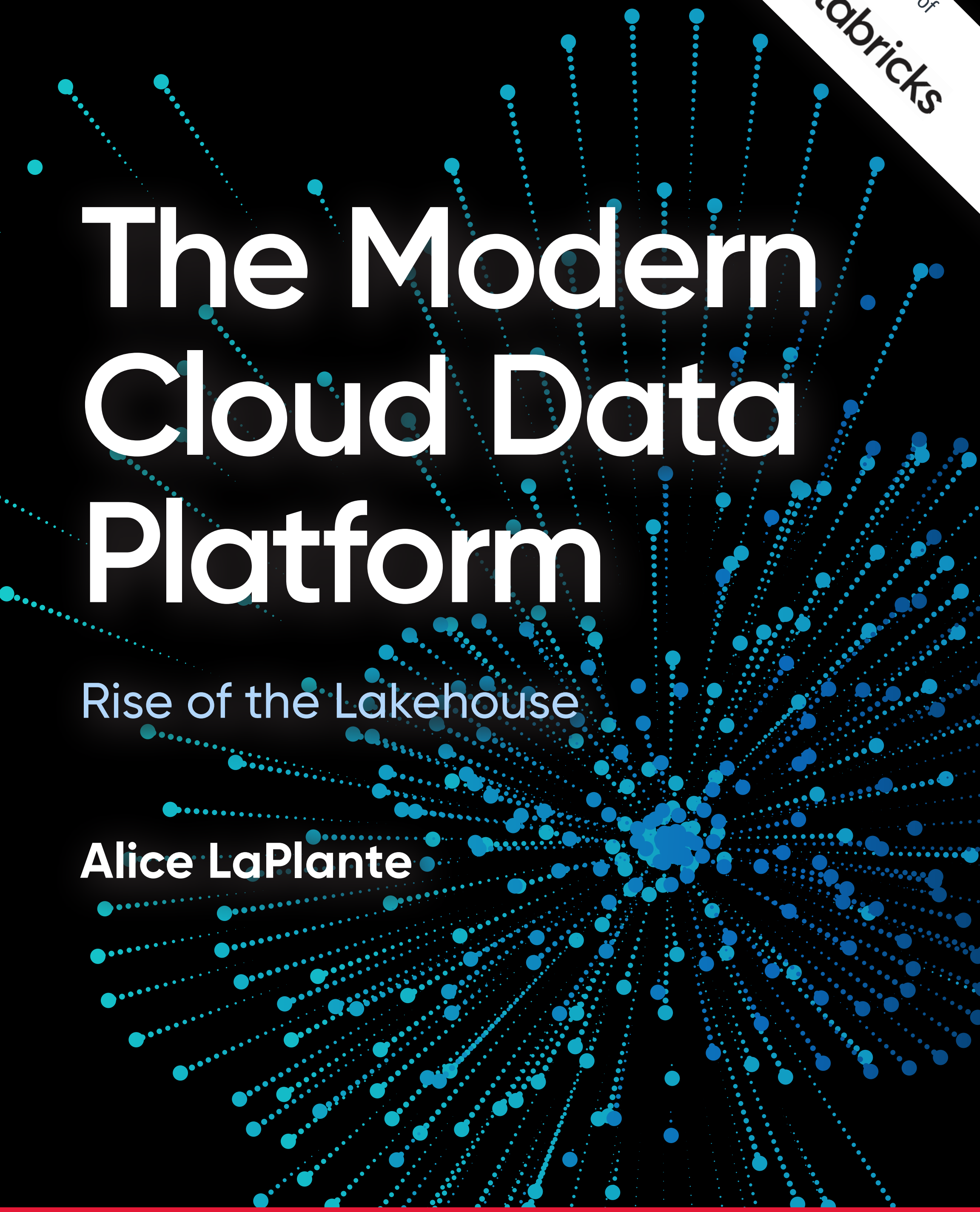
O'REILLY®



The Modern Cloud Data Platform

Rise of the Lakehouse

Alice LaPlante

REPORT

The Modern Cloud Data Platform

Rise of the Lakehouse

Alice LaPlante

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

The Modern Cloud Data Platform
by Alice LaPlante

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Acquisitions Editor: Jonathan Hassell
Development Editor: Amelia Blevins
Production Editor: Deborah Baker
Copyeditor: Rachel Head
Proofreader: Piper Editorial, LLC
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Kate Dullea

December 2020: First Edition

Revision History for the First Edition
2020-12-18: First Release

The views expressed in this work are those of the author, and do not represent the publisher's views. While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

This work is part of a collaboration between O'Reilly and Databricks. See our statement of editorial independence.

978-1-492-08792-2

[LSI]

Table of Contents

The Modern Cloud Data Platform: Rise of the Lakehouse

Leading organizations understand the importance of making high-quality data accessible, usable, and trusted. A 2019 McKinsey survey found that the companies with the greatest growth in earnings over the previous three years attributed at least 20% of that growth directly to their data initiatives.

How did they achieve this? These high-performing companies deploy a three-pronged strategy, according to McKinsey. First, they articulate clear, long-term data strategies. Second, they nurture a data-driven culture by making data an integral part of employees' jobs and educating them on proper data governance. And third, they deploy modern data platforms to support all their data activities at scale.

But what is a "modern data platform"? Is it a data warehouse? A data lake? Can all or part of it be on premises, or must it involve the cloud (or even multiple clouds)? What are the benefits and challenges of these various approaches? And if there were an ideal data platform architecture, what would it look like?

In 2020, O'Reilly Media, in collaboration with Databricks, performed a global survey of more than three thousand data professionals to determine the state of modern cloud data platform architectures. Respondents were asked to assess their current data platform architectures—especially the challenges they had with them—and how those challenges impact business and team success. They were also asked to recommend criteria that would be important to consider when evaluating new types of data architectures.

In this report, we'll first talk about the people responsible for ensuring that businesses are advancing in their data journeys, and how, as teams rather than individuals, they are foundational for transforming siloed data processes into integrated ones that give businesses a broader understanding of their data landscapes. We'll then talk about the various data architectures being used today, and their relative strengths and weaknesses. Finally, we will introduce the idea of a new, unified data architecture that aggregates the advantages and mitigates the disadvantages of traditional data models, and wrap up with advice about five steps you can take to succeed in your data initiatives today.

Data Teams and Their Challenges

As data becomes more important to businesses, the need to know what data they have and how to monetize it has become fundamental to successful growth. This is true for all businesses, regardless of whether they are traditional tech organizations or not—and this need is only going to become more important in the coming months and years as even more data is generated, and using it effectively becomes a competitive necessity.

The focus is increasingly on data professionals working together as a team to make the most out of their organization's data. Why? Because it's not enough for data professionals to work autonomously. The synergy between the different functions is too important. They have to unite as a *community*.

Traditionally, the key roles in data existed separately in four silos of data professionals: data scientists, data engineers, data architects, and data analysts. Although they're all part of the same organization, these data roles have historically depended on completely different sets of tools and processes, and each worked with the data in their own silo. This pattern leads to increased organizational complexity and cost. It also takes longer and is more difficult to get value from data when the data professionals operate separately.

Today, leading data-driven businesses are moving to a more unified architecture that meets the needs of all these data professionals.

The Importance of the Cloud to the Data-Driven Company

It turns out that the data world is as enamored with the cloud as the rest of the universe. A full 81% of the survey's respondents said their organizations had adopted cloud services and infrastructure as data architectures in at least some capacity. Only about 2 in 10 (19%) reported not having moved any data workloads to the cloud at all.

So what types of data platform architectures are enterprises currently using, whether in the cloud, on premises, or both? Respondents were encouraged to check all that applied. As Figure 1 shows, data warehouses are slightly ahead (57%), followed closely by data lakes (53%) and so-called "specialized systems" (54%), which encompass specialized databases such as SAP ERP and Oracle PeopleSoft on premises, and Salesforce and Workday in the cloud.

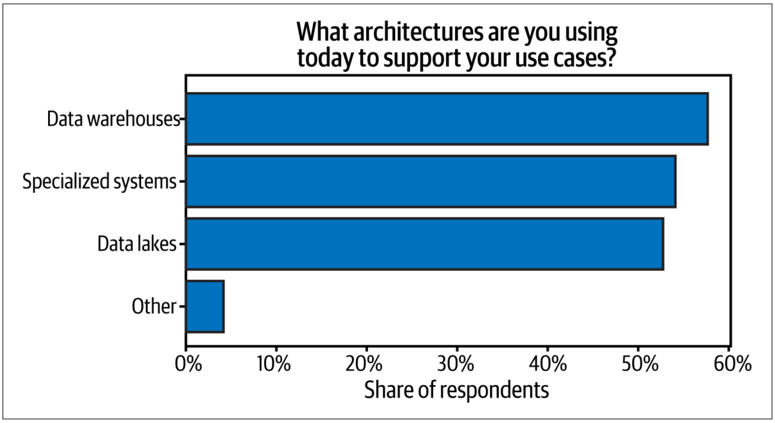


Figure 1. A close race between today's most popular data platform architectures

One of the most important facts revealed by responses to this question, of course, was that numerous organizations are running multiple data platform architectures. As we see in Figure 2, the challenges raised by this are prevalent.

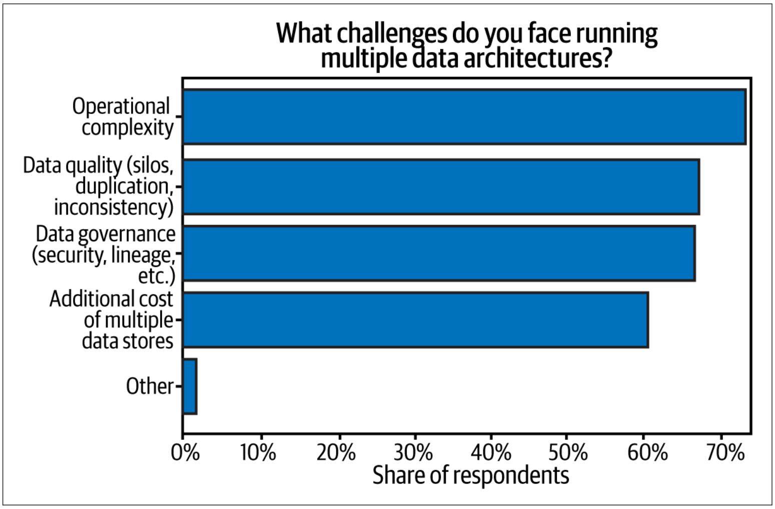


Figure 2. Challenges of running multiple data architectures

Chief among the challenges of running multiple data architectures is operational complexity—just keeping the infrastructure stable and data repositories running in such an environment was a problem for more than 70% of enterprises.

Data quality (framed as issues with data silos, duplication, and inconsistency) and data governance (including security and data lineage) were also identified as key challenges by 67% and 66% of respondents, respectively. Data quality is critical because business decisions will be made based on data stored in a data lake or data warehouse. Businesses need to have absolute trust that they're working from a single point of truth. Governance is also increasingly important, especially as it relates to data privacy, because of the growing number of regulations protecting sensitive data from access by unauthorized individuals.

Finally, a robust 60% of respondents identified the additional costs of having to support multiple data stores as an issue.

As a direct result of these challenges, a full 63% of enterprises are either currently implementing or actively evaluating new data architectures, as shown in Figure 3. Just 2 in 10 (19%) professed that they were happy with what they had—which could mean either that they were considerably ahead of the pack in getting to a better data platform, or that they are somehow coping with the level of pain involved in supporting multiple architectures.

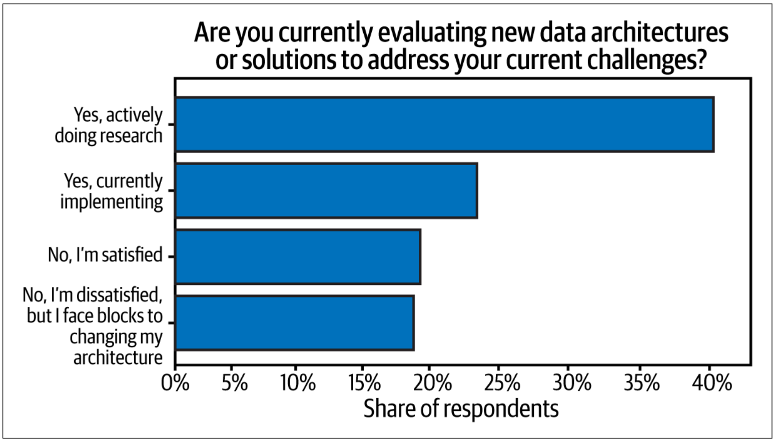


Figure 3. Many data professionals are currently seeking new data platform architecture options

To understand more precisely what enterprises currently have installed, and why they might be searching for more modern solutions, the next section will define each type of data platform architecture and its benefits and challenges.

Data Architecture Options

Until recently, enterprises have had three basic choices of data platform architectures: data warehouses, data lakes, and specialized data systems. All three of these can be deployed either on premises or in the cloud. Let's examine each in turn.

Data Warehouses

A data warehouse is a central repository where data from one or more different sources is integrated and used for business reporting and analyses. It is considered fundamental to an enterprise's ability to leverage its data for business intelligence (BI).

A key attribute of a data warehouse is that it is highly structured. Data stored in a data warehouse has been prepared and "transformed"—cleansed and deduplicated and formatted to meet established standards. Indeed, data typically isn't put into a data warehouse until data professionals are fairly sure they know how it will be used, and for what purposes. Most data warehouses, whether on premises or in the cloud, still follow the guidelines and frameworks defined by Ralph Kimball and Bill Inmon in the mid-1980s.

Data warehousing fundamentally changed the way businesses analyzed data and made strategic decisions. Prior to its emergence, transactional and operational data were locked up in different silos, making it difficult to guarantee that there was consistency throughout the enterprise in how data was defined, to put data directly in the hands of business users who needed it to get their jobs done, and to get a full picture of an organization's business. Today, data warehouses are very popular—indeed, as Figure 2 shows, they're currently the leading data platform.

Although traditional data warehouses are located on premises, cloud-based data warehouses are rapidly gaining favor, for cost and scalability reasons as well as the fact that they free organizations from having to procure, deploy, and maintain the necessary infrastructure to support the warehouses (more on that later).

Benefits of data warehouses

The data warehouse's prime directive is to help the organization make better business decisions. In helping data professionals and data consumers reach this goal, other benefits accrue. Data warehouses:

Deliver business intelligence

Putting data from different sources into a single data warehouse and making that accessible to authorized business users within an organization means businesses no longer have to rely upon employees' and executives' instincts to make critical decisions. Instead, these decisions can be supported by data.

Improve query performance

The constant queries from business users can push analytics infrastructure such as data marts and legacy databases to the limits. A data warehouse can be more efficient at managing queries, relieving the burden on the overall ecosystem.

Improve data and decision quality

Data is transformed before it is placed in the data warehouse. This means data from multiple sources is put into a standardized format, and users from across the organization can see—and access—consistent information that allows them to steer the business in a common and consistent direction.

Democratize data

More recently, thanks to advances in the databases themselves as well as analytics and visualization tools, advanced data-driven organizations are attempting to fully democratize data throughout the organization by allowing more and more of their users access to their data warehouses. But as you'll see, this benefit is linked to one of the biggest challenges the survey respondents had with data warehouses: scalability.

Challenges of data warehouses

Although the benefits of using data warehouses are significant, challenges also exist. A data warehouse is, almost by definition, a super-large database. Successfully designing and deploying one is an enormous undertaking. Planning, collaboration, and coordination of both people and resources are required.

When we asked data professionals about the challenges they faced with their data warehouses, as Figure 4 shows, cost topped the list (50%), followed by complex operations (47%) and scalability (46%).

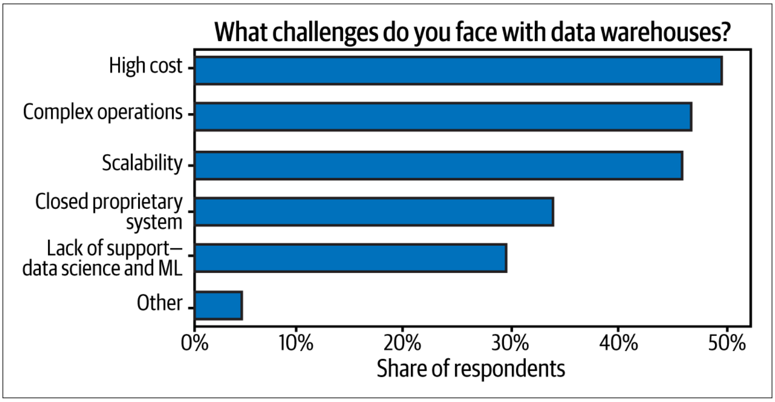


Figure 4. Challenges with data warehouses

Let's take the top five challenges one by one.

Potentially high cost. Half of all respondents said the high cost of a data warehouse was its biggest drawback. This can be a problem with both on-premises and cloud data warehouses.

For on-premises data warehouses, the high licensing fees create an expensive cost foundation. Then there are the operational costs: all the responsibility for procuring, installing, and maintaining the data warehouse infrastructure falls on the organization. And as the data warehouse grows—as it inevitably will—more people and resources have to be committed to the initiative. To ensure that there is always enough compute and storage, many organizations overprovision their infrastructure to allow for unexpected spikes or peaks in workloads; for example, retailers must prepare for this sort of thing during the holiday season. This can mean that for much of the year they have underutilized capacity.

On the other hand, the price structure for cloud data warehouses, which is more commonly a rental model, allows you to pay for as much or as little as you need. Although this eliminates the risk of paying for unused resources, it can still be an expensive endeavor as the data warehouse grows.

Complex operations. Almost half (48%) of respondents said the operations of data warehouses are too complex.

For on-premises data warehouses, IT and, frequently, the data team are deeply involved in deployment, upgrades, security implementation, and ongoing operations. This is not trivial. Data platforms need to be regularly tuned to ensure consistently high performance over time, especially as data volumes scale. Otherwise, the data warehouse can become painfully slow, ineffective, or even nonfunctional.

Operations can also be complex with cloud data warehouses, with regard to both pricing and infrastructure support. Like other cloud services and solutions, this market is still maturing. Different vendors use different cost models. Some charge a flat monthly fee while others offer pay-as-you-go pricing schemes. Different vendors also approach infrastructure support differently. Some cloud data warehouse products require you to provision and manage your cloud resources, while others take a serverless approach, in which the burden of provisioning and managing cloud servers is completely abstracted from the enterprise. Then there are the operational tasks of enforcing service-level agreements (SLAs), integrating the data warehouse with existing processes—both cloud-based and on premises—and ensuring that security and disaster recovery measures are robust.

Finally, given that many organizations first land data in a data lake, numerous data pipelines need to be maintained to move data out of the data lake and into one or more data warehouses. In the cases where those data warehouses make changes, data pipelines are also required to move the changed data back into the data lake.

Scalability. Scalability is a very real problem for on-premises data warehouses, with 47% of respondents saying it is a key concern. The IT department must be vigilant to ensure that there are sufficient resources at all times—especially to deal with unexpected bumps in traffic. Scaling up is a time-consuming and resource-intensive task, as it usually entails purchasing and installing new hardware.

For cloud data warehouses this is not a problem, as organizations can procure more compute or storage at any time they need it, even for "bursty" traffic. However, scalability is still an issue here, as it is difficult to maintain the hundreds or thousands of data pipelines needed to feed all the different reports large data warehouses must serve up. This concern is exacerbated by two factors: most customers have multiple data warehouse vendors in their architectures, and the data architectures are increasingly split across multiple cloud providers.

Closed proprietary systems. One-third (33%) of data professionals find this issue to be a critical one. Unfortunately, many on-premises data warehouses do not play well with others. Lock-in is real, and it can be an expensive annoyance when you want to move to a different data warehouse solution.

Even the cloud doesn't escape this challenge, as different cloud providers have different functions and capabilities, and moving a data warehouse from, say, Google Cloud to Microsoft Azure is not a seamless process.

When asked about their attitudes toward lock-in caused by closed data formats or proprietary software, a strong majority (86%) of data professionals expressed being at least “somewhat” concerned, with 58% being “concerned” or “very concerned,” as shown in Figure 5. Vendor lock-in, of course, has been a challenge to organizations since the dawn of the digital era, as makers of both hardware and software try to make it difficult for customers to extricate themselves after committing to a particular platform. This challenge is continuing to bother data professionals in the data-driven age.

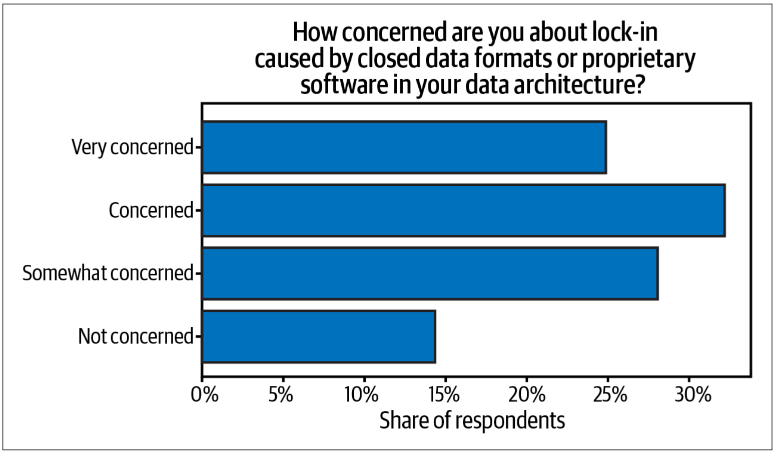


Figure 5. Concerns about lock-in abound

Lack of support for data science and machine learning. The challenge here for 29% of data professionals is that data warehouses are based on 40-year-old technology that was never designed to handle anything but structured data. Audio, video, text in natural language, and other unstructured data types don’t fit into data warehouse schemas. The increasing prevalence of this kind of data as fuel for data science and machine learning is what drove the rise of data lakes and the subsequent complexity of trying to maintain both data lakes and data warehouses in an enterprise data architecture.

Data Lakes

A critical difference between data lakes and data warehouses is the type of data and use cases they can handle. Still, many think that data lakes and data warehouses are the same thing. Here are the attributes they share:

- They are repositories for storing data.
- They can be cloud-based or on premises.
- They are deployed when organizations seek to democratize data.

But that's where the similarities between data lakes and data warehouses generally end. Table 1 outlines the differences.

Table 1. Main differences between data lakes and data warehouses

	Data lakes	Data warehouses
Types of data that can be stored and processed	Structured, semi-structured, and unstructured data	Structured and semi-structured data only
Purpose of data	Undefined purpose for data	Data defined for specific use cases
Types of users	Users are data scientists and data engineers	Users are nontechnical business users
Structure	Flexible and easy to change	Rigid and difficult to change

So what is a data lake? It's a system or repository of data stored in its natural/raw format, usually files or object blobs. It can contain both structured and unstructured data in its raw form—including structured data from relational or transactional databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs), and binary data (images, audio, video).

The goal of a data lake is to make all this data available for transformation and mining to generate reports, perform visualizations, and conduct advanced analytics and machine learning in order to, ultimately, gain a competitive business edge.

Benefits of data lakes

Data lakes offer some significant benefits over data warehouses:

Data lakes can ingest and keep all enterprise data.

Much of the work in building a data warehouse revolves around understanding and making decisions about data. Where does it come from? What needs to be done to it? How will it inform business processes? The goal is to build a highly structured data model designed for reporting, with some capability for ad hoc querying by more advanced users. Typically, if data isn't needed to answer specific questions or be included in a particular report, it will be excluded from the warehouse. There are two reasons for this: to simplify the data model and, specifically for on-premises data warehouses, to avoid filling up expensive disk storage. Storage is costly because in on-premises data warehouses, it is coupled with compute. Increasing storage means you have to buy more compute, too, even if you don't need it. The opposite is also true.

In contrast, a data lake retains all enterprise data. All data generated or collected by the company can be, and usually is, put into the lake. The reason? It's impossible to predict in advance what data will be useful for exploratory data science and machine learning, or even future BI needs. Data lakes give you that flexibility. Data is also conserved indefinitely, so businesses can check and recheck old data as needed.

Data lakes can store and process all data types.

Whereas data warehouses deal with structured data, such as data from traditional transactional systems, data lakes can absorb both structured and unstructured data. This includes streaming data like web server logs, sensor data, social network activity, text, and images. Whereas in the past this kind of data was difficult and expensive to store and analyze, the data lake accepts it all.

Data lakes make all data accessible to all users.

There are three general types of data users: business users, data analysts, and data scientists. The data warehouse is well structured, easy to use, and purpose-built to answer questions that the first type of user—business users—might have. Data analysts are the builders of the reports and dashboards the business users consume. Finally, the third category of users—data scientists, data engineers, and other data professionals—usually bypass the data warehouse as being too limiting. They are interested in doing deep statistical analysis, often using artificial intelligence (AI) tools. Data lakes serve all these types of data consumers equally.

Data lakes can be easily changed.

Because organizations put so much effort up front into designing data warehouses and data structures to be the way they want, changing them requires allotting a lot of development resources—and time—to the task. Conversely, the data lake stores all data in its raw form and makes it accessible to anyone who needs it, to use in any way they like. And importantly, data lakes have a schema-on-read framework (rather than the schema-on-write practice of data warehouses) and use the extract, load raw data, and transform as needed (ELT) process instead of the conventional extract, transform, and load (ETL) process that data warehouses follow. This allows users to build models and explore data and schemas as they desire. Any results of data experiments that aren't useful can simply be thrown away, with no changes to the data structures and no need to involve IT developers for help. This makes data lakes infinitely more flexible—they don't require structural changes to answer new questions.

Data lakes can deliver actionable insights more swiftly.

Precisely because data lakes contain all data and data types, and because they allow users of all kinds to access data before it has been structured and transformed, users get their results faster than if they had to wait for data professionals to cleanse and standardize the data for them. Unfortunately, data lakes can also turn into data swamps, precisely because they can become dumping grounds for data that doesn't conform to any standards (more on this in the next section).

In summary, the benefits of data lakes are substantial, especially when it comes to larger, unstructured data stores for which enterprises may have numerous—as yet undetermined—end purposes.

Challenges of data lakes

However, there are also many challenges to implementing data lakes. Not surprisingly, governance is the number one concern, as shown in Figure 6, with more than 60% of respondents saying it was a challenge. And lack of governance inevitably results in messy, unreliable data, which was the number two challenge cited by more than half (52%) of data professionals surveyed. Complex operations came in a close third, with 51% of data professionals listing this as a major concern.

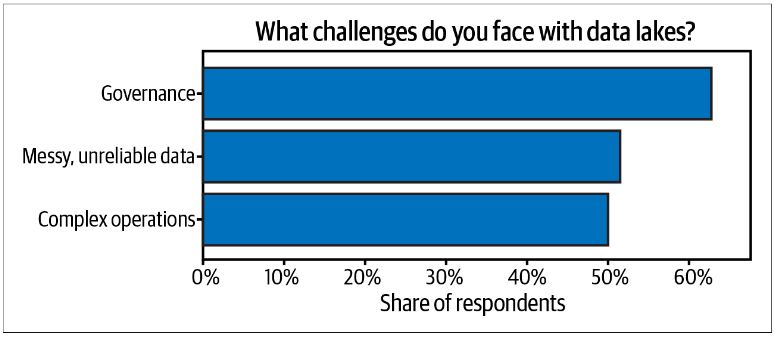


Figure 6. Challenges with data lakes

Let's dig a little more deeply into these leading issues:

Governance

Precisely because of the huge volume of data sitting in a data lake, with users of all types dipping into, querying, consuming, or reporting on data at will, it's a considerable challenge to ensure that data remains both secure and private. Complying with the growing list of privacy rules—from industry, state, federal, and even international governing bodies—is important to avoid hefty fines and public embarrassment. It's essential that a senior data professional (perhaps even the chief data officer) set the right policies for the data throughout its entire life cycle. Data must remain secure and private in its raw state, and stay that way even as data scientists explore it, analysts curate it, and business users analyze it.

Messy, unreliable data

As data lakes continue to accumulate more and more data in different structures and formats, keeping it consistent and clean can be a formidable task. Data lake architecture is more distributed and has fewer constraints on the format and scale of data stored than a data warehouse. Data lakes also take time to reflect writes to their clients. This results in queries showing inconsistent data until all the nodes of the lake become consistent. Additionally, data lakes lack any mechanisms to alert users when writes fail. It can take weeks or months to discover that subsets of data are corrupted or incomplete.

Complex operations

On-premises data lakes have the same operational issues as on-premises data warehouses. Performance and security are top of mind, and IT must keep the data lakes up and functional, running at optimal performance at all times, for them to be successful. If a data lake gets the reputation of being slow or unresponsive, the team won't use it. A different way to think of this is to simply invert the data warehouse scalability problem: with data lakes you don't necessarily have a lot of pipelines to maintain, but you do need to do a lot of engineering to scale response times with increased usage.

Building, transitioning to, or maintaining a cloud data lake can also be operationally challenging, especially when an organization has both on-premise and cloud data to manage. Additionally, today multi-cloud solutions are becoming more common, for three main reasons. First, companies often need to diversify their infrastructure to comply with regulations or to mitigate risk. Second, independent decision making in large enterprises frequently results in different departments investing in different cloud providers. And third, merger and acquisition (M&A) activity forces the acquirer to absorb the technology investments of the acquiree.

In summary, there are significant challenges involved when using data lakes that are different from those faced with data warehouses. Despite this, companies are increasingly moving to data lakes because of their flexibility and ability to make all data more readily accessible.

Specialized Systems

The third and final type of data platform architecture falls into a category called "specialized systems." These are applications that tend to be large data repositories for specific types of data. For example, Salesforce is a large data repository that many enterprises use to manage customer relationship management (CRM) data. Workday is another example, which houses data that is particularly related to HR.

The main benefit of specialized systems is that the data is tightly controlled and organized according to the application vendor's specification. There are typically well-established ways to query the system and get reports on common topics.

The challenges arise when you want to integrate data from one of these specialty systems, such as combining billing data sitting in an on-premises data warehouse with customer data that's most likely in the cloud in Salesforce.

Enterprises that possess specialty system data platform architectures are the only ones that put complex operations first when asked about challenges (see Figure 7).

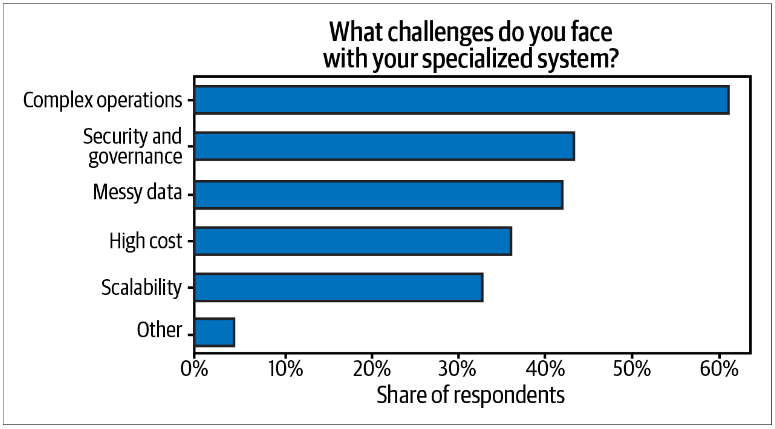


Figure 7. Challenges data professionals face with specialized systems

The chief challenge comes down to one word: integration. Although APIs are available to make common integrations fairly straightforward, it's tricky business to eliminate many of the silos that can arise in these specialty systems. Data mapping, mastering, deduplication, and—perhaps most importantly—migrating data from proprietary formats can be major headaches.

The Business Impact of Data

Data makes a difference. Almost three-fourths (67%) of the data professionals surveyed said that improved business decision making drives their data use cases. More than half also indicated that customer experiences (60%) and key business processes (56%) were critical drivers. Indeed, these were the top three drivers regardless of use case (see Figure 8).

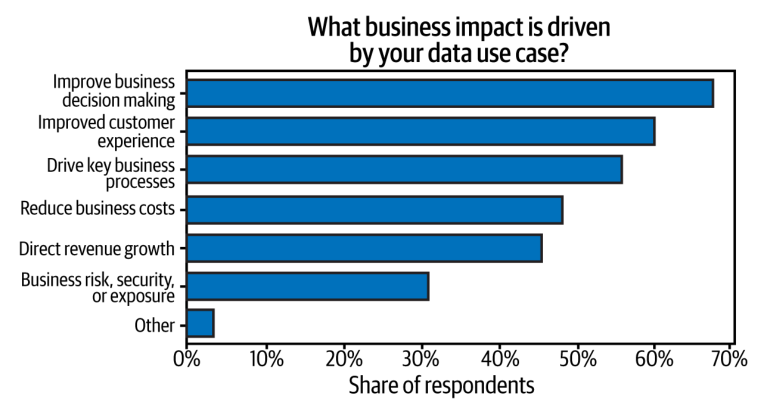


Figure 8. The business impact of data initiatives

This data suggests that most organizations favor a long-term view on their data investments rather than only seeking near-term rewards like reductions in business cost (49%) or increases in revenue growth (45%). This reflects the generally accepted view in the market that the ability to effectively and efficiently make use of data is going to determine which organizations succeed into the foreseeable future.

Impact of Managing Complex Data Architectures

Despite the fact that the majority of enterprises seek business value from their data activities, it appears that the complexity of data platforms, infrastructures, and architectures is holding data teams back from fully realizing their goals (see Figure 9).

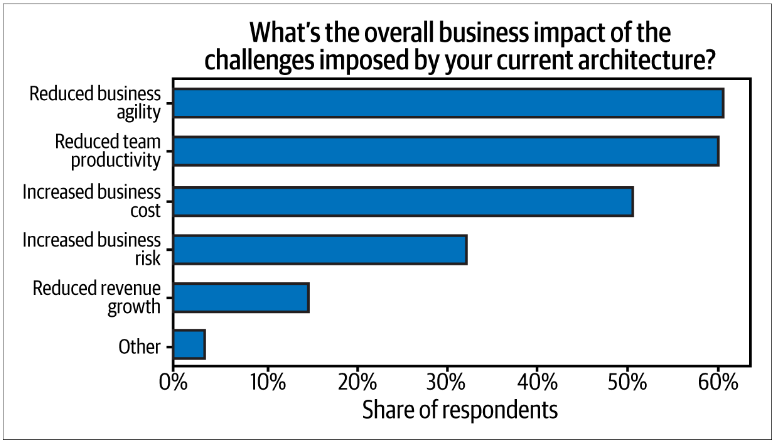


Figure 9. Overall impact of architectural challenges

More than 60% of respondents said that having to manage a complex data infrastructure reduced business agility and team productivity. Only slightly more than half (51%) said that the high cost of complex data architectures impacted their business. This again reinforces the finding that reducing cost isn't necessarily the prime driver for simplifying a data architecture. Softer drivers such as business agility and productivity seem to matter more when it comes to getting value from data.

Envisioning a Better Data Environment

Not surprisingly, then, 89% of the survey respondents felt that simplifying data architecture while still supporting all necessary use cases would be valuable, with the plurality (47%) saying it would be "very valuable," as shown by Figure 10.

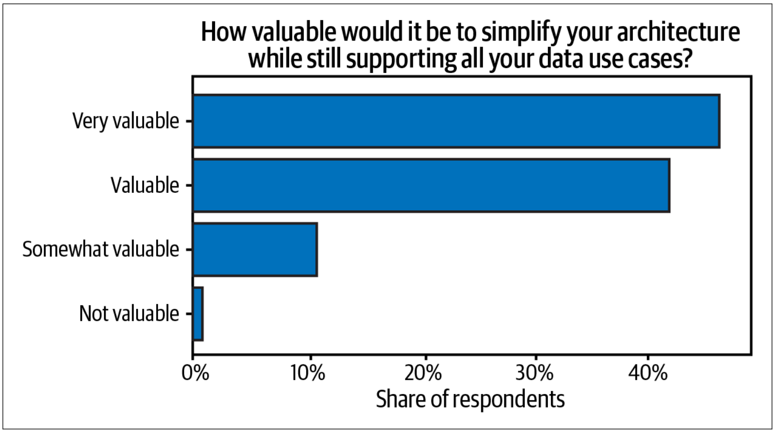
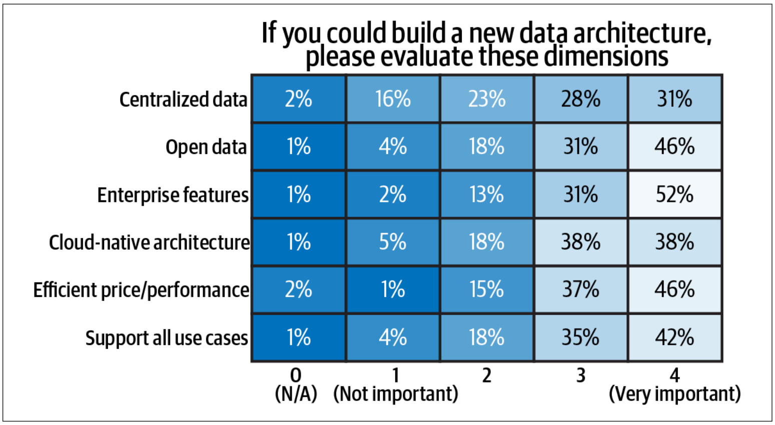


Figure 10. Data professionals are eager to simplify their data architecture platforms

And when asked which qualities were important when envisioning a new data architecture, as depicted in Figure 11, a large majority of respondents said the following were either "important" or "very important" features for an ideal data environment: centralized data, open data, enterprise features, cloud-native architecture, an efficient price/performance ratio, and one that supports all use cases.



	0 (N/A)	1 (Not important)	2	3	4 (Very important)
Centralized data	2%	16%	23%	28%	31%
Open data	1%	4%	18%	31%	46%
Enterprise features	1%	2%	13%	31%	52%
Cloud-native architecture	1%	5%	18%	38%	38%
Efficient price/performance	2%	1%	15%	37%	46%
Support all use cases	1%	4%	18%	35%	42%

Figure 11. Importance of proposed dimensions for a new type of data architecture

Let's take a closer look at these "must-have" qualities:

Centralized data

Having centralized data is an important or "very important" aspect of being data-driven for approximately 6 out of 10 data professionals (59%). Centralizing your data gives you a single source of truth. It ensures that your data teams and indeed everyone throughout the organization is using the same data, leading to more aligned reporting across departments and, ultimately, better decision making.

Having a centralized data store is critical for helping businesses ascertain which business processes are working efficiently and which aren't. It is also essential for making intelligent data-driven decisions that'll ensure business relevance. It's the best approach to ensuring that data remains an asset to and not a limiter of business success.

Open data

Even more data professionals (77%) want a platform that supports storage formats that are open and standardized, such as Apache Parquet. Platforms that offer open and standardized formats should also provide APIs, so that any type of tool or engine—including libraries for machine learning and Python or R—can access the data directly in an efficient manner. Overall, data professionals increasingly believe that the future of data and AI depends on cloud-agnostic, open platforms.

Enterprise features

Data professionals also want a platform with enterprise features, with 83% of respondents saying this is an important or "very important" need. Enterprise-grade systems require such things as tools for security and access control. Enterprise-class table stakes include data governance capabilities like auditing, retention, and lineage. Tools that enable data discovery, such as data catalogs and data usage metrics, are also needed.

Cloud-native architecture

More than three-quarters (76%) of respondents thought a cloud-native architecture was important or "very important," because it is necessary to get the most out of cloud-native applications and cloud-native data warehouses and data lakes. Cloud-native architectures help organizations deliver the agile, automated, scalable, and highly available digital solutions that are foundational to a data-driven organization. Also, as enterprises increasingly deploy across more than one cloud provider, this means adopting technology that can provide a consistent experience regardless of where data is stored.

Efficient price/performance

According to the survey, having an efficient price/performance ratio is important or "very important" to 83% of data professionals. This invariably means moving to a cloud architecture, which enables the transformation of the heavy capital expense (CAPEX) requirements for an on-premise data platform into operational expenses (OPEX) that are aligned with what's going on in the business. Exponential growth in data does not justify exponential growth in data infrastructure costs. And when you have no visibility into who is doing what with what data, it results in uncontrolled costs—including infrastructure costs, data costs, and labor costs. Businesses need a platform that avoids this excess expense.

Support for all use cases

Finally, more than three-quarters of respondents (77%) said that having a platform that supports all use cases is important or "very important." This means the platform needs to support data engineering, data analytics, data science, and machine learning use cases involving structured and unstructured data, as well as supporting both batch and streaming data. In short, whatever your organization wants to do with data, the data platform should be able to handle it.

Addressing the Limits of Current Data Architectures

Over the past several years, support for a new kind of data management archetype has grown. This report covered *data warehouses* earlier, and how they've evolved since their beginnings in the late 1980s to accommodate the changing decision support and business intelligence needs of businesses. It also discussed how, although data warehouses have been great for structured data, today's businesses have tremendous amounts of semi-structured and unstructured data that they want to make use of. That's why, in the early 2010s, organizations began to build *data lakes*—large data repositories for raw data that support both structured and unstructured data.

But here, too, there have been limitations. Data lakes can store massive amounts of data, but they cannot support transactions, they are weak on data governance, and their lack of consistency and isolation make it very difficult to mix appends and reads and to perform both batch and streaming jobs.

For these and other reasons, data lakes haven't lived up to their promise. But the desire of businesses for a flexible and high-performing data system remains strong. As the survey has shown, companies seek systems that can handle a broad array of diverse use cases, encompassing SQL analytics, real-time monitoring, data science, and AI and machine learning.

With regard to the latter, many recent AI innovations center on processing unstructured data such as text, images, and video. Data warehouses can't store these types of data, and data lakes are not optimal for other use cases, for the reasons stated above. So in many cases, businesses deploy and manage multiple systems: perhaps a data lake, several data warehouses, and other specialized systems such as streaming, time-series, graph, or image databases.

But, again, as the survey showed, running multiple systems introduces complexity, with all its attendant difficulties. What solution is up for the challenge?

Recently, a new architecture called the *lakehouse* has emerged as an alternative to the legacy architectures of the past.

What Is a Lakehouse ?

A lakehouse combines the best elements of data lakes and data warehouses to build something new. Lakehouses have similar data structures and data management features to data warehouses, but use the low-cost, flexible storage of data lakes. In other words, they're what data warehouses would be like if they were designed today, in a world where cheap and highly reliable storage, in the form of object stores, is available (see Figure 12).

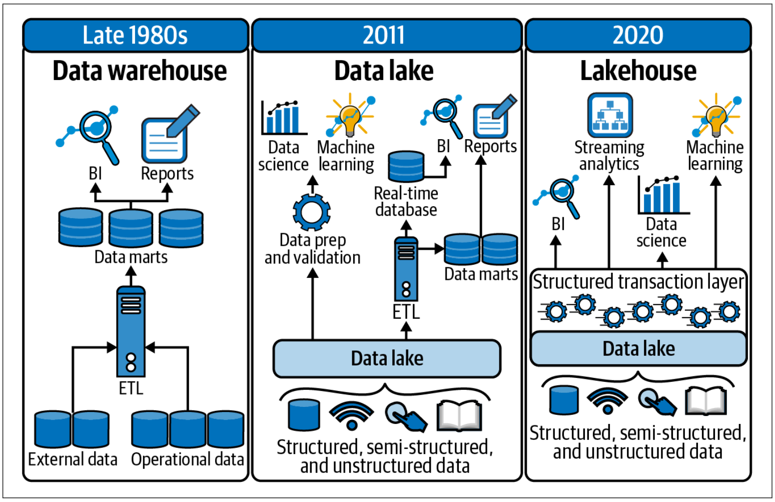


Figure 12. Each type of data repository has strengths and challenges

A lakehouse has the following key features:

Transaction support
: In a lakehouse, multiple data pipelines will frequently read and write data concurrently, usually using SQL. Support for atomicity, consistency, isolation, and durability (ACID transactions) ensures consistency when this happens.

Schema enforcement and governance
: The lakehouse should support schema enforcement and evolution, including data warehouse schema paradigms such as star/snowflake schemas. The system should be able to enforce data integrity as well as possessing robust governance and auditing mechanisms.

BI support

With a data lakehouse, you can use your BI tools directly on your data lake. This improves data freshness, reduces latency, and cuts the expense of having to place and support copies of the data in both a data lake and a data warehouse.

Storage decoupled from compute

Because storage and compute use separate clusters, lakehouses are able to scale to many more concurrent users and larger data sizes.

Openness

The storage formats that lakehouses use (such as Apache Parquet, Delta Lake, and Apache HUDI) are open and standardized, and they provide APIs so that a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data directly.

Support for diverse data types, ranging from unstructured to structured data

A lakehouse can be used to store, refine, analyze, and access data types needed for many new data applications that require semi-structured and unstructured data, including images, video, audio, and text.

Support for diverse workloads

A lakehouse supports all use cases and workloads, including data science, machine learning, and SQL and analytics. Multiple tools might be needed to support all these workloads, but they all rely on the same data repository.

End-to-end streaming

Real-time reports are the norm in many businesses. Lakehouses support streaming, which eliminates the need for separate systems dedicated to serving real-time data applications.

All these key features add up to make data lakehouses considerably more attractive than either data lakes or data warehouses alone.

Conclusion: Dive Into a Lakehouse

The most successful companies in the coming decades will be—underneath the surface trappings of specific industries—data companies.

To enable the massive data transformation discussed in this report, you need to bring all your users and all your data together, and give your users the tools and infrastructure they need to draw insights from the data. You need a single enterprise data platform built on open standards that scales across every department and every team.

Enterprises new to these challenges may take an incremental approach, or take on-premises solutions and move them to the cloud. But without a holistic, cloud-native approach, you are setting yourself up for replacing an outdated architecture with another that won't deliver the goods over the long term.

The following five steps can ensure that you are progressing toward a system that can stand the test of time.

Step 1: Bring all your data together

Businesses for decades have depended on data warehouses to aggregate structured business data and make decisions by creating BI dashboards using visualization tools. When data lakes debuted in the early 2010s, they finally made data science, AI, and machine learning feasible for businesses. Today, the lakehouse model combines the reliability of data warehouses with the scalability of data lakes using an open format such as Delta Lake or Apache HUDI. Regardless of your specific architecture choices, you need to choose a platform that can store all your data—structured and unstructured alike—in open formats suitable for both data analytics and data science workloads, which allows you to maintain long-term control over your data.

Step 2: Enable all users to securely access the data they require to do their jobs

Make sure every member of your data team, across various roles and business units—data engineers, data scientists, data architects, and data analysts—has access to all the data they need, and none of the data they're not authorized to access. This means complying with regulations such as GDPR, CCPA, HIPAA, and PCI, depending on your industry and geographic location.

It is also critical that all of your data remains together, in a centralized place. If you are fragmenting the data by copying it into a new system for a subset of users—for example, into a data warehouse for a certain set of your BI users—you will suffer from "data drift," which leads to issues in step 3. It also means you will have truth drift, where some information in your organization will be stale or of substandard quality, leading to (at best) organizational mistrust of data, or, more likely, poor decisions that lead to bad business results.

Step 3: Manage your data platform like you manage your business

When you onboard a new employee, you set them up for success. They get the right computer, access to the right systems, and so on. Your data platform should be the same—it should be set up to succeed.

Since all of your data is in one place, every employee will see a different facet of that data, according to each one's defined role and responsibilities. This kind of data access should be automated based on your onboarding processes, and it should be transparent so it can be easily audited.

Step 4: Leverage cloud-native security

As the cloud has become the de facto location for massive data processing and machine learning, the traditional "demilitarized zone" (DMZ) and perimeter security of "on-premise" security are being replaced with zero-trust and software-defined networking. Businesses must accordingly ensure that their data processing platforms are designed for the cloud and that they leverage best-in-class cloud-native controls.

Moreover, since every user accesses the data they need with their own online credentials, cloud auditing and telemetry capabilities give you a record of data access and modification through cloud-native tools. This makes step 3 possible.

Step 5: Automate for scale

Whether you're rolling out your data platform to hundreds of business units or many thousands of customers, this process needs to be completely automated. This means that your data platform should be deployed with zero human intervention.

Further, for each workspace (environment for each business unit), data access, machine learning models, and other templates must also be configured automatically.

Powering this scale demands powerful controls. With the compute power of millions of machines at your fingertips, it is easy to run up massive cloud bills. To deploy to departments across the enterprise, the right spend policies and chargebacks need to be designed to ensure power is being deployed as the business expects. APIs can automate everything from provisioning users and team workspaces to running production pipelines, controlling costs, and measuring business outcomes. A fully automatable platform is necessary to power your enterprise.

It's time to begin your journey to compete as a data-driven company. The survey results presented here have clearly identified both the benefits and challenges of existing data platform architectures, and how traditional models such as data warehouses and data lakes limit a business's ability to make the most of its data. By following these five best practices and moving to a single data architecture—represented by the lakehouse—companies like yours will finally be able to reap the very real competitive advantages your data offers.

About the Author

Alice LaPlante is an award-winning writer, editor, and teacher of writing, both fiction and nonfiction. A Wallace Stegner Fellow and Jones Lecturer at Stanford University, Alice taught creative writing at both Stanford and in San Francisco State's MFA program for more than 20 years. A *New York Times* best-selling author, Alice has published four novels and five nonfiction books, as well as edited best-selling books for many other writers of fiction and nonfiction. She regularly consults with Silicon Valley firms such as Google, Salesforce, HP, and Cisco on their content marketing strategies. Alice lives with her family in Palo Alto, California, and Mallorca, Spain.