451 Research
Vanguard Report

January 2025

Cloud object storage drives all your data lake workloads

Commissioned by

aws

S&P Global
Market Intelligence

Introduction

For more than a decade, enterprises have deployed data lakes to store their most important data. Today, the data lake is the enterprise's foundational data layer, storing structured, unstructured and semi-structured data. Enterprise data varies greatly in form and content — it may consist of sales, customer and product data, as well as device, social and employee data, among other types. More than half of enterprise respondents to 451 Research's Voice of the Enterprise (VotE): Data & Analytics, Data Platforms & Real-Time Analytics 2023 study have a data lake in use or proof of concept (PoC). Another 22% plan to implement a data lake in the next three years. The initial benefits of a data lake seem to have struck a chord with enterprises as an environment in which they can store a variety of data types, specifically tabular data, from multiple sources while providing governed access to multiple users.

Enterprises overwhelmingly place significant importance on their data lakes to store large volumes of data, which they query to drive decision-making. Nearly 60% of respondents in our Data & Analytics survey cited above say they use data to make most or all of their strategic business decisions. Data lakes are especially supportive of data-driven decision-making because they not only store historical data but also large volumes of incoming data, allowing enterprises to query both the past and the present for strategic analysis. Furthermore, 82% of study respondents indicate that data will become even more important to their company's decision-making over the next 12 months.

The Take

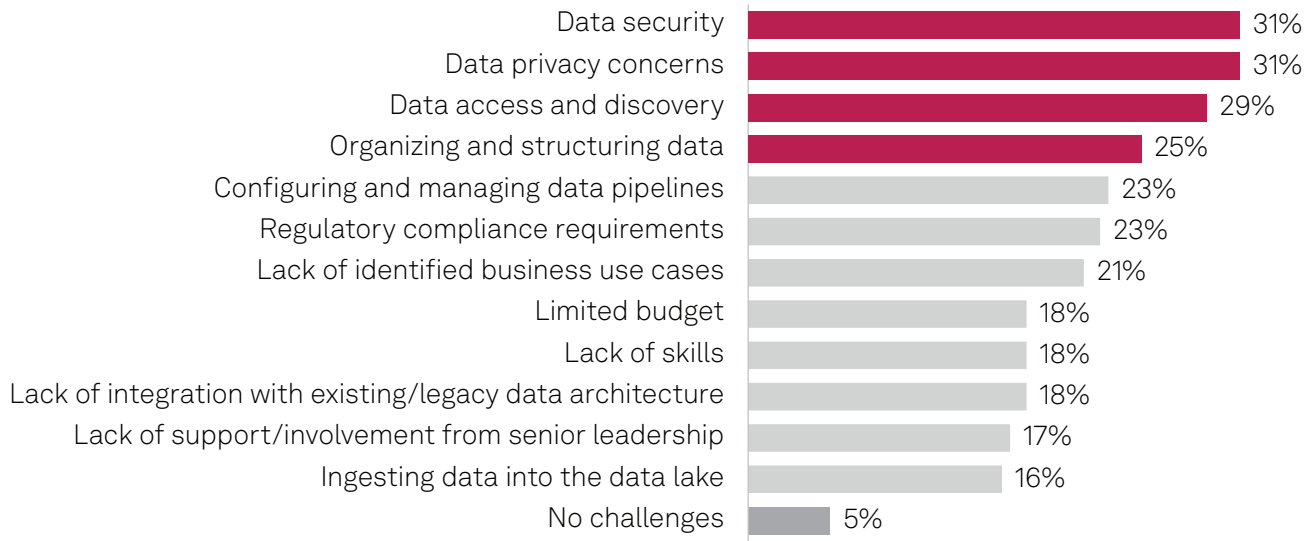
Generating business value from data lake projects is easier said than done, and key challenges remain: Survey respondents report that they still experience difficulty aligning technologically driven PoC projects with longer-term business goals, integrating the data lake with existing data and analytics infrastructure, and operationalizing multiple workloads. Other concerns include data quality, data curation and data governance requirements.

In more than a decade, the data lake has evolved to encompass a sophisticated set of tools and capabilities that collectively function as the foundational data repository supporting a variety of enterprise workloads. Much of this evolution has occurred at the storage layer, specifically around cloud object storage. Many of the data lake's enabling capabilities derive from innovations built into the cloud object storage platform. Cloud object storage plays a significant role in choosing a data lake because it often determines the performance, scalability, durability and elasticity of the data lake, among other factors. It is important that as market demands change, object storage platforms keep pace by supporting new workloads such as generative AI, as well as emerging and evolving open-source projects. Many enterprise customers, for instance, use open-source Parquet files to store tabular data in Apache Iceberg tables, which allows enterprises to analyze their data using a SQL-based query engine. While Iceberg tables often require regular maintenance, some cloud providers incorporate this as part of their object storage capabilities.

Challenges with data lakes

Early data lakes were primarily on-premises environments; today's data lakes are significantly more advanced. They are based on cloud object storage and incorporate data durability, elasticity and massive scalability as well as enable integration with multiple analytics engines that drive business decision-making. However, despite these benefits, enterprises continue to face challenges (see Figure 1).

Figure 1: Challenges in gaining insight from data lakes



Q. And what are the most significant challenges your organization faces in generating insight from your data lake environment? Please select all that apply.
Base: Respondents whose organization's primary approach is to keep existing data lake deployment(s) (n=191).
Source: 451 Research's Voice of the Enterprise: Data & Analytics, Data Platforms & Real-Time Analytics 2023.

Our survey results show that data security and privacy remain top challenges for enterprises: As long as enterprises store valuable business data, security will continue to be a priority. Public cloud environments possess strong security, often greater than other environments, which negates some of these perceived concerns.

But security is not an organization's only concern. Ultimately, enterprises want to access and leverage their data to drive business decisions, create new products and services, and differentiate themselves from their peers. At the same time, they don't want to be burdened with structuring data — for example, they can benefit from object storage that automatically generates metadata for querying newly added data. As such, 29% of respondents cite challenges with accessing data and data discovery, while a quarter of respondents report that organizing and structuring their data is a significant challenge in gaining insight from it.

Benefits of data lakes

With the majority of respondent enterprises already implementing a data lake and another 20% expecting to do so in the next three years, it stands to reason that data lakes provide significant benefits. According to 451 Research survey data, the top three benefits of a data lake are the ability to improve or develop products and services (38% of respondents), improve business agility (36%) and improve regulatory compliance (34%).

Figure 2: Data lake benefits

Q. What are the most significant benefits your organization expects from your data lake environment? Please select all that apply.
Base: Respondents whose organizations primary approach is to keep existing data lake deployment(s) (n=190).
Source: 451 Research's Voice of the Enterprise: Data & Analytics, Data Platforms & Real-Time Analytics 2023.

Certainly, most would agree that driving new revenue streams from new products or services is a welcome benefit of data lakes, and a big part of that benefit is due to the agility that data lakes provide. In short, business agility allows enterprises to adjust to changing market and environmental conditions. Data lakes' ability to store a variety of data types in their raw format enables faster data analysis and insight, and it allows businesses to get a full picture of their operations without constraints on data type.

An often-overlooked benefit of data lakes is that they can help improve regulatory compliance. Early on, data compliance was a challenge for data lake deployments given their use of on-premises Hadoop Distributed File System storage. The move to cloud object storage has significantly improved the data lake's ability to help enterprises with data compliance. However, not all cloud object platforms are created equal. Beyond security and privacy controls, enabling data durability and data elasticity are features that can help enterprises meet regulatory requirements. Further, having control over incoming writes — such as the ability to perform checks on incoming data or avoid overwriting existing data — is likewise important to ensure enterprises can meet compliance regulations.

Use cases

An advantage of data lakes is that they provide a platform for running multiple workloads. These workloads may include business analytics, data processing, machine learning, data engineering, data cleansing, real-time analytics, streaming analytics, data warehousing, operational reporting and text mining. While enterprises perform a variety of workloads on their data lake, all of these workloads fall into one of four broad use cases:

- Analytics (business intelligence, data warehousing)
- Data science (machine learning, generative AI)
- Operational reporting
- Migration and staging

The flexibility enabled by a data lake is largely based on its architecture, which separates compute processing from the underlying storage layer. A host of processing engines drive various workloads, including those targeted at tabular data; some of the most common processing engines (many are available as open source) are Spark, Trino and Flink, as well as many SQL query engines. While each compute engine has workload-specific advantages and disadvantages, the object storage layer remains the same regardless of the compute engine used.

The ability to use a variety of computing engines on the same data is an organizational benefit and a driving force behind the use of data lakes. For example, enterprises often use data lakes for analytics, but operational reporting — generating real-time insights — is also becoming a popular use case. It requires a fast processing engine and a cloud object storage platform that can handle incoming data quickly for immediate querying.

But a multiple-engine approach is not without its challenges. For instance, when enterprises run data warehousing workloads for business intelligence, there can be transactional correctness issues when multiple engines attempt to operate on data tables (Apache Iceberg, for instance) at the same time. This could be a challenge if the cloud object storage platform lacks capabilities to account for data correctness issues.

Conclusion

Data lakes have evolved from largely on-premises deployments to operating mostly in the cloud. Today's data lakes function as the organization's foundational data repository, enabling a multitude of workload types. Data lake flexibility is driven by the transition from on-premises storage to cloud object storage, which allows multiple teams to access data quickly for a variety of purposes. Enterprises can use multiple compute engines, yet the underlying storage layer remains a constant — this is the enabling power of the data lake. The object storage platform must not only accommodate multiple processing engines, but also directly integrate with multiple services within the cloud platform where it resides, including analytical, machine learning and AI services. Due to the enterprise reality of hybrid and multicloud infrastructure, the data lake, and specifically the object storage platform, must also integrate with other cloud environments and systems regardless of location.

When enterprises contemplate their data environment options, a data lake should be in the list of considerations. However, the more impactful decision is not whether to deploy a data lake, but where to store the enterprise data, because the real power of a data lake rests with the cloud object storage platform and its ability to integrate with other cloud services, processing engines and open-source tools.

About the author

James Curtis

Senior Research Analyst, Data, AI & Analytics

James Curtis is a senior research analyst at S&P Global Market Intelligence 451 Research, leading the database and data platforms vertical within the Data, AI & Analytics channel. Previously, he covered business intelligence, analytics and reporting, along with machine learning and data science. James' current areas of concentration include database and related technology analysis, real-time analytics, cloud computing and cloud-native technologies for database as a service, database optimization tooling, generative AI and retrieval augment generation (RAG) environments, including vector stores and semantic search.

James arrived at S&P Global Market Intelligence through its 2019 acquisition of 451 Research, which he joined in 2015. Prior to 451 Research, he held several senior roles in technology, marketing and communications. He served as VP of a large BPO firm where he oversaw marketing for analytic solutions. He held senior technical marketing roles at Netezza and later at IBM with responsibility for data warehousing, analytics and big-data products. He has managed global programs at HPE and worked as a case editor at Harvard Business School.

James holds a bachelor's degree in English from Utah State University, a master's degree in writing from Northeastern University in Boston and an MBA from Texas A&M University.

About S&P Global Market Intelligence

At S&P Global Market Intelligence, we understand the importance of accurate, deep and insightful information. Our team of experts delivers unrivaled insights and leading data and technology solutions, partnering with customers to expand their perspective, operate with confidence, and make decisions with conviction.

S&P Global Market Intelligence is a division of S&P Global (NYSE: SPGI). S&P Global is the world's foremost provider of credit ratings, benchmarks, analytics and workflow solutions in the global capital, commodity and automotive markets. With every one of our offerings, we help many of the world's leading organizations navigate the economic landscape so they can plan for tomorrow, today. For more information, visit www.spglobal.com/marketintelligence.

CONTACTS

Americas: +1 800 447 2273
Japan: +81 3 6262 1887
Asia-Pacific: +60 4 291 3600
Europe, Middle East, Africa: +44 (0) 134 432 8300

www.spglobal.com/marketintelligence
www.spglobal.com/en/enterprise/about/contact-us.html