451 Research

S&P Global
Market Intelligence **Business Impact Brief**

The Data Lakehouse as a Unified Architecture for Modern Data and Analytics

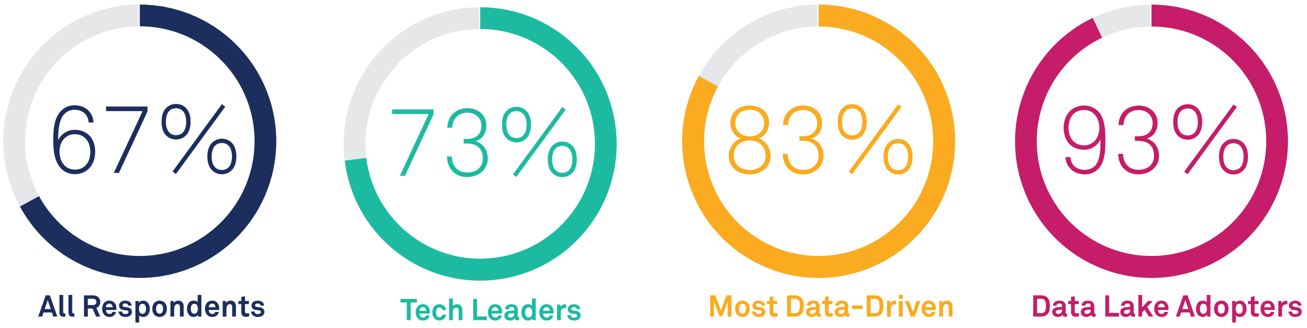
The 451 Take

The concept of the data lakehouse became reality during 2020 as open source and commercial products evolved to combine the data structure and data management features of a data warehouse with the openness and low-cost storage advantages of a data lake. When the data lake emerged a decade ago, it promised primarily to provide a single data environment that could be used for multiple analytics projects. Delivering on that promise proved to be easier said than done: while the data lake provides a more agile environment than a traditional data warehouse, delivering efficient and performant analytics on a data lake requires appropriate data engineering processes, as well as data management and governance functionality.

Enter the data lakehouse, which can be seen as delivering on the original promise of the data lake by combining the cost and flexibility advantages of persisting data in cloud storage while delivering ACID (atomic, consistent, isolated, durable) transactions for data reliability and schema enforcement for curated subsets of data, where appropriate. This notion of appropriateness separates the data lakehouse from more traditional data warehouse architecture, in which the enforcement of schema is a prerequisite. Lakehouse platforms are built on open and standardized file formats, such as Parquet, and provide direct access to data with APIs for a variety of tools including machine learning and Python/R libraries. While structure is applied in the lakehouse to serve traditional analytics workloads that rely on structured data, the lakehouse also retains the flexibility of the data lake to support data science and machine workloads that more commonly use semi-structured and unstructured data. This coexistence of use cases applied to the same data foundation is one of the defining attributes of a data lakehouse.

Interest in the Data Lakehouse Is Growing

Organizations that are using or planning to adopt data lakehouse architecture within the next 12 months



Source: 451 Research's Voice of the Enterprise: Data & Analytics, Data Platforms 2021

Demand for the combination of functionality delivered by the lakehouse is the inevitable result of the increased volume of data being stored in cloud storage services: the majority of enterprises surveyed by 451 Research see object storage becoming a primary data platform for data processing and analytics.

Interest in the data lakehouse concept is also growing rapidly. More than two thirds (67%) of enterprises are currently using or piloting a data lakehouse environment, or plan to do so within the next 12 months, according to respondents to 451 Research's Voice of the Enterprise: Data & Analytics, Data Platforms 2021. That figure rises to 73% among tech leaders (earlier adopters); it rises to 83% among the most data-driven, and 93% among enterprises that already have a data lake in production.

Business Impact

Reduction of data silos and data integration overhead. The data lakehouse provides a single unified environment for storing and analyzing structured, semi-structured and unstructured data in low-cost cloud storage, reducing data silos and integration challenges.

Support for multiple use cases, applications and business units. The ability to apply structure when it is appropriate means that the data lakehouse fulfills the advantage promised by the data lake by better enabling multiple use cases, applications and business units to be served by a single unified data repository.

Enable different users to apply their preferred tools to the same data. Supporting multiple use cases, applications and business units enables users in different roles (such as business decision-makers, data analysts, data engineers and data scientists) to bring their respective tools and skills to the same data without the need for data duplication and the risk of data fragmentation.

Provide flexibility to respond to evolving data. Unlike a traditional data warehouse, in which structure and schema are enforced as the data is ingested and stored, the lakehouse environment enables structure to be applied only to the data and use cases for which it is required, providing flexibility to adapt as the source data evolves.

Support business agility by accelerating new analytics initiatives. Applying schema and structure to data as required enables the data lakehouse to accelerate the development of new analytics applications and use cases by avoiding the need to configure and deploy dedicated data processing workloads for each new initiative.

Looking ahead

The data lake has become an integral component of many enterprise data and analytics strategies, and the use of cloud object storage services as the basis of a data lake environment to store large volumes of structured, semi-structured and unstructured data continues to gain momentum. We see wisdom in the desire to bring the structured analytics advantages of data warehousing – specifically support for ACID transactions, updates and deletes, and schema enforcement – to data stored in low-cost cloud-based data lakes, especially for data types and workloads that do not lend themselves naturally to relational databases.

While data lakes and data warehouses can be complementary, we also believe there are clear performance and efficiency advantages in bringing structured data processing concepts and functionality to data in the data lake, rather than having to export it into external data warehousing environments for analysis. The starting point for a data lake was low-cost object storage combined with the ability to query data using interactive query engines. The addition of a structured transactional layer that turns the data lake into a data lakehouse provides additional functionality that will enable the lakehouse to serve as an integral platform for accelerating business insight in the years to come.

databricks

Databricks is the data and AI company. Built on lakehouse architecture, Databricks combines the best of data warehouses and data lakes to offer a simple, open, and collaborative platform for all data workloads. More than five thousand organizations worldwide — including Shell, Comcast, CVS Health, HSBC, T-Mobile, and Regeneron — rely on Databricks for massive-scale data engineering, exploratory data science, full-lifecycle machine learning, and business analytics. With a global presence and hundreds of partners, including Microsoft, Amazon, Google, and Tableau, Databricks is on a mission to help data teams solve their toughest problems.