BRIEF

Not your ex-boss's data lake

7 things to look for in a data lakehouse solution

Starburst

In today's business environment, enterprises that can leverage more data faster than competitors for analysis and insights will have a clear edge. Yet, as the volume, velocity, and variety of data grow, it can be challenging to manage it all.

Cloud-based data lakes and warehouses offer basic data storage, but they are increasingly inadequate for today's enterprise needs. Companies must typically supplement a data lake with multiple data warehouses and other specialized systems to manage their data adequately. This introduces additional complexity, delays, and governance issues as data is copied from one silo to another.

A data lakehouse combines the capabilities of data lakes and data warehouses, providing a single platform for cloud data, kept in a common storage medium across the whole architecture, both for warehousing and machine learning (ML). Data lands in the lakehouse, where users can structure and consume it all in one location.

What to look for in a data lakehouse solution

Although there are many advantages to a data lakehouse architecture, it's important to understand that not all data lakehouses are created equal. The best choice is one that allows you to easily access and analyze data from disparate sources, without having to move or copy data to a separate data warehouse. This approach gives you complete control over your data and frees you from vendor lock-in.

Here are 7 things to look for in a data lakehouse solution:

1

True separation of compute and storage

Freedom of choice requires the separation of storage and compute, which can bring a new level of cost and performance efficiency. Store all your data at the source and/or in ADLS, and simply scale your compute up or down to meet your needs. Your application should be able to act on data independent of its location so that you can choose the least expensive storage option and have the flexibility to apply any application you need. This looser approach lets you leverage different engines to access the same data.

2

Data federation

Although a single cloud data repository is ideal, data decentralization remains a reality, particularly for complex global organizations. Choose a solution that can access raw data from multiple data stores within a single query. Make sure you can run analytic queries across cloud data sources and on-prem systems.

3

Connectivity

Look for a comprehensive set of enterprise connectors for common data stores to minimize data duplication, control the dial on cost per performance, and use multiple platforms to access the same data. Beware of proprietary query engines and data storage formats, which can limit flexibility in terms of data sources and integrations and lead to vendor lock-in.

4

Query fault tolerance

Be sure that you can enable use cases on the lakehouse — such as building large rollup tables — preparing datasets for machine learning models and wrangling data that feeds into data applications

5

Security and compliance

Look for role-based access control, authentication engine integration, data encryption and masking, and query auditing. Your solution should not only ensure that the right people have access to the right data, but also provide flexibility in adopting built-in and third-party security protocols. It should also let you easily configure and apply access rights to catalogs, individual schemas, tables, and views. Make sure your solution is SOC 2 Type 2 compliant, and look for ISO 27001 certification, a globally recognized security standard with a heavy focus on risk management based on controls, tracking, and addressing concerns.

6

Partner ecosystem

Make sure your solution has a rich partner ecosystem, including client connections and integrations with your favorite BI platforms — as well as ML/AI/workflow management integrations, so your data teams can continue to use tools they know and love.

7

Data products

Reusable data products increase efficiency and benefit any analytics journey — large or small, on-premises or multiple clouds, in a warehouse or a lake, federated across all locations. Whatever solution you choose should let your data engineers create, manage, and share data products, data catalogs, and materialized views to speed the discovery and consumption of high-quality data. They should also be able to leverage those data products easily, using existing tools. Look for the ability to create, discover, publish, manage, and share data products based on multiple datasets. In addition, choose a solution that offers curated, high-quality datasets that can shorten the path from raw data to trusted insights.

Starburst Galaxy: A fully managed cloud data lakehouse

A data lakehouse architecture can revolutionize yourorganization's data strategy making it faster and more cost-effective — resulting in actionable analytics and insights.

Starburst Galaxy is a fully managed service designed for running fast, interactive analytic queries against distributed data sources like data lakes and data warehouses ranging in size from terabytes to petabytes. Starburst Galaxy enables immediate analysis of siloed data without expensive data warehouse appliances.

Connect your data sources, start a cluster, and start querying using the analytics tools you already know in minutes. The solution easily autoscales up and down with demand and automatically shuts down clusters after a set period. By implementing a data lakehouse architecture incorporating Starburst Galaxy, you can supercharge your cloud data management.

Get Started with a $500 credit on Starburst Galaxy

Learn More

Starburst

Need a Cloud Data Lakehouse Solution? Look for these 7 Capabilities