BRIEF

How to Implement a Data Lakehouse on Azure

Starburst | Microsoft

As enterprises continue to work toward extracting value from growing volumes of data as quickly as possible, they're finding that Azure Data Lake (ADL) doesn't offer the breadth of capabilities they need. Companies must typically supplement ADL with several 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. As a result, more and more organizations are turning to an approach to their data strategy on Azure with a data lakehouse at its center. A data lakehouse combines the capabilities of data lakes and data warehouses. Think of them as data lakes on steroids, bringing the functionality of the warehouse to end users — with the scalability and lower cost of a data lake.

What to look for in a data lakehouse solution

Although there are many advantages to a data lakehouse solution, it's important to understand that not all data lakehouses are created equal. The best choice is one that allows you complete control over your data and frees you from vendor lock-in. To begin with, your solution 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. By the same token, data decentralization remains a reality, particularly for complex global organizations, so they need a solution that can access raw data from multiple data stores within a single query. **In addition to these primary capabilities, best-in-class data lakehouse solutions should offer:**

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.

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.

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.

Data products

Reusable data products increase efficiency and benefit any analytics journey. 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.

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.

Best practices for data lakehouse implementation on Azure

Azure can be an ideal environment for your data lakehouse solution, but to get the most out of it, you should plan your architecture with a few best practices in mind:

Separate compute and storage: While some solutions bind compute and storage, separating the two can lower cost, increase efficiency and prevent vendor lock-in. Store all your data at the source and/or in ADLS, and simply scale your compute up or down to meet your needs. This decoupled approach lets you leverage different engines to access the same data.

Base layer: Build "data gravity" in ADLS: Although a single cloud data repository is ideal, data decentralization remains a reality, particularly for complex global organizations. Start by building data gravity in ADLS, while at the same time retaining access to raw data sources. This will give you the speed of centralized data with the flexibility of a federated structure. It also serves as the base of a three-tiered architecture that will provide your organization with cost savings, flexibility, and the ability to leverage multiple products to achieve a tailored solution. The base layer is storage – which ADLS serves well.

Middle layer: Maintain data in open table format: The middle tier standardizes the data and makes it available to an upper tier of applications which can safely work with the same tables at the same time. Open table formats such as Apache Iceberg allow you to perform data warehouse-like queries directly in a data lake. Users can interact with the data lake as easily as they would a database using SQL. This reduces the need for data movement migration, which provides substantial cost and time savings. Additionally, these open table formats provide increased performance benefits over traditional formats.

Top layer: Combine applications to optimize management and analytics: With your data in an open table format, you're free to plug in different applications to process the data as needed. This is where data and compute separation come into play. Rather than each application working on its own data set, you can have as many applications as you need working on the same data at the same time. You now have access to best-in-class ingestion, ETL, ML, BI, and query engines to get the most out of your data.

Starburst Galaxy: A fully managed data lakehouse for Azure

A data lakehouse architecture on Azure can revolutionize your organization'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 Azure data management.

To learn more about Starburst Galaxy on Azure, find us on the Microsoft Azure Marketplace.

Learn More

Starburst | Microsoft