Unity Catalog (Data Lakehouse)

Last updated by | Vamsi Kode | Sep 11, 2023 at 2:02 PM EDT

What is a Data Lakehouse?
A data lakehouse is a new, open data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data.

Simple Overview
Data lakehouses are enabled by a new, open system design: implementing similar data structures and data management features to those in a data warehouse, directly on the kind of low-cost storage used for data lakes. Merging them together into a single system means that data teams can move faster as they are able to use data without needing to access multiple systems. Data lakehouses also ensure that teams have the most complete and up-to-date data available for data science, machine learning, and business analytics projects.
Evolution of data storage, from data warehouses to data lakes to data lakehouses

Metadata layers for data lakes
new query engine designs providing high-performance SQL execution on data lakes
optimized access for data science and machine learning tools.
Metadata layers, like the open source Delta Lake, sit on top of open file formats (e.g. Parquet files) and track which files are part of different table versions to offer rich management features like ACID-compliant transactions. The metadata layers enable other features common in data lakehouses, like support for streaming I/O (eliminating the need for message buses like Kafka), time travel to old table versions, schema enforcement and evolution, as well as data validation. Performance is key for data lakehouses to become the predominant data architecture used by businesses today as it's one of the key reasons that data warehouses exist in the two-tier architecture. While data lakes using low-cost object stores have been slow to access in the past, new query engine designs enable high-performance SQL analysis. These optimizations include caching hot data in RAM/SSDs (possibly transcoded into more efficient formats), data layout optimizations to cluster co-accessed data, auxiliary data structures like statistics and indexes, and vectorized execution on modern CPUs. Combining these technologies together enables data lakehouses to achieve performance on large datasets that rivals popular data warehouses, based on TPC-DS benchmarks. The open data formats used by data lakehouses (like Parquet), make it very easy for data scientists and machine learning engineers to access the data in the lakehouse. The other features of a data lakehouse, like audit history and time travel, also help with improving reproducibility in machine learning.

How does Unity Catalog helps with our current Data Warehouse, Data Governance and Metastore requirements ?

https://007gc.sharepoint.com/sites/DataHubsCommitteeComitHubsdesdonnes/Shared Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataHubsCommitteeComitHubsdesdonnes%2FShared Documents%2FEnterprise Data Hubs%2F04-Design%2FUnity Catalog-Data Lakehouse%2F[ECCC] Unity Catalog Adoption Guide.pdf&viewid=ac731a40-5bff-46ac-8dc8-d0e99223e9b0&parent=%2Fsites%2FDataHubsCommitteeComitHubsdesdonnes%2FShared Documents%2FEnterprise Data Hubs%2F04-Design%2FUnity Catalog-Data Lakehouse

Unity Catalog Setup.
https://007gc.sharepoint.com/:w:/r/sites/DataHubsCommitteeComitHubsdesdonnes/Shared Documents/Enterprise Data Hubs/04-Design/Unity Catalog-Data Lakehouse/unity_catalog_walk_through.docx?d=wea43ed0b6513421fb77cf68702502c19&csf=1&web=1&e=NAvgk1

Reference documents:
https://007gc.sharepoint.com/:b:/r/sites/DataHubsCommitteeComitHubsdesdonnes/Shared Documents/Enterprise Data Hubs/04-Design/Unity Catalog-Data Lakehouse/cidr2021_paper17.pdf?csf=1&web=1&e=6O45Ke

https://007gc.sharepoint.com/:b:/r/sites/DataHubsCommitteeComitHubsdesdonnes/Shared Documents/Enterprise Data Hubs/04-Design/Unity Catalog-Data Lakehouse/The-Data-Lakehouse.pdf?csf=1&web=1&e=ycI2LL

Public Documentation:

https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/

https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/get-started

Create a storage account for Azure Data Lake Storage Gen2 - Azure Storage | Microsoft Learn

Use Azure managed identities in Unity Catalog to access storage - Azure Databricks | Microsoft Learn

https://learn.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/how-manage-user-assigned-managed-identities

Unity Catalog best practices - Azure Databricks | Microsoft Learn
https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/

Using Unity Catalog with Structured Streaming - Azure Databricks | Microsoft Learn
https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/get-started

Tutorial: Unity Catalog metastore admin tasks for Databricks SQL - Azure Databricks | Microsoft Learn

Unity Catalog privileges and securable objects - Azure Databricks | Microsoft Learn

Working with Unity Catalog in Azure Databricks - Microsoft Community Hub