Building a Lake House on AWS

Myles Brown | Friday, October 28, 2022

Organizations have been moving to the cloud steadily for well over ten years. In that time, they have amassed quite a bit of data on customers and internal operations, but that data is not always stored in a way that easily can be accessed by all parties. With advancements in data analysis and the ability to use datasets for machine learning (ML) and artificial intelligence (AI), some organizations have made great strides by moving to a modern architecture that allows them to store and analyze both structured and unstructured data: a Lake House. This article will explain the benefits of building a Lake House on AWS.

Both data warehouses and data lakes have their pros and cons. Traditional data warehouses only store structured data, but they query quite well. Data lakes store both structured and unstructured data, but because there's no built-in query analytics, you need a query engine and a data catalog to make use of your data. Wouldn't it be great if you could marry these two into one system? You basically can with a Lake House architecture, which allows you to store and query all types of data. A Lake House on AWS connects your data lake, your data warehouse, and all your other purpose-built services into one shared catalog. Once you build your Lake House in AWS, you can store, secure and analyze your data, and control its access.

A Lake House architecture allows you to store your data in an easy-to-access data lake and to manage and analyze that data in the same quick reliable way found in data warehouses, providing quick performance and high reliability and integrity to support all your workloads.

Benefits of a Lake House on AWS

The heart of the Lake House architecture is the data lake on Amazon S3. But data also can be stored in an Amazon Redshift data warehouse or other purpose-built database services like Amazon Aurora or Amazon DynamoDB. This means that you can store any kind of data and choose the best place for it to live, based on storage costs, size, expected throughput, and the type of tools you want to use to access that data. No matter where that data lives in Amazon S3, it will have universal governance and access control.

A Lake House provides the following abilities:

- To integrate data stored in data warehouses, data lakes and purpose-built storage services so they can be accessed from one location to run analytics across all data
- To house all data, regardless of its format, in a single virtual location
- To apply linear, non-linear, tree and hash structures that are typically used in a warehouse to deliver analytics to the type of unstructured data usually stored in a lake. This allows users to access insights faster and put them to work without delay.
- To deploy intelligent metadata, allowing you to identify and extract features that enable the data to be cataloged and indexed in the same way structured data is processed.

Below is a list of the primary characteristics of a Lake House:

- Uses structured, semi-structured and unstructured data
- Scales automatically with data load
- Ingests data rapidly
- Provides a unified interface for processing and consumption
- Prevents the need to move data between warehouses and lakes

How a Lake House Functions

The power of a Lake House comes from its ability to integrate data stored in warehouses and lakes using a unified interface that brings your multiple systems together. It's easy to add new data sources, support new use cases and develop new methods for data analytics. However, adopting any new technology can be cumbersome, and a Lake House is no exception. You need to determine how to best architect your Lake House and the platforms and tools that will best suit your needs.

Building a Lake House on AWS

AWS offers an excellent platform on which to build your Lake House. As the Lake House architecture develops, you'll need to choose between various standards for the data framework.

Governed Tables is the leading proprietary framework for a Lake House on AWS. This feature of AWS Lake Formation provides the data lake with advanced features such as ACID (atomic, consistent, isolated and durable) transactions, data compaction and time-travel queries. Governed Tables comes with the disadvantage of vendor lock-in.

However, you don't have to use Governed Tables. Although they can be a little more administrative effort, using an open source framework instead of Governed Tables can provide much of the same functionality but avoid vendor lock-in. You can use one of the following three open source solutions:

- **Apache Hudi:** An open source data management framework, originally developed at Uber, that simplifies incremental data processing and data pipeline development
- **Apache Iceberg:** An open platform, originally developed by Netflix and Apple, to provide a highly scalable and flexible analytic engine and services without vendor lock-in for Lake House, hybrid and multicloud environments
- **Delta Lake:** An open source storage framework, originally developed by Databricks, for building a Lake House architecture using existing data lakes such as AWS

These open source options offer some advantages, such as a lower risk of vendor lock-in and the chance to continue using your preferred tools. They can be used to help manage a data lake in AWS or in any other cloud. Still, their lack of integration with AWS can require extra administration.

AWS Lake House Analytics

AWS has two distinct services, Amazon Athena and Amazon Redshift Spectrum, which enable unified access to data wherever it's stored in the Lake House. These applications allow federated queries across different storage options and are built for specific use cases, depending on the amount of data involved and the desired results.

- **Athena:** This serverless interactive query function was originally designed for data stored in Amazon S3 and uses standard SQL. Athena is designed to be easy to use and enables anyone with SQL skills to analyze large-scale data sets with rapid results. Athena integrates with AWS Glue Catalog to manage data across various services, including S3, Amazon Redshift, and other purpose-built data storage services.
- **Redshift Spectrum:** This serverless data warehouse solution is designed to support online analytical processing for petabyte-scale enterprise structured datasets. The natively integrated tool can query multiple datasets in data lake and warehouse storage, as well as other purpose-built storage services.

The primary difference between Athena and Redshift Spectrum is that Athena uses a general SQL engine that supports ANSI standard SQL, while Spectrum queries run within Redshift and uses its query engine. Therefore, Spectrum queries can easily and quickly join tables directly with data already ingested into the Redshift data warehouse. Users with most of their data in an S3 data lake should consider Athena, while users who need queries closely tied to a Redshift data warehouse should opt for Redshift Spectrum.

Preparing to Build a Data Lake House

ExitCertified delivers courses on some of the most in-demand AWS cloud computing solutions. The main Data Analytics on AWS course is **Building Modern Data Analytics Solutions on AWS**, which is a four-day bundle of the following one-day classes:

- Building Data Lakes on AWS
- Building Batch Data Analytics Solutions on AWS
- Building Streaming Data Analytics Solutions on AWS
- Building Data Analytics Solutions Using Amazon Redshift

Introduction to ML and Amazon SageMaker | View Course

Myles Brown
Change is the only constant in life. Heraclitus wrote that in 500 BC but it has never been truer than in my last twenty years of IT. In that time I have worked as a developer, architect, and trainer of Oracle, Java, Spring, Cloudera, AWS, and Google Cloud. Technologies come and go, which really scratches the itch for a lifelong learner like me. Cloud and DevOps have been my obsessions for the past five years, and it brings me great satisfaction to help an organization make a successful digital transformation.

Previous Article | Next Article

RELATED COURSES

Data Warehousing on AWS

POPULAR LINKS

Getting Started with Red Hat OpenShift Service on AWS and Red Hat Training [Webinar]

TAGGED TOPICS

Article | AWS | Data Warehousing

SUBSCRIBE TO OUR NEWSLETTER

Name
Email
SUBSCRIBE

Popular Posts | Latest Posts | Recommended | Search Articles

5 Tips for Going Cloud Native and Getting the

https://www.exitcertified.com/training-resources/white-papers/accelerate-your-modernization-efforts-with-a-

Cloud

Top Microsoft Azure Blogs to Follow in 2022

ExitCertified ranked the top 10 Microsoft Azure blogs you should follow in 2022 to stay up to date on the latest

News

How Microsoft Azure Works

Learn how the Microsoft Azure platform works and its key benefits for day-to-day business operations, including it's

Cloud