First International Workshop on Composable Data Management Systems, 2022

Delta Lake: The one storage format for all your analytics

(Invited Talk)
Tathagata Das
Staff Software Engineer
Databricks, Apache Spark PMC
tdas@databricks.com

Abstract

Delta Lake is an open-source storage format that brings ACID transactions to big data workloads on cloud object stores. Traditionally, there has been two kinds of data management frameworks:

- Databases that provide transactional guarantees but are not scalable in a cost-effective way
- Data lakes that provide cost-effective scalability, but limited transactional and data quality guarantees

The Delta Lake format combines the best of both worlds for analytical workloads:

- Scales in a cost-efficient by storing data - both data and metadata scale to 10s of TBs of data
- Provides data quality guarantees - ACID transaction, schema enforcement, data constraints
- Simplifies modern workloads on ever changing data - Schema evolution when appending or upserting data

Delta Lake allows you to store data on blob stores like HDFS, S3, Azure Data Lake, GCS, query from many processing engines including Spark, PrestoDB, Trino, Hive, Flink, and provides APIs for Scala, Java, Python, Rust, and Ruby. In this talk we are going to discuss:

- Architecture - Why does it scale so well?
- Features - What makes it so unique?
- Roadmap - What is in the future?
- Connector ecosystem - What can you read / write from?
- Community - How to contribute and engage?

Biography

Tathagata Das is a Staff Software engineer at Databricks, and a member of Apache Spark Project Management Committee (PMC). He has been involved with the Apache Spark project for the last 12 years. He developed the original Spark Streaming (DStreams) in his grad student days in AMPLab, UC Berkeley. He was one of the core developers of Structured Streaming, and since 2018, one of the core developers of the Delta Lake product. Currently, he leads the development of the Delta Lake open source project and the ecosystem around it.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License and appears in CDMS 2022, 1st International Workshop on Composable Data Management Systems, September 9, 2022, Sydney, Australia.