

STARBURST ENTERPRISE PRESTO: NATIVE DELTA LAKE READER

The secure, enterprise-grade distribution of the open source Presto SQL query engine now includes a Native Delta Lake Reader.

Open-sourced by Databricks in 2019, Delta Lake enables data modification and optimizations in data lakes. The Native Delta Lake Reader helps Databricks customers take advantage of Presto's speed, concurrency, and scalability to query their Delta Lake. Starburst and Databricks share many of the same enterprise customers, and this new tool provides enterprises with greater cost control, flexibility, and speed of access to the data in their data lakes.

Databricks Delta Lake

As object storage became increasingly popular over the last decade, a frustrating flaw became apparent. Updating data such as customer or product information was a very difficult, time-consuming process. Database engineers were constantly forced to modify, join, and overwrite tables. Databricks changed this with its Delta Lake, a storage platform that lets users easily update and modify data stored in a cloud data lake. Additionally, Delta Lake provides performance and file management optimizations which didn't exist in cloud data lakes.

Starburst Native Delta Lake Reader

On a mission to power analytics anywhere, Starburst recognized the need to support the leading technology that enables ACID transactions and performance optimizations on top of object storage. Starburst's Native Delta Lake Reader was written from scratch, specifically for Delta Lake, to make it as efficient as possible. Features include:

- Fast, efficient reads of Delta Lake transaction logs
- Support for data skipping to enhance performance
- Optimization of queries using Delta Lake file statistics

Starburst always supports the best platforms first – and Delta Lake is the industry leader and a favorite tool of some of our top enterprise customers.



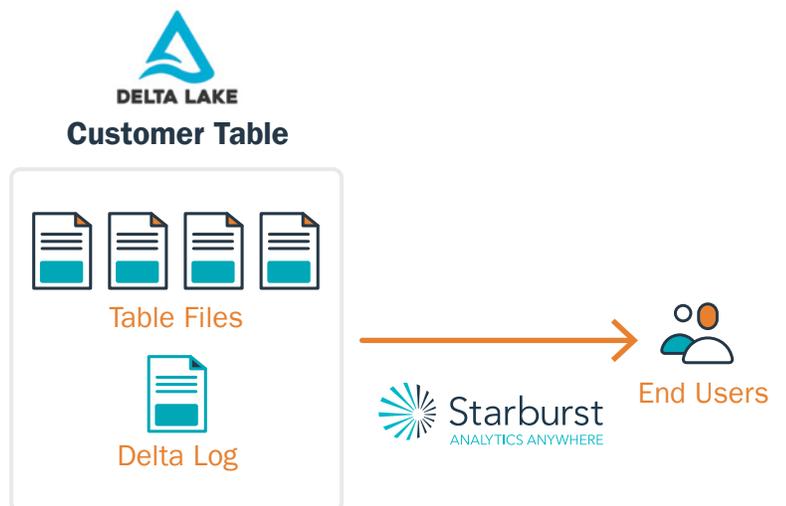
Starburst & Databricks: Complementary Platforms



Databricks and Starburst share many of the same customers, and these large global enterprises have been asking for a Native Delta Lake Reader. Our shared customers overturn a common misconception in the industry – the idea that Databricks and Starburst are competitors. Instead, the two are complementary. Enterprises use Spark and Databricks for Machine Learning, AI, ETL, and streaming ingestion, while Starburst Enterprise Presto is their high-concurrency SQL query engine, providing a single point of access to all of their data.

How the Native Delta Lake Reader Works

If your enterprise has distributed storage, whether in S3, ADLS, or another system, you probably have tables logically defined on top of these storage platforms. Files reside inside the tables, and in Delta Lake, these files are stored in an open-source format called Parquet. In Delta Lake, each time there is a change to a file – when customer information is modified, for example – this is added to the transaction log, along with an associated timestamp. Starburst’s new tool reads the Delta Log and files (when needed) and extends multiple benefits to end users.





ACID Transactions

Previously you'd have to rewrite an entire table to add a new customer address or some other additional piece of information – now you can simply run an update or merge statement.



Governance

GDPR sometimes requires companies to remove specific customer data. This was not possible with Hadoop and previous storage platforms, but Delta Lake makes it as simple as a delete or update statement. As part of our commitment to fine-grained global security, Starburst's Native Delta Lake Reader supports this functionality.



Data Skipping

Min, max, nulls, counts and other high-level statistics are stored on each file. Starburst took advantage of this when designing the Native Delta Lake Reader. Users can quickly narrow which files they actually need to query. The Starburst connector feeds this information into the Presto Cost Based Optimizer, which greatly improves performance by reducing the amount of files that actually need to be read for a query.



Vacuum

The process of updating data does leave you with many small files, and although object storage can be pretty affordable, it's not that cheap. Delta Lake allows you to run a vacuum command that clears up the table and gets rid of older files.



Optimize

The well-known small file problem can be a drag on performance, so Delta Lake added an optimize command that combines small files into larger ones, which greatly increases performance. Starburst's Native Delta Lake Reader allows users to take advantage of this as well.



Z-Ordering

After you optimize files, and are left with, say, 10,000 files instead of 100,000, you can order them by the columns of your choosing. Included in the optimize command, you can optionally choose to z-order these files by one or more columns. This increases performance for queries that include the z-ordered columns in their predicate.

To operate the Native Delta Lake Reader in Starburst Enterprise Presto, you can either keep separate metastores or utilize a shared metastore.

The Reader allows you to:

- Create tables in Starburst Presto that point to Delta Lake tables
- Execute queries on Delta Lake data as you would with normal tables
- Optimize and z-order your tables to improve performance

All of this is seamless to the end user. Overall, the Native Delta Lake Reader adds to the 27+ connectors included with the Starburst Enterprise platform, and aligns with Starburst's mission to give its customers a single point of access to all of their data – no matter where it resides.