Starburst
ANALYTICS ANYWHERE

DATA SHEET

Starburst Delta Lake Connector

Closing the gap between data lake cost efficiencies and analytic capabilities

Starburst Enterprise, the secure enterprise-grade distribution of the open source MPP SQL engine Trino, includes the Starburst Delta Lake connector.

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

Databricks Delta Lake

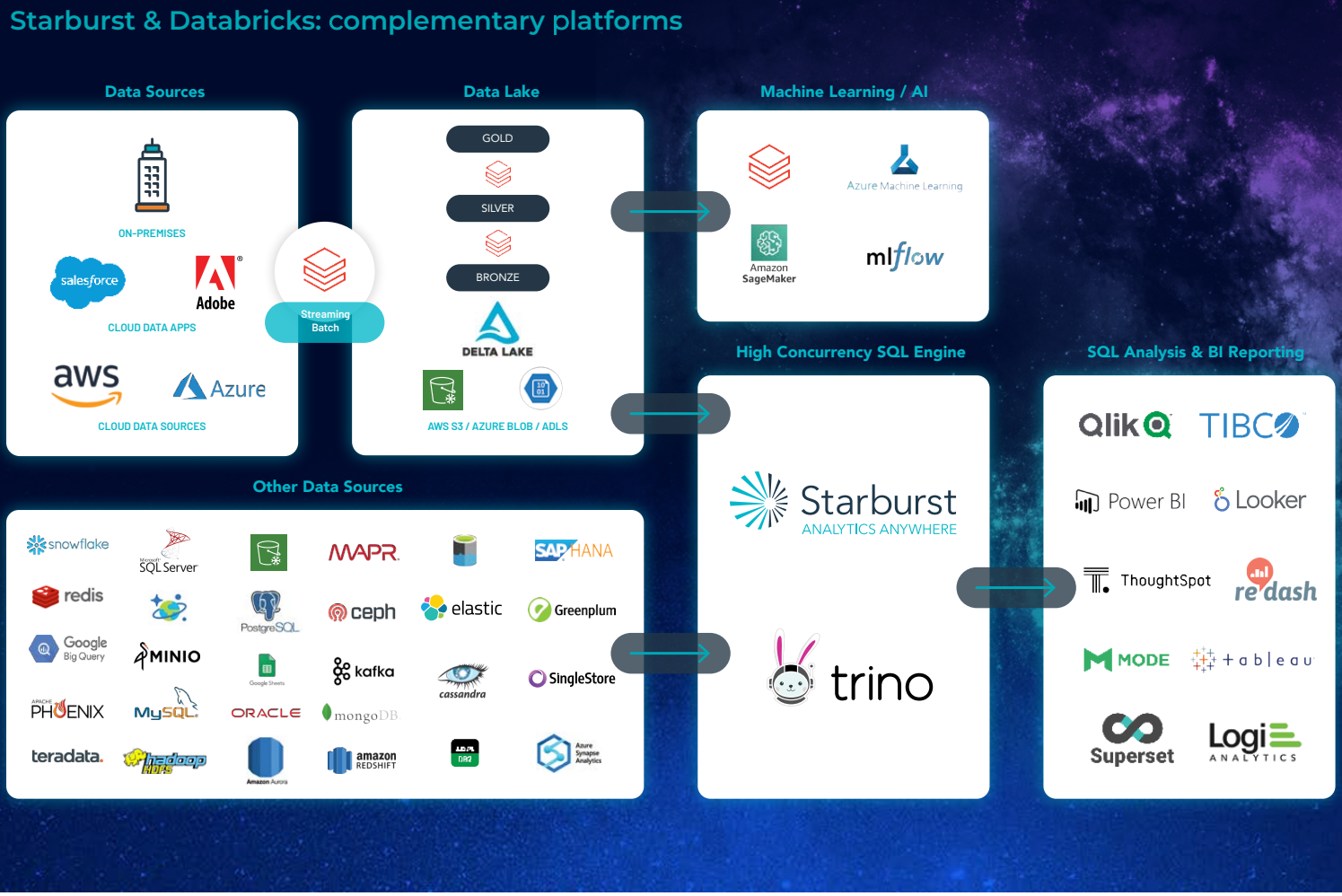
As object storage became increasingly popular over the last decade, a frustrating flaw became apparent. Updating customer or product data was a very difficult, time-consuming process. Database engineers were constantly forced to modify, join, and overwrite tables. Databricks changed this with its Delta 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 Delta Lake connector

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. The Starburst Delta Lake connector was written from scratch, specifically for Delta Lake, to make it as efficient as possible.

Features include:

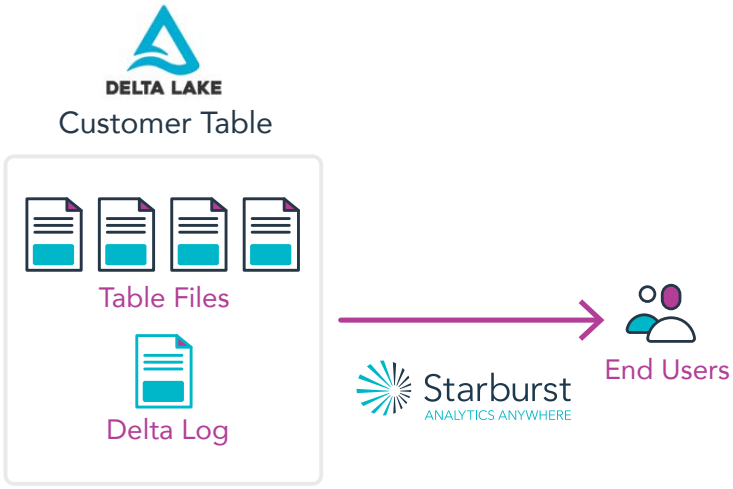
- Fast, efficient reads of Delta Lake transaction logs, with support for Amazon S3, HDFS, and Azure Storage
- Support for data skipping to enhance performance
- Optimization of queries using Delta Lake file statistics
- Parallelism that distributes processing for better performance
- Table statistics support for the Trino cost-based optimizer
- Query special columns of metadata
- Robust security features including Apache Ranger and Privacera platform integration
- Data Manipulation Language (DML) support that includes INSERT/DELETE/UPDATE



Our shared customers invalidate a common misconception in the industry — the idea that Databricks and Starburst are not complementary enterprise solutions. Enterprises use Spark and Databricks for Machine Learning, AI, ETL, and streaming ingestion, while Starburst Enterprise's distribution of Trino is their high-concurrency SQL query engine, providing a single point of access for ad-hoc analytics to all of their data.

How the Starburst Delta Lake connector works

If your enterprise has distributed storage, whether in S3, ADLS, or another system, you most likely 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 reads the Delta Log and files (when neededand 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. Starburst detects changes to a Delta table as soon as a transaction is committed, without the need to generate a manifest file or update the Hive metastore.

Governance

GDPR may require 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, the Starburst Delta Lake connector 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 Starburst Delta Lake connector. Users can quickly narrow which files they actually need to query. The Starburst connector feeds this information into our 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

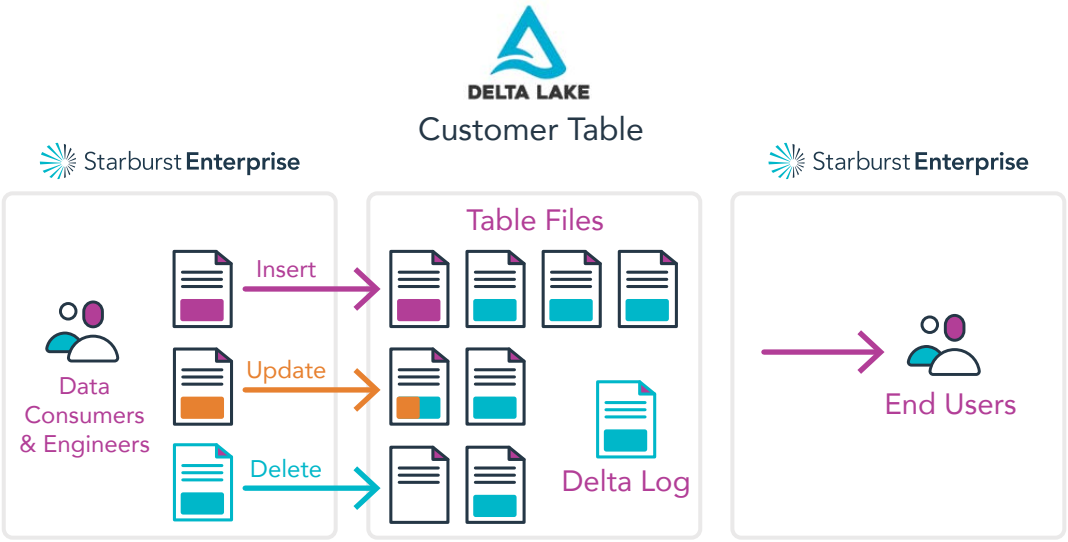
The well-known small file problem can be a drag on performance, so Delta Lake added an optimize command that combines small files into larger ones, which greatly increases performance. The Starburst Delta Lake connector allows users to take advantage of this as well.

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.

Write support

Starburst's Delta Lake connector includes support for Data Manipulation Language (DML) within AWS and Azure to empower data analysts, platform admins, and engineers with Delta Lake write capability for updating the Delta Lake themselves, saving valuable time and resources. Historically, organizations duplicated data in their data lakes across a variety of data warehouses and operational systems, presenting compliance and data quality challenges. The Starburst Delta Lake connector transforms data lake reliability and contextualizing data to deliver meaningful operational efficiencies and cost savings. The added DML feature to the Starburst Delta Lake connector simplifies data management, significantly reduces costs and complexity and accelerates data team productivity. By allowing companies to read and write to Delta Lake, Starburst can help with a path to a Lakehouse architecture. Starburst customers benefit from a fine grained ability to modify data directly in the lake, maximizing their ROI in Delta Lake



To operate the Starburst Delta Lake connector, you can either keep separate metastores or utilize a shared metastore.

The connector allows you to:

- Create tables in Starburst Enterprise that point to Delta Lake tables and create new tables with CTAS statements
- Execute queries on Delta Lake data as you would with normal tables
- Improve query performance for optimized and z-ordered tables
- Perform write operations on Delta Lake

All of this is seamless to the end user. Overall, the Starburst Delta Lake connector adds to the 40+ 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.