Starburst

Data sheet

Starburst and Delta Lake

Enhancing data lake efficiency and analytics

Closing the gap between data lake cost efficiencies and analytic capabilities

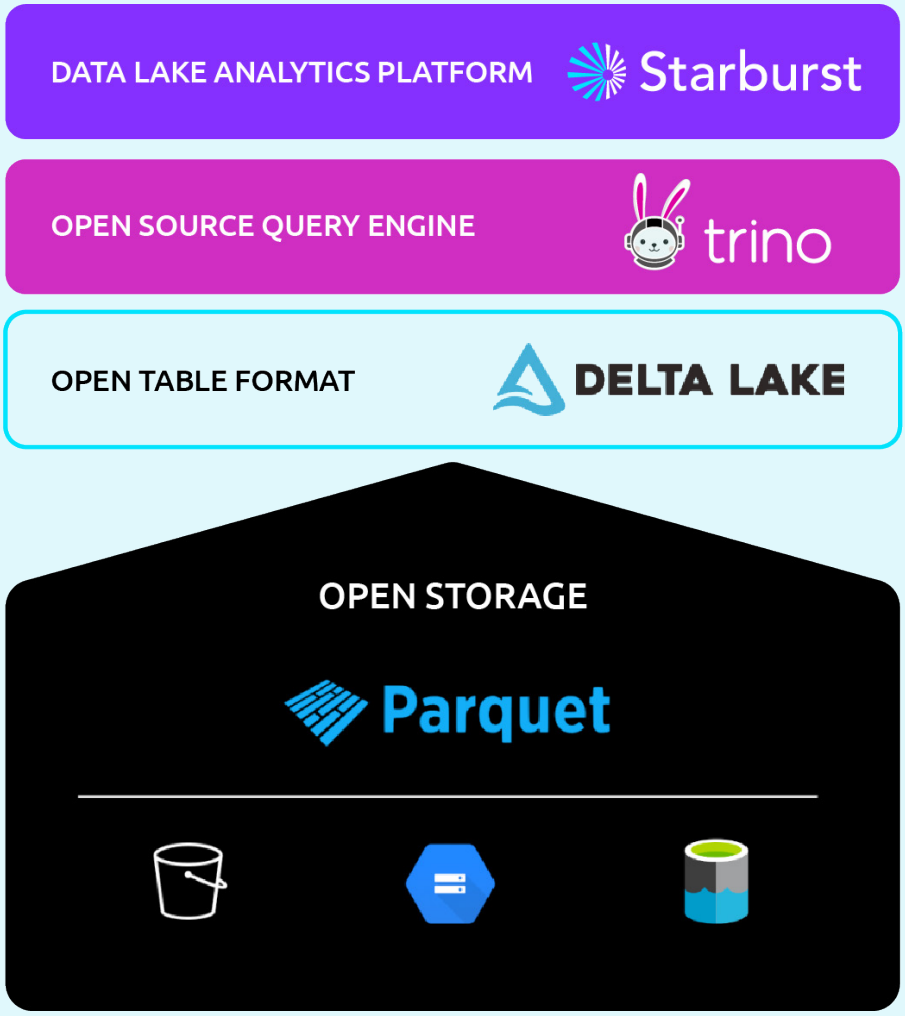
In the quest to bridge the gap between data lake cost efficiencies and advanced analytic capabilities, Starburst offers a secure enterprise-grade distribution of the powerful open source MPP SQL engine, Trino. Complementing this, the Starburst Delta Lake connector provides seamless connectivity to Delta Lake, the innovative data modification and optimization platform introduced by Databricks in 2019. Together, Starburst and Delta Lake empower enterprises with enhanced cost control, flexibility, and accelerated access to data within their data lakes.

Starburst Delta Lake connector: Unlocking Trino's potential

The Starburst Delta Lake connector unleashes the true potential of Trino by enabling customers to leverage Trino's speed, concurrency, and scalability in both Starburst Enterprise and Starburst Galaxy. This connector facilitates seamless querying and write operations on Delta Lake, which is known for its ability to support ACID transactions and performance optimizations on object storage.

The key features include:

- Fast, efficient reads of Delta Lake transaction logs with support for Amazon S3, HDFS, Azure Storage, and Google Cloud Storage
- Support for data skipping to enhance query performance
- Optimization of queries using Delta Lake file statistics
- Parallel processing for improved performance
- Table statistics support for the Trino cost-based optimizer
- Querying special columns of metadata
- Robust fine-grained access control security integrations
- Data Manipulation Language (DML) support, including INSERT/DELETE/UPDATE/MERGE
- Dynamic filtering and table scan redirections for improved performance
- Fault-tolerant execution for memory intensive queries

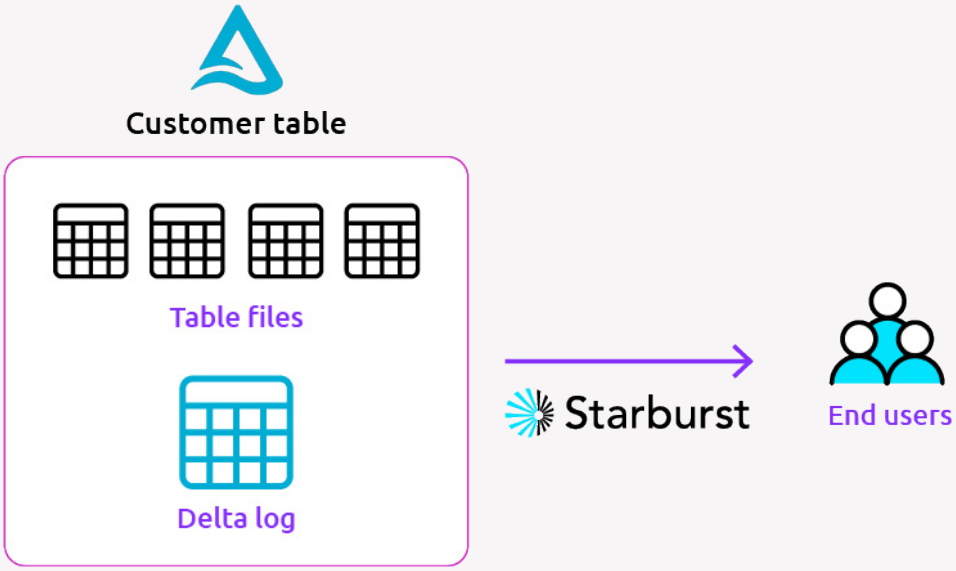


Delta Lake: Revolutionizing data lake efficiency

Delta Lake addresses a significant challenge faced by data engineers and analysts - the difficulty and time-consuming nature of updating customer or product data in object storage. This storage platform allows users to easily modify and update data in a cloud data lake, providing performance and file management optimizations that were previously unavailable.

How Starburst's Delta Lake connectivity works

The Starburst Delta Lake connector works seamlessly with distributed storage systems like S3, ADLS, GCS, and MinIO. It leverages the Parquet format, storing files in Delta Lake, and intelligently reads the Delta Log and files, delivering multiple benefits to end users.



ACID transactions

Users can perform updates and modifications to Delta tables without rewriting the entire table. This ensures efficient and precise data management without the need for manifest files or Hive metastore updates.

Governance

The connector supports GDPR compliance by allowing easy removal of specific customer data through delete or update statements.

Data skipping

Users can quickly narrow down the files needed for querying, leveraging high-level statistics like min, max, nulls, and counts stored in each file. This improves query performance by reducing unnecessary file reads.

VACUUM

Running the VACUUM command clears up the table and removes older files, ensuring optimal storage utilization.

OPTIMIZE

The connector supports Delta Lake's OPTIMIZE command, which combines small files into larger ones, enhancing overall performance.

Write Support

Support for Data Manipulation Language (DML) empowers data teams to perform write operations directly on Delta Lake, eliminating the need for data duplication across various systems.

MERGE

The connector supports efficient data upserts using the MERGE operation, streamlining the process of synchronizing data between different datasets and enhancing data management workflows.

Dynamic filtering

Optimize query performance by pushing predicates down to the data source, reducing data movement and speeding up query execution.

Table scan redirection

Enables users to offload data access to tables accessed in one catalog, to equivalent tables accessed in another catalog, shifting data access to a more performant system.

Operational excellence with shared metastores

Users can either maintain separate metastores or utilize a shared metastore, ensuring a seamless experience for end users.

The Starburst Delta Lake connector, in conjunction with Trino and Delta Lake, opens up new possibilities for enterprises seeking to maximize the potential of their data lakes. By providing a single point of access for high-concurrency SQL queries, seamless write operations, and advanced data management capabilities, Starburst and Delta Lake deliver unparalleled efficiency and analytics to businesses across various industries.

About Starburst

For data-drive companies, Starburst offersa full-featured analytics platform built on open-source Trino.

Our platform includes the capabilities needed to discover, organize, and consume data without the need for time-consuming and costly migrations. We believe the lake should be the center of gravity and be the starting point for querying disparate data.

With Starburst, teams can access more complete data, lower the cost of infrastructure, use the tools best suited to their specific needs, and avoid vendor lock-in.

Trusted by companies like Novant Health, Assurance, Optum, Pfizer, Sophia Genetics, EMIS Health, Gilead and Genius, Starburst helps companies make better decisions faster on all their data.

To learn more, visit **www.starburst.io** and follow Starburst on Twitter and Linkedin.

A single point of access to all your data

Learn more at www.starburst.io