Starburst Academy

Optimizing data lakehouses with Starburst v3.1.0

This 3-day course comprises instructor-led discussions, demonstrations, and hands-on exercises designed to build a working knowledge of the Starburst query engine. Participants will gain a more thorough awareness of Starburst architecture, focusing on best practices for data lake based schemas, including table formats and partitioning, file formats and sizes, and other optimization techniques.

Objectives

Upon completion of this course, you will be able to:

- Use Starburst as a single point of access for multiple data sources and federating queries across them
- Evaluate and describe how queries are executed within a Starburst cluster
- Use Hive and Iceberg table formats; construct, populate, query, and modify partitioned tables
- Employ file size/format/hierarchy strategies to improve query performance
- Understand the role of the Cost-based optimizer and read query plans to ensure optimizations are occurring as expected and to identify possible issues
- Create role-based access control policies for table operations
- Build a data engineering pipeline with Starburst Galaxy

Audience & Prerequisites

This course is designed for data engineers, data architects, and experienced data analysts and data scientists. Intermediate experience with SQL is assumed. techniques.

Agenda

1. Starburst features
- Overview & architecture
- Web UI
- Connectors & catalogs
- Client tools integrations

2. Data lake tables
- Separation of storage & compute
- Schema on read

3. Data lake performance
- Limit Data Exchanges
- File format options
- Small files problem
- Partitioning & bucketing

4. Table formats
- Moving beyond Hive
- Compare/contrast alternatives
- Explore Delta Lake

5. Apache Iceberg
- Table format architecture
- Creating tables
- Insert, update & delete

6. Advanced Iceberg
- CDC with merge
- Schema & partition evolution
- Snapshots & compaction

7. Parallel processing
- Divide & conquer
- Beyond single-stage queries

8. Cost-based optimizer
- Benefits of statistics
- Query plan analysis

9. Access control
- Configuration options
- Role-based access control

10. Data pipelines
- Definition & differentiation
- Reference architecture

Optimizing data lakehouses for Starburst - Labs

Section 1: Starburst features

- Set up your student account in Starburst Galaxy
- Execute queries in Starburst Galaxy
- Exploring federated queries

Section 2: Data lake tables

- Create a schema and tables
- Investigate Hive's special columns

Section 3: Data lake performance

- Create tables with multiple file formats
- Using columnar file formats and eliminating small files
- Exploring table partitioning and bucketing

Section 4: Table formats

- Exploring the Delta Lake table format

Section 5. Apache Iceberg

- Create and populate Iceberg tables
- Explore partitions with Iceberg
- Data modifications and snapshots with Iceberg

Section 6. Advanced Iceberg

- Utilizing Iceberg's MERGE statement
- Exercise advanced features of Iceberg

Section 8: Cost-based optimizer

- The EXPLAIN command
- The EXPLAIN ANALYZE command
- Explore the impact of statistics on query plans

Section 9: Access control

- Creating and validating RBAC policies with Starburst Galaxy
- Creating and validating ABAC policies with Starburst Galaxy

Section 10: Data pipelines

- Construct a pipeline with insert only transactions
- Construct a pipeline with the MERGE statement