Optimizing Apache Spark™ on Databricks

Course ID #: 7000-807-ZZ-Z
Hours: 14

Course Content

Course Description:

In this course, you will cover five key problems that represent the vast majority of performance problems in an Apache Spark application: skew, spill, shuffle, storage, and serialization. With each of these topics, we explore coding examples based on 100 GB to 1+ TB datasets that demonstrate how these problems are introduced, how to diagnose these problems with tools like the Spark UI, and conclude by discussing mitigation strategies for each of these problems.

We continue the conversation by looking at a series of key ingestion concepts that promote strategies for processing terabytes of data including managing Spark partition sizes, disk-partitioning, bucketing, z-ordering, and more. With each of these topics, we explore when and how each of these techniques should be implemented, new challenges that productionalizing these solutions might provide along with corresponding mitigation strategies.

Finally, we introduce a couple of other key topics such as issues with data locality, IO caching and spark caching, pitfalls of broadcast Joins, and new features like Spark 3's Adaptive Query Execution and Dynamic Partition Pruning. We then conclude the course with discussions and exercises on designing and configuring clusters for optimal performance given specific use cases, personas, the divergent needs of various teams, and cross-team security concerns.

Course Objectives:

Prerequisites:

- Intermediate to advanced programming experience in Python or Scala
- Hands-on experience developing Apache Spark applications

Target Audience:

- Data engineers
- Data architects

Optimizing Apache Spark™ on Databricks

Course ID #: 7000-807-ZZ-Z
Hours: 14

Topics:

Lesson 1: Introduction and the 5 Most Common Performance Problems

- Setup, Introductions
- Spark Architecture Review
- Spark UI Review, Lab
- Skew, Labs

Lesson 2: The 5 Most Common Performance Problems Continued

- Spill, Labs
- Shuffle
- Storage, Labs
- Serialization, Labs

Lesson 3: Key Ingestion Concepts

- Ingestion Basics, Labs
- Predicate Push Downs, Labs
- Disk Partitioning, Lab
- Z-Ordering, Labs
- Bucketing, Labs

Lesson 4: Optimizing with AQE and High Performance; Designing Clusters for High Performance

- Tuning Shuffle Partitions, Lab
- Join Optimizations, Lab
- Skew Join Optimizations, Lab
- Dynamic Partition Pruning, Lab
- Designing Clusters for High Performance
- Cluster Configurations Scenarios
- Designing Clusters Breakout