Cloudera - Advanced Spark Application Performance Tuning

Code:	SPARK-PT
Length:	3 days
URL:	View Online

This three-day hands-on training course delivers the key concepts and expertise developers need to improve the performance of their Apache Spark applications. During the course, participants will learn how to identify common sources of poor performance in Spark applications, techniques for avoiding or solving them, and best practices for Spark application monitoring.

Apache Spark Application Performance Tuning presents the architecture and concepts behind Apache Spark and underlying data platform, then builds on this foundational understanding by teaching students how to tune Spark application code. The course format emphasizes instructor-led demonstrations illustrate both performance issues and the techniques that address them, followed by hands-on exercises that give students an opportunity to practice what they've learned through an interactive notebook environment. The course applies to Spark 2.4, but also introduces the Spark 3.0 Adaptive Query Execution framework.

Skills Gained

- Understand Apache Spark's architecture, job execution, and how techniques such as lazy execution and pipelining can improve runtime performance
- Evaluate the performance characteristics of core data structures such as RDD and DataFrames
- Select the file formats that will provide the best performance for your application
- Identify and resolve performance problems caused by data skew
- Use partitioning, bucketing, and join optimizations to improve SparkSQL performance
- Understand the performance overhead of Python-based RDDs, DataFrames, and user-defined functions
- Take advantage of caching for better application performance
- Understand how the Catalyst and Tungsten optimizers work
- Understand how Workload XM can help troubleshoot and proactively monitor Spark applications performance
- Learn about the new features in Spark 3.0 and specifically how the Adaptive Query Execution engine improves performance

Who Can Benefit

This course is designed for software developers, engineers, and data scientists who have experience developing Spark applications and want to learn how to improve the performance of their code. This is not an introduction to Spark.

Prerequisites

Spark examples and hands-on exercises are presented in Python and the ability to program in this language is required. Basic familiarity with the Linux command line is assumed. Basic knowledge of SQL is helpful.

Course Details

Course Outline

Spark Architecture Data Sources and Formats Inferring Schemas Dealing With Skewed Data Catalyst and Tungsten Overview Mitigating Spark Shuffles Partitioned and Bucketed Tables Improving Join Performance Pyspark Overhead and UDFs Caching Data for Reuse Workload XM (WXM) Introduction What's New in Spark 3.0? Appendix A: Partition Processing Appendix B: Broadcasting Appendix C: Scheduling

---

Schedule (as of 3)

Date	Location