€TRAINING

Apache Spark Application Performance Tuning

18 - 22 November 2024
Madrid (Spain)

Apache Spark Application Performance Tuning

REF: B1739 DATE: 18 - 22 November 2024 Venue: Madrid (Spain) - Fee: 5850 Euro

Introduction

The training equips developers with the knowledge and skills necessary to enhance the performance of their Apache Spark applications. Participants will get knowledge of the best practices for monitoring Spark applications as well as how to recognize typical causes of poor performance in Spark applications.

Course Objectives

At the end of the course, participants will be able to:

- Learn Apache Spark's architecture, job execution, and performance-enhancing methods like lazy execution and pipelining work.
- Analyze the performance traits of fundamental data structures like RDDs and DataFrames.
- Choose the file types that will run your applications most effectively.
- Determine and fix performance issues brought on by data skew.
- Use join improvements, bucketing, and partitioning to boost SparkSQL's speed.
- Recognize the performance overhead of RDDs, DataFrames, and user-defined functions based on Python.
- Utilize caching for improved application performance.
- Recognize the operation of the Tungsten and Catalyst optimizers.
- Learn how Workload XM may be used to proactively monitor and troubleshoot Spark application performance.
- Discover the improvements in performance brought by the Adaptive Query Execution engine as well as other new features in Spark 3.0.

Targeted Audience

- Software developers
- Engineers
- Data scientists who have experience developing Spark applications and want to learn how to improve the performance of their code.

Course Outline

Unit 1: Spark Architecture & Data Sources and Formats

- RDDs
- DataFrames and Datasets
- Lazy Evaluation
- Pipelining
- Available Formats Overview
- Impact on Performance
- The Small Files Problem

Unit 2: Inferring Schemas & Dealing With Skewed Data

- The Cost of Inference

- Mitigating Tactics
- Recognizing Skew
- Mitigating Tactics

Unit 3: Catalyst and Tungsten Overview & Mitigating Spark Shuffles

- Catalyst Overview
- Tungsten Overview
- Denormalization
- Broadcast Joins
- Map-Side Operations
- Sort Merge Joins

Unit 4: Partitioned and Bucketed Tables & Improving Join Performance

- Partitioned Tables
- Bucketed Tables
- Impact on Performance
- Skewed Joins
- Bucketed Joins
- Incremental Joins

Unit 5: Pyspark Overhead and UDFs & Caching Data for Reuse

- Pyspark Overhead
- Scalar UDFs
- Vector UDFs using Apache Arrow
- Scala UDFs
- Caching Options
- Impact on Performance
- Caching Pitfalls

Unit 6: Workload XM WXM Introduction & What's New in Spark 3.0

- WXM Overview
- WXM for Spark Developers
- Adaptive Number of Shuffle Partitions
- Skew Joins
- Convert Sort Merge Joins to Broadcast Joins
- Dynamic Partition Pruning
- Dynamic Coalesce Shuffle Partitions