CLOUDERA DEVELOPER TRAINING FOR SPARK & HADOOP

Course Outline

1. Introduction

2. Introduction to Apache Hadoop and the Hadoop Ecosystem

- Apache Hadoop Overview
- Data Processing
- Introduction to the Hands-On Exercises

3. Apache Hadoop File Storage

- Apache Hadoop Cluster Components
- HDFS Architecture
- Using HDFS

4. Distributed Processing on an Apache Hadoop Cluster

- YARN Architecture
- Working With YARN

5. Apache Spark Basics

- What is Apache Spark?
- Starting the Spark Shell
- Using the Spark Shell
- Getting Started with Datasets and DataFrames
- DataFrame Operations

6. Working with DataFrames and Schemas

- Creating DataFrames from Data Sources
- Saving DataFrames to Data Sources
- DataFrame Schemas
- Eager and Lazy Execution

7. Analyzing Data with DataFrame Queries

- Querying DataFrames Using Column Expressions
- Grouping and Aggregation Queries
- Joining DataFrames

8. RDD Overview

- RDD Overview
- RDD Data Sources
- Creating and Saving RDDs
- RDD Operations

9. Transforming Data with RDDs

- Writing and Passing Transformation Functions
- Transformation Execution
- Converting Between RDDs and DataFrames

10. Aggregating Data with Pair RDDs

- Querying Tables in Spark Using SQL
- Querying Files and Views
- The Catalog API
- Comparing Spark SQL, Apache Impala, and Apache Hive-on-Spark

11. Querying Tables and Views with SQL

- Querying Tables in Spark Using SQL
- Querying Files and Views
- The Catalog API

12. Working with Datasets in Scala

- Datasets and DataFrames
- Creating Datasets
- Loading and Saving Datasets
- Dataset Operations

13. Writing, Configuring, and Running Spark Applications

- Writing a Spark Application
- Building and Running an Application
- Application Deployment Mode
- The Spark Application Web UI
- Configuring Application Properties

14. Spark Distributed Processing

- Review: Apache Spark on a Cluster
- RDD Partitions
- Example: Partitioning in Queries
- Stages and Tasks
- Job Execution Planning
- Example: Catalyst Execution Plan
- Example: RDD Execution Plan

15. Distributed Data Persistence

- DataFrame and Dataset Persistence
- Persistence Storage Levels
- Viewing Persisted RDDs

16. Common Patterns in Spark Data Processing

- Common Apache Spark Use Cases
- Iterative Algorithms in Apache Spark
- Machine Learning
- Example: k-means

17. Introduction to Structured Streaming

- Apache Spark Streaming Overview
- Creating Streaming DataFrames
- Transforming DataFrames
- Executing Streaming Queries

18. Structured Streaming with Apache Kafka

- Overview
- Receiving Kafka Messages
- Sending Kafka Messages

19. Aggregating and Joining Streaming DataFrames

- Streaming Aggregation
- Joining Streaming DataFrames

20. Conclusion

A. Message Processing with Apache Kafka

- What Is Apache Kafka?
- Apache Kafka Overview
- Scaling Apache Kafka
- Apache Kafka Cluster Architecture
- Apache Kafka Command Line Tools