Course Outline

Cloudera Developer Training for Spark & Hadoop

Course Summary

Description

This four-day hands-on training course delivers the key concepts and expertise developers need to use Apache Spark to develop high-performance parallel applications. Participants will learn how to use Spark SQL to query structured data and Spark Streaming to perform real-time processing on streaming data from a variety of sources. Developers will also practice writing applications that use core Spark to perform ETL processing and iterative algorithms. The course covers how to work with “big data” stored in a distributed file system, and execute Spark applications on a Hadoop cluster. After taking this course, participants will be prepared to face real-world challenges and build applications to execute faster decisions, better decisions, and interactive analysis, applied to a wide variety of use cases, architectures, and industries.

Objectives

At the end of this course, students will be able to understand:

- How the Apache Hadoop ecosystem fits in with the data processing lifecycle
- How data is distributed, stored, and processed in a Hadoop cluster
- How to write, configure, and deploy Apache Spark applications on a Hadoop cluster
- How to use the Spark shell and Spark applications to explore, process, and analyze distributed data
- How to query data using Spark SQL, DataFrames, and Datasets
- How to use Spark Streaming to process a live data stream

Topics

- Introduction
- Introduction to Apache Hadoop and the Hadoop Ecosystem
- Apache Hadoop Overview
- Apache Hadoop File Storage
- Distributed Processing on an Apache Hadoop Cluster
- Apache Spark Basics
- Working with DataFrames and Schemas
- Analyzing Data with DataFrame Queries
- RDD Overview
- Transforming Data with RDDs
- Aggregating Data with Pair RDDs
- Querying Tables and Views with SQL
- Working with Datasets in Scala
- Writing, Configuring, and Running Spark Applications
- Spark Distributed Processing
- Distributed Data Persistence
- Introduction to Structured Streaming
- Structured Streaming with Apache Kafka
- Aggregating and Joining Streaming DataFrames
- Conclusion

Audience

This course is designed for developers and engineers who have programming experience, but prior knowledge of Hadoop and/or Spark is not required.

Prerequisites

This course is designed for developers and engineers who have programming experience, but prior knowledge of Spark and Hadoop is not required. Apache Spark examples and hands-on exercises are presented in Scala and Python. The ability to program in one of those languages is required. Basic familiarity with the Linux command line is assumed. Basic knowledge of SQL is helpful.

Duration

Four days

Course Outline

Cloudera Developer Training for Spark & Hadoop

Course Outline

I. Introduction

II. Introduction to Apache Hadoop and the Hadoop Ecosystem

III. Apache Hadoop Overview
- A. Data Processing
- B. Introduction to the Hands-On Exercises

IV. Apache Hadoop File Storage
- A. Apache Hadoop Cluster Components
- B. HDFS Architecture
- C. Using HDFS

V. Distributed Processing on an Apache Hadoop Cluster
- A. YARN Architecture
- B. Working With YARN

VI. Apache Spark Basics
- A. What is Apache Spark?
- B. Starting the Spark Shell
- C. Using the Spark Shell
- D. Getting Started with Datasets and DataFrames
- E. DataFrame Operations

VII. Working with DataFrames and Schemas
- A. Creating DataFrames from Data Sources
- B. Saving DataFrames to Data Sources
- C. DataFrame Schemas
- D. Eager and Lazy Execution

VIII. Analyzing Data with DataFrame Queries
- A. Querying DataFrames Using Column Expressions
- B. Grouping and Aggregation Queries
- C. Joining DataFrames

IX. RDD Overview
- A. RDD Overview
- B. RDD Data Sources
- C. Creating and Saving RDDs
- D. RDD Operations

X. Transforming Data with RDDs
- A. Writing and Passing Transformation Functions
- B. Transformation Execution
- C. Converting Between RDDs and DataFrames

XI. Aggregating Data with Pair RDDs
- A. Querying Tables in Spark Using SQL
- B. Querying Files and Views
- C. The Catalog API
- D. Comparing Spark SQL, Apache Impala, and Apache Hive-on-Spark

XII. Querying Tables and Views with SQL
- A. Querying Tables in Spark Using SQL
- B. Querying Files and Views
- C. The Catalog API

XIII. Working with Datasets in Scala
- A. Datasets and DataFrames
- B. Creating Datasets
- C. Loading and Saving Datasets
- D. Dataset Operations

XIV. Writing, Configuring, and Running Spark Applications
- A. Writing a Spark Application
- B. Building and Running an Application
- C. Application Deployment Mode
- D. The Spark Application Web UI
- E. Configuring Application Properties

XV. Spark Distributed Processing
- A. Review: Apache Spark on a Cluster
- B. RDD Partitions
- C. Example: Partitioning in Queries
- D. Stages and Tasks
- E. Job Execution Planning
- F. Example: Catalyst Execution Plan
- G. Example: RDD Execution Plan

Cloudera Developer Training for Spark & Hadoop

Course Outline (cont'd)

XVI. Distributed Data Persistence

A. DataFrame and Dataset Persistence
B. Persistence Storage Levels
C. Viewing Persisted RDDs
D. Common Patterns in Spark Data Processing
E. Common Apache Spark Use Cases
F. Iterative Algorithms in Apache Spark
G. Machine Learning
H. Example: k-means

XVII. Introduction to Structured Streaming

A. Apache Spark Streaming Overview
B. Creating Streaming DataFrames
C. Transforming DataFrames
D. Executing Streaming Queries

XVIII. Structured Streaming with Apache Kafka

A. Overview
B. Receiving Kafka Messages
C. Sending Kafka Messages

XIX. Aggregating and Joining Streaming DataFrames

A. Streaming Aggregation
B. Joining Streaming DataFrames

XX. Conclusion

A. Message Processing with Apache Kafka
B. What Is Apache Kafka?
C. Apache Kafka Overview
D. Scaling Apache Kafka
E. Apache Kafka Cluster Architecture
F. Apache Kafka Command Line Tools