Developer Training for Spark & Hadoop

Course Summary

Description

This course delivers the key concepts and expertise developers need to develop high-performance parallel applications with Apache Spark 2. 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 large datasets 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

After taking this course, students will be able to:

- Distribute, store, and process data in a Hadoop cluster
- Write, configure, and deploy Spark applications on a cluster
- Use the Spark shell for interactive data analysis
- Process and query structured data using Spark SQL
- Use Spark Streaming to process a live data stream

Topics

- Introduction to Apache Hadoop and the Hadoop Ecosystem
- 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 Apache Spark SQL
- Working with Datasets in Scala
- Writing, Configuring, and Running Apache Spark Applications
- Distributed Processing
- Distributed Data Persistence
- Common Patterns in Apache Spark Data Processing
- Apache Spark Streaming: Introduction to DStreams
- Apache Spark Streaming: Processing Multiple Batches
- Apache Spark Streaming: Data Sources

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

The prerequisites for this course are:

- 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

Developer Training for Spark & Hadoop

Course Outline

I. Introduction to Apache Hadoop and the Hadoop Ecosystem

A. Apache Hadoop Overview
B. Data Ingestion and Storage
C. Data Processing
D. Data Analysis and Exploration
E. Other Ecosystem Tools
F. Introduction to the Hands-On Exercises

II. Apache Hadoop File Storage

A. Apache Hadoop Cluster Components
B. HDFS Architecture
C. Using HDFS

III. Distributed Processing on an Apache Hadoop Cluster

A. YARN Architecture
B. Working With YARN

IV. 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

V. 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

VI. Analyzing Data with DataFrame Queries

A. Querying DataFrames Using Column Expressions
B. Grouping and Aggregation Queries
C. Joining DataFrames

VII. RDD Overview

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

VIII. Transforming Data with RDDs

A. Writing and Passing Transformation Functions
B. Transformation Execution
C. Converting Between RDDs and DataFrames

IX. Aggregating Data with Pair RDDs

A. Key-Value Pair RDDs
B. Map-Reduce
C. Other Pair RDD Operations

X. Querying Tables and Views with Apache Spark SQL

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

XI. Working with Datasets in Scala

A. Datasets and DataFrames
B. Creating Datasets
C. Loading and Saving Datasets
D. Dataset Operations

XII. Writing, Configuring, and Running Apache 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

XIII. 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

Developer Training for Spark & Hadoop

Course Outline (cont.)

XIV. Distributed Data Persistence

A. DataFrame and Dataset Persistence
B. Persistence Storage Levels
C. Viewing Persisted RDDs

XV. Common Patterns in Apache Spark Data Processing

A. Common Apache Spark Use Cases
B. Iterative Algorithms in Apache Spark
C. Machine Learning
D. Example: k-means

XVI. Apache Spark Streaming: Introduction to DStreams

A. Apache Spark Streaming Overview
B. Example: Streaming Request Count
C. DStreams
D. Developing Streaming Applications

XVII. Apache Spark Streaming: Processing Multiple Batches

A. Multi-Batch Operations
B. Time Slicing
C. State Operations
D. Sliding Window Operations
E. Preview: Structured Streaming

XVIII. Apache Spark Streaming: Data Sources

A. Streaming Data Source Overview
B. Apache Flume and Apache Kafka Data Sources
C. Example: Using a Kafka Direct Data Source