Blended Learning Outline: Developer Training for Apache Spark and Hadoop (180404a)

Cloudera's Developer Training for Apache Spark and Hadoop delivers the key concepts and expertise need to develop high-performance parallel applications with Apache Spark. 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.

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.

Format:

Participants enrolling in blended learning will be provided with:

- OnDemand access of Developer Training for Spark and Hadoop.
    - Including 20-hours of cloud-based lab access
    - Access will be available one week prior to the first live session through the Friday following the final live session
- Five three hour live-virtual sessions with a senior Cloudera Instructor.
    - Live-virtual sessions will be focused on demonstrating labs and covering select topics from the weekly lessons. The live sessions allow time for students to ask questions, but assume participants have already completed the Ondemand lessons for that particular week.

Week 1:

Introduction

In this session students will ensure they have access to the courseware materials, can connect to the lab environment and be given details about the structure of the upcoming sessions.

Week 2:

Introduction to Apache Hadoop and the Hadoop Ecosystem

- Apache Hadoop Overview
- Data Ingestion and Storage
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to the Hands-On Exercises
- Essential Points

Apache Hadoop File Storage

- Apache Hadoop Cluster Components
- HDFS Architecture
- Using HDFS
- Essential Points

Distributed Processing on an Apache Hadoop Cluster

- YARN Architecture
- Working With YARN
- Essential Points

Apache Spark Basics

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

Working with DataFrames and Schemas

- Creating DataFrames from Data Sources
- Saving DataFrames to Data Sources
- DataFrame Schemas

- Eager and Lazy Execution
- Essential Points

Analyzing Data with DataFrame Queries
- Querying DataFrames Using Column Expressions
- Grouping and Aggregation Queries
- Joining DataFrames
- Essential Points

Week 3:

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

Transforming Data with RDDs
- Writing and Passing Transformation Functions
- Transformation Execution
- Converting Between RDDs and DataFrames
- Essential Points

Aggregating Data with Pair RDDs
- Key-Value Pair RDDs
- Map-Reduce
- Other Pair RDD Operations
- Essential Points

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

Working with Datasets in Scala
- Datasets and DataFrames
- Creating Datasets
- Loading and Saving Datasets
- Dataset Operations
- Essential Points

Week 4:

Writing, Configuring, and Running Apache Spark Applications

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

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
- Essential Points

Distributed Data Persistence

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

Common Patterns in Apache Spark Data Processing

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

Week 5:

Apache Spark Streaming: Introduction to DStreams

- Apache Spark Streaming Overview
- Example: Streaming Request Count
- DStreams
- Developing Streaming Applications
- Essential Points

Apache Spark Streaming: Processing Multiple Batches

- Multi-Batch Operations

- Time Slicing
- State Operations
- Sliding Window Operations
- Preview: Structured Streaming
- Essential Points

Apache Spark Streaming: Data Sources

- Streaming Data Source Overview
- Apache Flume and Apache Kafka Data Sources
- Example: Using a Kafka Direct Data Source
- Essential Points

Conclusion