CLOUDERA

Educational Services

Outline: Custom Svenska Krafnat Developer Training for Apache Spark and Hadoop with Airflow and Spark Perf (based on devsh 200417)

Course 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.

Introduction and Scala specific chapters removed (red) and replaced (green) with Custom Airflow hands on demo/exercises and switch to the Spark performance course and VM for 4 chapters.

Prerequisites

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

Course Contents

1. **Introduction**

2. **Introduction to Apache Hadoop and the Hadoop Ecosystem**
   - Apache Hadoop Overview
   - Data Processing
   - Introduction to the Hands-On Exercises
   - Hands-On Exercise: Starting the Exercise Environment

3. **Apache Hadoop File Storage**
    - Apache Hadoop Cluster Components
    - HDFS Architecture
    - Using HDFS
    - Hands-On Exercise: Working with HDFS

4. **Distributed Processing on an Apache Hadoop Cluster**
    - YARN Architecture
    - Working With YARN
    - Hands-On Exercise: Running and Monitoring a YARN Job

5. **Apache Spark Basics**
    - What is Apache Spark?
    - Starting the Spark Shell
    - Using the Spark Shell
    - Getting Started with Datasets and DataFrames
    - DataFrame Operations
    - Hands-On Exercise: Exploring DataFrames Using the Apache Spark Shell

6. **Working with DataFrames and Schemas**
    - Creating DataFrames from Data Sources
    - Saving DataFrames to Data Sources
    - DataFrame Schemas
    - Eager and Lazy Execution
    - Hands-On Exercise: Working with DataFrames and Schemas

7. **Analyzing Data with DataFrame Queries**
    - Querying DataFrames Using Column Expressions
    - Grouping and Aggregation Queries
    - Joining DataFrames
    - Hands-On Exercise: Analyzing Data with DataFrame Queries

8. **RDD Overview**
    - RDD Overview
    - RDD Data Sources
    - Creating and Saving RDDs
    - RDD Operations
    - Hands-On Exercise: Working With RDDs

9. **Transforming Data with RDDs**
    - Writing and Passing Transformation Functions
    - Transformation Execution
    - Converting Between RDDs and DataFrames
    - Hands-On Exercise: Transforming Data Using RDDs

10. **Aggregating Data with Pair RDDs**
    - Key-Value Pair RDDs
    - Map-Reduce
    - Other Pair RDD Operations
    - Hands-On Exercise: Joining Data Using Pair RDDs

11. **Querying Tables and Views with SQL**
    - Querying Tables in Spark Using SQL
    - Querying Files and Views
    - The Catalog API

- Hands-On Exercise: Querying Tables and Views with SQL

12. **Working with Datasets in Scala**
    - Datasets and DataFrames
    - Creating Datasets
    - Loading and Saving Datasets
    - Dataset Operations
    - Hands-On Exercise: Using Datasets in Scala

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
    - Hands-On Exercise: Writing, Configuring, and Running a Spark Application

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
    - Hands-On Exercise: Exploring Query Execution

15. **Distributed Data Persistence**
    - DataFrame and Dataset Persistence
    - Persistence Storage Levels
    - Viewing Persisted RDDs
    - Hands-On Exercise: Persisting DataFrames

16. **Common Patterns in Spark Data Processing**
    - Common Apache Spark Use Cases
    - Iterative Algorithms in Apache Spark
    - Machine Learning
    - Example: k-means
    - Hands-On Exercise: Implementing an Iterative Algorithm

17. **Introduction to Structured Streaming**
    - Apache Spark Streaming Overview
    - Creating Streaming DataFrames
    - Transforming DataFrames
    - Executing Streaming Queries
    - Hands-On Exercise: Processing Streaming Data

18. **Structured Streaming with Apache Kafka**
    - Overview
    - Receiving Kafka Messages
    - Sending Kafka Messages
    - Hands-On Exercise: Working with Apache Kafka Streaming Messages

19. **Aggregating and Joining Streaming DataFrames**
    - Streaming Aggregation
    - Joining Streaming DataFrames

- Hands-On Exercise: Aggregating and Joining Streaming DataFrames

20: Message Processing with Apache Kafka

- What Is Apache Kafka?
- Apache Kafka Overview
- Scaling Apache Kafka
- Apache Kafka Cluster Architecture
- Apache Kafka Command Line Tools
- Hands-On Exercise: Producing and Consuming Apache Kafka Messages

21: Using Apache Airflow

CDE Airflow

- **CDE Airflow UI tour**
- **CDE Airflow deploy exercises**

Apache Airflow

- **Airflow UI and deploying applications**
- **Airflow Operators Bash and Kafka**
- **Airflow with Hive SQL**
- **Airflow with Spark**

22: Spark Performance - Change VM environment.

First chapters from Spark performance course and skipping past the intro and spark refresher chapters.

23: Data Sources and Formats

- Available formats overview
- Impact on performance
- The small files problem

24: Inferring Schemas

- The cost of inference
- Mitigating tactics

25: Dealing With Skewed Data

- Recognizing skew
- Mitigating tactics

26: Catalyst and Tungsten Overview

- Catalyst overview
- Tungsten overview

27: Mitigating Spark Shuffles

- Denormalization
- Broadcast joins
- Map side operations
- Sort Merge joins