CLOUDERA DEVELOPER TRAINING(SPARK-HADOOP-OD)

Introduction to Hadoop and the Hadoop Ecosystem

- Problems with Traditional Large-Scale Systems
- Hadoop!
- Data Storage and Ingest
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to the Hands-On Exercises

Hadoop Architecture and HDFS

- Distributed Processing on a Cluster
- Storage: HDFS Architecture
- Storage: Using HDFS
- Resource Management: YARN Architecture
- Resource Management: Working with YARN

Importing Relational Data with Apache Sqoop

- Sqoop Overview
- Basic Imports and Exports
- Limiting Results
- Improving Sqoop's Performance
- Sqoop 2

Introduction to Impala and Hive

- Introduction to Impala and Hive
- Why Use Impala and Hive?
- Querying Data With Impala and Hive
- Comparing Hive and Impala to Traditional Databases

Modeling and Managing Data with Impala and Hive

- Data Storage Overview
- Creating Databases and Tables
- Loading Data into Tables
- HCatalog
- Impala Metadata Caching

Data Formats

- Selecting a File Format
- Hadoop Tool Support for File Formats
- Avro Schemas
- Using Avro with Impala, Hive, and Sqoop
- Avro Schema Evolution
- Compression

Data File Partitioning

- Partitioning Overview
- Partitioning in Impala and Hive

Capturing Data with Apache Flume

- What is Apache Flume?
- Basic Flume Architecture
- Flume Sources
- Flume Sinks
- Flume Channels
- Flume Configuration

Spark Basics

- What is Apache Spark?
- Using the Spark Shell
- RDDs (Resilient Distributed Datasets)
- Functional Programming in Spark

Working with RDDs in Spark

- Creating RDDs
- Other General RDD Operations

Writing and Deploying Spark Applications

- Spark Applications vs. Spark Shell
- Creating the SparkContext
- Building a Spark Application (Scala and Java)
- Running a Spark Application
- The Spark Application Web UI
- Configuring Spark Properties
- Logging

Parallel Processing in Spark

- Review: Spark on a Cluster
- RDD Partitions
- Partitioning of File-Based RDDs

- HDFS and Data Locality
- Executing Parallel Operations
- Stages and Tasks

Spark RDD Persistence

- RDD Lineage
- RDD Persistence Overview
- Distributed Persistence

Common Patterns in Spark Data Processing

- Common Spark Use Cases
- Iterative Algorithms in Spark
- Graph Processing and Analysis
- Machine Learning
- Example: k-means

DataFrames and Spark SQL

- Spark SQL and the SQL Context
- Creating DataFrames
- Transforming and Querying DataFrames
- Saving DataFrames
- DataFrames and RDDs
- Comparing Spark SQL, Impala, and Hive-on-Spark