[DEV-203]: HDP Operations: Apache Pig and Hive

Length : 4 Days
Delivery Method : Instructor-led (Classroom)

Course Overview

This 4 days training course is designed for developers who need to create applications to analyze Big Data stored in Apache Hadoop using Pig and Hive. Topics include: Hadoop, YARN, HDFS, MapReduce, data ingestion, workflow definition, using Pig and Hive to perform data analytics on Big Data and an introduction to Spark Core and Spark SQL.

Audience Profile

Software developers who need to understand and develop applications for Hadoop.

Pre-Requisites

Participants should be familiar with programming principles and have experience in software development. SQL knowledge is also helpful. No prior Hadoop knowledge is required.

Formats

Lecture/Discussion	50%
Hands-on Labs	50%

Course Outline

Module 1: Understanding Hadoop and the Hadoop Distributed File System (HDFS)
Lessons

- List the Three “V”s of Big Data
- List the Six Key Hadoop Data Types
- Describe Hadoop, YARN and Use Cases for Hadoop
- Describe Hadoop Ecosystem Tools and Frameworks
- Describe the Differences Between Relational Databases and Hadoop
- Describe What is New in Hadoop 2.x
- Describe the Hadoop Distributed File System (HDFS)
- Describe the Differences Between HDFS and an RDBMS
- Describe the Purpose of NameNodes and DataNodes
- List Common HDFS Commands
- Describe HDFS File Permissions
- List Options for Data Input
- Describe WebHDFS
- Describe the Purpose of Sqoop and Flume
- Describe How to Export to a Table

- Describe the Purpose of MapReduce
- Define Key/Value Pairs in MapReduce
- Describe the Map and Reduce Phases
- Describe Hadoop Streaming

Labs

- Starting an HDP Cluster
- Demonstration: Understanding Block Storage
- Using HDFS Commands
- Importing RDBMS Data into HDFS
- Exporting HDFS Data to an RDBMS
- Importing Log Data into HDFS Using Flume
- Demonstration: Understanding MapReduce
- Running a MapReduce Job

Module 2: Pig Programming

Lessons

- Describe the Purpose of Apache Pig
- Describe the Purpose of Pig Latin
- Demonstrate the Use of the Grunt Shell
- List Pig Latin Relation Names and Field Names
- List Pig Data Types
- Define a Schema
- Describe the Purpose of the GROUP Operator
- Describe Common Pig Operators, Including
  - ORDER BY
  - CASE
  - DISTINCT
  - PARALLEL
  - FLATTEN
  - FOREACH
- Perform an Inner, Outer and Replicated Join
- Describe the Purpose of the DataFu Library

Labs

- Demonstration: Understanding Apache Pig
- Getting Starting with Apache Pig
- Exploring Data with Apache Pig
- Splitting a Dataset
- Joining Datasets with Apache Pig
- Preparing Data for Apache Hive
- Demonstration: Computing Page Rank
- Analyzing Clickstream Data
- Analyzing Stock Market Data Using Quantiles

Module 3: DHive Programming

Lessons

- Describe the Purpose of Apache Hive
- Describe the Differences Between Apache Hive and SQL
- Describe the Apache Hive Architecture
- Demonstrate How to Submit Hive Queries
- Describe How to Define Tables
- Describe How to Load Date Into Hive
- Define Hive Partitions, Buckets and Skew
- Describe How to Sort Data
- List Hive Join Strategies
- Describe the Purpose of HCatalog
- Describe the HCatalog Ecosystem
- Define a New Schema
- Demonstrate the Use of HCatLoader and HCatStorer with Apache Pig
- Perform a Multi-table/File Insert
- Describe the Purpose of Views
- Describe the Purpose of the OVER Clause
- Describe the Purpose of Windows
- List Hive Analytics Functions
- List Hive File Formats
- Describe the Purpose of Hive SerDe

Labs

- Understanding Hive Tables
- Understanding Partition and Skew
- Analyzing Big Data with Apache Hive
- Demonstration: Computing NGrams
- Joining Datasets in Apache Hive
- Computing NGrams of Emails in Avro Format
- Using HCatalog with Apache Pig

Module 4: Advanced Hive Programming, Hadoop 2 and YARN, Introduction to Spark Core

Lessons

- Describe the Purpose HDFS Federation
- Describe the Purpose of HDFS High Availability (HA)
- Describe the Purpose of the Quorum Journal Manager
- Demonstrate How to Configure Automatic Failover
- Describe the Purpose of YARN
- List the Components of YARN
- Describe the Lifecycle of a YARN Application
- Describe the Purpose of a Cluster View
- Describe the Purpose of Apache Slider
- Describe the Origin and Purpose of Apache Spark

- List Common Spark Use Cases
- Describe the Differences Between Apache Spark and MapReduce
- Demonstrate the Use of the Spark Shell
- Describe the Purpose of an Resilient Distributed Dateset (RDD)
- Demonstrate How to Load Data and Perform a Word Count
- Define Lazy Evaluation
- Describe How to Load Multiple Types of Data
- Demonstrate How to Perform SQL Queries
- Demonstrate How to Perform DataFrame Operations
- Describe the Purpose of the Optimization Engine
- Describe the Purpose of Apache Oozie
- Describe Apache Pig Actions
- Describe Apache Hive Actions
- Describe MapReduce Actions
- Describe How to Submit an Apache Oozie Workflow
- Define an Oozie Coordinator Job

Labs

- Advanced Apache Hive Programming
- Running a YARN Application
- Getting Started with Apache Spark
- Exploring Apache Spark SQL
- Defining an Apache Oozie Workflow