Course Outline

Hortonworks HDP Developer Quick Start

Course Summary

Description

This four day training course is designed for developers who need to create applications to analyze Big Data stored in Apache Hadoop using Apache Pig and Apache Hive, and developing applications on Apache Spark.

Topics include: Essential understanding of HDP and its capabilities, Hadoop, YARN, HDFS, MapReduce/Tez, data ingestion, using Pig and Hive to perform data analytics on Big Data and an introduction to Spark Core, Spark SQL, Apache Zeppelin, and additional Spark features.

Topics

- An Introduction to Apache Hadoop and HDFS
- Advanced Apache Pig Programming
- Advanced Apache Pig Programming
- Working with Pair RDDS and Building Yarn Applications

Audience

This course is designed for developers and data engineers who need to understand and develop applications on HDP.

Prerequisites

Before taking this course, students should be familiar with programming principles and have experience in software development. SQL and light scripting knowledge is also helpful. No prior Hadoop knowledge is required.

Duration

Four days

Course Outline

Hortonworks HDP Developer Quick Start

Course Outline

I. An Introduction to Apache Hadoop and HDFS

A. Describe the Case for Hadoop
B. Describe the Trends of Volume, Velocity and Variety
C. Discuss the Importance of Open Enterprise Hadoop
D. Describe the Hadoop Ecosystem Frameworks Across the Following Five Architectural Categories:
   1. Data Management
   2. Data Access
   3. Data Governance & Integration
   4. Security
   5. Operations
E. Describe the Function and Purpose of the Hadoop Distributed File System (HDFS)
F. List the Major Architectural Components of HDFS and their Interactions
G. Describe Data Ingestion
H. Describe Batch/Bulk Ingestion Options
I. Describe the Streaming Framework Alternatives
J. Describe the Purpose and Function of MapReduce
K. Describe the Purpose and Components of YARN
L. Describe the Major Architectural Components of YARN and their Interactions
M. Define the Purpose and Function of Apache Pig
N. Work with the Grunt Shell
O. Work with Pig Latin Relation Names and Field Names
P. Describe the Pig Data Types and Schema

Labs and Demonstrations

- Starting an HDP Cluster
- Using HDFS Commands
- Demonstration: Understanding Apache Pig
- Getting Started with Apache Pig
- Exploring Data with Pig

II. Advanced Apache Pig Programming

A. Demonstrate Common Operators Such as:
   1. Order by
   2. Case
   3. Distinct
   4. Parallel
   5. Foreach
B. Understand how Hive Tables are Defined and Implemented
C. Use Hive to Explore and Analyze Data Sets
D. Explain and Use the Various Hive File Formats
E. Create and Populate a Hive Table that Uses ORC File Formats
F. Use Hive to Run SQL-like Queries to Perform Data Analysis
G. Use Hive to Join Datasets Using a Variety of Techniques
H. Write Efficient Hive Queries
I. Explain the Uses and Purpose of HCatalog
J. Use HCatalog with Pig and Hive

Labs and Demonstrations

- Splitting a Dataset
- Joining Datasets
- Preparing Data for Apache Hive
- Understanding Apache Hive Tables
- Demonstration: Understanding Partitions 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

III. Advanced Apache Pig Programming

A. Describe How to Perform a Multi-Table/File Insert
B. Define and Use Views
C. Define and Use Clauses and Windows
D. List the Hive File Formats Including:
   1. Text Files
   2. SequenceFile
   3. RCFile
   4. ORC File

Hortonworks HDP Developer Quick Start

Course Outline (cont'd)

E. Define Hive Optimization
F. Use Apache Zeppelin to Work with Spark
G. Describe the Purpose and Benefits of Spark
H. Define Spark REPLs and Application Architecture
I. Explain the Purpose and Function of RDDs
J. Explain Spark Programming Basics
K. Define and Use Basic Spark Transformations
L. Define and Use Basic Spark Actions
M. Invoke Functions for Multiple RDDs, Create Named Functions and Use Numeric Operations

Labs

- Advanced Apache Hive Programming
- Introduction to Apache Spark REPLs and Apache Zeppelin
- Creating and Manipulating RDDs
- Creating and Manipulating Pair RDDs

IV. Working with Pair RDDS and Building Yarn Applications

A. Define and Create Pair RDDs
B. Perform Common Operations on Pair RDDs
C. Name the Various Components of Spark SQL and Explain their Purpose
D. Describe the Relationship Between DataFrames, Tables and Contexts
E. Use Various Methods to Create and Save DataFrames and Tables
F. Understand Caching, Persisting and the Different Storage Levels
G. Describe and Implement Checkpointing
H. Create an Application to Submit to the Cluster
I. Describe Client vs Cluster Submission with YARN
J. Submit an Application to the Cluster
K. List and Set Important Configuration Items

Labs

- Creating and Saving DateFrames and Tables
- Working with DataFrames
- Building and Submitting Applications to YARN