HDP Developer: Apache Pig and Hive

Overview

This 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 and using Pig and Hive to perform data analytics on Big Data. Labs are executed on a 7-node HDP cluster.

Duration

4 days

Target Audience

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

Course Objectives

- Describe Hadoop, YARN and use cases for Hadoop
- Describe Hadoop ecosystem tools and frameworks
- Describe the HDFS architecture
- Use the Hadoop client to input data into HDFS
- Transfer data between Hadoop and a relational database
- Explain YARN and MaoReduce architectures
- Run a MapReduce job on YARN
- Use Pig to explore and transform data in HDFS
- Use Hive to explore Understand how Hive tables are defined and implementedand analyze data sets
- Use the new Hive windowing functions
- Explain and use the various Hive file formats
- Create and populate a Hive table that uses ORC file formats
- Use Hive to run SQL-like queries to perform data analysis
- Use Hive to join datasets using a variety of techniques, including Map-side joins and Sort-Merge-Bucket joins
- Write efficient Hive queries
- Create ngrams and context ngrams using Hive
- Perform data analytics like quantiles and page rank on Big Data using the DataFu Pig library
- Explain the uses and purpose of HCatalog
- Use HCatalog with Pig and Hive
- Define a workflow using Oozie
- Schedule a recurring workflow using the Oozie Coordinator

Hands-On Labs

- Lab: Starting and HDP 2.3 Cluster
- Demo: Block Stprage
- Lab: Using HDFS commands
- Lab: Importing and Exporting Data in HDFS
- Lab: Using Flume to import log files into HDFS
- Demo: MapReduce
- Lab: Running a MapReduce Job
- Demo: Apache Pig
- Lab: Getting started with Apache Pig
- Lab: Exploring data with Apache Pig
- Lab: Splitting a datasetUse Sqoop to transfer data between HDFS and a RDBMS
- Run MapReduce and YARN application jobs
- Explore and transform data using Pig
- Split and join a dataset using Pig
- Use Pig to transform and export a dataset for use with Hive
- Use HCatLoader and HCatStorer
- Use Hive to discover useful information in a dataset
- Describe how Hive queries get executed as MapReduce jobs
- Perform a join of two datasets with Hive
- Use advanced Hive features: windowing, views, ORC files
- Use Hive analytics functions
- Write a custom reducer in Python
- Analyze and sessionize clickstream data
- Compute quantiles of NYSE stock prices
- Use Hive to compute ngrams on Avro-formatted files
- Lab: Exploring Spark SQL
- Lab: Defining an Oozie workflow

Prerequisites

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

Format

50% Lecture/Discussion
50% Hands-on Labs

Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit

HORTONWORKS®