ProTech "Charting the Course ...

protechtraining.com

... to Your Success!"

Hortonworks HDP Developer Apache Pig and Hive

Course Summary

Description

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, using Pig and Hive to perform data analytics on Big Data and an introduction to Spark Core and Spark SQL.

Objectives

By the end of this course, students will be able to:

- 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
- Understand how Hive tables are defined and implemented
- Use Hive to explore and 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
- Write efficient Hive queries
- Create ngrams and context ngrams using Hive
- Perform data analytics using the DataFu Pig library
- Explain the uses and purpose of HCatalog
- Use HCatalog with Pig and Hive
- Define and schedule an Oozie workflow
- Present the Spark ecosystem and high-level architecture
- Perform data analysis with Spark's Resilient Distributed Dataset API
- Explore Spark SQL and the DataFrame API

Topics

- Use HDFS commands to add/remove files and folders
- Use Sqoop to transfer data between HDFS and a RDBMS
- Run MapReduce and YARN application jobs
- Explore, transform, split and join datasets 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 clickstream data and compute quantiles with DataFu
- Use Hive to compute ngrams on Avro-formatted files
- Define an Oozie workflow
- Use Spark Core to read files and perform data analysis
- Create and join DataFrames with Spark SQL

Hortonworks HDP Developer Apache Pig and Hive

Course Summary (cont'd)

Audience

This course is designed for software developers who need to understand and develop applications for Hadoop.

Prerequisites

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

Duration

Four days