Hortonworks University

Hortonworks CERTIFIED DEVELOPER

Data Sheet

Developing Apache™ Hadoop© 2.0 Solutions for Data Analysts

This 4-day hands-on training course teaches students how to develop applications and analyze Big Data stored in Apache Hadoop 2.0 using Apache Pig™ and Apache Hive™. Students will learn the details of Hadoop 2.0, YARN, the Hadoop Distributed File System (HDFS™), an overview of MapReduce, and a deep dive into using Pig and Hive to perform data analytics on Big Data, including finding trends and detecting patterns.

Duration:

4 days

Prerequisites

Students should be familiar with SQL and have a minimal understanding of programming principles.
No prior Hadoop knowledge is required

Target Audience:

Data Analysts, BI Analysts, BI Developers, SAS Developers and other types of analysts who need to answer questions and analyze Big Data stored in a Hadoop cluster.

Course Objectives

At the completion of the course students will be able to:

- Explain Hadoop 2.0 and YARN
- Explain how HDFS Federation works in Hadoop 2.0
- Explain the various tools and frameworks in the Hadoop 2.0 ecosystem
- Explain the architecture of the Hadoop Distributed File System (HDFS)
- Use the Hadoop client to input data into HDFS
- Explain the architecture of MapReduce
- Run a MapReduce job on Hadoop
- Use Sqoop to transfer data between Hadoop and a relational database
- Write a Pig script to explore and transform data in HDFS
- Define advanced Pig relations
- Use Pig to apply structure to unstructured Big Data
- Invoke a Pig User-Defined Function
- Write a Hive query
- Understand how Hive tables are defined and implemented
- Use Hive to run SQL-like queries to perform data analysis
- Perform a multi-table select in Hive
- Design a proper schema for Hive
- Explain the uses and purpose of HCatalog™
- Use HCatalog with Pig and Hive
- Use Pig to organize and analyze Big Data
- Define a workflow using Oozie

Agenda

Day 1

- Understanding Hadoop 2.0 and YARN
- The Hadoop Distributed File System (HDFS)
- Inputting Data into HDFS
- The MapReduce Framework

Day 2

- Introduction to Pig
- Advanced Pig Programming

Day 3

- Hive Programming
- Using HCatalog

Day 4

- Advanced Hive Programming
- Data Analysis and Statistics
- Defining Workflow with Oozie

Lab Content

Students will work through the following lab exercises using the Hortonworks Data Platform 2.0:

- Using HDFS commands
- Using Sqoop to transfer data between HDFS and a RDBMS
- Running a MapReduce job
- Monitoring a MapReduce job
- Exploring data with Pig
- Splitting a dataset with Pig
- Joining datasets with Pig
- Using Pig to prepare data for Hive
- Understanding Hive tables
- Analyzing Big Data with Hive
- Understanding MapReduce in Hive
- Joining datasets with Hive
- A Multi-table select with Hive
- Streaming data with Hive and Python
- Using HCatalog with Pig
- Computing Quantiles with Pig
- Computing ngrams with Hive
- Defining Workflow with Oozie

Hortonworks certification identifies you as an expert in the Apache Hadoop ecosystem. Hortonworks offers a comprehensive certification program for students that attend a Hortonworks public or private on-site training course. Please visit hortonworks.com/training for more information.

Hortonworks University is your expert source for Apache Hadoop training and certification. Public and private on-site courses are available for developers, administrators, data analysts and other IT professionals involved in implementing big data solutions. Classes combine presentation material with industry-leading hands-on labs that fully prepare students for real-world Hadoop scenarios.

About Hortonworks
Hortonworks develops, distributes and supports the only 100-perecent open source distribution of Apache Hadoop explicitly architected, built and tested for enterprise-grade deployments.

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

3460 West Bayshore Road
Palo Alto, CA 94303 USA