Hewlett Packard Enterprise

Hortonworks Data Platform Developer—Apache Pig and Hive (EDU-PRIV-DEV-PIGHIVE-200) H7G68S

HPE course number	H7G68S
Course length	4 days
Delivery mode	ILT
View schedule, local pricing, and register	**View now**
View related courses	**View now**

This 4-day hands-on training course teaches students how to develop applications and analyze Big Data stored in Apache Hadoop 2.0 using Pig and 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. Other topics covered include data ingestion using Sqoop and Flume, and defining workflow using Oozie.

Note: this course was formerly named: Developing Apache Hadoop 2.0 Solutions for Data Analysts.

Why HPE Education Services?

- IDC MarketScape leader 4 years running for IT education and training*
- Recognized by IDC for leading with global coverage, unmatched technical expertise, and targeted education consulting services*
- Key partnerships with industry leaders OpenStack®, VMware®, Linux®, Microsoft®, ITIL, PMI, CSA, and (ISC)²
- Complete continuum of training delivery options—self-paced eLearning, custom education consulting, traditional classroom, video on-demand instruction, live virtual instructor-led with hands-on lab, dedicated onsite training
- Simplified purchase option with HPE Training Credits

*Realize Technology Value with Training, IDC Infographic 2037, Sponsored by HPE, January 2016

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

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

Course objectives

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

- Explain Hadoop 2.0 and YARN
- Explain use cases for Hadoop
- 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
- Use Sqoop to transfer data between Hadoop and a relational database
- Explain the architecture of MapReduce
- Explain the architecture of YARN
- Run a MapReduce job on YARN
- 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
- Use Pig to organize and analyze Big Data
- Understand how Hive tables are defined and implemented
- Use the new Hive windowing functions
- Explain and use the various Hive file formats
- Create and populate a Hive table that uses the new ORC file format
- 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

Benefits to you

- This course will provide in depth explanation on how to perform Hadoop 2.0 application development

Detailed course outline

Day 1	• Understanding Hadoop 2.0 • The Hadoop Distributed File System (HDFS) • Inputting Data into HDFS • The MapReduce Framework and YARN
Day 2	• Introduction to Pig • Advanced Pig Programming
Day 3	• Hive Programming • Using HCatalog • Advanced Hive Programming
Day 4	• Advanced Hive Programming (cont.) • 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: • Use HDFS commands to add/remove files and folders from HDFS • Use Sqoop to transfer data between HDFS and a RDBMS • Run a MapReduce job • Run a YARN application • Explore and transform data using Pig • Split a dataset using Pig • Join two datasets using Pig • Use Pig to transform and export a dataset for use with Hive • Use HCatLoader and HCatStorer to retrieve HCatalog schemas from within a Pig script • Understand how a Hive table is stored in HDFS • Use Hive to discover useful information in a dataset • Understand how Hive queries get executed as MapReduce jobs • Perform a join of two datasets with Hive • Use advanced Hive features like windowing, views and ORC files • Use the Hive analytics functions (rank, dense_rank, cume_dist, row_number) • Write a custom reducer in Python that reduces the number of underlying MapReduce jobs generated from a Hive query • Analyze and sessionize clickstream data using the Pig DataFu library • Compute quantiles of NYSE stock prices • Use Hive to compute ngrams on Avro-formatted files • Define an Oozie workflow

Learn more at
hpe.com/ww/learnbigdata

Follow us:

Hewlett Packard Enterprise

c04577739, December 2016, Rev. 1