HORTONWORKS
POWERING THE FUTURE OF DATA™

TRAINING AGENDA

HORTONWORKS DATA PLATFORM (HDP®) BATCH WORKLOAD DEVELOPMENT

2 DAYS INTERMEDIATE DEVELOPER

This 2 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 & 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

PREREQUISITES

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.

TARGET AUDIENCE

Developers and data engineers who need to understand and develop applications on HDP

AGENDA SUMMARY

Day 1 Morning: HDP Essentials
Day 1 Afternoon: Pig, Hive & Sqoop
Day 2: Spark

Day 1 Intro: HDP Essentials

- Describe the Case for Hadoop
- Identify the Hadoop Ecosystem via architectural categories
  - Data Management — HDFS, YARN

  - Data Access — Pig, Hive, Apache HBase, Apache Storm, Apache Solr, Spark
  - Data Governance & Integration — Apache Falcon, Apache Flume, Apache Sqoop, Apache Kafka, Apache Atlas
- Detail the HDFS architecture
- Describe data ingestion options and frameworks for batch and real-time streaming
- Explain the fundamentals of parallel processing
- Detail the architecture and features of YARN
- Optional: Describe how to secure Hadoop

Day 1: Pig & Hive

- Use Pig to explore and transform data in HDFS
- Transfer data between Hadoop and a relational database
- Understand how Hive tables are defined and implemented
- Use Hive to explore and analyze data sets
- Use Hive to run SQL-like queries to perform data analysis
- Explain the uses and purpose of HCatalog
- Use HCatalog with Pig and Hive
- Understand Sqoop for importing and exporting data to HDFS and a RDBMS

Hands-On Labs

- Use Sqoop to transfer data between HDFS and a RDBMS
- Use HDFS commands to add/remove files and folders
- 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

About Hortonworks

Hortonworks is a leading innovator at creating, distributing and supporting enterprise-ready open data platforms. Our mission is to manage the world's data. We have a single-minded focus on driving innovation in open source communities such as Apache Hadoop, NiFi, and Spark. Our open Connected Data Platforms power Modern Data Applications that deliver actionable intelligence from all data: data in-motion and data-at-rest. Along with our 1600+ partners, we provide the expertise, training and services that allows our customers to unlock the transformational value of data across any line of business. We are Powering the Future of Data™.

Contact

For further information visit
www.hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405

Day 2: Spark

- Describe Spark and Spark specific use cases
- Explore data interactively through the spark shell utility
- Explore and manipulate data using Zeppelin
- Explain the RDD concept
- Use the Python or Scala Spark APIs
- Create all types of RDDs: Pair, Double, and Generic
- Use RDD type-specific functions
- Deploy applications to the cluster using YARN
- Create applications using the Spark SQL library
- Create/transform data using Spark DataFrames

Spark Python or Scala Hands-On Labs

- Create a Spark “Hello World” word count application
- Use advanced RDD programming to perform sort, join, pattern matching and regex tasks
- Build/package a Spark application using Maven
- Create a data frame and perform analysis
- Load/transform/store data using Spark with Hive tables

About Hortonworks

Hortonworks is a leading innovator at creating, distributing and supporting enterprise-ready open data platforms. Our mission is to manage the world’s data. We have a single-minded focus on driving innovation in open source communities such as Apache Hadoop, NiFi, and Spark. Our open Connected Data Platforms power Modern Data Applications that deliver actionable intelligence from all data: data in-motion and data-at-rest. Along with our 1600+ partners, we provide the expertise, training and services that allows our customers to unlock the transformational value of data across any line of business. We are Powering the Future of Data™.

Contact

For further information visit
www.hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405