



Hortonworks  
**UNIVERSITY**

# Training Catalog

## Apache Hadoop Training from the Experts



February 2016






# Hortonworks University

Hortonworks University provides an immersive and valuable real world experience with scenario-based training Courses in public, private on site and virtual led courses, s self-paced learning library, and an Academic program. All courses include Industry-leading lecture and hands-on labs.

## Individualized Learning Paths



### I'm a Developer


Interested in:  
Architecture & Fundamentals  
MapReduce Programming  
Real-time Analytics

**Developer Classes:**

- [HDP Developer: Java](#)  
4 Days
- [HDP Developer: Apache Pig and Hive](#)  
4 Days
- [HDP Developer: Windows](#)  
4 Days
- [HDP Developer: Custom YARN Applications](#)  
2 Days
- [HDP Developer: Storm and Trident](#)  
2 Days

**COMING SOON**

- [HDP Developer: Apache Spark Using Python](#)  
3 Days
- [HDP Developer: Apache Spark using Scala](#)  
3 Days



### I'm a System Admin


Interested in:  
Cluster Monitoring  
Security & Governance  
Certification

**Admin Classes**

- [HDP Operations: Migrating to the Hortonworks Data Platform](#)  
2 Days
- [HDP Operations: Hadoop Administration I](#)  
4 Days
- [HDP Operations: Apache HBase Advanced Management](#)  
4 Days

**COMING SOON**

- [HDP Operations: Hortonworks Data Flow](#)  
3 Days
- [HDP Operations: Security](#)  
4 Days
- [HDP Operations: Hadoop Administration 2](#)  
4 Days



### I'm a Data Analyst

Interested in:  
SQL & Scripting Languages  
Large Scale Data Sets  
Creating Value & Opportunity

**Data Classes:**

- [HDP Analyst: Data Science](#)  
3 Days
- [HDP Analyst: Apache HBase Essentials](#)  
2 Days

## Hadoop Certification

Join an exclusive group of professionals with demonstrated skills and the qualifications to prove it. Hortonworks certified professionals are recognized as leaders in the field. Hortonworks certified professionals are recognized as leaders in the field.

### Hortonworks Certified Developer:

- HDP Certified Developer (HDPCD)
- HDP Certified Developer: Java (HDPCD: Java)

### Hortonworks Certified Administrator:

- HDP Certified Administrator (HDPCA)



## Table of Contents

<b>Hortonworks University Self-Paced Learning Library</b>	<b>7</b>
<b>HDP Overview: Apache Hadoop Essentials</b>	<b>8</b>
<b>HDP Analyst: Apache HBase Essentials</b>	<b>9</b>
<b>HDP Analyst: Data Science</b>	<b>10</b>
<b>HDP Developer: Apache Pig and Hive</b>	<b>11</b>
<b>HDP Developer: Java</b>	<b>12</b>
<b>HDP Developer: Windows</b>	<b>13</b>
<b>HDP Developer: Custom YARN Applications</b>	<b>14</b>
<b>HDP Developer: Apache Spark using Python</b>	<b>15</b>
<b>HDP Developer: Apache Spark using Scala</b>	<b>16</b>
<b>HDP Developer: Storm and Trident Fundamentals</b>	<b>17</b>
<b>HDP Operations: Hadoop Administration 1</b>	<b>18</b>
<b>HDP Operations: Hadoop Administration 2</b>	<b>19</b>
<b>HDP Operations: Apache HBase Advanced Management</b>	<b>20</b>
<b>HDP Operations: Hortonworks Data Flow</b>	<b>21</b>
<b>HDP Operations: Security</b>	<b>22</b>
<b>HDP Operations: Migrating to the Hortonworks Data Platform</b>	<b>23</b>
<b>HDP Certified Administrator (HDPCA)</b>	<b>24</b>
<b>HDP Certified Developer (HDPCD)</b>	<b>27</b>
<b>HDP Certified Java Developer (HDPCD_Java)</b>	<b>31</b>
<b>Hortonworks University Academic Program</b>	<b>32</b>





## Hortonworks University Self-Paced Learning Library

### Overview

Hortonworks University Self-Paced Learning Library is an on-demand, online, learning repository that is accessed using a Hortonworks University account. Learners can view lessons anywhere, at any time, and complete lessons at their own pace. Lessons can be stopped and started, as needed, and completion is tracked via the Hortonworks University Learning Management System.

This learning library makes it easy for Hadoop Administrators, Data Analysts, and Developers to continuously learn and stay up-to-date on Hortonworks Data Platform.

Hortonworks University courses are designed and developed by Hadoop experts and provide an immersive and valuable real world experience. In our scenario-based training courses, we offer unmatched depth and expertise. We prepare you to be an expert with highly valued, practical skills and prepare you to successfully complete Hortonworks Technical Certifications.

The Self-Paced learning library accelerates time to Hadoop competency. In addition, the learning library content is constantly being expanded with new content being added on an ongoing basis.

### Duration

Access to the Hortonworks University Self Paced Learning Library is provided for a 12-month subscription period per individual named user. The subscription includes access to over 400 hours of individual lessons.

### Target Audience

The Hortonworks University Self-Paced Learning Library is designed for architects, developers, analysts, data scientists, and IT decision makers – as well as those new to Hadoop...essentially anyone with a need or desire to learn more about Apache Hadoop and the Hortonworks Data Platform framework.

### Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit [hortonworks.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Self-Paced Learning Content

- **HDP Overview:** HDP Essentials
- **HDP Developer Learning Path**
  - Apache Pig and Hive
  - Windows
  - Developing Applications with Java
  - Developing Custom YARN Applications
  - Storm and Trident Fundamentals
  - Apache Spark using Python\*\*
  - Apache Spark using Scala\*\*
- **HDP Operations**
  - Hadoop Administration I
  - Hortonworks Data Flow\*\*
  - Apache HBase Advanced Management
  - Migrating to HDP
  - Hadoop Administration II\*\*
  - Hadoop Security\*\*
- **HDP Analyst**
  - Apache Pig and Hive
  - Apache HBase Essentials
  - Data Science

\*\* Coming soon!

### Accessing the Self-Paced Learning Library

Access to Hortonworks Self Paced Learning Library is included as part of the Hortonworks Enterprise, Enterprise Plus & Premiere Subscriptions for each named Support Contact. Additional Self Paced Learning Library subscriptions can be purchased on a per-user basis for individuals who are not named Support Contacts.

### Prerequisites

None.

### Hortonworks University

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 deployments.

For more information contact: [trainingops@hortonworks.com](mailto:trainingops@hortonworks.com)



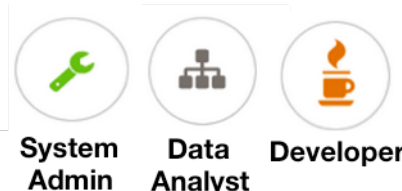
#### About Hortonworks

Hortonworks develops, distributes and supports the only 100 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Overview: Apache Hadoop Essentials

### Overview

This course provides a technical overview of Apache Hadoop. It includes high-level information about concepts, architecture, operation, and uses of the Hortonworks Data Platform (HDP) and the Hadoop ecosystem. The course provides an *optional* primer for those who plan to attend a hands-on, instructor-led course

### Course Objectives

- Describe what makes data “Big Data”
- List data types stored and analyzed in Hadoop
- Describe how Big Data and Hadoop fit into your current infrastructure and environment
- Describe fundamentals of:
  - the Hadoop Distributed File System (HDFS)
  - YARN
  - MapReduce
  - Hadoop frameworks: (Pig, Hive, HCatalog, Storm, Solr, Spark, HBase, Oozie, Ambari, ZooKeeper, Sqoop, Flume, and Falcon)
  - Recognize use cases for Hadoop
  - Describe the business value of Hadoop
  - Describe new technologies like Tez and the Knox Gateway

### Hands-On Labs

- There are no labs for this course.

### Duration

8 Hours, On Line.

### Target Audience

Data architects, data integration architects, managers, C-level executives, decision makers, technical infrastructure team, and Hadoop administrators or developers who want to understand the fundamentals of Big Data and the Hadoop ecosystem.

### Prerequisites

No previous Hadoop or programming knowledge is required. Students will need browser access to the Internet.

### Format

- 100% self-paced, online exploration (for employees, partners or support subscription customers)  
or
- 100% instructor led discussion

### Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit [hortonworks.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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

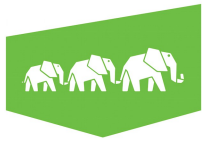
Hortonworks develops, distributes and supports the only 100 percent open source distribution of Apache Hadoop explicitly architected, built and tested for enterprise-grade deployments.

**US:** 1.855.846.7866

**International:** +1.408.916.4121  
[www.hortonworks.com](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA





## HDP Analyst: Apache HBase Essentials

### Overview

This course is designed for big data analysts who want to use the HBase NoSQL database which runs on top of HDFS to provide real-time read/write access to sparse datasets. Topics include HBase architecture, services, installation and schema design.

### Course Objectives

- How HBase integrates with Hadoop and HDFS
- Architectural components and core concepts of HBase
- HBase functionality
- Installing and configuring HBase
- HBase schema design
- Importing and exporting data
- Backup and recovery
- Monitoring and managing HBase
- How Apache Phoenix works with HBase
- How HBase integrates with Apache ZooKeeper
- HBase services and data operations
- Optimizing HBase Access

### Hands-On Labs

- Using Hadoop and MapReduce
- Using HBase
- Importing Data from MySQL to HBase
- Using Apache ZooKeeper
- Examining Configuration Files
- Using Backup and Snapshot
- HBase Shell Operations
- Creating Tables with Multiple Column Families
- Exploring HBase Schema
- Blocksize and Bloom filters
- Exporting Data
- Using a Java Data Access Object Application to Interact with HBase

### Duration

2 days

### Target Audience

Architects, software developers, and analysts responsible for implementing non-SQL databases in order to handle sparse data sets commonly found in big data use cases.

### Prerequisites

Students must have basic familiarity with data management systems. Familiarity with Hadoop or databases is helpful but not required. Students new to Hadoop are encouraged to attend the *HDP Overview: Apache Hadoop Essentials* course.

### Format

35% Lecture/Discussion  
65% Hands-on Labs

### Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit [hortonworks.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Analyst: Data Science

### Overview

This course Provides instruction on the processes and practice of data science, including machine learning and natural language processing. Included are: tools and programming languages (Python, IPython, Mahout, Pig, NumPy, pandas, SciPy, Scikit-learn), the Natural Language Toolkit (NLTK), and Spark MLlib.

### Duration

3 days

### Target Audience

Architects, software developers, analysts and data scientists who need to apply data science and machine learning on Hadoop.

### Course Objectives

- Recognize use cases for data science on Hadoop
- Describe the Hadoop and YARN architecture
- Describe supervised and unsupervised learning differences
- Use Mahout to run a machine learning algorithm on Hadoop
- Describe the data science life cycle
- Use Pig to transform and prepare data on Hadoop
- Write a Python script
- Describe options for running Python code on a Hadoop cluster
- Write a Pig User-Defined Function in Python
- Use Pig streaming on Hadoop with a Python script
- Use machine learning algorithms
- Describe use cases for Natural Language Processing (NLP)
- Use the Natural Language Toolkit (NLTK)
- Describe the components of a Spark application
- Write a Spark application in Python
- Run machine learning algorithms using Spark MLlib
- Take data science into production

### Prerequisites

Students must have experience with at least one programming or scripting language, knowledge in statistics and/or mathematics, and a basic understanding of big data and Hadoop principles. Students new to Hadoop are encouraged to attend the *HDP Overview: Apache Hadoop Essentials* course.

### Hands-On Content

- Lab: Setting Up a Development Environment
- Demo: Block Storage
- Lab: Using HDFS Commands
- Demo: MapReduce
- Lab: Using Apache Mahout for Machine Learning
- Demo: Apache Pig
- Lab: Getting Started with Apache Pig
- Lab: Exploring Data with Pig
- Lab: Using the IPython Notebook
- Demo: The NumPy Package
- Demo: The pandas Library
- Lab: Data Analysis with Python
- Lab: Interpolating Data Points
- Lab: Defining a Pig UDF in Python
- Lab: Streaming Python with Pig
- Demo: Classification with Scikit-Learn
- Lab: Computing K-Nearest Neighbor
- Lab: Generating a K-Means Clustering
- Lab: POS Tagging Using a Decision Tree
- Lab: Using NLTK for Natural Language Processing
- Lab: Classifying Text using Naive Bayes
- Lab: Using Spark Transformations and Actions
- Lab Using Spark MLlib
- Lab: Creating a Spam Classifier with MLlib

### 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## 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 MapReduce 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 implemented 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, 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 Storage
- 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 dataset Use 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



#### About Hortonworks

Hortonworks develops, distributes and supports the only 100 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Developer: Java

### Overview

This advanced course provides Java programmers a deep-dive into Hadoop application development. Students will learn how to design and develop efficient and effective MapReduce applications for Hadoop using the Hortonworks Data Platform, including how to implement combiners, partitioners, secondary sorts, custom input and output formats, joining large datasets, unit testing, and developing UDFs for Pig and Hive. Labs are run on a 7-node HDP 2.1 cluster running in a virtual machine that students can keep for use after the training.

### Duration

4 days

### Target Audience

Experienced Java software engineers who need to develop Java MapReduce applications for Hadoop.

### Course Objectives

- Describe Hadoop 2 and the Hadoop Distributed File System
- Describe the YARN framework
- Develop and run a Java MapReduce application on YARN
- Use combiners and in-map aggregation
- Write a custom partitioner to avoid data skew on reducers
- Perform a secondary sort
- Recognize use cases for built-in input and output formats
- Write a custom MapReduce input and output format
- Optimize a MapReduce job
- Configure MapReduce to optimize mappers and reducers
- Develop a custom RawComparator class
- Distribute files as LocalResources
- Describe and perform join techniques in Hadoop
- Perform unit tests using the UnitMR API
- Describe the basic architecture of HBase
- Write an HBase MapReduce application
- List use cases for Pig and Hive
- Write a simple Pig script to explore and transform big data
- Write a Pig UDF (User-Defined Function) in Java
- Write a Hive UDF in Java
- Use JobControl class to create a MapReduce workflow
- Use Oozie to define and schedule workflows

### Hands-On Labs

- Configuring a Hadoop Development Environment
- Putting data into HDFS using Java
- Write a distributed grep MapReduce application
- Write an inverted index MapReduce application
- Configure and use a combiner
- Writing custom combiners and partitioners
- Globally sort output using the TotalOrderPartitioner
- Writing a MapReduce job to sort data using a composite key
- Writing a custom InputFormat class
- Writing a custom OutputFormat class
- Compute a simple moving average of stock price data
- Use data compression
- Define a RawComparator
- Perform a map-side join
- Using a Bloom filter
- Unit testing a MapReduce job
- Importing data into HBase
- Writing an HBase MapReduce job
- Writing User-Defined Pig and Hive functions
- Defining an Oozie workflow

### Prerequisites

Students must have experience developing Java applications and using a Java IDE. Labs are completed using the Eclipse IDE and Gradle. 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Developer: Windows

### Overview

This course is designed for developers who create applications and analyze Big Data in Apache Hadoop on Windows using Pig and Hive. Topics include: Hadoop, YARN, the Hadoop Distributed File System (HDFS), MapReduce, Sqoop and the HiveODBC Driver.

### Duration

4 days

### Target Audience

Software developers who need to understand and develop applications for Hadoop 2.x on Windows.

### Course Objectives

- Describe Hadoop and Hadoop and YARN
- Describe the Hadoop ecosystem
- List Components & deployment options for HDP on Windows
- Describe the HDFS architecture
- Use the Hadoop client to input data into HDFS
- Transfer data between Hadoop and Microsoft SQL Server
- Describe the MapReduce and YARN architecture
- Run a MapReduce job on YARN
- Write a Pig script
- 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
- Describe how Hive tables are defined and implemented
- Use Hive windowing functions
- Define and use Hive file formats
- Create Hive tables that use the ORC file format
- Use Hive to run SQL-like queries to perform data analysis
- Use Hive to join datasets
- Create ngrams and context ngrams using Hive
- Perform data analytics
- Use HCatalog with Pig and Hive
- Install and configure HiveODBC Driver for Windows
- Import data from Hadoop into Microsoft Excel
- Define a workflow using Oozie

### Hands-On Labs

- Start HDP on Windows
- Add/remove files and folders from HDFS
- Transfer data between HDFS and Microsoft SQL Server
- Run a MapReduce job
- Using Pig to analyze data
- Retrieve HCatalog schemas from within a Pig script
- Using Hive tables and queries
- Advanced Hive features like windowing, views and ORC files
- Hive analytics functions using the Pig DataFu library
- Compute quantiles
- Use Hive to compute ngrams on Avro-formatted files
- Connect Microsoft Excel to Hadoop with HiveODBC Driver
- Run a YARN application
- Define an Oozie workflow

### Prerequisites

Students should be familiar with programming principles and have experience in software development. SQL knowledge and familiarity with Microsoft Windows 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA





## HDP Developer: Custom YARN Applications

### Overview

This course is designed for developers who want to create custom YARN applications for Apache Hadoop. It will include: the YARN architecture, YARN development steps, writing a YARN client and ApplicationMaster, and launching Containers. The course uses Eclipse and Gradle connected remotely to a 7-node HDP cluster running in a virtual machine.

### Course Objectives

- Describe the YARN architecture
- Describe the YARN application lifecycle
- Write a YARN client application
- Run a YARN application on a Hadoop cluster
- Monitor the status of a running YARN application
- View the aggregated logs of a YARN application
- Configure a ContainerLaunchContext
- Use a LocalResource to share application files across a cluster
- Write a YARN ApplicationMaster
- Describe the differences between synchronous and asynchronous ApplicationMasters
- Allocate Containers in a cluster
- Launch Containers on NodeManagers
- Write a custom Container to perform specific business logic
- Explain the job schedulers of the ResourceManager
- Define queues for the Capacity Scheduler

### Hands-On Labs

- Run a YARN Application
- Setup a YARN Development Environment
- Write a YARN Client
- Submit an ApplicationMaster
- Write an ApplicationMaster
- Requesting Containers
- Running Containers
- Writing Custom Containers

### Duration

2 days

### Target Audience

Java software engineers who need to develop YARN applications on Hadoop by writing YARN clients and ApplicationMasters.

### Prerequisites

Students should be experienced Java developers who have attended *HDP Developer: Java* **OR** *HDP Developer: Pig and Hive* **OR** are experienced with Hadoop and MapReduce development.

### 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Developer: Apache Spark Using Python

### Overview

This course is designed for developers who need to create applications to analyze Big Data stored in Apache Hadoop using Spark. Topics include: Hadoop, YARN, HDFS, using Spark for interactive data exploration, building and deploying Spark applications, optimization of applications, creating Spark pipelines with multiple libraries, working with different filetypes, building data frames, exploring the Spark SQL API, using Spark Streaming and an introduction to Spark MLlib.

### Duration

3 days

### Target Audience

Software engineers that are looking to develop time sensitive applications for Hadoop.

### Course Objectives

- Describe Hadoop, HDFS, YARN, and uses cases for Hadoop
- Describe Spark and Spark specific use cases
- Understand the HDFS architecture
- Use the HDFS commands to insert and retrieve data
- Explain the differences between Spark and MapReduce
- Explore data interactively through the spark shell utility
- Explain the RDD concept
- Understand concepts of functional programming
- Use the Python or Scala Spark APIs
- Create all types of RDDs: Pair, Double, and Generic
- Use RDD type-specific functions
- Explain interaction of components of a Spark Application
- Explain the creation of the DAG schedule
- Build and package Spark applications
- Use application configuration items
- Deploy applications to the cluster using YARN
- Use data caching to increase performance of applications
- Implement advanced features of spark
- Learn general application optimization guidelines/tips
- Create applications using the Spark SQL library
- Create/transform data using dataframes
- Read, use, and save to different Hadoop file formats
- Understand the concepts of Spark Streaming
- Create a streaming application
- Use Spark MLlib to gain insights from data

### Hands-On Labs

- Create a Spark “Hello World” word count application
- Use HDFS commands to add and remove files and folders
- Use advanced RDD programming to perform sort, join, pattern matching and regex tasks
- Explore partitioning and the Spark UI
- Increase performance using data caching
- Checkpoint iterative applications
- Build/package a Spark application using Maven
- Use a broadcast variable to efficiently join a small dataset to a massive dataset
- Use an accumulator for reporting data quality issues
- Create a data frame and perform analysis
- Load/transform/store data using Spark with Hive tables
- Create a point-in-time spark stream application
- Create a spark stream application using window functions
- Create a Spark MLlib application using K-Means

### Prerequisites

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

### 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)  
5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Developer: Apache Spark Using Scala

### Overview

This course is designed for developers who need to create applications to analyze Big Data stored in Apache Hadoop using Spark. Topics include: Hadoop, YARN, HDFS, using Spark for interactive data exploration, building and deploying Spark applications, optimization of applications, creating Spark pipelines with multiple libraries, working with different filetypes, building data frames, exploring the Spark SQL API, using Spark Streaming and an introduction to Spark MLlib.

### Duration

3 days

### Target Audience

Software engineers that are looking to develop time sensitive applications for Hadoop.

### Course Objectives

- Describe Hadoop, HDFS, YARN, and use cases for Hadoop
- Describe Spark and Spark specific use cases
- Understand the HDFS architecture
- Use the HDFS commands to insert and retrieve data
- Explain the differences between Spark and MapReduce
- Explore data interactively through the spark shell utility
- Explain the RDD concept
- Understand concepts of functional programming
- Use the Python or Scala Spark APIs
- Create all types of RDDs: Pair, Double, and Generic
- Use RDD type-specific functions
- Explain interaction of components of a Spark Application
- Explain the creation of the DAG schedule
- Build and package Spark applications
- Use application configuration items
- Deploy applications to the cluster using YARN
- Use data caching to increase performance of applications
- Implement advanced features of spark
- Learn general application optimization guidelines/tips
- Create applications using the Spark SQL library
- Create/transform data using dataframes
- Read, use, and save to different Hadoop file formats
- Understand the concepts of Spark Streaming
- Create a streaming application
- Use Spark MLlib to gain insights from data

### Hands-On Labs

- Create a Spark “Hello World” word count application
- Use HDFS commands to add and remove files and folders
- Use advanced RDD programming to perform sort, join, pattern matching and regex tasks
- Explore partitioning and the Spark UI
- Increase performance using data caching
- Checkpoint iterative applications
- Build/package a Spark application using Maven
- Use a broadcast variable to efficiently join a small dataset to a massive dataset
- Use an accumulator for reporting data quality issues
- Create a data frame and perform analysis
- Load/transform/store data using Spark with Hive tables
- Create a point-in-time spark stream application
- Create a spark stream application using window functions
- Create a Spark MLlib application using K-Means

### Prerequisites

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

### 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)  
5470 Great America Parkway  
Santa Clara, CA 95054 USA





## HDP Developer: Storm and Trident Fundamentals

### Overview

This course provides a technical introduction to the fundamentals of Apache Storm and Trident that includes the concepts, terminology, architecture, installation, operation, and management of Storm and Trident. Simple Storm and Trident code excerpts are provided throughout the course. The course also includes an introduction to, and code samples for, Apache Kafka. Apache Kafka is a messaging system that is commonly used in concert with Storm and Trident.

### Duration

Approximately 2 days

### Target Audience

Hadoop developers who need to be able to design and build Storm and Kafka applications using Java and the Trident API.

### Prerequisites

Students must have experience developing Java applications and using a Java IDE. Labs are completed using the Eclipse IDE and Gradle. Students should have a basic understanding of Hadoop.

### Format

Self-paced, online exploration or  
Instructor led exploration and discussion

### Hortonworks University

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.

### Course Objectives

- Recognize differences between batch and real-time data processing
- Define Storm elements including tuples, streams, spouts, topologies, worker processes, executors, and stream groupings
- Explain and install Storm architectural components, including Nimbus, Supervisors, and ZooKeeper cluster
- Recognize/interpret Java code for a spout, bolt, or topology
- Identify how to develop and submit a topology to a local or remote distributed cluster
- Recognize and explain the differences between reliable and unreliable Storm operation
- Manage and monitor Storm using the command-line client or browser-based Storm User Interface (UI)
- Define Kafka topics, producers, consumers, and brokers
- Publish Kafka messages to Storm or Trident topologies
- Define Trident elements including tuples, streams, batches, partitions, topologies, Trident spouts, and operations
- Recognize and interpret the code for Trident operations, including filters, functions, aggregations, merges, and joins
- Recognize the differences between the different types of Trident state
- Identify how Trident state supports exactly-once processing semantics and idempotent operation
- Recognize the differences in fault tolerance between different types of Trident spouts
- Recognize and interpret the code for Trident state-based operations

### Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit [hortonworks.com/training/certification](http://hortonworks.com/training/certification) for more information.



#### About Hortonworks

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

**US:** 1.855.846.7866 Hortonworks  
**International:** +1.408.916.4121  
[www.hortonworks.com](http://www.hortonworks.com) Hadoop

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Operations: Hadoop Administration 1

### Overview

This course is designed for administrators who will be managing the Hortonworks Data Platform (HDP) 2.3 with Ambari. It covers installation, configuration, and other typical cluster maintenance tasks.

### Duration

4 days

### Target Audience

IT administrators and operators responsible for installing, configuring and supporting an HDP 2.3 deployment in a Linux environment using Ambari.

### Hands-On Labs

- Introduction to the Lab Environment
- Performing an Interactive Ambari HDP Cluster Installation
- Configuring Ambari Users and Groups
- Managing Hadoop Services
- Using HDFS Files and Directories
- Using WebHDFS
- Configuring HDFS ACLs
- Managing HDFS
- Managing HDFS Quotas
- Configuring HDFS Transparent Data Encryption
- Configuring and Managing YARN
- Non-Ambari YARN Management
- Configuring YARN Failure Sensitivity, Work Preserving Restarts, and Log Aggregation Settings
- Submitting YARN Jobs
- Configuring Different Workload Types
- Configuring User and Groups for YARN Labs
- Configuring YARN Resource Behavior and Queues
- User, Group and Fine-Tuned Resource Management
- Adding Worker Nodes
- Configuring Rack Awareness
- Configuring HDFS High Availability
- Configuring YARN High Availability
- Configuring and Managing Ambari Alerts
- Configuring and Managing HDFS Snapshots
- Using Distributed Copy (DistCP)

### Course Objectives

- Summarize and enterprise environment including Big Data, Hadoop and the Hortonworks Data Platform (HDP)
- Install HDP
- Manage Ambari Users and Groups
- Manage Hadoop Services
- Use HDFS Storage
- Manage HDFS Storage
- Configure HDFS Storage
- Configure HDFS Transparent Data Encryption
- Configure the YARN Resource Manager
- Submit YARN Jobs
- Configure the YARN Capacity Scheduler
- Add and Remove Cluster Nodes
- Configure HDFS and YARN Rack Awareness
- Configure HDFS and YARN High Availability
- Monitor a Cluster
- Protect a Cluster with Backups

### Prerequisites

Attendees should be familiar with with Hadoop and Linux environments.

### Format

60% Lecture/Discussion  
40% Hands-on Labs

### Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit [hortonworks.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA

## HDP Operations: Hadoop Administration 2

### Overview

This course is designed for experienced administrators who manage Hortonworks Data Platform (HDP) 2.3 clusters with Ambari. It covers upgrades, configuration, application management, and other common tasks.

### Duration

4 days

### Target Audience

IT administrators and operators responsible for configuring, managing, and supporting an Hadoop 2.3 deployment in a Linux environment using Ambari.

### Course Objectives

- Execute automated installation of and upgrades to HDP clusters
- Configure HDFS for NFS integration and centralized caching
- Control application behavior using node labels
- Deploy applications using Slider
- Understand how to configure HDP for optimum Hive performance
- Understand how to manage HDP data compression
- Integrate Ambari with an existing LDAP environments to manage users and groups
- Configure high availability for Hive and Oozie
- Ingest SQL tables and log files into HDFS
- Support scalable and automated HDP application best practices
- Configure automated HDP data replication

### Prerequisites

Attendees should have attended HDP Operations: Hadoop Administration 1 or possess equivalent knowledge and experience. Attendees should be familiar with basic HDP administration and Linux environments.

### Hands-On Labs

- Introduction to the Lab Environment
- Perform an HDP Rolling Upgrade
- Configure HDFS NFS Gateway
- Configure HDFS Centralized Cache
- Configure YARN Node Labels
- Deploy HBase using Slider
- Integrate Ambari with LDAP
- Configure Hive High Availability
- Move Data Using Flume and Sqoop
- Run an Oozie Workflow
- Configure Oozie High Availability
- Configure Data Mirroring using Falcon
- Install HDP using Ambari Blueprints

### Additional Topics

- Hive Performance Tuning
- Managing Data Compression

### Format

60% Lecture/Discussion

40% Hands-on Labs

### Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit [hortonworks.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent open source distribution of Apache Hadoop explicitly architected, built and tested for enterprise-grade deployments.

US: 1.855.846.7866

International: +44(0) 20 3826 1405  
[www.hortonworks.com](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Operations: Apache HBase Advanced Management

### Overview

This course is designed for administrators who will be installing, configuring and managing HBase clusters. It covers installation with Ambari, configuration, security and troubleshooting HBase implementations. The course includes an end-of-course project in which students work together to design and implement an HBase schema.

### Course Objectives

- Hadoop Primer
  - Hadoop, Hortonworks, and Big Data
  - HDFS and YARN
- Discussion: Running Applications in the Cloud
- Apache HBase Overview
- Provisioning the Cluster
- Using the HBase Shell
- Ingesting Data
- Operational Management
- Backup and Recovery
- Security
- Monitoring HBase and Diagnosing Problems
- Maintenance
- Troubleshooting

### Hands-On Labs

- Installing and Configuring HBase with Ambari
- Manually Installing HBase (Optional)
- Using Shell Commands
- Ingesting Data using ImportTSV
- Enabling HBase High Availability
- Viewing Log Files
- Configuring and Enabling Snapshots
- Configuring Cluster Replication
- Enabling Authentication and Authorization
- Diagnosing and Resolving Hot Spotting
- Region Splitting
- Monitoring JVM Garbage Collection
- End-of-Course Project: Designing an HBase Schema

### Duration

4 days

### Target Audience

Architects, software developers, and analysts responsible for implementing non-SQL databases in order to handle sparse data sets commonly found in big data use cases.

### Prerequisites

Students must have basic familiarity with data management systems. Familiarity with Hadoop or databases is helpful but not required. Students new to Hadoop are encouraged to take the *HDP Overview: Apache Hadoop Essentials* course.

### 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Operations: Hortonworks Data Flow

### Overview

This course is designed for 'Data Stewards' or 'Data Flow Managers' who are looking forward to automate the flow of data between systems. Topics Include Introduction to NiFi, Installing and Configuring NiFi, Detail explanation of NiFi User Interface, Explanation of its components and Elements associated with each. How to Build a dataflow, NiFi Expression Language, Understanding NiFi Clustering, Data Provenance, Security around NiFi, Monitoring Tools and HDF Best practices.

### Duration

3 days

### Format

50% Lecture/Discussion  
50% Hands-on Labs

### Target Audience

Data Engineers, Integration Engineers and Architects who are looking to automate Data flow between systems.

### Course Objectives

- Describe HDF, Apache NiFi and its use cases.
- Describe NiFi Architecture
- Understand NiFi Features and Characteristics.
- Understand System requirements to run NiFi.
- Understand Installing and Configuring NiFi
- Understand NiFi user interface in depth.
- Understand how to build a DataFlow using NiFi
- Understand Processor and its Elements
- Understand Connection and its Elements
- Understand Processor Group and its elements
- Understand Remote Processor Group and its Elements
- Learn how to optimize a DataFlow
- Learn how to use NiFi Expression language and its use.
- Learn about Attributes and Templates in NiFi
- Understand Concepts of NiFi Cluster
- Explain Data Provenance in NiFi
- Learn how to Secure NiFi
- Learn How to effectively Monitor NiFi
- Learn about HDF Best Practices

### Hands-On Labs

- Manual Installation of NiFi
- Building a WorkFlow
- Working with Processor Groups
- Working with Remote Processor Groups
- Using NiFi Expression Language.
- Understanding and using Templates.
- Installing and Configuring NiFi Cluster
- Securing NiFi
- Monitoring NiFi
- End Of the Course Project.

### Demos

- Getting Familiar to NiFi User Interface
- Anatomy of a Processor
- Anatomy of a Connection
- Data Provenance

### Prerequisites

Students should be familiar with programming principles and have previous experience in software development. Experience with Linux and a basic understanding of DataFlow tools would be helpful. No prior Hadoop experience required, but is very helpful.

### Certification

Hortonworks offers a comprehensive certification program that identifies you as an expert in Apache Hadoop. Visit [hortonworks.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA





## HDP Operations: Security

### Overview

This course is designed for experienced administrators who will be implementing secure Hadoop clusters using authentication, authorization, auditing and data protection strategies and tools.

### Duration

4 days

### Target Audience

IT administrators and operators responsible for installing, configuring and supporting an Apache Hadoop 2.3 deployment in a Linux environment.

### Course Objectives

- Describe the 5 pillars of a secure environment
- List the reasons why a secure environment is needed
- Describe how security is integrated within Hadoop
- Choose which security tool is best for specific use cases
- List security prerequisites
- Configure Ambari security
- Set up Ambari Views for controlled access
- Describe Kerberos use and architecture
- Install Kerberos
- Configure Ambari for Kerberos
- Configure Hadoop for Kerberos
- Enable Kerberos
- Install and configure Apache Knox
- Install and configure Apache Ranger
- Install and configure Ranger Key Management Services
- Use Ranger to assure secure data access
- Describe available partner security solutions

### 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hands-On Labs

- Setting up Active Directory/Operating System Integration
- Configuring Ambari for Non-Root
- Encrypting Ambari Database and Passwords
- Setting up Ambari for LDAP/Active Directory Authentication
- Setting up HTTPS/SSL Server for Ambari
- Setting up Two-Way SSL for Ambari Server and Agents
- Enabling SPNEGO Authentication for Hadoop
- Setting up Standalone Ambari Views Server
- Configuring Ambari Views for Kerberos
- Setting up Primary Kerberos KDC
- Setting up Backup Kerberos KDC
- Configuring One Way Trust Relationship Between Kerberos and Active Directory
- Setting up Ambari for Kerberos
- Setting up/Enabling HDP for Kerberos
- Installing Knox Service via Ambari
- Configuring Knox Gateway
- Configuring Knox to Authenticate via Active Directory/LDAP
- Configuring Knox Topology to Connect to Hadoop Cluster Service
- Installing Ranger Service via Ambari
- Configuring Ranger Repository Manager, Policy Manager, User Groups and Auditing
- Installing Ranger KMS via Ambari
- Configuring HDFS for Encryption
- Configuring Hive to utilize Encrypted HDFS
- Enabling Ranger KMS Audit
- Using Ranger KMS Service
- Testing Secure Access with HDFS, Hive, Pig and Sqoop

### Prerequisites

Students should be experienced in the management of Hadoop using Ambari and Linux environments. Completion of the **Hadoop Administration I** course is highly recommended.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Operations: Migrating to the Hortonworks Data Platform

### Overview

This course is designed for administrators who are familiar with administering other Hadoop distributions and are migrating to the Hortonworks Data Platform (HDP). It covers installation, configuration, maintenance, security and performance topics.

### Course Objectives

- Install and configure an HDP 2.x cluster
- Use Ambari to monitor and manage a cluster
- Mount HDFS to a local filesystem using the NFS Gateway
- Configure Hive for Tez
- Use Ambari to configure the schedulers of the ResourceManager
- Commission and decommission worker nodes using Ambari
- Use Falcon to define and process data pipelines
- Take snapshots using the HDFS snapshot feature
- Implement and configure NameNode HA using Ambari
- Secure an HDP cluster using Ambari
- Setup a Knox gateway

### Hands-On Labs

- Install HDP 2.x using Ambari
- Add a new node to the cluster
- Stop and start HDP services
- Mount HDFS to a local file system
- Configure the capacity scheduler
- Use WebHDFS
- Dataset mirroring using Falcon
- Commission and decommission a worker node using Ambari
- Use HDFS snapshots
- Configure NameNode HA using Ambari
- Secure an HDP cluster using Ambari
- Setting up a Knox gateway

### Duration

2 days

### Target Audience

Experienced Hadoop administrators and operators responsible for installing, configuring and supporting the Hortonworks Data Platform.

### Prerequisites

Attendees should be familiar with Hadoop fundamentals, have experience administering a Hadoop cluster, and installation of configuration of Hadoop components such as Sqoop, Flume, Hive, Pig and Oozie.

### 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.com/training/certification](http://hortonworks.com/training/certification) for more information.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## Hortonworks University Academic Program

### Overview

The Big Data skills gap is real.

Every minute there are more than two million Google searches, roughly 685,000 Facebook updates, 200 million sent emails and 48 hours of video uploaded to YouTube. Companies collect this data about their customers, but many struggle to implement meaningful ways to analyze and process it. And while there are emerging big data solutions and tools to better understand business problems, there are not enough candidates in today's employment pool with appropriate skills to implement them.

- More than 85% of Fortune 500 organizations will be unable to exploit big data analytics as a competitive advantage. (Gartner)
- 46% of companies report inadequate staffing for big data analytics (TDWI Research)
- By 2018 the US could face a shortfall of as many as 1.5 million analysts skilled in big data (McKinsey)

### Academic Partners

Becoming an Academic Partner is easy:

- There is no cost to join
- Student materials are purchased directly from our book vendor
- Instructors may prepare to teach our materials at their own pace using our materials.

### A Win-Win Situation

#### For Students

The Hortonworks University Academic Program enables students to obtain authorized training that will prepare them for certification, bolstering their employment opportunities with firms seeking skilled Hadoop professionals.

Students receive worldwide access to high quality educational content, certification opportunities, and experience with Hortonworks technologies.

#### For Academic Institutions

Hortonworks University partners with accredited colleges and universities to meet those needs.

Academic partners receive support from Hortonworks for the inclusion of Hortonworks technologies in their course catalog.

### Academic Partner Responsibilities

- Each academic institution is responsible for meeting classroom set-up requirements
- Students must be currently enrolled in the college or university
- Instructional hours must spread across an entire semester
- Hortonworks course materials may not be altered, but institutions are free to add supplemental content.



### About Hortonworks

Hortonworks develops, distributes and supports the only 100 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA. 95054 USA



## HDP Certified Administrator (HDPCA)

### Certification Overview

Hortonworks has redesigned its certification program to create an industry-recognized certification where individuals prove their Hadoop knowledge by performing actual hands-on tasks on a Hortonworks Data Platform (HDP) cluster, as opposed to answering multiple-choice questions. The HDP Certified Administrator (HDPCA) exam is designed for Hadoop system administrators and operators responsible for installing, configuring and supporting an HPD cluster.

### Purpose of the Exam

The purpose of this exam is to provide organizations that use Hadoop with a means of identifying suitably qualified staff to install, configure, secure and troubleshoot a Hortonwork Data Platform cluster using Apache Ambari.

### Exam Description

The exam has five main categories of tasks that involve:

- Installation
- Configuration
- Troubleshooting
- High Availability
- Security

The exam is based on the Hortonworks Data Platform 2.3 installed and managed with Ambari 2.1.

### Exam Objectives

View the complete list of objectives below, which includes links to the corresponding documentation and/or other resources.

### Language

The exam is delivered in English.

### Take the Exam Anytime, Anywhere

The HDPCA exam is available from any computer, anywhere, at any time. All you need is a webcam and a good Internet connection.

### How to Register

Candidates need to create an account at [www.examslocal.com](http://www.examslocal.com). Once you are registered and logged in, select "Schedule an Exam", and then enter "Hortonworks" in the "Search Here" field to locate and select the HDP Certified Administrator exam. The cost of the exam is \$250 USD.

### Duration

2 hours

### Description of the Minimally Qualified Candidate

The Minimally Qualified Candidate (MQC) for this certification has hands-on experience installing, configuring, securing and troubleshooting a Hadoop cluster, and can perform the objectives of the HDPCA exam.

### Prerequisites

Candidates for the HPDCA exam should be able to perform each of the tasks in the list of exam objectives below. Candidates are also encouraged to attempt the practice exam. Visit [www.hortonworks.com/training/class/hdp-certified-administrator-hdpca-exam/](http://www.hortonworks.com/training/class/hdp-certified-administrator-hdpca-exam/) for more details.

### Hortonworks University

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 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA

## HDP Certified Administrator (HDP-CA) Exam Objectives

Candidates for the HPDCA exam should be able to perform each of the tasks below:

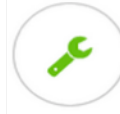
Category	Task/Resources
Installation	Configure a local HDP repository <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.1.0.0/bk_Installing_HDP_AMB/content/_using_a_local_repository.html">http://docs.hortonworks.com/HDPDocuments/Ambari-2.1.0.0/bk_Installing_HDP_AMB/content/_using_a_local_repository.html</a>
	Install ambari-server and ambari-agent <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-f123c19f-f2b8-429f-bdaf-3535df363080">http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-f123c19f-f2b8-429f-bdaf-3535df363080</a> <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-848b09cc-db45-4a2c-b8f7-3617987c53f2">http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-848b09cc-db45-4a2c-b8f7-3617987c53f2</a>
	Install HDP using the Ambari install wizard <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.1.0.0/bk_Installing_HDP_AMB/content/ch_Deploy_and_Configure_a_HDP_Cluster.html">http://docs.hortonworks.com/HDPDocuments/Ambari-2.1.0.0/bk_Installing_HDP_AMB/content/ch_Deploy_and_Configure_a_HDP_Cluster.html</a>
	Add a new node to an existing cluster <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-d745870f-2b0a-47ad-9307-8c01b440589b">http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-d745870f-2b0a-47ad-9307-8c01b440589b</a>
	Decommission a node <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Sys_Admin_Guides/content/ch_slave_nodes.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Sys_Admin_Guides/content/ch_slave_nodes.html</a>
	Add an HDP service to a cluster using Ambari <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-556d8737-67b1-43da-8331-bccb6ff28ac6">http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-556d8737-67b1-43da-8331-bccb6ff28ac6</a>
	Define and deploy a rack topology script <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hdfs_admin_tools/content/ch05.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hdfs_admin_tools/content/ch05.html</a>
Configuration	Change the configuration of a service using Ambari <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.2.4/index.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.2.4/index.html</a>
	Configure the Capacity Scheduler <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_yarn_resource_mgt/content/ch_capacity_scheduler.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_yarn_resource_mgt/content/ch_capacity_scheduler.html</a>
	Configure the location of log files for services <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-a3d954f2-00e6-4f5c-8a3e-f7fca7e566cc">http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-a3d954f2-00e6-4f5c-8a3e-f7fca7e566cc</a>
	Create a home directory for a user and configure permissions <a href="http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html">http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html</a>
	Configure the include and exclude DataNode files <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Sys_Admin_Guides/content/ch_slave_nodes.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Sys_Admin_Guides/content/ch_slave_nodes.html</a>

### About Hortonworks

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



Hortonworks®  
**UNIVERSITY**



**System  
Admin**



**Data  
Analyst**



**Developer**

<b>Troubleshooting</b>	Restart an HDP service <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-fc5734c8-1957-4406-8fe8-71a1de0193c1">http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-fc5734c8-1957-4406-8fe8-71a1de0193c1</a>
	View an application's log file <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.1.5/bk_using-apache-hadoop/content/log_files.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.1.5/bk_using-apache-hadoop/content/log_files.html</a>  <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_yarn_resource_mgt/content/ch_log_aggregation.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_yarn_resource_mgt/content/ch_log_aggregation.html</a>
	Configure and manage alerts <a href="http://hortonworks.com/blog/announcing-apache-ambari-2-0/#alerts">http://hortonworks.com/blog/announcing-apache-ambari-2-0/#alerts</a>
	Troubleshoot a failed job <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.1.5/bk_using-apache-hadoop/content/mrv2_troubleshooting.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.1.5/bk_using-apache-hadoop/content/mrv2_troubleshooting.html</a>
<b>High Availability</b>	Configure NameNode HA <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hadoop-ha/content/ch_HA-NameNode.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hadoop-ha/content/ch_HA-NameNode.html</a>
	Configure ResourceManager HA <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hadoop-ha/content/ch_HA-ResourceManager.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hadoop-ha/content/ch_HA-ResourceManager.html</a>
	Copy data between two clusters using distcp <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Sys_Admin_Guides/content/ch_distcp.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Sys_Admin_Guides/content/ch_distcp.html</a>
	Create a snapshot of an HDFS directory <a href="http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsSnapshots.html">http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsSnapshots.html</a>
	Recover a snapshot <a href="http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsSnapshots.html">http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsSnapshots.html</a>
	Configure HiveServer2 HA <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hadoop-ha/content/ch_HA-HiveServer2.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hadoop-ha/content/ch_HA-HiveServer2.html</a>
<b>Security</b>	Install and configure Knox <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Knox_Gateway_Admin_Guide/content/ch01.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Knox_Gateway_Admin_Guide/content/ch01.html</a>
	Install and configure Ranger <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Ranger_Install_Guide/content/ch_overview_ranger_ambari_install.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_Ranger_Install_Guide/content/ch_overview_ranger_ambari_install.html</a>
	Configure HDFS ACLs <a href="http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hdfs_admin_tools/content/ch01.html">http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.0/bk_hdfs_admin_tools/content/ch01.html</a>
	Configure Hadoop for Kerberos <a href="http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-1097012d-c999-4a81-960b-7473ed584db3">http://docs.hortonworks.com/HDPDocuments/Ambari-2.0.0.0/Ambari_Doc_Suite/ADS_v200.html#ref-1097012d-c999-4a81-960b-7473ed584db3</a>



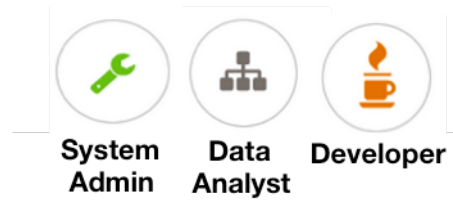
#### About Hortonworks

Hortonworks develops, distributes and supports the only 100 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



## HDP Certified Developer (HDP CD) Exam

### Certification Overview

Hortonworks has redesigned its certification program to create an industry-recognized certification where individuals prove their Hadoop knowledge by performing actual hands-on tasks on a Hortonworks Data Platform (HDP) cluster, as opposed to answering multiple-choice questions. The HDP Certified Developer (HDP CD) exam is the first of our new hands-on, performance-based exams designed for Hadoop developers working with frameworks like Pig, Hive, Sqoop and Flume.

### Purpose of the Exam

The purpose of this exam is to provide organizations that use Hadoop with a means of identifying suitably qualified staff to develop Hadoop applications for storing, processing, and analyzing data stored in Hadoop using the open-source tools of the Hortonworks Data Platform (HDP), including Pig, Hive, Sqoop and Flume.

### Exam Description

The exam has three main categories of tasks that involve:

- Data ingestion
- Data transformation
- Data analysis

The exam is based on the Hortonworks Data Platform 2.2 installed and managed with Ambari 1.7.0, which includes Pig 0.14.0, Hive 0.14.0, Sqoop 1.4.5, and Flume 1.5.0. Each candidate will be given access to an HDP 2.2 cluster along with a list of tasks to be performed on that cluster.

### Exam Objectives

View the complete list of objectives below, which includes links to the corresponding documentation and/or other resources.

### Duration

2 hours

### Description of the Minimally Qualified Candidate

The Minimally Qualified Candidate (MQC) for this certification can develop Hadoop applications for ingesting, transforming, and analyzing data stored in Hadoop using the open-source tools of the Hortonworks Data Platform, including Pig, Hive, Sqoop and Flume.

### Prerequisites

Candidates for the HPDCD exam should be able to perform each of the tasks in the list of exam objectives below.

### Language

The exam is delivered in English

### Hortonworks University

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.



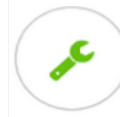
#### About Hortonworks

Hortonworks develops, distributes and supports the only 100 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



**System  
Admin**



**Data  
Analyst**



**Developer**

## HDP Certified Developer (HDPD) Exam Objectives

Candidates for the HPDCE exam should be able to perform each of the tasks below:

Category	Task/Resources
<b>Data Ingestion</b>	Input a local file into HDFS using the Hadoop file system shell <a href="http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html#put">http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html#put</a>
	Make a new directory in HDFS using the Hadoop file system shell <a href="http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html#mkdir">http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html#mkdir</a>
	Import data from a table in a relational database into HDFS <a href="http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html#_literal_sqoop_import_literal">http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html#_literal_sqoop_import_literal</a>
	Import the results of a query from a relational database into HDFS <a href="http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html#_free_form_query_imports">http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html#_free_form_query_imports</a>
	Import a table from a relational database into a new or existing Hive table <a href="http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html#_importing_data_into_hive">http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html#_importing_data_into_hive</a>
	Insert or update data from HDFS into a table in a relational database <a href="http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html#_literal_sqoop_export_literal">http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html#_literal_sqoop_export_literal</a>
	Given a Flume configuration file, start a Flume agent <a href="https://flume.apache.org/FlumeUserGuide.html#starting-an-agent">https://flume.apache.org/FlumeUserGuide.html#starting-an-agent</a>
	Given a configured sink and source, configure a Flume memory channel with a specified capacity <a href="https://flume.apache.org/FlumeUserGuide.html#memory-channel">https://flume.apache.org/FlumeUserGuide.html#memory-channel</a>

<b>Data Transformation</b>	Write and execute a Pig script <a href="https://pig.apache.org/docs/r0.14.0/start.html#run">https://pig.apache.org/docs/r0.14.0/start.html#run</a>
	Load data into a Pig relation without a schema <a href="https://pig.apache.org/docs/r0.14.0/basic.html#load">https://pig.apache.org/docs/r0.14.0/basic.html#load</a>
	Load data into a Pig relation with a schema <a href="https://pig.apache.org/docs/r0.14.0/basic.html#load">https://pig.apache.org/docs/r0.14.0/basic.html#load</a>
	Load data from a Hive table into a Pig relation <a href="https://cwiki.apache.org/confluence/display/Hive/HCatalog+LoadStore">https://cwiki.apache.org/confluence/display/Hive/HCatalog+LoadStore</a>
	Use Pig to transform data into a specified format <a href="https://pig.apache.org/docs/r0.14.0/basic.html#foreach">https://pig.apache.org/docs/r0.14.0/basic.html#foreach</a>
	Transform data to match a given Hive schema <a href="https://pig.apache.org/docs/r0.14.0/basic.html#foreach">https://pig.apache.org/docs/r0.14.0/basic.html#foreach</a>
	Group the data of one or more Pig relations <a href="https://pig.apache.org/docs/r0.14.0/basic.html#group">https://pig.apache.org/docs/r0.14.0/basic.html#group</a>
	Use Pig to remove records with null values from a relation <a href="https://pig.apache.org/docs/r0.14.0/basic.html#filter">https://pig.apache.org/docs/r0.14.0/basic.html#filter</a>
	Store the data from a Pig relation into a folder in HDFS <a href="https://pig.apache.org/docs/r0.14.0/basic.html#store">https://pig.apache.org/docs/r0.14.0/basic.html#store</a>
	Store the data from a Pig relation into a Hive table <a href="https://cwiki.apache.org/confluence/display/Hive/HCatalog+LoadStore">https://cwiki.apache.org/confluence/display/Hive/HCatalog+LoadStore</a>

### About Hortonworks

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

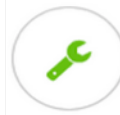
**US:** 1.855.846.7866

**International:** +1.408.916.4121  
[www.hortonworks.com](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA



Hortonworks®  
**UNIVERSITY**



**System  
Admin**



**Data  
Analyst**



**Developer**

	Sort the output of a Pig relation <a href="https://pig.apache.org/docs/r0.14.0/basic.html#order-by">https://pig.apache.org/docs/r0.14.0/basic.html#order-by</a>
	Remove the duplicate tuples of a Pig relation <a href="https://pig.apache.org/docs/r0.14.0/basic.html#distinct">https://pig.apache.org/docs/r0.14.0/basic.html#distinct</a>
	Specify the number of reduce tasks for a Pig MapReduce job <a href="https://pig.apache.org/docs/r0.14.0/perf.html#parallel">https://pig.apache.org/docs/r0.14.0/perf.html#parallel</a>
	Join two datasets using Pig <a href="https://pig.apache.org/docs/r0.14.0/basic.html#join-inner">https://pig.apache.org/docs/r0.14.0/basic.html#join-inner</a> <a href="https://pig.apache.org/docs/r0.14.0/basic.html#join-outer">https://pig.apache.org/docs/r0.14.0/basic.html#join-outer</a>
	Perform a replicated join using Pig <a href="https://pig.apache.org/docs/r0.14.0/perf.html#replicated-joins">https://pig.apache.org/docs/r0.14.0/perf.html#replicated-joins</a>
	Run a Pig job using Tez <a href="https://pig.apache.org/docs/r0.14.0/perf.html#tez-mode">https://pig.apache.org/docs/r0.14.0/perf.html#tez-mode</a>
	Within a Pig script, register a JAR file of User Defined Functions <a href="https://pig.apache.org/docs/r0.14.0/basic.html#register">https://pig.apache.org/docs/r0.14.0/basic.html#register</a> <a href="https://pig.apache.org/docs/r0.14.0/udf.html#piggybank">https://pig.apache.org/docs/r0.14.0/udf.html#piggybank</a>
	Within a Pig script, define an alias for a User Defined Function <a href="https://pig.apache.org/docs/r0.14.0/basic.html#define-udfs">https://pig.apache.org/docs/r0.14.0/basic.html#define-udfs</a>
	Within a Pig script, invoke a User Defined Function <a href="https://pig.apache.org/docs/r0.14.0/basic.html#register">https://pig.apache.org/docs/r0.14.0/basic.html#register</a>

<b>Data Analysis</b>	Write and execute a Hive query <a href="https://cwiki.apache.org/confluence/display/Hive/Tutorial">https://cwiki.apache.org/confluence/display/Hive/Tutorial</a>
	Define a Hive-managed table <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-Create/Drop/TruncateTable">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-Create/Drop/TruncateTable</a>
	Define a Hive external table <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-ExternalTables">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-ExternalTables</a>
	Define a partitioned Hive table <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-PartitionedTables">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-PartitionedTables</a>
	Define a bucketed Hive table <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-BucketedSortedTables">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-BucketedSortedTables</a>
	Define a Hive table from a select query <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-CreateTableAsSelect(CTAS)">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-CreateTableAsSelect(CTAS)</a>
	Define a Hive table that uses the ORCFile format <a href="http://hortonworks.com/blog/orcfile-in-hdp-2-better-compression-better-performance/">http://hortonworks.com/blog/orcfile-in-hdp-2-better-compression-better-performance/</a>
	Create a new ORCFile table from the data in an existing non-ORCFile Hive table <a href="http://hortonworks.com/blog/orcfile-in-hdp-2-better-compression-better-performance/">http://hortonworks.com/blog/orcfile-in-hdp-2-better-compression-better-performance/</a>



#### About Hortonworks

Hortonworks develops, distributes and supports the only 100 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA





Hortonworks®  
**UNIVERSITY**



**System  
Admin**



**Data  
Analyst**



**Developer**

	Specify the storage format of a Hive table <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-RowFormat,StorageFormat,andSerDe">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-RowFormat,StorageFormat,andSerDe</a>
	Specify the delimiter of a Hive table <a href="http://hortonworks.com/hadoop-tutorial/using-hive-data-analysis/">http://hortonworks.com/hadoop-tutorial/using-hive-data-analysis/</a>
	Load data into a Hive table from a local directory <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-Loadingfilesintotables">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-Loadingfilesintotables</a>
	Load data into a Hive table from an HDFS directory <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-Loadingfilesintotables">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-Loadingfilesintotables</a>
	Load data into a Hive table as the result of a query <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-InsertingdataintoHiveTablesfromqueries">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-InsertingdataintoHiveTablesfromqueries</a>
	Load a compressed data file into a Hive table <a href="https://cwiki.apache.org/confluence/display/Hive/CompressedStorage">https://cwiki.apache.org/confluence/display/Hive/CompressedStorage</a>
	Update a row in a Hive table <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-Update">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-Update</a>
	Delete a row from a Hive table <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-Delete">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-Delete</a>
	Insert a new row into a Hive table <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-InsertingvaluesintotablesfromSQL">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML#LanguageManualDML-InsertingvaluesintotablesfromSQL</a>
	Join two Hive tables <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Joins">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Joins</a>
	Run a Hive query using Tez
	Run a Hive query using vectorization <a href="http://hortonworks.com/hadoop-tutorial/supercharging-interactive-queries-hive-tez/">http://hortonworks.com/hadoop-tutorial/supercharging-interactive-queries-hive-tez/</a>
	Output the execution plan for a Hive query <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Explain">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Explain</a>
	Use a subquery within a Hive query <a href="https://cwiki.apache.org/confluence/display/Hive/LanguageManual+SubQueries">https://cwiki.apache.org/confluence/display/Hive/LanguageManual+SubQueries</a>
	Output data from a Hive query that is totally ordered across multiple reducers <a href="https://issues.apache.org/jira/browse/HIVE-1402">https://issues.apache.org/jira/browse/HIVE-1402</a>
	Set a Hadoop or Hive configuration property from within a Hive query <a href="http://hortonworks.com/wp-content/uploads/downloads/2013/08/Hortonworks.CheatSheet.SQLtoHive.pdf">http://hortonworks.com/wp-content/uploads/downloads/2013/08/Hortonworks.CheatSheet.SQLtoHive.pdf</a>



#### About Hortonworks

Hortonworks develops, distributes and supports the only 100 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA

## HDP Certified Java Developer (HDPCD:Java)

### Certification Overview

Hortonworks has redesigned its certification program to create an industry-recognized certification where individuals prove their Hadoop knowledge by performing actual hands-on tasks on a Hortonworks Data Platform (HDP) cluster, as opposed to answering multiple-choice questions. The HDP Certified Java Developer (HDPCD:Java) exam is designed for Hadoop developers who write Java MapReduce applications.

### Purpose of the Exam

The purpose of this exam is to provide organizations that use Hadoop with a means of identifying suitably qualified staff to develop custom Hadoop MapReduce applications in Java for the Hortonworks Data Platform (HDP).

### Exam Description

This exam consists of tasks associated with writing Java MapReduce jobs, including the development and configuring of combiners, partitions, custom keys, custom sorting, and the joining of datasets. The exam is based on the Hortonworks Data Platform 2.2 and candidates are provided with an Eclipse environment that is pre-configured and ready for the writing of Java classes.

### Exam Objectives

View the complete list of objectives below, which includes links to the corresponding documentation and/or other resources.

### Language

The exam is delivered in English.

### Take the Exam Anytime, Anywhere

The HDPCD:Java exam is available from any computer, anywhere, at any time. All you need is a webcam and a good Internet connection.

### How to Register

Candidates need to create an account at [www.examslocal.com](http://www.examslocal.com). Once you are registered and logged in, select "Schedule an Exam", and then enter "Hortonworks" in the "Search Here" field to locate and select the HDP Certified Developer:Java exam. The cost of the exam is \$250 USD.

### Duration

2 hours

### Description of the Minimally Qualified Candidate

The Minimally Qualified Candidate (MQC) for this certification has hands-on experience write Java MapReduce applications for Hadoop and can perform the objectives of the HDPCD:Java exam.

### Prerequisites

Candidates for the HDPCD:Java exam should be able to perform each of the tasks in the list of exam objectives below. Candidates are also encouraged to attempt the practice exam. Visit <http://hortonworks.com/training/class/hdp-certified-java-developer-exam/> for more details.

### Hortonworks University

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

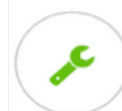
Hortonworks develops, distributes and supports the only 100 percent open source distribution of Apache Hadoop explicitly architected, built and tested for enterprise-grade deployments.

**US:** 1.855.846.7866

**International:** +1.408.916.4121  
[www.hortonworks.com](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA





**System  
Admin**



**Data  
Analyst**



**Developer**

## HDP Certified Java Developer (HDPJD:Java) Exam Objectives

Candidates for the HDPJD:Java exam should be able to perform each of the tasks below:

Task	Resource(s)
Write a Hadoop MapReduce application in Java	<a href="http://hadoop.apache.org/docs/r2.6.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html">http://hadoop.apache.org/docs/r2.6.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html</a>
Run a Java MapReduce application on a Hadoop cluster	<a href="http://hadoop.apache.org/docs/r2.6.0/hadoop-yarn/hadoop-yarn-site/YarnCommands.html#jar">http://hadoop.apache.org/docs/r2.6.0/hadoop-yarn/hadoop-yarn-site/YarnCommands.html#jar</a>
Write and configure a Combiner for a MapReduce job	<a href="https://developer.yahoo.com/hadoop/tutorial/module4.html#functionality">https://developer.yahoo.com/hadoop/tutorial/module4.html#functionality</a>
Write and configure a custom Partitioner for a MapReduce job	<a href="https://developer.yahoo.com/hadoop/tutorial/module5.html#partitioning">https://developer.yahoo.com/hadoop/tutorial/module5.html#partitioning</a>
Sort the output of a MapReduce job	<a href="http://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/mapreduce/Job.html#setGroupingComparatorClass(java.lang.Class)">http://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/mapreduce/Job.html#setGroupingComparatorClass(java.lang.Class)</a>
Write and configure a custom key class for a MapReduce job	<a href="http://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/io/WritableComparable.html">http://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/io/WritableComparable.html</a>
Configure the input and output formats of a MapReduce job	<a href="http://hadoop.apache.org/docs/r2.6.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Job_Input">http://hadoop.apache.org/docs/r2.6.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Job_Input</a>
Use a LocalResource instance to distribute files and resources for a MapReduce job	<a href="http://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/yarn/api/records/LocalResource.html">http://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/yarn/api/records/LocalResource.html</a>
Perform a join of two or more datasets	Many online examples and resources
Perform a map-side join of two datasets	Many online examples and resources



### About Hortonworks

Hortonworks develops, distributes and supports the only 100 percent 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](http://www.hortonworks.com)

5470 Great America Parkway  
Santa Clara, CA 95054 USA





---

Our courses are designed by the leaders and committers of Hadoop  
Hortonworks provides an immersive and valuable real world experience  
In scenario-based training Courses  
Offer unmatched depth and expertise  
Available both in classroom or online from anywhere in the world  
We prepare you to be an expert with highly valued, fresh skills  
And for Certification

---

**Visit us online:**  
[training.hortonworks.com](http://training.hortonworks.com)