



Hortonworks
UNIVERSITY

Training Catalog

Apache Hadoop Training from the Experts



September 2015





Hortonworks University

Hortonworks University provides an immersive and valuable real world experience In scenario-based training Courses

- Public, private on site and virtual led courses
- Self paced learning library
- Global delivery capability
- Performance based certifications
- Academic program
- Industry-leading lecture and hands-on labs
- High definition instructor led training
- Individualized learning paths

Learning Paths

		
I'm a Developer Architecture & Fundamentals MapReduce and YARN Programming	I'm a System Admin Cluster Management Installation and Configuration Security and Governance Certification	I'm a Data Analyst Real-Time Data Analysis NoSQL & Scripting Languages Sparse Data Sets/Big Data
Courses: HDP Developer: Java 4 Days HDP Developer: Apache Pig & Hive 4 Days HDP Developer: Windows 4 Days HDP Developer: Storm & Trident Fundamentals 2 Days HDP Developer: Custom YARN Applications 2 Days	Courses: HDP Operations: Hadoop Administration 1 4 Days HDP Operations: Migrating to the Hortonworks Data Platform 4 Days HDP Operations: APACHE HBase Advanced Management 4 Days	Courses: HDP Analyst: Data Science 3 Days HDP Analyst: 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 Developer:

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

Hortonworks Certified Administrator:

- HDP Certified Administrator (HDPCA)

Table of Contents

Hortonworks University Self-Paced Learning Library	7
HDP Overview: Essentials	8
HDP Analyst: Data Science	9
HDP Analyst: HBase Essentials	10
HDP Developer: Java	11
HDP Developer: Apache Pig and Hive	12
HDP Developer: Windows	13
HDP Developer: YARN	14
HDP Developer: Storm and Trident Fundamentals	15
HDP Operations: Hadoop Administration 1	16
HDP Operations: Migrating to the Hortonworks Data Platform HDP	17
Operations: Apache HBase Advanced Management	18
HDP Certified Administrator (HDPCA)	19
HDP Certified Developer (HDPCD)	20
Hortonworks University Academic Program	21
HDP Academic Analyst: Data Science	22
HDP Academic Developer: Apache Pig and Hive	23
HDP Academic Operations: Install and Manage with Apache Ambari	24



Hortonworks University Self-Paced Learning Library

Overview

Hortonworks University “Self-Paced” Learning Library is an on demand-learning library 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.

Target Audience

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

Prerequisites: None.

Self-Paced Learning Content includes:

- HDP Overview: Essentials
- HDP Developer: Apache Pig & Hive
- HDP Developer: Java
- HDP Developer: Windows
- HDP Developer: Developing Custom YARN Applications
- HDP Operations: Install and Manage with Apache Ambari
- HDP Operations: Migrating to the Hortonworks Data Platform
- HDP Analyst: Data Science
- HDP Analyst: HBase Essentials

Duration

Access to the Hortonworks University Self Paced Learning Library is provided for a 12-month period per individual named user. The subscription includes access to over 200 hours of learning lessons.

Access the Self Paced Learning Library today

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.

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 please contact

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

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 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

5470 Great America Parkway
Santa Clara, CA 95054 USA

HDP Analyst: Data Science

Overview

This course is designed for students preparing to become familiar with 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.

Target Audience

Computer science and data analytics students who need to apply data science and machine learning on Hadoop.

Course Objectives

- Recognize use cases for data science
- Describe the architecture of Hadoop and YARN
- Describe supervised and unsupervised learning differences
- List the six machine learning tasks
- Use Mahout to run a machine learning algorithm on Hadoop
- Use Pig to transform and prepare data on Hadoop
- Write a Python script
- Use NumPy to analyze big data
- Use the data structure classes in the pandas library
- Write a Python script that invokes SciPy machine learning
- 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
- Write a Python script that invokes scikit-learn
- Use the k-nearest neighbor algorithm to predict values
- Run a machine learning algorithm on a distributed data set
- Describe use cases for Natural Language Processing (NLP)
- Perform sentence segmentation on a large body of text
- Perform part-of-speech tagging
- 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

Hands-On Labs

- Setting Up a Development Environment
- Using HDFS Commands
- Using Mahout for Machine Learning
- Getting Started with Pig
- Exploring Data with Pig
- Using the IPython Notebook
- Data Analysis with Python
- Interpolating Data Points
- Define a Pig UDF in Python
- Streaming Python with Pig
- K-Nearest Neighbor and K-Means Clustering
- K-Means Clustering
- Using NLTK for Natural Language Processing
- Classifying Text using Naive Bayes
- Spark Programming and Spark MLlib
- Running Data Science Algorithms using Spark MLlib

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.

Certification

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

Hortonworks University

Hortonworks University is your expert source for Apache Hadoop training and certification. Courses are available for developers, data analysts and administrators. Classes combine presentation material with industry-leading hands-on labs that fully prepare students for real-world Hadoop scenarios.

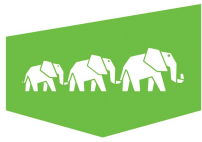
About Hortonworks

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

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 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

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 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

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

- Use HDFS commands to add/remove files and folders
- Use Sqoop to transfer data between HDFS and a RDBMS
- Run MapReduce and YARN application jobs
- Explore 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
- Define 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 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

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 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

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 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.

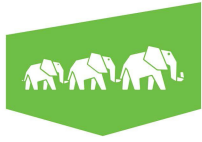
About Hortonworks

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

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

Data architects, data integration architects, technical infrastructure team, and Hadoop administrators or developers who want to understand the fundamentals of Storm and Trident.

Prerequisites

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

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 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 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 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

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 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

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 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

5470 Great American Parkway
Santa Clara, CA 95054 USA



Developer



System
Admin



Data Analyst

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 HDP 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 Hortonworks 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 2.2 installed and managed with Ambari 2.0.0.

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. 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 HDPCA 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/ 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

5470 Great America Parkway
Santa Clara, CA 95054 USA

HDP Certified Developer (HDPD) 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 (HDPD) 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 HPDCE 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.

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 in 2015 (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.

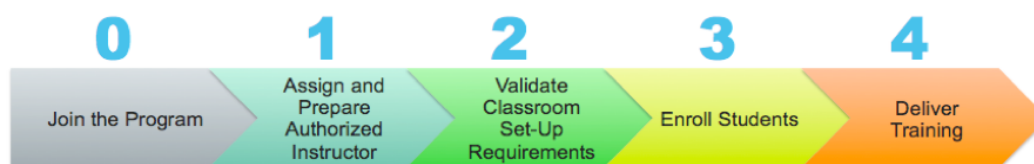
For Academic Institutions

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.



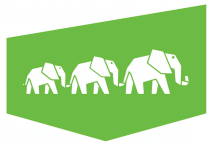
About Hortonworks

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

5470 Great America Parkway
Santa Clara, CA 95054 USA



HDP Academic Analyst: Data Science

Overview

This course is designed for students preparing to become familiar with 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.

Target Audience

Computer science and data analytics students who need to apply data science and machine learning on Hadoop.

Course Objectives

- Recognize use cases for data science
- Describe the architecture of Hadoop and YARN
- Describe supervised and unsupervised learning differences
- List the six machine learning tasks
- Use Mahout to run a machine learning algorithm on Hadoop
- Use Pig to transform and prepare data on Hadoop
- Write a Python script
- Use NumPy to analyze big data
- Use the data structure classes in the pandas library
- Write a Python script that invokes SciPy machine learning
- 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
- Write a Python script that invokes scikit-learn
- Use the k-nearest neighbor algorithm to predict values
- Run a machine learning algorithm on a distributed data set
- Describe use cases for Natural Language Processing (NLP)
- Perform sentence segmentation on a large body of text
- Perform part-of-speech tagging
- 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

Hands-On Labs

- Setting Up a Development Environment
- Using HDFS Commands
- Using Mahout for Machine Learning
- Getting Started with Pig
- Exploring Data with Pig
- Using the IPython Notebook
- Data Analysis with Python
- Interpolating Data Points
- Define a Pig UDF in Python
- Streaming Python with Pig
- K-Nearest Neighbor and K-Means Clustering
- K-Means Clustering
- Using NLTK for Natural Language Processing
- Classifying Text using Naive Bayes
- Spark Programming and Spark MLlib
- Running Data Science Algorithms using Spark MLlib

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.

Certification

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

Hortonworks University

Hortonworks University is your expert source for Apache Hadoop training and certification. Courses are available for developers, data analysts and administrators. 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

5470 Great America Parkway
Santa Clara, CA 95054 USA



HDP Academic Developer: Apache Pig and Hive

Overview

This course is designed for students preparing to become familiar with Big Data application development 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.

Target Audience

Computer Science students 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
- 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

- Use HDFS commands to add/remove files and folders
- Use Sqoop to transfer data between HDFS and a RDBMS
- Run MapReduce and YARN application jobs
- Explore 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
- Define 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.

Certification

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

Hortonworks University

Hortonworks University is your expert source for Apache Hadoop training and certification. Courses are available for developers, data analysts and administrators. 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

5470 Great America Parkway
Santa Clara, CA 95054 USA

HDP Academic Operations: Install and Manage with Apache Ambari

Overview

This course is designed for students preparing to become administrators for the Hortonworks Data Platform (HDP). It covers installation, configuration, maintenance, security and performance topics.

Target Audience

Computer science students who need to learn about installing, configuring and supporting an Apache Hadoop 2.0 deployment in a Linux environment.

Course Objectives

- Describe tools and frameworks in the Hadoop ecosystem
- Describe the Hadoop Distributed File System (HDFS)
- Install and configure an HDP cluster
- Ensure data integrity
- Deploy and configure YARN on a cluster
- Schedule YARN jobs
- Configure and troubleshoot MapReduce jobs
- Describe enterprise data movement
- Use HDFS web services
- Configure a Hiveserver
- Transfer data with Sqoop
- Transfer data with Flume
- Data processing and management with Falcon
- Monitor HDP2 services
- Perform backups and recovery
- Providing high availability through rack awareness
- Providing high availability of through NameNode HA
- Use Apache Knox gateway as a single authentication point
- Describe security requirements in an HDP cluster
- Commission and decommission nodes

Hands-On Labs

- Setting up the HDP environment
- Installing an HDP cluster
- Adding a new host to the an HDP cluster
- Managing HDP services
- Using HDFS commands
- Demo: Understanding block storage
- Verifying data with block scanner and fsck
- Using WebHDFS
- Demo: Understanding MapReduce
- Using Hive tables
- Using Sqoop
- Installing and testing Flume
- Running an Oozie Workflow
- Defining and processing the data pipeline with Falcon
- Using HDFS Snapshot
- Configuring rack awareness
- Implementing NameNode HA
- Securing an HDP Cluster
- Setting up a Knox gateway

Prerequisites

Attendees should be familiar with Hadoop and Linux environments.

Certification

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

Hortonworks University

Hortonworks University is your expert source for Apache Hadoop training and certification. Courses are available for developers, data analysts and administrators. Classes combine presentation material with industry-leading hands-on labs that fully prepare students for real-world Hadoop scenarios.

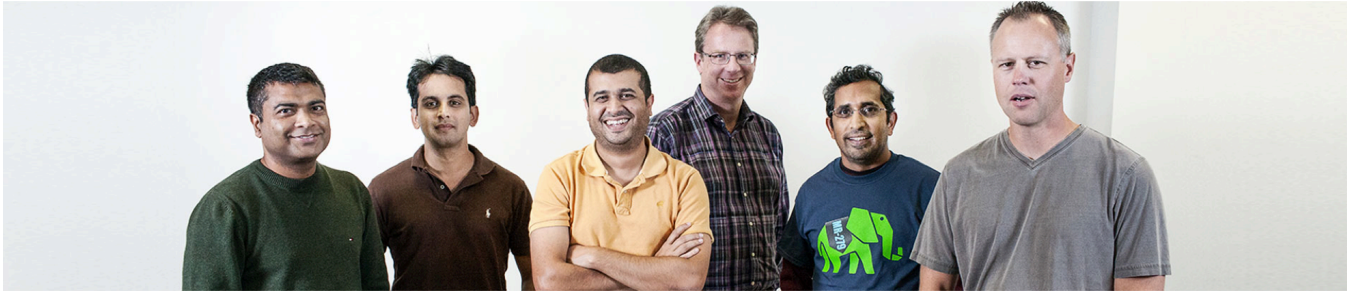
About Hortonworks

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

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