Hewlett Packard Enterprise

HPE Digital Learner Hadoop EcoSystem Content Pack

HPE Content Pack number	CP010
Content Pack length	22 Hours
Content Pack category	Category 2
Learn more	**View now**

Why HPE Education Services?

- IDC MarketScape leader 5 years running for IT education and training*
- Recognized by IDC for leading with global coverage, unmatched technical expertise, and targeted education consulting services*
- Key partnerships with industry leaders OpenStack®, VMware®, Linux®, Microsoft®, ITIL, PMI, CSA, and SUSE
- Complete continuum of training delivery options—self-paced eLearning, custom education consulting, traditional classroom, video on-demand instruction, live virtual instructor-led with hands-on lab, dedicated onsite training
- Simplified purchase option with HPE Training Credits

*Realize Technology Value with Training, IDC Infographic 2037, Sponsored by HPE, October 2017

Hadoop is an open-source, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It facilitates running applications on large hardware clusters and incorporates features similar to those of the MapReduce computing paradigm.

This 22-hour collection of self-paced eLearning content provides both a broad and deep overview of Hadoop. Topics explored include reviewing the full Hadoop eco-system while also discussing key data management techniques and data flow concepts. The use of data repositories, data factories and data refineries in the context of Hadoop is explored, as is the critical role Hadoop and its eco-system can play in driving better enterprise analytics and business success.

Audience

- Anyone interested in learning more about Hadoop, its usage and core components, and how Hadoop can be used/tuned for enterprise business success.

Content Pack objectives

This Content Pack consists of 6 courses, which are:

- Introduces users to Hadoop, its core open source components and general usage/ configurations
- Offers a primer in Hadoop-related data management and modeling concepts
- Provides basic instruction regarding the "Cloudcloud-based" use of Apache Hadoop
- Explores in greater depth the entire Hadoop eco-system – including but not limited to HFDS, Sqoop, Hive, Flume, YARN, Pig and Oozie
- Examines the server architecture for Hadoop and the Hadoop Distributed File System (HDFS) – including daemon functions and configurations
- Demonstrates how Sqoop and Hive can be used with Hadoop to flow and fuse data – including data pre-processing, data partitioning, and data joining
- Details how Sqoop is used to extract and load structured data
- Details how Flume is used to extract and load unstructured data
- Shows how YARN is used as a parallel processing framework for Hadoop
- Demonstrates the use of Hive as an SQL-like tool for interfacing with Hadoop
- Explains how Pig is used as a data flow scripting tool with Hadoop
- Describes how Oozie is used as a workflow tool to manage multiple stage tasks in Hadoop

Detailed Content Pack outline

Introduction to Hadoop

Hadoop is an open-source, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. This course will introduce Hadoop and its key tools and their applications.

Outline

- Start the course
- Recognize what big data is, sources and types of data, evolution and characteristics of big data, and use cases of big data
- Identify big data infrastructure issues, and explain benefits of Hadoop
- Recognize basics of Hadoop, history, milestones, and core components
- Set up a virtual machine
- Install Linux on a virtual machine
- Recognize basic and most useful UNIX commands
- Identify Hadoop components
- Define HDFS components
- Recognize how to read and write in HDFS
- Use HDFS
- Recognize basics of YARN
- Define basics of MapReduce
- Identify how MapReduce processes information
- Use code that runs on Hadoop
- Define Pig, HIVE, and HBase
- Define Sqoop, Flume, Mahout, and Oozie
- Recognize storing and modeling data in Hadoop
- Identify available commercial distributions for Hadoop
- Recognize Spark and its benefits over traditional MapReduce
- Filter information in Hadoop

Introduction to Data Modeling in Hadoop

This course covers various data genres and management tools, the reasons behind the evolving plethora of new big data platforms from the perspective of big data management systems, and analytical tools.

Outline

- Start the course
- Define data management
- Recognize important data modeling concepts in Hadoop
- Identify important issues for storing data in Hadoop
- Recognize important considerations when designing HDFS schema
- Recognize important points when designing HDFS schema
- Identify basic concepts of data movement in Hadoop
- List important factors that need to be considered for importing data into Hadoop
- Identify tools and methods for moving data into Hadoop
- Recognize characteristics of a data stream
- Define how data lakes enable batch processing
- Define data security management and its major domains
- Define Kerberos
- Define basics of authentication in Hadoop using Kerberos
- Identify central issues in processing and management of big data
- Identify important points in Hadoop data modeling

Apache Hadoop

Apache Hadoop is a set of algorithms for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. This course will introduce the basic concepts of cloud computing using Apache Hadoop, cloud computing, big data, and the development tools applied

Outline

- Start the course
- Describe the basics of Hadoop
- Identify the major users of Hadoop, the end-user application, and the result
- Identify the characteristics of big data
- Compare and contrast the traditional data sources and big data sources
- Describe the clustering and distributed computing concepts of Hadoop
- Specify low cost commodity servers in big data and its configurations as nodes in small and large scale Hadoop installations
- Describe Hadoop installation requirements
- Troubleshoot Hadoop installation issues
- Configure Hadoop installation
- Identify the features of third party Hadoop distributions
- Describe the creation and evolution of Hadoop and its related projects
- Describe the use of YARN in Hadoop cluster management
- Describe the components and functions of Hadoop
- Compare and contrast the different types of Hadoop data
- Describe the four different types of cloud databases in NoSQL Databases
- Describe the basics of the Hadoop Distributed File System
- Describe HDFS and basic HDFS navigation operations
- Perform file operations such as add and delete within HDFS
- Describe the basic principles of MapReduce and general mapping issues
- Specify the use of Pig and Hive in Hadoop Map Reduce jobs
- Describe the use of MapReduce, MapReduce lifecycle, job client, job tracker, task tracker, map tasks, and reduce tasks
- Describe Hadoop MapReduce handles, data processes data, and vocabulary of the MapReduce dataflow process
- Describe the process of mapping and reducing
- Describe the basic principles and uses of Hadoop

Ecosystem for Hadoop

Hadoop's HDFS is a highly fault-tolerant distributed file system and, like Hadoop in general, designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications that have large data sets. This course examines the Hadoop ecosystem by demonstrating all of the commonly used open source software components. You will explore a big data model to understand how these tools combine to create a supercomputing platform. You will also learn how the principles of supercomputing apply to Hadoop and how this yields an affordable supercomputing environment. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam.

Outline

- Start the course
- Describe supercomputing
- Recall three major functions of data analytics
- Define big data
- Describe the two different types of data
- Describe the components of the big data stack
- Identify the data repository components
- Identify the data refinery components
- Identify the data factory components
- Recall the design principles of Hadoop
- Describe the design principles of sharing nothing
- Describe the design principles of embracing failure
- Describe the components of the Hadoop Distributed File System (HDFS)
- Describe the four main HDFS daemons
- Describe Hadoop YARN
- Describe the roles of the Resource Manager daemon
- Describe the YARN NodeManager and ApplicationMaster daemons
- Define MapReduce and describe its relations to YARN
- Describe data analytics
- Describe the reasons for the complexities of the Hadoop Ecosystem
- Describe the components of the Hadoop ecosystem

Data Flow for the Hadoop Ecosystem

Hadoop is a framework written in Java for running applications on large clusters of commodity hardware and incorporates features similar to those of the GFS and of the MapReduce computing paradigm. You will explore a demonstration of the use of Sqoop and Hive with Hadoop to flow and fuse data. The demonstration includes pre-processing data, partitioning data and joining data. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam.

Outline

- Start the course
- Describe the data lifecycle management
- Recall the parameters that must be set in the Sqoop import statement
- Create a table and load data into MySQL
- Use Sqoop to import data into Hive
- Recall the parameters that must be set in the Sqoop export statement
- Use Sqoop to export data from Hive
- Recall the three most common date datatypes and which systems support each
- Use casting to import datetime stamps into Hive
- Export datetime stamps from Hive into MySQL
- Describe dirty data and how it should be pre-processed
- Use Hive to create tables outside the warehouse
- Use pig to sample data
- Recall some other popular components for the Hadoop ecosystem
- Recall some best practices for pseudo-mode implementation
- Write custom scripts to assist with administrative tasks
- Troubleshoot classpath errors
- Create complex configuration files
- To use Sqoop and Hive for data flow and fusion in the Hadoop ecosystem

Data Repository with HDFS and Hbase

Hadoop is an open source Java framework for processing and querying vast amounts of data on large clusters of commodity hardware. It relies on an active community of contributors from all over the world for its success. In this course, you will explore the server architecture for Hadoop and learn about the functions and configuration of the daemons making up the Hadoop Distributed File System. You will also learn about the command line interface and common HDFS administration issues facing all end users. Finally, you will explore the theory of HBase as another data repository built alongside, or on top of, HDFS and basic HBase commands. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam.

Outline

- Start the course
- Configure the replication of data blocks
- Configure the default file system scheme and authority
- Describe the functions of the NameNode
- Recall how the NameNode operates
- Recall how the DataNode maintains data integrity
- Describe the purpose of the CheckPoint Node
- Describe the role of the Backup Node
- Recall the syntax of the file system shell commands
- Use shell commands to manage files
- Use shell commands to provide information about the file system
- Perform common administration functions
- Configure parameters for NameNode and DataNode
- Troubleshoot HDFS errors
- Describe key attributes of NoSQL databases
- Describe the roles of HBase and ZooKeeper
- Install and configure ZooKeeper
- Use the HBase command line to create tables and insert datall and configure HBase
- Manage tables and view the web interface
- Create and change HBase data
- Provide a basic understanding of how Hadoop Distributed File System functions

Data Repository with Flume

Hadoop is an open source software project that enables distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer. In this course, you will learn about the theory of Flume as a tool for dealing with extraction and loading of unstructured data. You will explore a detailed explanation of the Flume agents and a demonstration of the Flume agents in action. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam..

Outline

- Start the course
- Describe the three key attributes of Flume
- Recall some of the protocols cURL supports
- Use cURL to download web server data
- Recall some best practices for the Agent Conf files
- Install and configure Flume
- Create a Flume agent
- Describe a Flume agent in detail
- Use a Flume agent to load data into HDFS
- Identify popular sources
- Identify popular sinks
- Describe Flume channels
- Describe what is happening during a file roll
- Recall that Avro can be used as both a sink and a source
- Use Avro to capture a remote file
- Create multiple-hop Flume agents
- Describe interceptors
- Create a Flume agent with a TimeStampInterceptor
- Describe multifunction Flume agents
- Configure Flume agents for multiflow
- Create multi-source Flume agents
- Compare replicating to multiplexing
- Create a Flume agent for multiple data sinks
- Recall some common reasons for Flume failures
- Use the logger to troubleshoot Flume agents
- Configure the various Flume agents

Data Repository with Sqoop

Hadoop is an open-source software framework for storing and processing big data in a distributed fashion on large clusters of commodity hardware. Essentially, it accomplishes two tasks: massive data storage and faster processing. This course explains the theory of Sqoop as a tool for dealing with extraction and loading of structured data from an RDBMS. You will explore an explanation of Hive SQL statements and a demonstration of Hive in action. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam.

Outline

- Start the course
- Describe MySQL
- Install MySQL
- Create a database in MySQL
- Create MySQL tables and load data
- Describe Sqoop
- Describe Sqoop's architecture
- Recall the dependencies for Sqoop installation
- Install Sqoop
- Recall why it's important for the primary key to be numeric
- Perform a Sqoop import from MySQL into HDFS
- Recall what concerns the developers should be aware of
- Perform a Sqoop export from HDFS into MySQL
- Recall that you must execute a Sqoop import statement for each data element
- Perform a Sqoop import from MySQL into HBase
- Recall how to use chain troubleshooting to resolve Sqoop issues
- Use the log files to identify common Sqoop errors and their resolutions
- To use Sqoop to extract data from a RDBMS and load the data into HDFS

Data refinery with YARN and MapReduce

The core of Hadoop consists of a storage part, HDFS, and a processing part, MapReduce. Hadoop splits files into large blocks and distributes the blocks among the nodes in the cluster. To process the data, Hadoop and MapReduce transfer code to nodes that have the required data, which the nodes then process in parallel. This approach takes advantage of data locality to allow the data to be processed faster and more efficiently via distributed processing than by using a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking. In this course, you will learn about the theory of YARN as a parallel processing framework for Hadoop. You will also learn about the theory of MapReduce as the backbone of parallel processing jobs. Finally, this course demonstrates MapReduce in action by explaining the pertinent classes and then walking through a MapReduce program step-by-step. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam.

Outline

- Start the course
- Describe parallel processing in the context of supercomputing
- List the components of YARN and identify their primary functions
- Diagram YARN Resource Manager and identify its key components
- Diagram YARN Node Manager and identify its key components
- Diagram YARN ApplicationMaster and identify its key components
- Describe the operations of YARN
- Identify the standard configuration parameters to be changed for YARN
- Define the principle concepts of key-value pairs and list the rules for key-value pairs
- Describe how MapReduce transforms key-value pairs
- Load a large text book and then run WordCount to count the number of words in the text book
- Label all of the functions for MapReduce on a diagram
- Match the phases of MapReduce to their definitions
- Set up the classpath and test WordCount
- Build a JAR file and run WordCount
- Describe the base Mapper class of the MapReduce Java API and describe how to override its methods
- Describe the base Reducer class of the MapReduce Java API and describe how to override its methods
- Describe the function of the MapReduceDriver Java class
- Set up the classpath and test a MapReduce job
- Identify the concept of streaming for MapReduce
- Stream a Python job
- Understand YARN features and components, as well as MapReduce and its classes

Data Factory with Hive

Apache Hadoop is a set of algorithms for distributed storage and distributed processing of big data on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are commonplace and thus should be automatically handled in software by the framework. In this course, you will explore Hive as an SQL-like tool for interfacing with Hadoop. The course demonstrates the installation and configuration of Hive, followed by a demonstration of Hive in action. Finally, you will learn about extracting and loading data between Hive and an RDBMS. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam.

Outline

- Start the course
- Recall the key attributes of Hive
- Describe the configuration files
- Install and configure Hive
- Create a table in Derby using Hive
- Create a table in MySQL using Hive
- Recall the unique delimiter that Hive uses
- Describe the different operators in Hive
- Use basic SQL commands in Hive
- Use SELECT statements in Hive
- Use more complex HiveQL
- Write and use Hive scripts
- Recall what types of joins Hive can support
- Use Hive to perform joins
- Recall that a Hive partition schema must be created before loading the data
- Write a Hive partition script
- Recall how buckets are used to improve performance
- Create Hive buckets
- Recall some best practices for user defined functions
- Create a user defined function for Hive
- Recall the standard error code ranges and what they mean
- Use a Hive explain plan
- Understand configuration option, data loading and querying

Data Factory with Pig

Hadoop is an open source software for affordable supercomputing. It provides the distributed file system and the parallel processing required to run a massive computing cluster. This course explains Pig as a data flow scripting tool for interfacing with Hadoop. You will learn about the installation and configuration of Pig and explore a demonstration of Pig in action. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam.

Outline

- Start the course
- Describe Pig and its strengths
- Recall the minimal edits needed to be made to the configuration file
- Install and configure Pig
- Recall the complex data types used by Pig
- Recall some of the relational operators used by Pig
- Use the Grunt shell with Pig Latin
- Set parameters from both a text file and with the command line
- Write a Pig script
- Use a Pig script to filter data
- Use the FOREACH operator with a Pig script
- Set parameters and arguments in a Pig script
- Write a Pig script to count data
- Perform data joins using a Pig script
- Group data using a Pig script
- Cogroup data with a Pig script
- Flatten data using a pig script
- Recall the languages that can be used to write user defined functions
- Create a user defined function for Pig
- Recall the different types of error categories
- Use explain in a Pig script
- Install Pig, use Pig operators and Pig Latin, and retrieve and group records

Data Factory with Oozie and Hue

The Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly available service on top of a cluster of computers, each of which may be prone to failures. This course explains Oozie as a workflow tool used to manage multiple stage tasks in Hadoop. Additionally, you will learn how to use Hue, a front end tool which is browser based. This learning path can be used as part of the preparation for the Cloudera Certified Administrator for Apache Hadoop (CCA-500) exam.

Outline

- Start the course
- Describe metastore and HiveServer2
- Install and configure metastore
- Install and configure HiveServer2
- Describe HCatalog
- Install and configure WebHCat
- Use HCatalog to flow data
- Recall the Oozie terminology
- Recall the two categories of environmental variables for configuring Oozie
- Install Oozie
- Configure Oozie
- Configure Oozie to use MySQL
- Enable the Oozie Web Console
- Describe Oozie workflows
- Submit an Oozie workflow job
- Create an Oozie workflow
- Run an Oozie workflow job
- Describe Hue
- Recall the configuration files that must be edited
- Install Hue
- Configure the hue.ini file
- Install and configure Hue on MySQL
- Use the Hue File Browser and Job Scheduler
- Configure Hive daemons, Oozie, and Hue

Learn more at

www.hpe.com/ww/digitallearner
www.hpe.com/ww/digitallearner-contentpack

Follow us:

CP010 A.00 , December 2018

Hewlett Packard Enterprise