Big Data Hadoop & Spark Certification Training

intelliPaat

About Intellipaat

Intellipaat is a fast growing professional training provider that is offering training in over 150 most sought-after tools and technologies. We have a learner base of 700,000 in over 32 countries and growing. For job assistance and placement we have direct tie-ups with 80+ MNCs.

Key Features of IntellipaatTraining :

24X7 Life Time Support and Assistance	Real Time Projects	Life Time Access and Free Upgrade	Job Assistance	Industry Recognised Certification

About the Course

It is a comprehensive Hadoop Big Data training course designed by industry experts considering current industry job requirements to provide in-depth learning on big data and Hadoop Modules. This is an industry recognized Big Data certification training course that is a combination of the training courses in Hadoop developer, Hadoop administrator, Hadoop testing, and analytics. This Cloudera Hadoop training will prepare you to clear big data certification

Instructor Led
Duration – 60 Hrs
Weekend Batch – 3 Hrs/Session
Weekday Batch – 2Hrs/Session

Self Paced
Duration – 85 Hrs

Why Take This Course ?

Big Data is fastest growing and most promising technology for handling large volumes of data for doing data analytics. This Big Data Hadoop training will help you tobe up and running in the most demanding professional skills. Almost all the top MNC are trying to get into Big Data Hadoop hence there is a huge demand for Certified Big Data professionals. Our Big Data online training will help you to upgrade your career in big data domain.

❖ Shortage of 1.4 -1.9 million Hadoop Data Analysts in US alone by 2018– Mckinsey
❖ Hadoop Administrator in the US can get a salary of $123,000 – indeed.com

50% Jump in Salary

IntelliPaat

Course Contents

Hadoop Installation & setup	Hands on Exercises
- Hadoop 2.x Cluster Architecture - Federation and High Availability - A Typical Production Cluster setup - Hadoop Cluster Modes - Common Hadoop Shell Commands - Hadoop 2.x Configuration Files - Cloudera Single node cluster, Hive, Pig, Sqoop, Flume, Scala and Spark.	Setup of Pseudo node Hadoop Cluster in Local Machine
Introduction to Big Data Hadoop. Understanding HDFS & Mapreduce	**Hands on Exercises**
- Introducing Big Data & Hadoop - What is Big Data and where does Hadoop fits in - two important Hadoop ecosystem componentsnamely Map Reduce and HDFS - In-depth Hadoop Distributed File System - Replications, Block Size, Secondary Name node, High Availability, in-depth YARN - Resource Manager, Node Manager.	Working with HDFS,Replicating the data,Determining block size, Familiarizing with Namenode and Datanode.
Deep Dive in MapReduce	**Hands on Exercises**
- Detailed understanding of the working of MapReduce - The mapping and reducing process - Working of Driver, Combiners, Partitioners, Input Formats, Output Formats, Shuffle and Sort.	- The detailed methodology for writing the Word Count Program in MapReduce Writing custom partitioner, MapReduce with Combiner - Local Job Runner Mode, Unit Test, ToolRunner, MapSide Join, Reduce Side Join,Using Counters Joining two datasets using Map-Side Join &Reduce-Side Join.
Introduction to Hive	**Hands on Exercises**
- Introducing Hadoop Hive - Detailed architecture of Hive - Comparing Hive with Pig and RDBMS - Working with Hive Query Language - Creation of database, table, Group by and other clauses - Various types of Hive tables, Hcatalog, storing the Hive Results, Hive partitioning and Buckets.	- Creating of Hive database, how to drop database - Changing the database, - Creating of Hive table, loading of data - Dropping the table and altering it, writing hive queries to pull data using filter conditions - Group by clauses, partitioning Hive tables
Advance Hive & Impala	**Hands on Exercises**
- The indexing in Hive - The Map side Join in Hive - Working with complex data types	Working with Hive queries,Writing indexes, joining table, deploying external table, sequence table and storing data in another table.

- ❖ Hive User-defined Functions
- ❖ Introduction to Impala
- ❖ Comparing Hive with Impala
- ❖ Detailed architecture of Impala

Introduction to Pig	Hands on Exercises
❖ Apache Pig introduction ❖ Various features ❖ Various data types and schema in Hive ❖ Available functions in Pig, Hive Bags, Tuples and Fields.	❖ Working with Pig in MapReduce and local mode ❖ Loading of data, limiting data to 4 rows ❖ Storing the data into fileworking with Group By,FilterBy,Distinct,Cross,Split in Hive.
Flume, Sqoop &HBase	**Hands on Exercises**
❖ Introduction to Apache Sqoop ❖ Sqoop overview, basic imports and exports ❖ How to improve Sqoop performance ❖ Limitation of Sqoop ❖ Introduction to Flume and its Architecture ❖ Introduction to HBase ❖ CAP theorem.	❖ Working with Flume to generating of Sequence Number and consuming it ❖ using the Flume Agent to consume the Twitter data, using AVRO to create Hive Table, AVRO with Pig ❖ Creating Table in HBase, deploying Disable, Scan and Enable Table.
Writing Spark Applications using Scala	**Hands on Exercises**
❖ Using Scala for writing Apache Spark applications ❖ Detailed study of Scala ❖ Need for Scala ❖ Concept of object oriented programming ❖ Executing the Scala code ❖ Various classes in Scala like Getters,Setters, Constructors, Abstract ,Extending Objects, Overriding Methods ❖ Java and Scala interoperability ❖ Concept of functional programming and anonymous functions ❖ Bobsrockets package, comparing the mutable and immutable collections.	❖ Writing Spark application using Scala, understanding the robustness of Scala for Spark real-time analytics operation.
Spark framework	**Hands on Exercises**
❖ Detailed Apache Spark ❖ its various features ❖ Comparing with Hadoop ❖ Various Spark components ❖ Combining HDFS with Spark, Scalding ❖ Introduction to Scala, importance of Scala and RDD.	❖ Resilient Distributed Dataset in Spark and how it helps to speed up big data processing.
RDD in Spark	**Hands on Exercises**
❖ The RDD operation in Spark ❖ The Spark transformations, actions, data loading	❖ How to deploy RDD with HDFS, using the in-memory dataset, using file for RDD, how to

❖ Comparing with MapReduce, Key Value Pair.	define the base RDD from external file, deploying RDD via transformation, using the Map and Reduce functions, working on word count and count log severity.
Data Frames and Spark SQL ❖ The detailed Spark SQL ❖ The significance of SQL in Spark for working with structured data processing ❖ Spark SQL JSON support, working with XML data, and parquet files ❖ Creating HiveContext, writing Data Frame to Hive, reading of JDBC files ❖ The importance of Data Frames in Spark, creating Data Frames, schema manual inferring, working with CSV files, reading of JDBC tables, converting from Data Frame to JDBC ❖ The user-defined functions in Spark SQL, shared variable and accumulators ❖ How to query and transform data in Data Frames ❖ How Data Frame provides the benefits of both Spark RDD and Spark SQL ❖ Deploying Hive on Spark as the execution engine.	**Hands on Exercises** ❖ Data querying and transformation using Data Frames ❖ Finding out the benefits of Data Frames over Spark SQL and Spark RDD.
Machine Learning using Spark (Mlib) ❖ Different Algorithms, the concept of iterative algorithm in Spark ❖ Analyzing with Spark graph processing ❖ Introduction to K-Means and machine learning ❖ Various variables in Spark like shared variables ❖ Broadcast variables ❖ Learning about accumulators.	**Hands on Exercises** ❖ Writingspark code using Mlib.
Spark Streaming ❖ Introduction to Spark streaming ❖ The architecture of Spark Streaming ❖ Working with the Spark streaming program ❖ Processing data using Spark streaming, requesting count and Dstream, multi-batch and sliding window operations and working with advanced data sources.	**Hands on Exercises** ❖ Deploying Spark streaming for data in motion and checking the output is as per the requirement.

Hadoop Administration - Cluster Configuration	**Hands on Exercises**
❖ The overview of Hadoop configuration ❖ The importance of Hadoop configuration file ❖ The various parameters and values of configuration ❖ The HDFS parameters and MapReduce parameters ❖ Setting up the Hadoop environment ❖ The Includeand Exclude configuration files ❖ The administration and maintenance of Name node ❖ Data node directory structures and files ❖ File system image and Edit log	❖ The method to do performance tuning of MapReduce program.
Hadoop Administration - Maintenance, Monitoring and Troubleshooting ❖ Introduction to the Checkpoint Procedure ❖ Name node failure and how to ensure the recovery procedure, Safe Mode, Metadata and Data backup ❖ The various potential problems and solutions ❖ What to look for ❖ How to add and remove nodes.	**Hands on Exercises** ❖ How to go about ensuring the MapReduce File system Recovery for various different scenarios, JMX monitoring of the Hadoop cluster. ❖ How to use the logs and stack traces for monitoring and troubleshooting ❖ Using the Job Scheduler for scheduling jobs in the same cluster ❖ Getting the MapReduce job submission flow, FIFO schedule, getting to know the Fair Scheduler and its configuration.
Securing Hadoop Cluster with Kerberos and other Advance topics ❖ Advanced Hadoop administration functions ❖ Using the Quorum Journal Manager ❖ Configuring the Hadoop federation and security ❖ Fundamentals of the Hadoop Platform Security ❖ Working with Kerberos authentication ❖ Configuring Kerberos on Hadoop cluster.	**Hands on Exercises** ❖ Detailed procedure for configuring the Kerberos authentication with the Hadoop cluster and checking the results of the configuration.
ETL Connectivity with Hadoop Ecosystem ❖ How ETL tools work in Big data Industry ❖ Introduction to ETL and Data warehousing ❖ Working with prominent use cases of Big data in ETL industry ❖ End to End ETL PoC showing big data integration with ETL tool.	**Hands on Exercises** ❖ Connecting to HDFS from ETL tool and moving data from Local system to HDFS, Moving Data from DBMS to HDFS, Working with Hive with ETL Tool, Creating Map Reduce job in ETL tool
IBM Project Solution Discussion and Cloudera Certification Tips & Tricks ❖ Working towards the solution of the Hadoop IBM project solution ❖ Its problem statements and the possible solution outcomes ❖ Preparing for the Cloudera Certifications ❖ Points to focus for scoring the highest marks	**Hands on Exercises** ❖ The IBM project of a real-world high value Big Data Hadoop application and getting the right solution based on the criteria set by the IBM team.

	Hands on Exercises
❖ Tips for cracking Hadoop interview questions.	

Hadoop Application Testing	**Hands on Exercises**
❖ Why testing is important ❖ Unit testing, Integration testing, Performance testing ❖ Diagnostics, Nightly QA test ❖ Benchmark and end to end tests ❖ Functional testing ❖ Release certification testing ❖ Security testing, Scalability Testing ❖ Commissioning and Decommissioning of Data Nodes Testing, Reliability testing, Release testing	**NA**

Roles and Responsibilities of Hadoop Testing Professional	**Hands on Exercises**
❖ Understanding the Requirement, preparation of the Testing Estimation ❖ Test Cases, Test Data, Test bed creation, Test Execution ❖ Defect Reporting, Defect Retest, Daily Status report delivery, Test completion ❖ ETL testing at every stage (HDFS, HIVE, HBASE) while loading the input (logs/files/records etc) using sqoop/flume which includes but not limited to data verification ❖ Reconciliation ❖ User Authorization and Authentication testing (Groups, Users, Privileges etc) ❖ Report defects to the development team or manager and driving them to closure ❖ Consolidate all the defects and create defect reports ❖ Validating new feature and issues in Core Hadoop.	**NA**

Framework called MR Unit for Testing of Map-Reduce Programs	**Hands on Exercises**
❖ Report defects to the development team or manager and driving them to closure ❖ Consolidate all the defects and create defect reports ❖ Responsible for creating a testing Framework called MR Unit for testing of Map-Reduce programs.	❖ Working with MR unit and creating small test cases

Unit Testing	**Hands on Exercises**
❖ Automation testing using the OOZIE ❖ Data validation using the query surge tool	❖ Automation testing using the OOZIE ❖ Data validation using the query surge tool

Test Execution	**Hands on Exercises**
❖ Test plan for HDFS upgrade, Test automation and result	❖ Test automation exercises

Test Plan Strategy and writing Test Cases for testing Hadoop Application ❖ How to test install and configure	**Hands on Exercises** Test Planning and configuration

Big Data Hadoop and Spark Projects

Project 1 - Working with MapReduce, Hive, Sqoop

Topics : As part of this Big data Hadoop certification training, you will undergo the project which involves working on the various Hadoop components like MapReduce, Apache Hive and Apache Sqoop. Work with Sqoop to import data from relational database management system like MySQL data into HDFS. Deploy Hive for summarizing data, querying and analysis. Convert SQL queries using HiveQL for deploying MapReduce on the transferred data. You will gain considerable proficiency in Hive, and Sqoop after completion of this project.

Project 2 - Work on MovieLens data for finding top records

Data – MovieLens dataset

Topics : In this project you will work exclusively on data collected through MovieLens available rating data sets. The project involves the following important components:

- ❖ You will write a MapReduce program in order to find the top 10 movies by working in the data file
- ❖ Learn to deploy Apache Pig create the top 10 movies list by loading the data
- ❖ Work with Apache Hive and create the top 10 movies list by loading the

Project 3 - Hadoop YARN Project – End to End PoC

Topics : In this Big Data project you will work on a live Hadoop YARN project. YARN is part of the Hadoop 2.0 ecosystem that lets Hadoop to decouple from MapReduce and deploy more competitive processing and wider array of applications. You will work on the YARN central Resource Manager. The salient features of this project include:

- ❖ Importing of Movie data
- ❖ Appending the data
- ❖ Using Sqoop commands to bring the data into HDFS
- ❖ End to End flow of transaction data
- ❖ Processing data using MapReduce program in terms of the movie data, etc.

Project 4 - Partitioning Tables in Hive

Topics : This project involves working with Hive table data partitioning. Ensuring the right partitioning helps to read the data, deploy it on the HDFS, and run the MapReduce jobs at a much faster rate. Hive lets you partition data in multiple ways like:

- ❖ Manual Partitioning
- ❖ Dynamic Partitioning

❖ Bucketing

This will give you hands-on experience in partitioning of Hive tables manually, deploying single SQL execution in dynamic partitioning, bucketing of data so as to break it into manageable chunks.

Project 5 - Connecting Pentaho with Hadoop Ecosystem

Topics : This project lets you connect Pentaho with the Hadoop ecosystem. Pentaho works well with HDFS, HBase, Oozie and Zookeeper. You will connect the Hadoop cluster with Pentaho data integration, analytics, Pentaho server and report designer. Some of the components of this project include the following:

❖ Clear hands-on working knowledge of ETL and Business Intelligence
❖ Configuring Pentaho to work with Hadoop Distribution
❖ Loading, Transforming and Extracting data into Hadoop cluster

Project 6 - Multi-node cluster setup

Topics : This is a project that gives you opportunity to work on real world Hadoop multi-node cluster setup in a distributed environment. The major components of this project involve:

❖ Running a Hadoop multi-node using a 4 node cluster on Amazon EC2
❖ Deploying of MapReduce job on the Hadoop cluster

You will get a complete demonstration of working with various Hadoop cluster master and slave nodes, installing Java as a prerequisite for running Hadoop, installation of Hadoop and mapping the nodes in the Hadoop cluster.

❖ Hadoop Multi-Node Cluster Setup using Amazon ec2 – Creating 4 node cluster setup
❖ Running Map Reduce Jobs on Cluster

Project 7 - Hadoop Testing using MR

Topics : In this project you will gain proficiency in Hadoop MapReduce code testing using MRUnit. You will learn about real world scenarios of deploying MRUnit, Mockito, and PowerMock. Some of the important aspects of this project include:

❖ Writing JUnit tests using MRUnit for MapReduce applications
❖ Doing mock static methods using PowerMock&Mockito
❖ MapReduceDriver for testing the map and reduce pair

After completion of this project you will be well-versed in test driven development and will be able to write light-weight test units that work specifically on the Hadoop architecture.

Project 8 - Hadoop Weblog Analytics

Data – Weblogs

Topics : This project is involved with making sense of all the web log data in order to derive valuable insights from it. You will work with loading the server data onto a Hadoop cluster using various techniques. The various modules of this project include:

❖ Aggregation of log data
❖ Processing of the data and generating analytics

The web log data can include various URLs visited, cookie data, user demographics, location, date and time of web service access, etc. In this project you will transport the data using Apache Flume or Kafka, workflow and data cleansing using MapReduce, Pig or Spark. The insight thus derived can be used for analyzing customer behavior and predict buying patterns.

Project 9 -Hadoop Maintenance

Topics : This project is involved with working on the Hadoop cluster for maintaining and managing it. You will work on a number of important tasks like**:**

- ❖ Administration of distributed file system
- ❖ Checking the file system
- ❖ Working with name node directory structure
- ❖ Audit logging, data node block scanner, balancer
- ❖ Learning about the properties of safe mode
- ❖ Entering and exiting safe mode
- ❖ HDFS federation and high availability
- ❖ Failover, fencing, DISTCP, Hadoop file formats

Project 10 -Movie Recommendation

Topics : This is a hands-on Apache Spark project deployed for the real-world application of movie recommendations. This project helps you gain essential knowledge in Spark MLlib which is a machine learning library, you will know how to create collaborative filtering, regression, clustering and dimensionality reduction using Spark MLlib. Upon finishing the project you will have first-hand experience in the Apache Spark streaming data analysis, sampling, testing, statistics among other vital skills.

Project 11 - Twitter API Integration for tweet Analysis

Topics : This is a hands-on Twitter analysis project using the Twitter API for analyzing of tweets. You will integrate the Twitter API,do programming using the various scripting languages like Ruby, Python, PHP for developing the essential server side codes. Finally you will be able to read the results for various operations by filtering, parsing, and aggregating it depending on the tweet analysis requirement.

Project 12 –Data Exploration Using Spark SQL – Wikipedia data set

Topics : In this project you will be using the Spark SQL tool for analyzing the Wikipedia data. You will gain hands-on experience in integrating Spark SQL for various applications like batch analysis, machine learning, visualizing and processing of data, ETL processes along with real-time analysis of data.

What makes us who we are

"I have learnt various big data techniques, it only gave me the skills that i needed but also gave me confindence in those skills."….*Read More!*

Monica

VIEW ALL SUCCESS STORIES   **READ ALL REVIEWS**

"I wanted to align my career in data analytics field as it is booming in the industry, this course helped in going profound knowledge of big data concepts"…..Read More!

READ ALL REVIEWS   **VIEW ALL SUCCESS STORIES**

Sagar