Big Data Hadoop Certification Training

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 600,000 in over 32 countries and growing. For job assistance and placement we have direct tie-ups with 80+ MNCs.

Key Features of Intellipaat Training:

Instructor Led Training

60 Hrs of highly interactive instructor led training

Self-Paced Training

85 Hrs of Self-Paced session with Lifetime access

Exercise and project work

120 Hrs of real-time projects after every module

Lifetime Access

Lifetime access and free upgrade to latest version

Support

Lifetime 24*7 technical support and query resolution

Get Certified

Get global industry recognized certifications

Job Assistance

Job assistance through 80+ corporate tie-ups

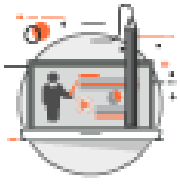
Flexi Scheduling

Attend multiple batches for lifetime & stay updated.

About the Course

It is a comprehensive Hadoop Big Data training course designed by industry experts considering current industry job requirements to help you learn Big Data Hadoop and Spark 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 with Apache Spark. This Cloud era Hadoop & Spark training will prepare you to clear Cloud era CCA 175 big data certification.

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



Self Paced
Duration – 85 Hrs

Big Data Hadoop Certification Training

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 to be up and run 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 learn big data and upgrade your career in the big data domain. Getting the big data certification from Intellipaat can put you in a different league when it comes to applying for the best jobs. Intellipaat big data online course has been created with a complete focus on the practical aspects of big data Hadoop.

- Global Hadoop Market to Reach $84.6 Billion by 2021 – Allied Market Research
- Shortage of 1.4 -1.9 million Hadoop Data Analysts in the US alone by 2018– Mckinsey
- Hadoop Administrator in the US can get a salary of $123,000 – indeed.com

Course Content

Module /Topic	Hands-on exercises
Hadoop Installation & setup ❖ 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 ❖ Cloud era Single node cluster, Hive, Pig, Sqoop, Flume, Scala, and Spark	
Introduction to Big Data Hadoop. Understanding HDFS & Map-reduce ❖ Introducing Big Data & Hadoop ❖ what is Big Data and where does Hadoop fits in ❖ Two important Hadoop ecosystem components namely 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 the block size ❖ Familiarizing with Name node and Data node

Big Data Hadoop Certification Training

Deep Dive in MapReduce ❖ Detailed understanding of the working of MapReduce ❖ The mapping and reducing process ❖ The working of Driver, Combiners, Partitioners, Input Formats, Output Formats, Shuffle and Sor	❖ 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 ❖ 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 ❖ The various types of Hive tables, Hcatalog, storing the Hive Results, Hive partitioning and Buckets	❖ Creating of Hive database, how to drop the 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 ❖ The indexing in Hive ❖ The Map side Join in Hive, working with complex data types ❖ The Hive User-defined Functions, ❖ Introduction to Impala, comparing Hive with Impala, the detailed architecture of Impala	❖ Working with Hive queries ❖ Writing indexes, joining table, deploying external table ❖ Sequence table and storing data in another table
Introduction to Pig ❖ Apache Pig introduction, its various features, the various data types and schema in Pig ❖ The available functions in Pig, Pig Bags, Tuples and Fields	❖ Working with Pig in MapReduce and local mode, loading of data, limiting data to 4 rows, storing the data into a file ❖ Working with Group By, Filter By, Distinct, Cross, Split in Pig
Flume, Sqoop & HBase	❖ Working with Flume to

❖ Introduction to Apache Sqoop, Sqoop overview, basic imports and exports, how to improve Sqoop performance, the limitation of Sqoop, ❖ Introduction to Flume and its Architecture ❖ Introduction to HBase, the CAP theorem	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 ❖ Using Scala for writing Apache Spark applications ❖ Detailed study of Scala, the need for Scala ❖ The concept of object-oriented programming, executing the Scala code ❖ The various classes in Scala like Getters, Setters, Constructors, Abstract, Extending Objects, Overriding Methods ❖ The Java and Scala interoperability, the 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 ❖ Detailed Apache Spark, its various features ❖ Comparing with Hadoop, the various Spark components ❖ Combining HDFS with Spark, Scalding ❖ Introduction to Scala, the importance of Scala and RDD	❖ The Resilient Distributed Dataset in Spark and how it helps to speed up big data processing
RDD in Spark ❖ The RDD operation in Spark ❖ The Spark transformations, actions, data loading ❖ Comparing with MapReduce, Key-Value Pair	❖ How to deploy RDD with HDFS, using the in-memory dataset, using the file for RDD ❖ How to define the base RDD from an 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 Hive Context, 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	❖ 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 the 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	❖ Writing sparks 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	❖ Deploying Spark streaming for data in motion and checking the output is as per the requirement
Hadoop Administration – Multi-Node Cluster Setup using azon EC2 ❖ Create a four-node Hadoop cluster setup ❖ Running the MapReduce Jobs on the Hadoop cluster,	❖ The method to build a multi-node Hadoop cluster using an Amazon EC2 instance ❖ Working with the Cloudera Manager

successfully running ❖ The MapReduce code, working with the Cloudera Manager setup	
Hadoop Administration – Cluster Configuration ❖ 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 Include and 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 ubleshooting ❖ 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	❖ 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
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	❖ 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 MapReduce job in ETL tool

Big Data Hadoop Certification Training

Project Solution Discussion and Cloudera Certification Tips ricks	
❖ Working towards the solution of the Hadoop project solution, its problem statements, and the possible solution outcomes ❖ Preparing for the Cloud era Certifications points to focus for scoring the highest marks, tips for cracking Hadoop interview questions	❖ The project of a real-world high value ❖ Big Data Hadoop application and getting the right solution based on the criteria set by the Intellipaat team

Project Work

Project 1: Working with MapReduce, Hive, Sqoop
Industry: General
Problem Statement: How to successfully import data using Sqoop into HDFS for data analysis.
Topics: As part of this project you will work 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.
Highlights:

- Sqoop data transfer from RDBMS to Hadoop
- Coding in Hive Query Language
- Data querying and analysis

Project 2: Work on MovieLens data for finding top movies
Industry: Media and Entertainment
Problem Statement: How to create the top ten movies list using the MovieLens data.
Topics: In this project, you will work exclusively on data collected through MovieLens available rating data sets. The project involves writing MapReduce program to analyze the MovieLens data and create a list of top ten movies. You will also work with Apache Pig and Apache Hive for working with distributed datasets and analyzing it.
Highlights:

- MapReduce program for working on the data file
- Apache Pig for analyzing data
- Apache Hive data warehousing and querying

Project 3: Hadoop YARN Project – End to End PoC
Industry: Banking
Problem Statement: How to bring the daily data (incremental data) into the Hadoop Distributed File System.
Topics: In this project, we have transaction data which is daily recorded/store in the RDBMS. Now, this data is transferred every day into HDFS for further Big Data Analytics. You will work on live Hadoop YARN cluster. YARN is part of the Hadoop 2.0 ecosystem that lets Hadoop decouple from MapReduce and deploy more competitive processing and a wider array of applications. You will work on the YARN central Resource Manager.

Highlights:

- Using Sqoop commands to bring the data into HDFS
- End to End flow of transaction data
- Working with the data from HDFS

Project 4: Table Partitioning in Hive
Industry: Banking
Problem Statement: How to improve the query speed using Hive data partitioning.
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. This will give you hands-on experience in the partitioning of Hive tables manually, deploying single SQL execution in dynamic partitioning, bucketing of data so as to break it into manageable chunks.
Highlights:

- Manual Partitioning
- Dynamic Partitioning
- Bucketing

Project 5: Connecting Pentaho with Hadoop Ecosystem
Industry: Social Network
Problem Statement: How to deploy ETL for data analysis activities.
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. This project will give you complete working knowledge of the Pentaho ETL tool.
Highlights:

- 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
Industry: General
Problem Statement: How to set up a Hadoop real-time cluster on Amazon EC2.
Topics: This is a project that gives you the opportunity to work on real-world Hadoop multi-node cluster setup in a distributed environment. 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.
Highlights:

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

Project 7: Hadoop Testing using MRUnit
Industry: General
Problem Statement: How to test MapReduce applications
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. This will give you hands-on experience in the various testing tools for Hadoop MapReduce. 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.
Highlights:

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

Project 8: Hadoop Weblog Analytics
Industry: Internet services
Problem Statement: How to derive insights from web log data
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 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.
Highlights:

- Aggregation of log data
- Apache Flume for data transportation
- Processing of data and generating analytics

Project 9: Hadoop Maintenance
Industry: General
Problem Statement: How to administer a Hadoop cluster
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 that include recovering of data, recovering from failure, adding and removing of machines from the Hadoop cluster and onboarding of users on Hadoop.
Highlights:

- Working with name node directory structure
- Audit logging, data node block scanner, balancer
- Failover, fencing, DISTCP, Hadoop file formats

Project 10: Twitter Sentiment Analysis
Industry: Social Media
Problem Statement: Find out what is the reaction of the people to the demonetization move by India by analyzing their tweets.
Description: This Project involves analyzing the tweets of people by going through what they are saying about the demonetization decision taken by the Indian government. Then you look for key phrases, words and analyze them using the dictionary and the value attributed to them based on the sentiment that it is conveying.
Highlights:

- Download the Tweets & Load into Pig Storage
- Divide tweets into words to calculate sentiment
- Rating the words from +5 to -5 on AFFIN dictionary
- Filtering the Tweets and analyzing sentiment

Project 11: Analyzing IPL T20 Cricket
Industry: Sports & Entertainment
Problem Statement: Analyze the entire cricket match and get answers to any question regarding the details of the match.
Description: This project involves working with the IPL dataset that has information regarding batting, bowling, runs scored, wickets are taken, and more. This dataset is taken as input and then it is processed so that the entire match can be analyzed based on the user queries or needs.
Highlights:

- Load the data into HDFS
- Analyze the data using Apache Pig or Hive
- Based on user queries give the right output

Intellipaat Job Assistance Program

Intellipaat is offering comprehensive job assistance to all the learners who have successfully completed the training. A learner will be considered to have successfully completed the training if he/she finishes all the exercises, case studies, projects and gets a minimum of 60% marks in the Intellipaat qualifying exam.
Intellipaat has exclusive tie-ups with over 80 MNCs for placement. All the resumes of eligible candidates will be forwarded to the Intellipaat job assistance partners. Once there is a relevant opening in any of the companies, you will get a call directly for the job interview from that particular company.

Frequently Asked Questions:

Q 1. What is the criterion for availing the Intellipaat job assistance program?

Ans. All Intellipaat learners who have successfully completed the training post April 2017 are directly eligible for the Intellipaat job assistance program.

Q 2. Which are the companies that I can get placed in?

Ans. We have exclusive tie-ups with MNCs like **Ericsson, Cisco, Cognizant, Sony, Mu Sigma, Saint-Gobain, Standard Chartered, TCS, Genpact, Hexaware**, and more. So you have the opportunity to get placed in these top global companies.

Q 3. Does Intellipaat help learners to crack the job interviews?

Ans. Intellipaat has an exclusive section which includes the top interview questions asked in top MNCs for most of the technologies and tools for which we provide training. Other than that our support and technical team can also help you in this regard.

Q 4. Do I need to have prior industry experience for getting an interview call?

Ans. There is no need to have any prior industry experience for getting an interview call. In fact, the successful completion of the Intellipaat certification training is equivalent to six months of industry experience. This is definitely an added advantage when you are attending an interview.

Q 5. What is the job location that I will get?

Ans. Intellipaat will try to get you a job in your same location provided such a vacancy exists in that location.

Q 6. Which is the domain that I will get placed in?

Ans. Depending on the Intellipaat certification training you have successfully completed, you will be placed in the same domain.

Q 7. Is there any fee for the Intellipaat placement assistance?

Ans. Intellipaat does not charge any fees as part of the placement assistance program.

Q 8. If I don't get a job in the first attempt, can I get another chance?

Ans. Definitely, yes. Your resume will be in our database and we will circulate it to our MNC partners until you get a job. So there is no upper limit to the number of job interviews you can attend.

Q 9. Does Intellipaat guarantee a job through its job assistance program?

Ans. Intellipaat does not guarantee any job through the job assistance program. However, we will definitely offer you full assistance by circulating your resume among our affiliate partners.

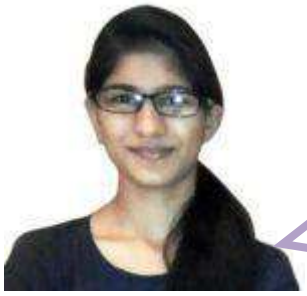
Q 10. What is the salary that I will be getting once I get the job?

Ans. Your salary will be directly commensurate with your abilities and the prevailing industry standards.

What makes us who we are?

"I am completely satisfied with the Intellipaat big data hadoop training. The trainer came with over a decade of industry experience. The entire big data online course was segmented into modules that were created with care so that the learning is complete and as per the industry needs."

-Bhuvana



"Full marks for the Intellipaat support team for providing excellent support services. Since Hadoop was new to me and I used to have many queries but the support team was very qualified and very patient in listening to my queries and resolve it to my highest expectations. The entire big data course was completely oriented towards the practical aspects."

- Bharati Jha