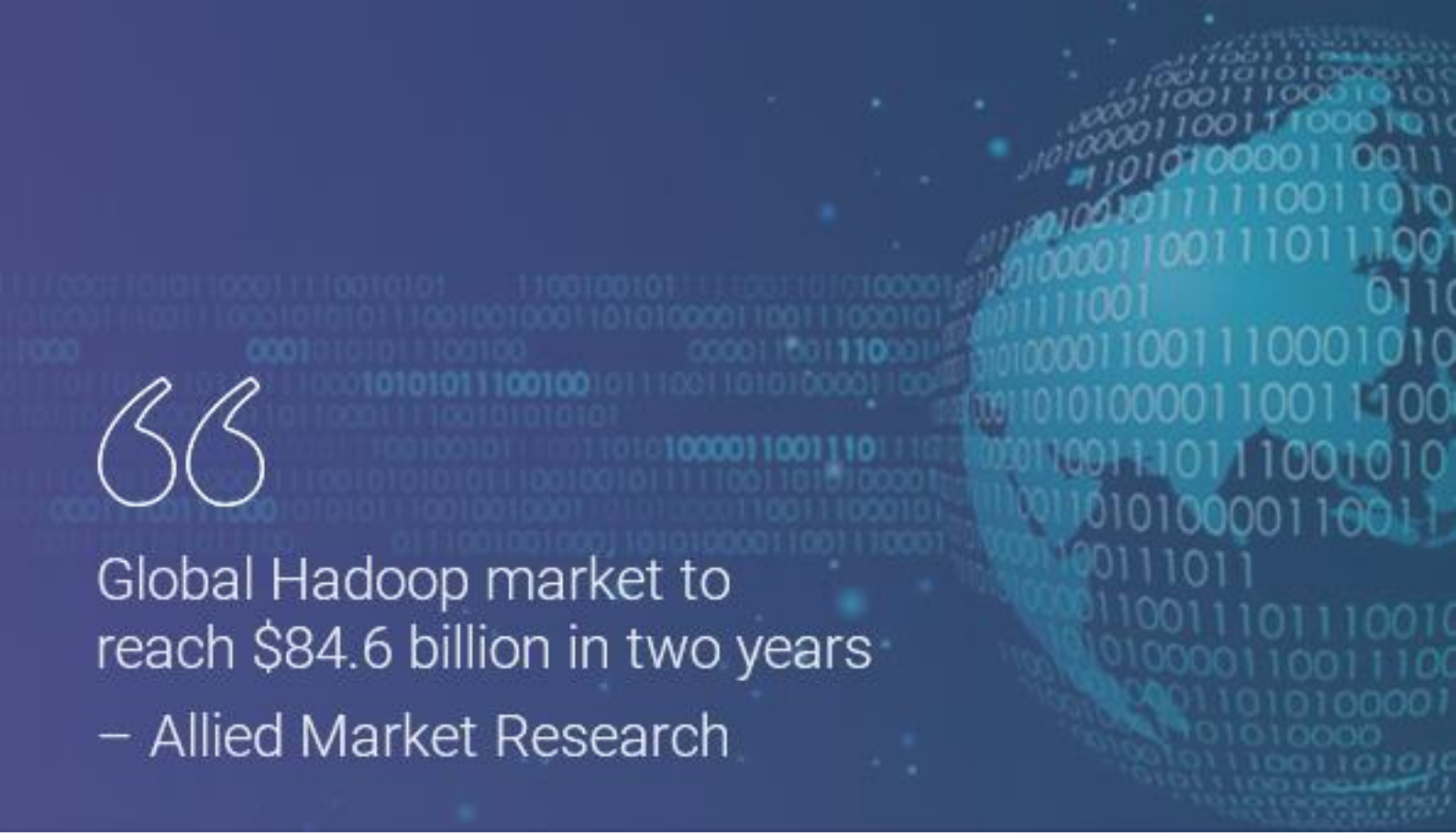
Big Data Hadoop & Spark

Certification Training

In Collaboration with **IBM**

IntelliPaat

Table of Contents



About the Program

Intellipaat's Big Data Hadoop training program helps you master Big Data Hadoop and Spark to get ready for the Cloudera CCA Spark and Hadoop Developer Certification (CCA175) exam, as well as to master Hadoop Administration, through 14 real-time industry-oriented case-study projects. In this Big Data course, you will master MapReduce, Hive, Pig, Sqoop, Oozie, and Flume and work with Amazon EC2 for cluster setup, Spark framework and RDDs, Scala and Spark SQL, Machine Learning using Spark, Spark Streaming, etc.

IBM

Collaborating with IBM

IBM is one of the leading innovators and the biggest player in creating innovative tools for Big Data Analytical tools. Top subject matter experts from IBM will share knowledge in the domain of Analytics and Big Data through this training program that will help you gain the breadth of knowledge and industry experience.

Benefits for students from IBM

- Industry-recognized IBM certificate
- Access to IBM Watson for hands-on training and practice
- Industry in-line case studies and project work

About Intellipaat

Intellipaat is one of the leading e-learning training providers with more than 600,000 learners across 55+ countries. We are on a mission to democratize education as we believe that everyone has the right to quality education.

Our courses are delivered by subject matter experts from top MNCs, and our world-class pedagogy enables learners to quickly learn difficult topics in no time. Our 24/7 technical support and career services will help them jump-start their careers in their dream companies.

Key Features

60 HRS INSTRUCTOR-LED TRAINING

80 HRS SELF-PACED TRAINING

120 HRS REAL-TIME PROJECT WORK

LIFETIME ACCESS

24/7 TECHNICAL SUPPORT

INDUSTRY-RECOGNIZED CERTIFICATION

JOB ASSISTANCE THROUGH 400+ CORPORATE TIE-UPS

FLEXIBLE SCHEDULING

Career Support

SESSIONS WITH INDUSTRY MENTORS

Attend sessions from top industry experts and get guidance on how to boost your career growth

MOCK INTERVIEWS

Mock interviews to make you prepare for cracking interviews by top employers

RESUME PREPARATION

Get assistance in creating a world-class resume from our career services team

Why take up this course?

- Global Hadoop market to reach US$84.6 billion in 2 years – Allied Market Research
- The number of jobs for all the US data professionals will increase to 2.7 million per year – IBM
- A Hadoop Administrator in the United States can get a salary of US$123,000 – Indeed

Big Data is the fastest growing and the most promising technology for handling large volumes of data for doing Data Analytics. This Big Data Hadoop training will help you be up and running in the most demanding professional skills. Almost all top MNCs 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 learn Big Data and upgrade your career in the domain.

Who should take up this course?

- Programming Developers and System Administrators
- Experienced working professionals and Project Managers
- Big Data Hadoop Developers eager to learn other verticals such as testing, analytics, and administration
- Mainframe Professionals, Architects, and Testing Professionals
- Business Intelligence, Data Warehousing, and Analytics Professionals
- Graduates and undergraduates eager to learn Big Data

Program Curriculum

BIG DATA HADOOP COURSE CONTENT

1. **HADOOP INSTALLATION AND SETUP**

   1.1 The architecture of Hadoop cluster

   1.2 What is high availability and federation?

   1.3 How to setup a production cluster?

   1.4 Various shell commands in Hadoop

   1.5 Understanding configuration files in Hadoop

   1.6 Installing a single node cluster with Cloudera Manager

   1.7 Understanding Spark, Scala, Sqoop, Pig, and Flume

2. **Introduction to Big Data Hadoop and Understanding HDFS and MapReduce**

   2.1 Introducing Big Data and Hadoop

   2.2 What is Big Data, and where does Hadoop fit in?

   2.3 Two important Hadoop ecosystem components, namely, MapReduce and HDFS

   2.4 In-depth Hadoop Distributed File System – Replications, Block Size, Secondary Name node, High Availability and in-depth YARN – resource manager and node manager

   Hands-on Exercise: *HDFS working mechanism, data replication process, how to determine the size of a block, and understanding a DataNode and a NameNode*

3. **DEEP DIVE IN MAPREDUCE**

   3.1 Learning the working mechanism of MapReduce

   3.2 Understanding the mapping and reducing stages in MR

   3.3 Various terminology in MR such as input format, output format, partitioners, combiners, shuffle, and sort

   Hands-on Exercise: *How to write a WordCount program in MapReduce? How to write a Custom Partitioner? What is a MapReduce Combiner? How to run a job in a local job*

runner? Deploying a unit test, What is a map-side join and reduce-side join?, What is a tool runner? How to use counters, and dataset joining with map-side and reduce-side joins?

4. INTRODUCTION TO HIVE

4.1 Introducing Hadoop Hive

4.2 Detailed architecture of Hive

4.3 Comparing Hive with Pig and RDBMS

4.4 Working with Hive Query Language

4.5 Creation of a database, table, group by, and other clauses

4.6 Various types of Hive tables and HCatalog

4.7 Storing Hive results, Hive partitioning, and buckets

***Hands-on Exercise**: Database creation in Hive, dropping a database, Hive table creation, how to change a database, data loading, dropping and altering a table, pulling data by writing Hive queries with filter conditions, table partitioning in Hive, and using the Group by clause*

5. ADVANCED HIVE AND IMPALA

5.1 Indexing in Hive

5.2 The map-side join in Hive

5.3 Working with complex data types

5.4 The Hive user-defined functions

5.5 Introduction to Impala

5.6 Comparing Hive with Impala

5.7 The detailed architecture of Impala

***Hands-on Exercise:** How to work with Hive queries, the process of joining a table and writing indexes, external table and sequence table deployment, and data storage in a different table*

6. INTRODUCTION TO PIG

6.1 Apache Pig introduction and its various features

6.2 Various data types and schema in Pig

6.3 The available functions in Pig, Hive bags, tuples, and fields

Hands-on Exercise: *Working with Pig in MapReduce and in a local mode, loading of data, limiting data to four rows, storing the data into files, and working with group by, filter by, distinct, cross, and split*

7. FLUME, SQOOP, AND HBASE

7.1 Apache Sqoop introduction

7.2 Importing and exporting data

7.3 Performance improvement with Sqoop

7.4 Sqoop limitations

7.5 Introduction to Flume and understanding the architecture of Flume

7.6 What are HBase and the CAP theorem?

Hands-on Exercise: *Working with Flume for generating a sequence number and consuming it, using Flume Agent to consume Twitter data, using AVRO to create a Hive table, AVRO with Pig, creating a table in HBase, and deploying Disable, Scan, and Enable table functions*

8. WRITING SPARK APPLICATIONS USING SCALA

8.1 Using Scala for writing Apache Spark applications

8.2 Detailed study of Scala

8.3 The need for Scala

8.4 The concept of object-oriented programming

8.5 Executing the Scala code

8.6 Various classes in Scala such as getters, setters, constructors, abstract, extending objects, and overriding methods

8.7 The Java and Scala interoperability

8.8 The concept of functional programming and anonymous functions

8.9 Bobsrockets package and comparing the mutable and immutable collections

8.10 Scala REPL, lazy values, control structures in Scala, directed acyclic graph (DAG), first Spark application using SBT/Eclipse, Spark Web UI, and Spark in Hadoop ecosystem

Hands-on Exercise*: Writing a Spark application using Scala and understanding the robustness of Scala for the Spark real-time analytics operation*

9. SPARK FRAMEWORK

9.1 Detailed Apache Spark and its various features

9.2 Comparing with Hadoop

9.3 Various Spark components

9.4 Combining HDFS with Spark and Scalding

9.5 Introduction to Scala

9.6 Importance of Scala and RDDs

Hands-on Exercise: *The resilient distributed dataset (RDD) in Spark, How does it help speed up Big Data processing?*

10. RDDS IN SPARK

10.1 Understanding Spark RDD operations

10.2 Comparison of Spark with MapReduce

10.3 What is a Spark transformation?

10.4 Loading data in Spark

10.5 Types of RDD operations, viz. transformation and action

10.6 What is a Key/Value pair?

Hands-on Exercise: *How to deploy RDDs with HDFS?, Using the in-memory dataset, using file for RDDs, how to define the base RDD from an external file? Deploying RDDs via transformation, using the Map and Reduce functions, and working on word count and count log severity*

11. DATAFRAMES AND SPARK SQL

11.1 The detailed Spark SQL

11.2 The significance of SQL in Spark for working with structured data processing

11.3 Spark SQL JSON support

11.4 Working with XML data and parquet files

11.5 Creating Hive Context

11.6 Writing a DataFrame to Hive

11.7 How to read a JDBC file?

11.8 Significance of a Spark DataFrames

11.9 How to create a DataFrame?

11.10 What is schema manual inferring?

11.11 Working with CSV files, JDBC table reading, data conversion from a DataFrame to JDBC, Spark SQL user-defined functions, shared variable, and accumulators

11.12 How to query and transform data in DataFrames?

11.13 How a DataFrame provides the benefits of both Spark RDDs and Spark SQL

11.14 Deploying Hive on Spark as the execution engine

Hands-on Exercise: *Data querying and transformation using DataFrames and finding out the benefits of DataFrames over Spark SQL and Spark RDDs*

12. MACHINE LEARNING USING SPARK (MLLIB)

12.1 Introduction to Spark MLlib

12.2 Understanding various algorithms

12.3 What is Spark iterative algorithm?

12.4 Spark graph processing analysis

12.5 Introducing Machine Learning

12.6 K-means clustering

12.7 Spark variables like shared and broadcast variables

12.8 What are accumulators?

12.9 Various ML algorithms supported by MLlib

12.10 Linear regression, logistic regression, decision tree, random forest, and k-means clustering techniques

Hands-on Exercise: *Building a recommendation engine*

13. INTEGRATING APACHE FLUME AND APACHE KAFKA

13.1 Why Kafka?

13.2 What is Kafka?

13.3 Kafka architecture

13.4 Kafka workflow

13.5 Configuring Kafka cluster

13.6 Basic operations

13.7 Kafka monitoring tools

13.8 Integrating Apache Flume and Apache Kafka

Hands-on Exercise: *Configuring single node single broker cluster, configuring single node multi broker cluster, producing and consuming messages, and integrating Apache Flume and Apache Kafka*

14. SPARK STREAMING

14.1 Introduction to Spark Streaming
14.2 The architecture of Spark Streaming
14.3 Working with the Spark Streaming program
14.4 Processing data using Spark Streaming
14.5 Requesting count and DStream
14.6 Multi-batch and sliding window operations
14.7 Working with advanced data sources
14.8 Features of Spark Streaming
14.9 Spark Streaming workflow
14.10 Initializing StreamingContext
14.11 Discretized Streams (DStreams)
14.12 Input DStreams and Receivers
14.13 Transformations on DStreams
14.14 Output operations on DStreams
14.15 Windowed operators and its uses
14.16 Important windowed operators and stateful operators

Hands-on Exercise: *Twitter Sentiment Analysis, streaming using netcat server, Kafka–Spark Streaming, and Spark–Flume Streaming*

15. HADOOP ADMINISTRATION – MULTI - NODE CLUSTER SETUP USING AMAZON EC2

15.1 Create a 4-node Hadoop cluster setup
15.2 Running the MapReduce Jobs on the Hadoop cluster
15.3 Successfully running the MapReduce code
15.4 Working with the Cloudera Manager setup

Hands-on Exercise: *Building a multi-node Hadoop cluster using an Amazon EC2 instance and Working with the Cloudera Manager*

16. HADOOP ADMINISTRATION – CLUSTER CONFIGURATION

16.1 Overview of Hadoop configuration

16.2 The importance of Hadoop configuration file

16.3 The various parameters and values of configuration

16.4 HDFS parameters and MapReduce parameters

16.5 Setting up the Hadoop environment

16.6 Include and exclude configuration files

16.7 The administration and maintenance of NameNode, DataNode, directory structures, and files

16.8 What is a File system image?

16.9 Understanding the edit log

Hands-on Exercise: *The process of performance tuning in MapReduce*

17. HADOOP ADMINISTRATION: MAINTENANCE, MONITORING, AND TROUBLESHOOTING

17.1 Introduction to the checkpoint procedure, NameNode failure

17.2 How to ensure the recovery procedure, safe mode, metadata and data backup, various potential problems and solutions, and what to look for and how to add and remove nodes

Hands-on Exercise: *How to go about ensuring the MapReduce File System Recovery for 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, and Getting to know the Fair Scheduler and its configuration*

18. ETL CONNECTIVITY WITH HADOOP ECOSYSTEM (SELF-PACED)

18.1 How do ETL tools work in Big Data industry?

18.2 Introduction to ETL and data warehousing

18.3 Working with prominent use cases of Big Data in the ETL industry

18.4 End-to-end ETL PoC showing Big Data integration with the ETL tool

Hands-on Exercise: *Connecting to HDFS from the ETL tool, moving data from a local system to HDFS, moving data from DBMS to HDFS, working with Hive with the ETL tool, and creating a MapReduce job in the ETL tool*

19. PROJECT SOLUTION DISCUSSION AND CLOUDERA CERTIFICATION TIPS AND TRICKS

19.1 Working toward the solution of the Hadoop project

19.2 Its problem statements and the possible solution outcomes

19.3 Preparing for the Cloudera certifications

19.4 Points to focus on scoring highest marks

19.5 Tips for cracking Hadoop interview questions

Hands-on Exercise: *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*

Following topics will be available only in self-paced mode:

20. HADOOP APPLICATION TESTING

20.1 Importance of testing

20.2 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, and release testing

21. ROLES AND RESPONSIBILITIES OF HADOOP TESTING PROFESSIONALS

21.1 Understanding the requirement

21.2 Preparation of the testing estimation

21.3 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, and 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.), reporting defects to the development team or manager, and driving them to closure

21.4 Consolidating all the defects and creating defect reports

21.5 Validating new features and issues in Core Hadoop

22. FRAMEWORK CALLED MRUNIT FOR TESTING OF MAPREDUCE PROGRAMS

22.1 Report defects to the development team or manager, and drive them to closure

22.2 Consolidate all the defects and create defect reports

22.3 Create a testing framework called MRUnit for testing of MapReduce programs

23. UNIT TESTING

23.1 Automation testing using the Oozie

23.2 Data validation using the query surge tool

24. TEST EXECUTION

24.1 Test plan for HDFS upgrade

24.2 Test automation and result

25. TEST PLAN STRATEGY AND WRITING TEST CASES FOR TESTING HADOOP APPLICATION

25.1 Test, install, and configure test cases

Big Data Hadoop Course Projects

Working with MapReduce, Hive, and Sqoop

In this project, you will successfully import data using Sqoop into HDFS for data analysis. The transfer will be done via Sqoop data transfer from RDBMS to Hadoop. You will code in the Hive query language and carry out data querying and analysis. You will acquire an understanding of Hive and Sqoop after the completion of this project.

Work on MovieLens Data for Finding the Top Movies

You will create the top-ten-movies list using the MovieLens data. For this project, you will use the MapReduce program to work on the data file, Apache Pig to analyze

data, and Apache Hive data warehousing and querying. You will be working with distributed datasets.

Hadoop YARN Project: End-to-End PoC

Bring the daily incremental data into the Hadoop Distributed File System. As part of the project, you will be using Sqoop commands to bring the data into HDFS, working with the end-to-end flow of transaction data, and the data from HDFS. You will work on a live Hadoop YARN cluster. You will also work on the YARN central resource manager.

Table Partitioning in Hive

In this project, you will learn how to improve the query speed using Hive data partitioning. You will get hands-on experience in partitioning Hive tables manually, deploying single SQL execution in dynamic partitioning, and bucketing of data to break it into manageable chunks.

Connecting Pentaho with Hadoop Ecosystem

You will deploy ETL for data analysis activities. In this project, you will challenge your working knowledge of ETL and Business Intelligence. You will configure Pentaho to work with Hadoop distribution and load, transform, and extract data into the Hadoop cluster.

Multi-node Cluster Setup

You will set up a Hadoop real-time cluster on Amazon EC2. The project will involve installing and configuring Hadoop. You will need to run a Hadoop multi-node using a 4-node cluster on Amazon EC2 and deploy a MapReduce job on the Hadoop cluster. Java will need to be installed as a prerequisite for running Hadoop.

Hadoop Testing Using MRUnit

In this project, you will be required to test MapReduce applications. You will write JUnit tests using MRUnit for MapReduce applications. You will also be doing mock static methods using PowerMock and Mockito and implementing MapReduce Driver for testing the map and reduce pair.

Hadoop Web Log Analytics

In this project, you will derive insights from web log data. The project involves the aggregation of the log data, implementation of Apache Flume for data transportation, and processing of data and generating analytics. You will learn to use workflow and do data cleansing using MapReduce, Pig, or Spark.

Hadoop Maintenance

Through this project, you will learn how to administer a Hadoop cluster for maintaining and managing it. You will be working with the NameNode directory structure, audit logging, DataNode block scanner, balancer, failover, fencing, DISTCP, and Hadoop file formats.

Twitter Sentiment Analysis

In this project, you will find out what is the reaction of the people to the demonetization move by India by analyzing their tweets. You will have to download the tweets, load them into Pig storage, divide the tweets into words to calculate sentiment, rate the words from +5 to −5 on the AFFIN dictionary, filter them, and then, analyze sentiment.

Analyzing IPL T20 Cricket

This project will require you to analyze an entire cricket match and get any details of it. You will need to load the IPL dataset into HDFS. You will then analyze the data using Apache Pig or Hive. Based on the user queries, the system will have to give the right output.

Movie Recommendation

In this project, you need to recommend the most appropriate movie to a user based on his taste. This is a hands-on Apache Spark project, which will include performing collaborative filtering, regression, clustering, and dimensionality reduction. You will need to make use of the Apache Spark MLlib component and statistical analysis.

Twitter API Integration for Tweet Analysis

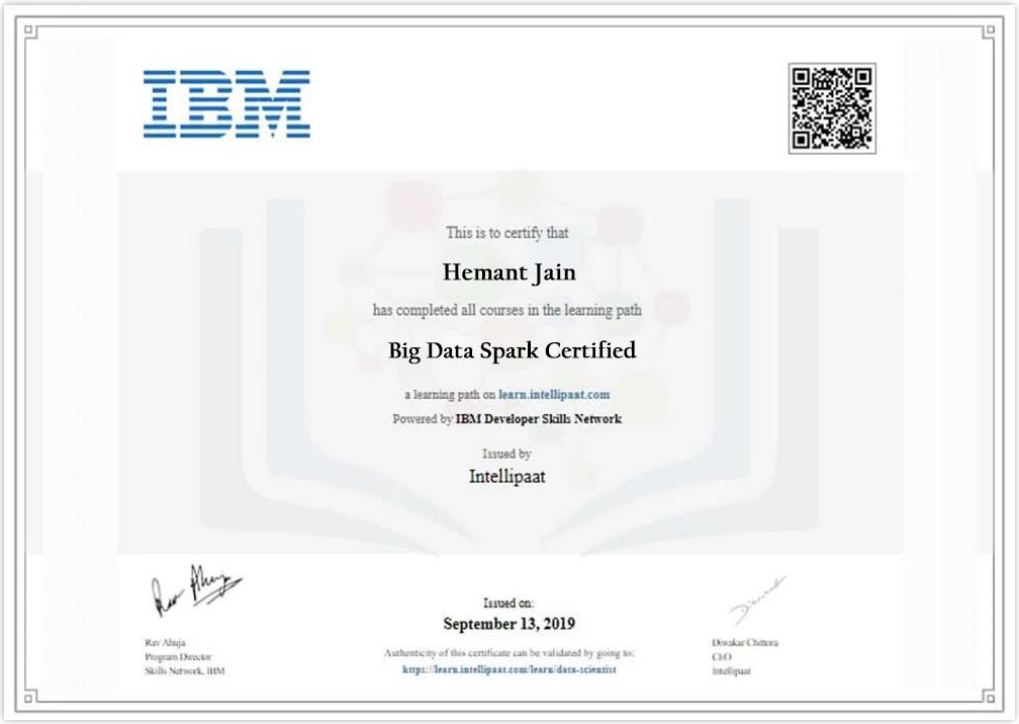
Here, you will analyze the user sentiment based on a tweet. In this Twitter analysis project, you will integrate the Twitter API and use Python or PHP for developing the essential server-side codes. You will carry out filtering, parsing, and aggregation, depending on the tweet analysis requirement.

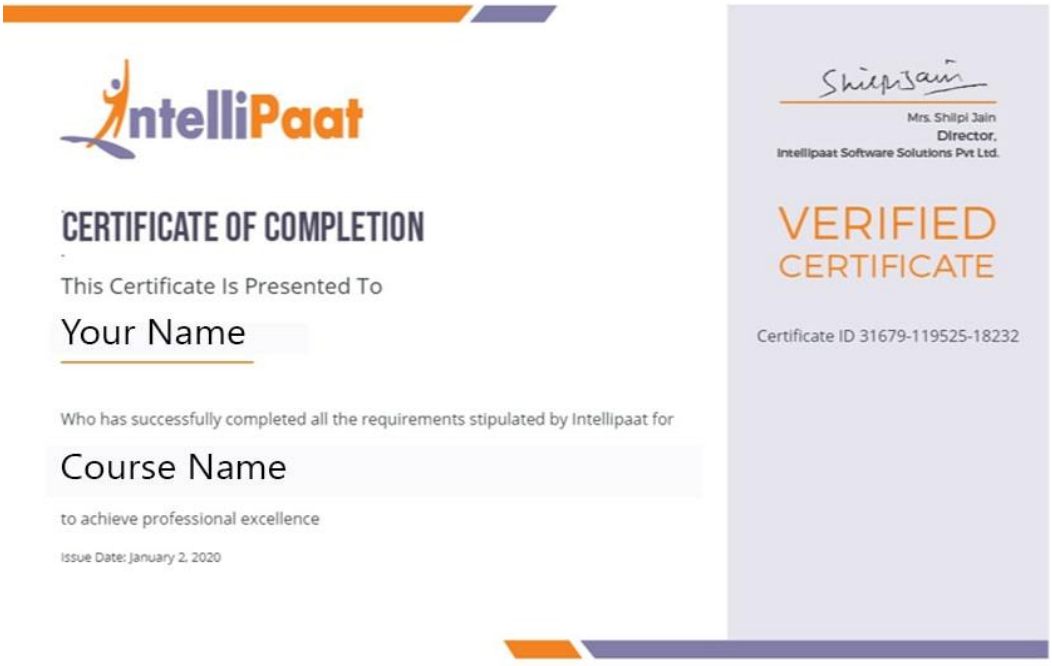
Data Exploration Using Spark SQL – Wikipedia Dataset

In this project, you will be making use of the Spark SQL tool for analyzing the Wikipedia dataset. You will be integrating Spark SQL for batch analysis, working with Machine Learning, visualizing, and processing data and ETL processes, along with real-time analysis of data.

Certification

After the completion of the course, you will get a certificate from IBM.





Success Stories

Joel Bassa

I am really thankful to Intellipaat for the Hadoop Architect course with Big Data certification. First of all, the team supported me in finding the best Big Data online course based on my experiences and current assignment. institution.

Kevin K Wada

Thank you very much for your top-class service. A special mention should be made for your patience in listening to my queries and giving me a solution, which was exactly what I was looking for. I am giving you a 10 on 10!

Sampson Basoah

The Intellipaat team helped me in selecting the perfect course that suits my profile. The whole course was practically oriented, and the trainers are always ready to answer any question. I found this course to be impactful. Thank you.

Paschal Ositadima

This regards to conveying my deepest gratitude to Intellipaat. The quality and methodology of this online Hadoop training were matchless. The self-study program for Big Data Hadoop training, for which I had enrolled, ticked off all the right boxes. I had access to free tutorials and videos to help me in my learning endeavor. A special mention must be made regarding the promptness and enthusiasm that Intellipaat showed when it comes to query resolution and doubt clearance. Kudos!

Rich Baker

Intellipaat's Hadoop tutorial delivered more than what they had promised to me. Since I have undergone a previous Hadoop training course, I was quite familiar with Big Data Hadoop concepts, but Intellipaat took it to a different level with their attention to detail and Hadoop domain expertise. I recommend this training

to everybody. You will learn everything from basic Hadoop concepts to advanced Big Data technology deployment. I am more than satisfied with this training. Thank you, Intellipaat!

CONTACT US

INTELLIPAAT SOFTWARE SOLUTIONS PVT. LTD.

Bangalore

AMR Tech Park 3, Ground Floor, Tower B,
Hongasandra Village, Bommanahalli,
Hosur Road, Bangalore – 560068

USA

1219 E. Hillsdale Blvd. Suite 205,
Foster City, CA 94404

If you have any further queries or just want to have a conversation with us, then do call us.

IND: +91-7022374614 | US: 1-800-216-8930