

# HADOOP COURSE CURRICULUM

## **Module 1: Introduction to Big Data**

**Learning Objectives** - In this module, you will understand about Big Data growth, the limitations of the existing solutions for Big Data problem, how Hadoop solves the Big Data problem, the common Hadoop ecosystem components, Hadoop Architecture, HDFS, how MapReduce Framework works.

**Topics** - What is Big Data and where it is produced? Rise of Big Data, Compare Hadoop vs traditional systems, Limitations and Solutions of existing Data Analytics Architecture, Attributes of Big Data, Types of data, other technologies vs Big Data.

## **Module 2: Hadoop Architecture and HDFS**

**Learning Objectives** - In this module, you will learn the Hadoop Cluster Architecture, Important Configuration files in a Hadoop Cluster, Data Loading Techniques, how to setup single node Hadoop cluster.

**Topics** - What is Hadoop? Hadoop History and Creation, Distributing Processing System, Core Components of Hadoop, HDFS Architecture, Hadoop Master – Slave Architecture, Daemon types - Learn Name node, Data node, Secondary Name node.

## **Module 3: Hadoop Configuration and installation**

**Learning Objectives** - In this module, you will installation of VM player and Hadoop, Important Configuration files in a Hadoop Cluster, Linux commands, Importing Hadoop Jars, Data Loading Techniques.

**Topics** - Installation of Hadoop, Hadoop terminal Commands, Hadoop Configuration Files, Hadoop Jars Import, Work with Cloudera version, VM player Installation.

## **Module 4: Hadoop Clusters and the Hadoop Ecosystem**

**Learning Objectives** - In this module, you will setup a single node cluster, Types of Distributed modes, Basics of Hadoop Ecosystem and Hadoop Architecture.

**Topics** - What is Hadoop Cluster? Pseudo Distributed mode, Type of clusters, Hadoop Ecosystem, Pig, Hive, Oozie, Flume, SQOOP.

# HADOOP COURSE CURRICULUM

## **Module 5: Hadoop MapReduce Framework**

**Learning Objectives** - In this module, you will learn the important processing tool in Hadoop with programming language like written in Java, Learn about the master – slave nodes.

**Topics** - Overview of MapReduce Framework, MapReduce Architecture, Learn about Job tracker and Task tracker, Use cases of MapReduce, Anatomy of MapReduce Program.

## **Module 6: MapReduce programs in Java**

**Learning Objectives** - In this module, you will understand the different classes in Java, Creating a Jar, Executing the program with data set, transferring the results into new File in HDFS.

**Topics** - Basic MapReduce API Concepts, Writing MapReduce Driver, Mappers, and Reducers in Java, Speeding up Hadoop Development by Using Eclipse, Unit Testing MapReduce Programs, and Demo on word count example.

## **Module 7: Advance MapReduce**

**Learning Objectives** - In this module, you will understand concepts like Input Splits in MapReduce, Combiner & Partitioners and Demos on MapReduce using different data sets, Comparison of MapReduce with Traditional ways.

**Topics** - Traditional way Vs MapReduce way, Combiners and Partitions, Joining datasets, Distributed Joins, Demo on Weather Dataset.

## **Module 8: Hive and HiveQL**

**Learning Objectives** - This module will help you in understanding the SQL concepts in Hadoop, Differences between SQL and Hive Hive concepts, Hive Data types, loading and Querying Data in Hive, running hive scripts and Hive UDF.

**Topics** - What is Hive?, Hive vs MapReduce, Hive DDL – Create/Show/Drop Tables, Internal and External Tables, Hive DML – Load Files & Insert Data, Hive Architecture & Components, Difference between Hive and RDBMS, Programming structure in UDF, Partitions and Buckets, Limitations of Hive.

# HADOOP COURSE CURRICULUM

## Module 9: PIG

**Learning Objectives** - In this module, you will learn the processing tool with scripting language, Pig, types of use case we can use Pig, tight coupling between Pig and MapReduce, and Pig Latin scripting, PIG running modes, PIG UDF, Pig Streaming, Testing PIG Scripts.

**Topics** - PIG vs MapReduce, PIG Architecture & Data types, Shell and Utility components, PIG Latin Relational Operators, PIG Latin: File Loaders and UDF, Programming structure in UDF, PIG Jars Import, limitations of PIG.

## Module 10: Apache SQOOP, Flume

**Learning Objectives** – the module will help you to understand working of multiple Hadoop ecosystem components together in a Hadoop implementation to solve Big Data problems, importing the data from databases and web log applications with different tools.

**Topics** - Why and what is SQOOP? SQOOP Architecture, Benefits of SQOOP, Importing Data Using SQOOP, Apache Flume Introduction, Flume Model and Goals, Features of Flume, Flume Use Case.

## Module 11: NoSQL Databases

**Learning Objectives** - This module will help you to understand the concepts of NoSQL Databases, HBase in Hadoop, HBase Region Servers, Column Families, One of the most demanding NoSQL in the market – MongoDB, Features of MongoDB.

**Topics** - What is HBase? HBase Architecture, HBase Components, Storage Model of HBase, HBase vs RDBMS, Introduction to Mongo DB, CRUD, Advantages of MongoDB over RDBMS, Use case.

## Module 12: Oozie and Zookeeper

**Learning Objectives** - In this module, you will understand the Oozie concept, Scheduling with Oozie, Zookeeper uses, Use of Zookeeper to HBase.

**Topics** - Oozie – Simple/Complex Flow, Oozie Workflow, Oozie Components, Demo on Oozie Workflow in XML, What is Zookeeper? Features of Zookeeper, Zookeeper Data Model.

# HADOOP COURSE CURRICULUM

## Module 13: Projects

**Learning Objectives** - In this module, you will work on multiple Hadoop ecosystem components together in a Hadoop implementation to solve the project, Real time data analyzing.

**Topics** - Demo of two sample projects, Analyzing Twitter Data with Apache Hadoop, Conclusion.

