# Big Data and Hadoop

**About The Course**

Become a Hadoop Expert by mastering MapReduce, Yarn, Pig, Hive, HBase, Oozie, Flume and Sqoop while working on industry based Use-cases and Projects. Also get an overview of Apache Spark for distributed data processing.

## Module 1

## Understanding Big Data and Hadoop

Learning Objectives - In this module, you will understand Big Data, 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, Anatomy of File Write and Read, how MapReduce Framework works

### Topics

✓ Big Data

✓ Limitations and Solutions of existing Data Analytics Architecture

✓ Hadoop

✓ Hadoop Features

✓ Hadoop Ecosystem

✓ Hadoop 2.x core components

✓ Hadoop Storage: HDFS, Hadoop Processing: MapReduce Framework

✓ Hadoop Different Distributions

## 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 and multi node Hadoop cluster

### Topics

✓ Hadoop 2.x Cluster Architecture - Federation and High Availability

✓ A Typical Production Hadoop Cluster

✓ Hadoop Cluster Modes

✓ Common Hadoop Shell Commands

✓ Hadoop 2.x Configuration Files

✓ Single node cluster

✓ Multi node cluster set up Hadoop Administration

## Module 3

## Hadoop MapReduce Framework

Learning Objectives - In this module, you will understand Hadoop MapReduce framework and the working of MapReduce on data stored in HDFS. You will understand concepts like Input Splits in MapReduce, Combiner & Partitioner and Demos on MapReduce using different data sets.

### Topics

✓ MapReduce Use Cases

✓ Traditional way Vs MapReduce way

✓ Why MapReduce

✓ Hadoop 2.x MapReduce Architecture

✓ Hadoop 2.x MapReduce Components ✓ YARN MR Application Execution Flow ✓ YARN Workflow, Anatomy of MapReduce Program

✓ Demo on MapReduce. Input Splits, Relation between Input Splits and HDFS Blocks

✓ MapReduce: Combiner & Partitioner

✓ Demo on de-identifying Health Care Data set

✓ Demo on Weather Data set.

## Module 4

## Advanced MapReduce

Learning Objectives - In this module, you will learn Advanced MapReduce concepts such as Counters, Distributed Cache, MRunit, Reduce Join, Custom Input Format, Sequence Input Format and XML parsing.

### Topics

✓ Counters

✓ Distributed Cache

✓ MRunit

✓ Reduce Join

✓ Custom Input Format

✓ Sequence Input Format

✓ Xml file Parsing using MapReduce

## Module 5

## Module 6

### Pig

Learning Objectives - In this module, you will learn 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. Demo on healthcare dataset.

**Topics**

- ✓ About Pig
- ✓ MapReduce Vs Pig
- ✓ Pig Use Cases
- ✓ Programming Structure in Pig
- ✓ Pig Running Modes
- ✓ Pig components
- ✓ Pig Execution
- ✓ Pig Latin Program
- ✓ Data Models in Pig
- ✓ Pig Data Types
- ✓ Shell and Utility Commands
- ✓ Pig Latin : Relational Operators File Loaders, Group Operator, COGROUP Operator, Joins and COGROUP, Union, Diagnostic Operators, Specialized joins in Pig
- ✓ Built In Functions ( Eval Function, Load and Store Functions, Math function, String Function, Date Function, Pig UDF, Piggybank
- ✓ Parameter Substitution ( PIG macros and Pig Parameter substitution )
- ✓ Pig Streaming
- ✓ Testing Pig scripts with Punit
- ✓ Aviation use case in PIG, Pig Demo on Healthcare Data set.

### Hive

Learning Objectives - This module will help you in understanding Hive concepts, Hive Data types, loading and Querying Data in Hive, running hive scripts and Hive UDF.

**Topics**

- ✓ Hive Background
- ✓ Hive Use Case
- ✓ About Hive ✓ Hive Vs Pig
- ✓ Hive Architecture and Components
- ✓ Metastore in Hive
- ✓ Limitations of Hive
- ✓ Comparison with Traditional Database
- ✓ Hive Data Types and Data Models
- ✓ Partitions and Buckets
- ✓ Hive Tables(Managed Tables and External Tables)
- ✓ Importing Data
- ✓ Querying Data
- ✓ Managing Outputs
- ✓ Hive Script
- ✓ Hive UDF
- ✓ Retail use case in Hive
- ✓ Hive Demo on Healthcare Data set.

### Module 7
### Advanced Hive and HBase

Learning Objectives - In this module, you will understand Advanced Hive concepts such as UDF, Dynamic Partitioning, Hive indexes and views, optimizations in hive. You will also acquire in-depth knowledge of HBase, HBase Architecture, running modes and its components.

**Topics**

- ✓ Hive QL: Joining Tables, Dynamic Partitioning, Custom Map/Reduce Scripts
- ✓ Hive Indexes and views
- ✓ Hive query optimizers
- ✓ Hive : Thrift Server, User Defined Functions
- ✓ HBase: Introduction to NoSQL Databases and HBase, HBase v/s RDBMS, HBase Components, HBase Architecture, HBase Cluster Deployment.

## Module 8
## Advance HBase

Learning Objectives - This module wil HBase concepts. We will see demo Filters. You will also learn what Zooke how it helps in monitoring a cluster Zookeeper.

**Topics**

- ✓ HBase Data Model
- ✓ HBase Shell
- ✓ HBase Client API
- ✓ Data Loading Techniques
- ✓ ZooKeeper Data Model
- ✓ Zookeeper Service
- ✓ Zookeeper
- ✓ Demos on Bulk Loading
- ✓ Getting and Inserting Data ✓ Filters in HBase

## Module 9
## Processing Distributed Data with Apa

Learning Objectives - In this module yo ecosystem and its components, how Spark, SparkContext. You will learn ho in Spark. Demo will be there on runni Spark Cluster, Comparing performanc and Spark.

**Topics**

- ✓ What is Apache Spark
- ✓ Spark Ecosystem
- ✓ Spark Components
- ✓ History of Spark and Spark Ver Spark a Polyglot ✓ What is Scala?
- ✓ Why Scala?
- ✓ SparkContext
- ✓ RDD

## Module 10
## Oozie and Hadoop Project

Learning Objectives - In this module, you will understand working of multiple Hadoop ecosystem components together in a Hadoop implementation to solve Big Data problems. We will discuss multiple data sets and specifications of the project. This module will also cover Flume & Sqoop demo, Apache Oozie Workflow Scheduler for Hadoop Jobs, and Hadoop Talend integration.

**Topics**

- ✓ Flume and Sqoop Demo
- ✓ Oozie
- ✓ Oozie Components
- ✓ Oozie Workflow
- ✓ Scheduling with Oozie
- ✓ Demo on Oozie Workflow
- ✓ Oozie Co-ordinator
- ✓ Oozie Commands
- ✓ Oozie Web Console
- ✓ Oozie for MapReduce ✓ PIG, Hive, and Sqoop,
- ✓ Combine flow of MR, PIG, Hive in Oozie
- ✓ Hadoop Project Demo
- ✓ Hadoop Integration with Talend

I so

,

w n

## Project Work

Towards the end of the course, you will be working on a live project where you will be using PIG, HIVE, HBase and MapReduce to perform Big Data analytics.
Here are the few Industry-wise Big Data case studies e.g. Finance, Retail, Media, and Aviation etc. which you can take up as your project work:

**Project #1: Analyze social bookmarking sites to find insights**
**Industry:** Social Media
**Data**: It comprises of the information gathered from sites like reddit.com, stumbleupon.com etc. which are bookmarking sites and allow you to bookmark, review, rate, search various links on any topic.reddit.com, stumbleupon.com, etc. A bookmarking site allows you to bookmark, review, and rate, search various links on any topic. The data is in XML format and contains various links/posts URL, categories defining it and the ratings linked with it. **Problem Statement**: Analyze the data in Hadoop Eco-system to:

1. Fetch the data into Hadoop Distributed File System and analyze it with the help of MapReduce, Pig and Hive to find the top rated links based on the user comments, likes etc.
2. Using MapReduce convert the semi-structured format (XML data) into structured format and categorize the user rating as positive and negative for each of the thousand links.
3. Push the output HDFS and then feed it into PIG, which splits the data into two parts: Category data and Ratings data.
4. Write a fancy Hive Query to analyze the data further and push the output is into relational database (RDBMS) using Sqoop.
5. Use a web server running on grails/java/ruby/python that renders the result in real time processing on a website.

**Project #2: Customer Complaints Analysis**
**Industry:** Retail
**Data**: Publicly available dataset, containing a few lakh observations with attributes like: Customer ID, Payment Mode, Product Details, Complaint, Location, Status of the complaint, etc.
**Problem Statement**: Analyze the data in Hadoop Eco-system to:
1. Get the number of complaints filed under each products
2. Get the total number of complaints filed from a particular location
3. Get the list of complaints grouped by location which has no timely response

**Project #3: Tourism Data Analysis**

**Industry:** Tourism

**Data**: The dataset comprises attributes like: City pair (Combination of from and to), Adults traveling, Seniors traveling, Children traveling, Air booking price, Car booking price, etc.

**Problem Statement**: Find the following insights from the data:

1. Top 20 destinations people travel most: Based on given data we can find the most popular destinations where people travel frequently, based on the specific initial number of trips booked for a particular destination
2. Top 20 locations from where most of the trips start based on booked trip count
3. Top 20 high air-revenue destinations i.e. which 20 cities generates high airline revenues for travel, so that the discount offers can be given to attract more bookings for these destinations

**Project #4: Airline Data Analysis**

**Industry:** Aviation

**Data**: Publicly available dataset which contains the flight details of various airlines like: Airport id, Name of the airport, Main city served by airport, Country or territory where airport is located, Code of Airport, Decimal degrees, Hours offset from UTC, Time zone, etc.

**Problem Statement**: Analyze the airlines data to:

1. Find list of Airports operating in the Country
2. Find the list of Airlines having zero stops
3. List of Airlines operating with code share
4. Which country (or) territory has the highest number of Airports
5. Find the list of Active Airlines in the United States

**Project #5: Analyze Loan Dataset**

**Industry:** Banking and Finance

**Data**: Publicly available dataset which contains complete details of all the loans issued, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information.

**Problem Statement**: Find the number of cases per location and categorize the count with respect to reason for taking loan and display the average risk score

**Project #6: Analyze Movie Ratings**

**Industry:** Media

**Data**: Publicly available data from sites like rotten tomatoes, imdb, etc.

**Problem Statement**: Analyze the movie ratings by different users to:

1. Get the user who has rated the most number of movies
2. Get the user who has rated the least number of movies
3. Get the count of total number of movies rated by user belonging to a specific occupation
4. Get the number of underage users

**Project #7: Analyze YouTube data**

**Industry:** Social Media

**Data**: It is about the YouTube videos and contains attributes like: Video ID, Uploader, Age, Category, Length, views, ratings, comments, etc.

**Problem Statement**: Identify out the top 5 categories in which the most number of videos are uploaded, the top 10 rated videos, the top 10 most viewed videos

Apart from these there are some twenty more use-cases to choose from: Market
data Analysis

Twitter Data Analysis

Big Data and Hadoop