edureka!
Discover Learning

Big Data Hadoop Certification Training

Course Curriculum : Your 0 module Learning Plan

https://www.edureka.co/big-data-hadoop-training-certification

About Edureka

Edureka is a leading e-learning platform providing live instructor-led interactive online training. We cater to professionals and students across the globe in categories like Big Data & Hadoop, Business Analytics, NoSQL Databases, Java & Mobile Technologies, System Engineering, Project Management and Programming. We have an easy and affordable learning solution that is accessible to millions of learners. With our students spread across countries like the US, India, UK, Canada, Singapore, Australia, Middle East, Brazil and many others, we have built a community of over 1 million learners across the globe.

About Course

Edureka's extensive Big Data Analytics certification is curated by Hadoop experts, and it covers in-depth knowledge on Big Data and Hadoop Ecosystem tools such as HDFS, YARN, MapReduce, Hive, and Pig. Throughout this online instructor-led Big Data Hadoop certification training, you will be working on real-life industry use cases in Retail, Social Media, Aviation, Tourism, and Finance domains using Edureka's Cloud Lab. Enroll now to learn Big Data from instructors with over 10+ years of experience, with hands-on demonstrations.

Curriculum

Understanding Big Data and Hadoop

Learning Objectives: In this module, you will understand what Big Data is, the limitations of the traditional solutions for Big Data problems, how Hadoop solves those Big Data problems, Hadoop Ecosystem, Hadoop Architecture, HDFS, Anatomy of File Read and Write & how MapReduce works.

Topics:

- Introduction to Big Data & Big Data Challenges
- Limitations & Solutions of Big Data Architecture
- Hadoop & its Features
- Hadoop Ecosystem
- Hadoop 2.x Core Components
- Hadoop Storage: HDFS (Hadoop Distributed File System)
- Hadoop Processing: MapReduce Framework
- Different Hadoop Distributions

Hadoop Architecture and HDFS

Learning Objectives: In this module, you will learn Hadoop Cluster Architecture, important configuration files of Hadoop Cluster, Data Loading Techniques using Sqoop & Flume, and how to setup Single Node and Multi-Node Hadoop Cluster.

Topics:

- Hadoop 2.x Cluster Architecture
- Federation and High Availability Architecture
- Typical Production Hadoop Cluster
- Hadoop Cluster Modes
- Common Hadoop Shell Commands
- Hadoop 2.x Configuration Files

- Single Node Cluster & Multi-Node Cluster set up
- Basic Hadoop Administration

Hadoop MapReduce Framework

Learning Objectives: In this module, you will understand Hadoop MapReduce framework comprehensively, the working of MapReduce on data stored in HDFS. You will also learn the advanced MapReduce concepts like Input Splits, Combiner & Partitioner.

Topics:

- Traditional way vs MapReduce way
- Why MapReduce
- YARN Components
- YARN Architecture
- YARN MapReduce Application Execution Flow
- YARN Workflow
- Anatomy of MapReduce Program
- Input Splits, Relation between Input Splits and HDFS Blocks
- MapReduce: Combiner & Partitioner
- Demo of Health Care Dataset
- Demo of Weather Dataset

Advanced Hadoop 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

Apache Pig

Learning Objectives: In this module, you will learn Apache Pig, types of use cases where we can use Pig, tight coupling between Pig and MapReduce, and Pig Latin scripting, Pig running modes, Pig UDF, Pig Streaming & Testing Pig Scripts. You will also be working on healthcare dataset.

Topics:

- Introduction to Apache Pig
- MapReduce vs Pig
- Pig Components & Pig Execution
- Pig Data Types & Data Models in Pig
- Pig Latin Programs
- Shell and Utility Commands
- Pig UDF & Pig Streaming
- Testing Pig scripts with Punit
- Aviation use-case in PIG
- Pig Demo of Healthcare Dataset

Apache 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:

- Introduction to Apache Hive

- Hive vs Pig
- Hive Architecture and Components
- Hive Metastore
- Limitations of Hive
- Comparison with Traditional Database
- Hive Data Types and Data Models
- Hive Partition
- Hive Bucketing
- 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 Dataset

Advanced Apache Hive and HBase

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

Topics:

- Hive QL: Joining Tables, Dynamic Partitioning
- Custom MapReduce Scripts
- Hive Indexes and views
- Hive Query Optimizers
- Hive Thrift Server
- Hive UDF
- Apache HBase: Introduction to NoSQL Databases and HBase

- HBase v/s RDBMS
- HBase Components
- HBase Architecture
- HBase Run Modes
- HBase Configuration
- HBase Cluster Deployment

Advanced Apache HBase

Learning Objectives: This module will cover advance Apache HBase concepts. We will see demos on HBase Bulk Loading & HBase Filters. You will also learn what Zookeeper is all about, how it helps in monitoring a cluster & why HBase uses Zookeeper.

Topics:

- HBase Data Model
- HBase Shell
- HBase Client API
- Hive Data Loading Techniques
- Apache Zookeeper Introduction
- ZooKeeper Data Model
- Zookeeper Service
- HBase Bulk Loading
- Getting and Inserting Data
- HBase Filters

Processing Distributed Data with Apache Spark

Learning Objectives: In this module, you will learn what is Apache Spark, SparkContext & Spark Ecosystem. You will learn how to work in Resilient Distributed Datasets (RDD) in Apache Spark. You will be running application on Spark Cluster & comparing the performance of MapReduce and Spark.

Topics:

- What is Spark
- Spark Ecosystem
- Spark Components
- What is Scala
- Why Scala
- SparkContext
- Spark RDD

Oozie and Hadoop Project

Learning Objectives: In this module, you will understand how multiple Hadoop ecosystem www.edureka.co © 2019 Brain4ce Education Solutions Pvt. Ltd. All rights Reserved. components work together to solve Big Data problems. This module will also cover Flume & Sqoop demo, Apache Oozie Workflow Scheduler for Hadoop Jobs, and Hadoop Talend integration.

Topics:

- Oozie
- Oozie Components
- Oozie Workflow
- Scheduling Jobs with Oozie Scheduler
- Demo of Oozie Workflow
- Oozie Coordinator
- Oozie Commands
- Oozie Web Console
- Oozie for MapReduce
- Combining flow of MapReduce Jobs
- Hive in Oozie
- Hadoop Project Demo
- Hadoop Talend Integration

Certification Project

Analyses of a Online Book Store

- Find out the frequency of books published each year. (Hint: Sample dataset will be provided)
- B. Find out in which year the maximum number of books were published
- Find out how many books were published based on ranking in the year 2002.

Sample Dataset Description

- The Book-Crossing dataset consists of 3 tables that will be provided to you.

Airlines Analysis

- Find list of Airports operating in Country India
- Find the list of Airlines having zero stops
- List of Airlines operating with codeshare
- Which country (or) territory having highest Airports
- Find the list of Active Airlines in United state

Sample Dataset Description

- In this use case, there are 3 data sets. Final_airlines, routes.dat, airports_mod.dat

Big Data Hadoop Course Projects

How will I execute projects in this Hadoop Training Course?

You will execute all your Big Data Hadoop Course Assignments/Case Studies on your Cloud LAB environment whose access details will be available on your LMS. You will be accessing your Cloud LAB environment from a browser. For any doubt, the 24*7 support team will promptly assist you.

What is CloudLab?

CloudLab is a cloud-based Hadoop and Spark environment that Edureka offers with the Hadoop Training course where you can execute all the in-class demos and work on real-life Big Data Hadoop projects in a fluent manner.

This will not only save you from the trouble of installing and maintaining Hadoop or Spark on a virtual machine, but will also provide you an experience of a real Big Data and Hadoop production cluster.

You'll be able to access the CloudLab via your browser which requires minimal hardware configuration. In case, you get stuck in any step, our support ninja team is ready to assist 24x7.

What are the system requirements for this Hadoop Training?

You don't have to worry about the system requirements as you will be executing your practicals on a Cloud LAB environment. This environment already contains all the necessary software that will be required to execute your practicals.

Which projects will be a part of this Big Data Hadoop Online Training Course?

Edureka's Big Data & Hadoop Training includes multiple real-time, industry-based projects, which will hone your skills as per current industry standards and prepare you for the upcoming Big Data roles & Hadoop jobs.

Project #1:

Industry: Stock Market

Problem Statement

TickStocks, a small stock trading organization, wants to build a Stock Performance System. You have been tasked to create a solution to predict good and bad stocks based on their history. You also have to build a customized product to handle complex queries such as calculating the covariance between the stocks for each month.

Project #2:

Industry: Health-Care

Problem statement

MobiHeal is a mobile health organization that captures patient's physical activities, by attaching various sensors on different body parts. These sensors measure the motion of diverse body parts like acceleration, the rate of turn, magnetic field orientation, etc. You have to build a system for effectively deriving information about the motion of different body parts like chest, ankle, etc.

Project #3:

Industry: Social Media

Problem Statement:

Socio-Impact is a social media marketing company which wants to expand its business. They want to find the websites which have a low rank web page. You have been tasked to find the low-rated links based on the user comments, likes etc.

Project #4:

Industry: Retail

Problem Statement:

A retail company wants to enhance their customer experience by analysing the customer reviews for different products. So that, they can inform the corresponding vendors and manufacturers about the product defects and shortcomings. You have been tasked to analyse the complaints filed under each product & the total number of complaints filed based on the geography, type of product, etc. You also have to figure out the complaints which have no timely response.

Project #5:

Industry: Tourism

Problem Statement:

A new company in the travel domain wants to start their business efficiently, i.e. high profit for low TCO. They want to analyse & find the most frequent & popular tourism destinations for their business. You have been tasked to analyse top tourism destinations that people frequently travel & top locations from where most of the tourism trips start. They also want you to analyze & find the destinations with costly tourism packages.

Project #6:

Industry: Aviation

Problem Statement:

A new airline company wants to start their business efficiently. They are trying to figure out the possible market and their competitors. You have been tasked to analyse & find the most active airports with maximum number of flyers. You also have to analyse the most popular sources & destinations, with the airline companies operating between them.

Project #7:

Industry: Banking and Finance

Problem Statement:

A finance company wants to evaluate their users, on the basis of loans they have taken. They have hired you to find the number of cases per location and categorize the count with respect to the reason for taking a loan. Next, they have also tasked you to display their average risk score.

Project #8:

Industry: Media & Entertainment

Problem Statement:

A new company in Media and Entertainment domain wants to outsource movie ratings & reviews. They want to know the frequent users who is giving review and rating consistently for most of the movies. You have to analyze different users, based on which user has rated the most number of movies, their occupations & their age-group.