BigData- Hadoop Course Contents:

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, Rack Awareness.

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, Anatomy of File Write and Read, Rack Awareness.

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.

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, Password-Less SSH, MapReduce Job Execution, Data Loading Techniques: Hadoop Copy Commands, FLUME, SQOOP.

3. > Hadoop MapReduce Framework - I

Learning Objectives - In this module, you will understand Hadoop MapReduce framework and the working of MapReduce on data stored in HDFS. You will learn about YARN concepts in MapReduce.

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.

4. > Hadoop MapReduce Framework - II

Learning Objectives - In this module, you will understand concepts like Input Splits in MapReduce, Combiner & Partitioner and Demos on MapReduce using different data sets.

Topics - Input Splits, Relation between Input Splits and HDFS Blocks, MapReduce Job Submission Flow, Demo of Input Splits, MapReduce: Combiner & Partitioner, Demo on de-identifying Health Care Data set, Demo on Weather Data set.

5. > Advance MapReduce

Learning Objectives - In this module, you will learn Advance MapReduce concepts such as Counters, Distributed Cache, MRunit, Reduce Join, Custom Input Format, Sequence Input Format and how to deal with complex MapReduce programs.

Topics - Counters, Distributed Cache, MRunit, Reduce Join, Custom Input Format, Sequence Input Format.

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.

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.

Pig Latin : Relational Operators, File Loaders, Group Operator, COGROUP Operator, Joins and COGROUP, Union, Diagnostic Operators, Pig UDF, Pig Demo on Healthcare Data set.

7. > Hive

Learning Objectives - This module will help you in understanding Hive concepts, Loading and Querying Data in Hive 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, Hive Demo on Healthcare Data set.

8. > Advance Hive and HBase

Learning Objectives - In this module, you will understand Advance Hive concepts such as UDF, dynamic Partitioning. You will also acquire in-depth knowledge of HBase, Hbase Architecture and its components.

Topics - Hive QL: Joining Tables, Dynamic Partitioning, Custom Map/Reduce Scripts, Hive : Thrift Server, User Defined Functions.

HBase: Introduction to NoSQL Databases and HBase, HBase v/s RDBMS, HBase Components, HBase Architecture, HBase Cluster Deployment.

9. > Advance HBase

Learning Objectives - This module will cover Advance HBase concepts. We will see demos on Bulk Loading , 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, Data Loading Techniques, ZooKeeper Data Model, Zookeeper Service, Zookeeper, Demos on Bulk Loading, Getting and Inserting Data, Filters in HBase.

10. > Oozie and Hadoop Project Explanation with Architecture

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 and Apache Oozie Workflow Scheduler for Hadoop Jobs.

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, Project Explanations with Architecture.

After the completion of the Big Data and Hadoop Course, you should be able to:

- Master the concepts of Hadoop Distributed File System and MapReduce framework

- Setup a Hadoop Cluster

- Understand Data Loading Techniques using Sqoop and Flume

- Program in MapReduce

- Learn to write Advance MapReduce programs

- Perform Data Analytics using Pig and Hive

- Implement HBase, MapReduce Integration, Advanced Usage and Advanced Indexing

- Implement best Practices for Hadoop Development and Debugging

- Implement a Hadoop Project

Who should go for this course?

This course is designed for professionals aspiring to make a career in Big Data Analytics using Hadoop Framework. Software Professionals, Analytics Professionals, ETL developers, Project Managers, Testing Professionals are the key beneficiaries of this course. Other professionals who are looking forward to acquire a solid foundation of Hadoop Architecture can also opt for this course.

Pre-requisites

Core java, basic linux commands, Basic sql

Why Learn Big Data and Hadoop?

BiG Data! A Worldwide Problem? According to Wikipedia, “Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.” In simpler terms, Big Data is a term given to large volumes of data that organizations store and process. However, It is becoming very difficult for companies to store, retrieve and process the ever-increasing data. If any company gets hold on managing its data well, nothing can stop it from becoming the next BIG success! The problem lies in the use of traditional systems to store enormous data. Though these systems were a success a few years ago, with increasing amount and complexity of data, these are soon becoming obsolete. The good news is - Hadoop, which is not less than a panacea for all those companies working with BIG DATA in a variety of applications has become an integral part for storing, handling, evaluating and retrieving hundreds or even petabytes of data. Apache Hadoop! A Solution for Big Data!

Hadoop is an open source software framework that supports data-intensive distributed applications. Hadoop is licensed under the Apache v2 license. It is therefore generally known as Apache Hadoop. Hadoop has been developed, based on a paper originally written by Google on MapReduce system and applies concepts of functional programming. Hadoop is written in the Java programming language and is the highest-level Apache project being constructed and used by a global community of contributors. Hadoop was developed by Doug Cutting and Michael J. Cafarella. And just don't overlook the charming yellow elephant you see, which is basically named after Doug's son's toy elephant! Big Data is a set of unstructured and structured data that is complex in nature and is growing exponentially with each passing day. Organizations are facing a major challenge in storing and utilizing this enormous data. This problem spans across the world because of a serious dearth of skilled programmers. "The United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data." - mckinsey.com

Some of the top companies using Hadoop: The importance of Hadoop is evident from the fact that there are many global MNCs that are using Hadoop and consider it as an integral part of their functioning, such as companies like Yahoo and Facebook! On February 19, 2008, Yahoo! Inc. established the world's largest Hadoop production application. The Yahoo! Search Webmap is a Hadoop application that runs on over 10,000 core Linux cluster and generates data that is now widely used in every Yahoo! Web search query. Facebook, a $5.1 billion company has over 1 billion active users in 2012, according to Wikipedia. Storing and managing data of such magnitude could have been a problem, even for a company like Facebook. But thanks to Apache Hadoop! Facebook uses Hadoop to keep track of each and every profile it has on it, as well as all the data related to them like their images, posts, comments, videos, etc. Opportunities for Hadoopers! Opportunities for Hadoopers are infinite - from a Hadoop Developer, to a Hadoop Tester or a Hadoop Architect, and so on. If cracking and managing BIG Data is your passion in life, then think no more and Join course and carve a niche for yourself