Big Data and Hadoop (4-5 Days)

About the Course

The Big Data and Hadoop training course is designed to enhance your knowledge and skills to become a successful Hadoop developer. In-depth knowledge of core concepts will be covered in the course along with implementation on varied industry use-cases.

Course Objectives

By the end of the course, you will:
1. Master the concepts of HDFS and MapReduce framework
2. Understand Hadoop 2.x Architecture
3. Setup Hadoop Cluster and write Complex MapReduce programs
4. Learn data loading techniques using Sqoop and Flume
5. Perform data analytics using Pig, Hive and YARN
6. Implement HBase and MapReduce integration
7. Implement Advanced Usage and Indexing
8. Schedule jobs using Oozie
9. Implement best practices for Hadoop development
10. Work on a real life Project on Big Data Analytics
11. Understand Spark and its Ecosystem
12. Learn how to work in RDD in Spark

Who should go for this Course?

Today, Hadoop has become a cornerstone of every business technology professional. To stay ahead in the game, Hadoop has become a must-know technology for the following professionals:
1. Analytics professionals
2. BI /ETL/DW professionals
3. Project managers
4. Testing professionals
5. Mainframe professionals
6. Software developers and architects
7. Graduates aiming to build a successful career around Big Data

What are the pre-requisites for this Course?

You can master Hadoop, irrespective of your IT background. While basic knowledge of Core Java and SQL might help, it is not a pre-requisite for learning Hadoop.

Which Case-Studies will be a part of the Course?

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-specific Big Data case studies e.g. Finance, Retail, Media, Aviation etc. which you can consider foryour project work:

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, 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.

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 and Multi node cluster set up Hadoop Administration.

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.

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.

5. 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.

6. 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.

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, Run Modes & Configuration, HBase Cluster Deployment.

8. Advanced HBase

Learning Objectives - This module will cover Advanced 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.

9. Processing Distributed Data with Apache Spark

Learning Objectives - In this module you will learn Spark ecosystem and its components, how scala is used in Spark, SparkContext. You will learn how to work in RDD in Spark. Demo will be there on running application on Spark Cluster, Comparing performance of MapReduce and Spark.

Topics - What is Apache Spark, Spark Ecosystem, Spark Components, History of Spark and Spark Versions/Releases, Spark a Polyglot, What is Scala?, Why Scala?, SparkContext, RDD.

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.