

Big Data and Hadoop

Module 1: Preface - Big Data and Hadoop

- Understanding Big Data and Hadoop
- Limitations to Big Data
- Hadoop features and Components
- Characteristics of Big Data (Known as 3Vs of Big Data)
- Why is Hadoop important – Key features
- Hadoop Ecosystem

Module 2: Hadoop 2.x Cluster – Architecture, Core Components and Workflow

- What is Hadoop Cluster Architecture
- Difference between Hadoop and Hadoop Cluster
- Core components of Hadoop Cluster : Client, Master
- Techniques of loading data
- How to configure important files to Hadoop Cluster
- What is Typical Workflow in HDFS: where are files metadata stored, how Hadoop recovers the file on failures? How files (data) are stored in Hadoop?

Module 3: Core Components of Hadoop – Hadoop Common, Hadoop Distributed file System (HDFS) and Hadoop MapReduce.

- Intro to Hadoop Common
- What is HDFS
- Hadoop 2.x cluster architecture
- Brief knowledge of Hadoop MapReduce
- Single node, name node, multi node cluster and set up

Module 4: Four Layers of Hadoop Ecosystem

- Data Storage Layer
- Data Processing Layer
- Data Access Layer
- Data Management Layer

Module 5: Deep Dive in Data Storage and its parts

- HBASE - Column DB Storage
- HBASE Architecture and its Components
- Joining Tables and Partitioning
- HBASE Cluster Deployment
- HDFS - Distributed File System

Module 6: What is Hadoop MapReduce Framework?

- Preface – MapReduce
- MapReduce - Cluster Management
- YARN - Cluster & Resource Management
- YARN Workflow & Demo
- Creating Relation between Input Splits & HDFS Blocks

Module 7: In-depth study of Hadoop Data Access

- Sqoop – RDBMS Connector
- Avro – RPC, Serialization
- Mahout – Machine Learning
- Pig (Data Flow) : Intro to Pig, Use Cases, Data Models, Pig Execution, Pig Latin Language, Pig vs. SQL, Relation between Pig and MapReduce etc.
- Hive (SQL)

Module 8: How Hadoop Data Management Layer Works?

- ZooKeeper – High Performance Management
- Flume – Monitoring
- Chukwa – Monitoring
- Oozie – Workflow Monitoring

Module 9: Understanding of Advanced HIVE & HBASE

- Dynamic Partitioning
- UDF MapReduce Scripts
- Understanding of Hive Indexes and Hive query Optimization
- HBASE – Preface to NoSQL, Databases & HBASE
- HBASE Architecture & Understanding Run nodes
- HBASE Cluster

Module 10: Oozie & Hadoop project Work

- Preface Oozie & Workflow Definitions
- Oozie components & Scheduling with Oozie
- Understanding of Flume & Sqoop
- Oozie Web Console & Oozie with MapReduce
- Oozie commands and Coordinator
- Understanding of PIG & Hive in Oozie
- Hadoop Demo Project & Integration of [Talend](#) with Hadoop

