# BIG DATA COURSE

## 80 HOURS

1) Introduction to Big Data & Hadoop – 1 Hrs.
2) Importance of Data & Data Analysis
    - What is Big Data?
    - Big Data & its hype
    - Big Data Users & Scenarios
    - Structured vs Unstructured Data
    - Challenges of Big Data
    - How to overcome the challenges?
    - Divide & Conquer philosophy
    - Overview of Hadoop
3) Hadoop and its file system – HDFS – 3 hrs.
    - History of Hadoop
    - Hadoop Ecosystem
    - Hadoop Animal Planet
    - What is Hadoop?
    - Key Distinctions of Hadoop
    - Hadoop Components
    - HDFS
    - Map Reduce
    - Why Distributed File System?
    - The Design of HDFS
    - Hadoop Distributed File System
    - What is a HDFS block?
    - Why HDFS block is so large in HDFS?
    - Name Node
    - Data Node
    - Secondary Name Node
    - A file in HDFS
    - Hadoop Components/Architecture
    - Name Node, Job Tracker, Data Node, TaskTracker & Secondary Namenode
    - Understanding Storage components(NameNode, DataNode & Secondary Namenode)

- Understanding Processing components(JobTracker & TaskTracker)
- How Secondary Namenode overcomes the failure of the primary Namenode
- Anatomy of a File Read
- Anatomy of a File Write

4) Understanding Hadoop Cluster – 1hr
    - Walkthrough of CDH VM setup
    - Hadoop Cluster modes
    - Standalone Mode
    - Pseudo-Distributed Mode
    - Distributed Mode
    - Hadoop Configuration files
    - core-site.xml
    - mapred-site.xml
    - hdfs-site.xml
    - yarn-site.xml
    - Understanding Cluster configuration

5) MapReduce – 5 hrs.
    - Meet MapReduce
    - Word Count algorithm – Traditional approach
    - Traditional approach on a Distributed system& it's drawbacks
    - MapReduce approach
    - Input & Output Forms of a MR program
    - Hadoop Data types
    - Map, Shuffle & Sort, Reduce Phases
    - Workflow & Transformation of Data
    - Word Count Code walkthrough
    - Input Split & HDFS Block
    - Relation between Split & Block
    - MR Flow with Single Reduce Task
    - MR flow with multiple Reducers
    - Data locality Optimization
    - Speculative Execution
    - Combiner
    - Partitioner

6) Pig – 10 hrs.
    - What is Pig?
    - Why Pig?
    - Pig vs Sql
    - Execution Types or Modes

- Running Pig
- Pig Data types
- Pig Latin relational Operators
- Multi Query execution
- Pig Latin Diagnostic Operators
- Pig Latin Macro & UDF statements
- Pig Latin Commands
- Pig Latin Expressions
- Schemas
- Pig Functions
- Pig Latin File Loaders
- Pig UDF & executing a Pig UDF
- Pig Use cases

7) Hive – 10hrs

- Introduction to Hive
- Pig vs. Hive
- Hive Limitations & Possibilities
- Hive Architecture
- Metastore
- Hive Data Organization
- Hive QL
- Sql vs. Hive QL
- Hive Data types
- Data Storage
- Managed & External Tables
- Partitions & Buckets
- Static Partitioning & Dynamic Partitioning
- Storage Formats
- File Formats – Sequence File & RC File
- Using Compression in Hive
- Built-in Serdes
- Importing Data (Using Load Data & Insert Into)
- Alter & Drop Commands
- Data Querying
- Using MR Scripts
- Hive Joins
- Sub Queries
- Views

8) HBase – 5 hrs.

- Introduction to NoSql & HBase
- HBase vs. RDBMS
- HBase Use cases
- Row & Column oriented storage
- Characteristics of a huge DB
- What is HBase?
- HBase Data-Model
- HBase logical model & physical storage
- HBase architecture
- HBase in operation (put, get, scan & delete)
- Loading Data into HBase
- HBase shell commands
- HBase operations through Java
- HBase operations through MR

9) ZooKeeper & Oozie – 10hrs

- Introduction to Zookeeper
- Distributed Coordination
- Zookeeper Data Model
- Zookeeper Service
- Introduction to Zookeeper
- Distributed Coordination
- Zookeeper Data Model
- Zookeeper Service

10) Sqoop – 5hrs

- Introduction to Sqoop
- Sqoop design
- Sqoop basic Commands
- Sqoop Table Import flow of execution
- Sqoop Import Commands – to HDFS, Hive & HBase tables
- Sqoop Incremental Import
- Incremental Append
- Incremental Last Modified
- Sqoop export flow of execution
- Sqoop Export Command

11) Flume – 5 hrs.

- Flume Architecture
- Flume Components
- Streaming live Twitter data with Flume

12) Spark – 25 hrs.

Module 1

- Introduction & Overview
- Architecture
- Installation of Spark-- Options
- Starting the Spark--- possibilities
- Amazon EMR
- EC2
- Maven
- Standalone mode
- With mesos
- With YARN
- HDinsight
- Spark context & Spark Session

Module 2

- Basics & Spark Shell Applications
- Various possibilities
- Eclipse with Maven
- Eclipse with SBT
- Zeppelin Notebook
- IntelliJ
- Spark Jobs & API’s.
- Spark Core
- RDD’s
- Transformations
- Actions
- Data Frame

Module 3

- Spark with External Data Sources
- From Local file system
- From HDFS
- From Amazon S3
- From Cassandra Spark SQL
- Schema
- Case Classes
- Joins
- Catalyst Optimizer

Module 4

- Spark Streaming
- Spark MLlib
- Spark GraphX
- PySpark