HADOOP SYLLABUS

Why Hadoop?

- Hadoop is open source
- Hadoop solves Big Data problem which is very difficult and sometime impossible to solve using SQL tools
- It can process Distributed data and no need to store entire data in centralized storage as it is required for SQL based tools.

Course Content

Week-1

(Introduction to Hadoop)

- Hadoop- Demo
- What is Bigdata
- When data becomes Bigdata
- 3V's of Bigdata
- Introduction to Hadoop Ecosystem
- Why Hadoop? If Existing Tools and Technologies are there in market since decades?
- How Hadoop is getting two categories Projects- New projects on Hadoop
- Clients want POC and migration of Existing tools and Technologies on Hadoop Technology
- How Open Source tool (HADOOP) is capable to run jobs in lesser time which take longer time in other tools in the market.
- Hadoop Storage – HDFS (Hadoop Distributed file system)
- Hadoop Processing Framework (Map Reduce) / YARN
- Alternates of Map Reduce
- Why NOSQL is in more demand now a days
- Distributed warehouse for DFS
- Most demanding tools which can run on the top of Hadoop Ecosystem for specific requirements in specific scenarios
- Data import/Export tools

Week-2

(Hadoop Setup Installation and Pig Basics)

- Hadoop installation
- Introduction to Hadoop FS and Processing Environment's UIs
- How to read and write files
- Basic Unix commands for Hadoop
- Hadoop's FS shell
- Hadoop's releases
- Hadoop's daemons
- Pig Introduction
- Why Pig if Map Reduce is there?
- How Pig is different from Programming languages
- Pig Data flow Introduction
- How Schema is optional in Pig
- Pig datatypes
- Pig- Basics
- Pig Use cases
- Pig Assignment

Week-3

(Pig Advanced, Hive Basic , Hive Advanced)

- Pig Advanced
- Complex Use cases on Pig
- Pig Advanced Assignment
- Real time scenarios of Pig
- Hive Introduction
- Hive Advanced
- Partitioning
- Bucketing
- External tables
- Complex Use cases in Hive
- Hive Advanced Assignment
- Real time scenarios of Hive

Week-4

(Map Reduce Basics and NOSQL Introduction, POC (Proof Of Concept))

- Introduction to NOSQL
- Why NOSQL if SQL is in market since several years
- Databases in market based on NOSQL
- How Map Reduce works as Processing Framework
- End to End execution flow of Map Reduce job
- Different tasks in Map Reduce job
- Why Reducer is optional while Mapper is mandatory?
- Introduction to Combiner
- Introduction to Partitioner
- Programming languages for Map Reduce
- Why Java is preferred for Map Reduce programming
- POC based on Pig, Hive, HDFS, MR

Week-5

(Map Reduce Advanced , HBase Basics)

- How to work on Map Reduce in real time
- Map Reduce complex scenarios
- Introduction to HBase
- Introduction to other NOSQL based data models
- Drawbacks of Hadoop
- Why Hadoop can't be used for real time processing
- How HBase or other NOSQL based tools made real time processing possible on the top of Hadoop
- HBase table and column family structure
- HBase versioning concept
- HBase flexible schema
- HBase Advanced

Week-6

(Zookeeper, Sqoop , Quick revision of previous classes)

- Introduction to Zookeeper
- How Zookeeper helps in Hadoop Ecosystem
- How to load data from Relational storage in Hadoop
- Sqoop basics Sqoop practical implementation
- Sqoop alternative What is connector in Sqoop
- Quick revision of previous classes to fill he gap in understanding and correct understandings

Week-7

(Flume , Oozie , Hadoop Releases, Introduction to YARN)

- How to load data in Hadoop that is coming from web server or other storage without fixed schema
- How to load unstructured and semi structured data in Hadoop
- Introduction to Flume
- Hands-on on Flume
- How to load Twitter data in HDFS using Hadoop
- Introduction to Oozie
- How to schedule jobs using Oozie
- What kind of jobs can be scheduled using Oozie
- How to schedule jobs which are time based
- Hadoop releases
- From where to get Hadoop and other components to install
- Introduction to YARN
- Significance of YARN

Week-8

(Introduction to Hue, Different vendors in market, Major Project discussion)

- Introduction to Hue
- How Hue is used in real time
- Real time Hadoop usage
- Real time cluster introduction
- Hadoop Release 1 vs Hadoop Release 2 in real time
- Hadoop real time project
- Major POC based on combination of several tools of Hadoop Ecosystem
- Datasets for practice purpose

Week-9

(Spark and Scala)

- Introduction to Spark
- Introduction to Scala
- Advantages of Spark over Hadoop
- Is Spark replacement of Hadoop?
- How Spark is Faster than Hadoop
- Real time scenarios examples of Spark where we prefer Spark over Hadoop
- How Spark is capable to process complex data sets in lesser time
- In Memory Processing Framework for Analytics
- Data Science on the top of Hadoop

Additional added values:

- 4 POCs
- 1 Real time project implementation
- 2 Projects Discussion
- 7 Domain Based Projects Available
- Interview Questions and Answers and Mock Interview