# Admin + Developer

## 1. The Motivation & Limitation for Hadoop

Motivation of Hadoop
Big data features and challenges

Problems with Traditional Large-Scale Systems Why Hadoop & Hadoop Fundamental Concepts Comparison between Hadoop and RDBMS
Is Hadoop replacing RDBMS?

History of Hadoop with Hadoopable problems Limitation of Hadoop

## 2. HDFS CONCEPTS

HDFS Design & Goals
Understand Blocks and Configuration of block size
Block replication and replication factor
Understand Hadoop Rack Awareness and configure racks in Hadoop
File read and writes anatomy in HDFS
Enable HDFS Trash
Configure HDFS Name and space Quota
Configure and use WebMD's ( Rest API For HDFS )

## 3. HADOOP ECOSYSTEM & CLUSTER

Available version Hadoop 1.x & 3
Available Distributions of Hadoop (Cloudera, Hortonworks)
Hadoop Projects & Components
Architecture of Hadoop & Planning for cluster
The Hadoop Distributed File System (HDFS)
Cluster Daemons & Its Functions

- Name Node
- Secondary Node
- Data Nodes
- Application Master and Task Tracker

YARN Responsibilities
Deployment of Hadoop Cluster

## 4. HIGH AVAILABILITY AND NAMESPACE FEDERATION

Introduction to HDFS Federation
Understand Name service ID and Block pools
Introduction to HDFS High Availability
Failover mechanisms in Hadoop 2.x
Concept of Active and StandBy NameNode
Configuring Journal Nodes and Split brain scenario
Automatic and manual failover techniques in HA
HDFS HAadmin commands
Understand NameNode Safemode, File system image and edits

## 5. YARN - YET ANOTHER RESOURCE NEGOTIATOR

YARN Architecture in Hadoop 2.x
Yarn Components

Resource Manager
Node Manager
Job History Server
Application Time Line Server
MR Application Master
YARN Application execution flow
Running and Monitoring YARN Applications
Understand and Configure Capacity / Fair Schedulers in YARN
YARN Rest API
Writing and executing YARN applications

## 6. LINUX INITIALS

Installation of Linux (Red Hat)
Basic Linux configurations
Basic Linux commands

- Password less ssh
- IP address and hostname
- Firewall and selinux
- Yum and creating yum repository
- NTP configurations

## 7. INSTALLATION & DEPLOYMENT OF HADOOP

Setting up Local Cloudera/Hortonworks repository
Installation for Cloudera Manager/Ambari
Overview of Cloudera Manager/Apache Ambari
Installing Multi (atlest 10 Machine) Node Hadoop (Cloudera/Hortonworks)
Setting up Cloudera/Hortonworks Hadoop environment
Specifying the Hadoop Configuration
Performing Initial HDFS Configuration
Performing Initial YARN and Map-Reduce Configuration
Hadoop Logging & Cluster Monitoring

## 8. SANDBOX / QUICK START VMS)

Overview of SandBox
Different flavours (Virtual Box / VMware) of SandBox
Installation of SandBox
Start Working with SandBox

## 9. HUE OR HADOOP UI

Introduction of HUE
Getting started with HUE
Deployment of Map-Reduce
Functional Execution of Hive/HBase
Design of work-flow using Job Designer
Data transfer in Sqoop and flume

## 10. MAPREDUCE CONCEPTS

Introduction to MapReduce
Architecture of Map-Reduce
Understanding the concept of Mappers & Reducers
Anatomy of MapReduce program
Phases of a MapReduce progam
Data-types in Hadoop MapReduce
Driver, Mapper and Reducer classes
InputSplit and RecordReader
Input format and Output format in Hadoop
Concepts of Combiner and Partitioner
Running and Monitoring MapReduce jobs
Writing your own MapReduce job using MapReduce API
Writing Mappers and Reducers with the Streaming API
Different interview questions raised for Map-Reduce

## 11. HADOOP SHELL AND COMMANDS

Hadoop Developer/Admin commands using shell
NameNode & Secondary NameNode Commands
HDFS DFSAdmin and File system shell commands
Hadoop NameNode / DataNode directory structure
HDFS permissions model
HDFS Offline Image Viewer
Map-Reduce job deployment
Oozie workflow design
Different Components Jobs design

## 12. APACHE ZOOKEEPER

Introduction to Apache Zookeeper
Zookeeper stand alone installation
Zookeeper Clustered installation
Understand Znodes and Ephemeral nodes
Manage Znodes using Java API
Zookeeper four letter word commands

## 13. HBASE: THE HADOOP DATABASE

Problems with RDBMS
Introduction to HBase
HBase components - Hbase master and Region servers
Non-RDBMS, Not-Only SQL or No-SQL
Installation HBase & Deployment Types
CRUD & Batch Operations
Filters, Counters, Pool
Rest Interface & Web-UI

## 14. HCATALOG OR METASTORE TABLES

Introduction of apache Hcatalog
Creating tables using Hcatalog
Bulk uploads using MetaStore Tables
Play with semi-structured data
Integration of Hcatalog with Hive
Hive SQL query analysis

## 15. HIVE

Problems with No-SQL Database
Introduction & Installation Hive
Hive Schema and Data Storage
Data Types & Introduction to SQL
Hive-SQL: DML & DDL
Hive-SQL: Views & Indexes
Explain and use the various Hive file formats
Use Hive to run SQL-like queries to perform data analysis
Use Hive to join data sets using a variety of techniques, including Map-side joins and Sort-Merge-Bucket joins
Integration to HBase & Cassandra
Sentiment Analysis and N-Grams
Hive Thrift Service

## 16. FLUME

Installation of Flume
Ingesting Data from External Sources with Flume
Configuration for flume
REST Interfaces
Best Practices for Importing Data

## 17. SQOOP

Installation of Sqoop
Ingesting Data from External (RDBMS) Sources with Sqoop
Ingesting Data from/to Relational Databases with Sqoop
Integration of Sqoop and Hbase
Integration of Sqoop and Hive
Best Practices for Importing Data

## 18. MANAGING AND SCHEDULING JOBS

Managing Running Jobs
Scheduling Hadoop Jobs
Configuring the FairScheduler

## 19. VISUALIZATION WITH EXCEL (GRAPH)

Introduction of data Visualization in excel
Hive integration with excel 2013
Real world examples with Hadoop and excel integration
Visualize Website Click stream Data
Analyze Machine and Sensor Data
Refine and Visualize Sentiment Data

## 20. APACHE DRILL – REPLACEMENT OF MAP-REDUCE

Installation of Drill
Query data using apache drill
Query data from Hadoop/HDFS file system
Drill & Hbase integration
Drill & Hive integration & Replacement

## 21. CONCLUSION & FAQS

Note:

- Every Topic has practical session
- Hadoop uses different components which discussed in required

# Advance Hadoop

## 1. WHY SPARK?

Problems with Traditional Large-Scale Systems
Introducing Spark
Spark Basics

## 2. WHAT IS APACHE SPARK?

Using the Spark Shell
Resilient Distributed Datasets (RDDs)
Functional Programming with Spark
Working with RDDs

## 3. RDD OPERATIONS

Key-Value Pair RDDs
MapReduce and Pair RDD Operations
The Hadoop Distributed File System

## 4. OVERVIEW

A Spark Standalone Cluster
The Spark Standalone Web UI
Parallel Programming with Spark

## 5. RDD PARTITIONS AND HDFS DATA LOCALITY

Working With Partitions
Executing Parallel Operations
Caching and Persistence

## 6. RDD LINEAGE

Caching Overview
Distributed Persistence
Writing Spark Applications

## 7. SPARK APPLICATIONS VS. SPARK SHELL

Creating the SparkContext
Configuring Spark Properties
Building and Running a Spark Application Logging
Spark, Hadoop, and the Enterprise Data Center

## 8. SPARK STREAMING OVERVIEW

Example: Streaming Word Count
Other Streaming Operations
Sliding Window Operations
Developing Spark Streaming Applications
Common Spark Algorithms

## 9. SHARK, SPARK SQL

Implement SparkSQL queries to perform several computations

Fees: 5,000 Rs/- (Extra)

Duration: 1 Month

# Machine Learning

## INTRODUCTION OF MACHINE LEARNING WITH TENSORFLOW

Motivation of Machine Learning?
Use Cases of Machine Learning
Future Scope of Machine Learning
Real World Domain using ML
Types of Machine Learning
Different tools/framework available for ML
Limitation of Machine Learning tensorflow

## BASICS OF MACHINE LEARNING

Understanding supervised & unsupervised learning
Understanding bases associated with any machine learning algorithm
Ways of reducing bias and increasing generalisation
Introduction to scikit-learn & SciPy in Python

## MACHINE LEARNING TECHNIQUES

Supervised & Unsupervised Learning
Recommender Systems

- User Based recommendation Engine
- Item Based recommendation Engine

Logistic Regression
Regularization
Linear Regression with One Variable
Linear Algebra Review
Linear Regression with Multiple Variables

## NAIVE BAYES

Use Naive Bayes with scikit learn in python/Mahout.
Splitting data between training sets
Testing sets with scikit learn (Python/Mahout).
Calculate the posterior probability
Pior probability of simple distributions

## SUPPORT VECTOR MACHINES (SVM)

Learn the simple intuition behind Support Vector Machines.
Implement an SVM classifier in SKLearn/scikit-learn
Choose the right kernel for your SVM
Learn about RBF and Linear Kernels

## NEURAL NETWORKS: REPRESENTATION

## NEURAL NETWORKS: LEARNING

## Advice for Applying Machine Learning

R & D

# PART 1: SETUP OF DATA CENTERS

## 1). CLOUDERA

CDH 5 Installation for Apache Hadoop developers and system administrators interested in Hadoop installation.
Describes installation and configuration of cloudera CDH 5.x on multiple machines.
Deploy all 21 components like HBase, Hive, Sqoop, Flume etc. in data centers machines
Learning Labs: Planning & Deployment, Monitoring, Performance tuning, Security using Kerberos,
HDFS High Availability using Quorum Journal Manager (QJM) and Oozie, Hcatalog/Hive Administration.

## 2). HORTONWORKS

HDP 2.x Installation for Apache Hadoop developers and system administrators interested in Hadoop installation.
Describes installation and configuration of
cloudera HDP 2.x on multiple machines.
Deploy components like HBase, Hive, Sqoop, Flume etc. in data centers machines
Learning Labs: Planning & Deployment, Monitoring,
Performance tuning, Security using Kerberos,
HDFS High Availability using Quorum Journal Manager (QJM) and Oozie, Hcatalog/Hive Administration.
Introduction of Apache Ambari for deploying and managing Apache Hadoop Securing your hadoop infrastructure with Apache Knox

Note: Hortonworks deployment is same as Cloudera but with different flavours.

## 3). APACHE HADOOP

Setup a minimum 3-4 Node Hadoop Cluster
Node 1 - Namenode, Other Master services
Node 2 – Secondary Name Node, Resource Manage
Node 3 - Data Node, Task Tracker
Node 4 - Data Node, Task Tracker)

HDFS High Availability using Quorum Journal Manager (QJM)
Hive installation on HDFS
Security implementation with Kerberos
Apache Ambari to add new nodes to your existing cluster

# PART 2: DEVELOPERS

## 1). FILE MANAGEMENT

Managing HDFS files with command line
Creating HDFS Snapshots to backup important Enterprise datasets
Installation and Configuration of Cloudera/Hortonworks ODBC driver on Windows/Mac
Sorted insert for circular linked list

## 2). DATA VISUALIZATION

Qlikview - Business Discovery and Visualizing Your Data Using QlikView
Tableau - Visualize Data with Tableau

## 3). SENTIMENT, PREDICTIVE, SENSOR DATA ANALYSIS

Learning Labs of Apache NiFi
Process Data with Apache Hive
Loading and Querying Data with Hadoop
Website Clickstream Data Analysis in Qlikview
Refine and Visualize Server Log Data
Social Media and Customer Sentiment Analysis
Analyze Machine and Sensor Data

# PART 3: DATA SCIENCE

## 1). INDEXING AND SEARCHING

Apache Solr
Setting up a Solr cluster
Creating and updating schemas
Indexing data & Searching
Tuning relevance
Extended features such as geospatial search, spell checking, highlighting, etc.
Analytics and visualizations using Solr

## 2). MAHOUT - MACHINE LEARNING

Supervised & Unsupervised Learning
Mahout - Recommendation
Mahout - Clustering
Mahout - Classification