HADOOP COURSE CONTENT

Module -1: Duration 75 Minutes

Basic Concepts

1) Big-Data:
    a. What is Data?
    b. What is Big-Data?
    c. Sources of Big-Data.
    d. Structured Vs Unstructured
    e. Big-Data Characteristics: 3Vs
    f. Common Use Cases.
    g. Project Discussion:
    h. Lab-1: Connectivity to cluster/Executing Unix commands / Winscp or Filezilla/ putty/Edge Node
2) Apache Hadoop
    a. Limitations of the Existing Solution on Big Data
    b. Compare Teradata with Hadoop
    c. How Hadoop provides the solution for Big Data?
    d. Apache Hadoop competitors in the market, then why Hadoop?
    e. What is Apache Hadoop?
    f. History of Apache Hadoop
    g. Why name Hadoop?
    h. Doug Cutting.
    i. Forecasting the job market across the globe
    j. Discuss Generation 1 Hadoop and Generation 2 Hadoop
    k. Do you know Hadoop is a Desktop application?
    l. YARN = 10 needs for Hadoop.
    m. Lab-2: Hortonworks, Cloudera, Pivotal distributions

Module -2: Duration: 120 Minutes

3) Distributed File System Advantages and Disadvantages.
4) Apache Hadoop Components
    a. Storage :
        i. Default File System
        ii. HDFS
        iii. Simple Storage Service
    b. Processing
        i. MapReduce
        ii. MPP

iii. Graph Processing

c. Lab-3 : Configuration of File System and Framework

5)

6) Introduction to Hadoop Distribute File System (HDFS)
   a. HDFS Components
   b. Name Node
   c. Secondary Name Node
   d. Data Node
   e. Lab-4: Demons in the cluster, starting of the demons.
7) Basic of HDFS
   a. HDFS Architecture
   b. Why Block?
   c. Why default block size is 64MB?
   d. Why default Replication factor 3?
   e. Block Management Service?
   f. Lab-5: HDFS basic commands.
8) Replication and Rack Awareness, Lab-6: Show replicated blocks in the cluster
9) Anatomy of File Write on HDFS, Lab-7: Show writing of file on HDFS
10) Anatomy of File Read on HDFS, Lap-8: Show reading of file on HDFS

Module -3: Duration: 75 MIN

11) Introduction to typical Hadoop Cluster
12) Secondary NameNode
   a. What is FsImage?
   b. What is edits.log?
   c. Usage of Secondary NameNode
   d. Lab-9: Show FsImage and edits.log in the cluster
13) NameNode is Single Point of Failure
   a. Generation 1 Hadoop SPOF
   b. Generation 2 Hadoop SPOF handled using HA.
      i. Failover Fencing
      ii. STONITH
      iii. Split Brain Disorder
14) NameNode Scalability:
   a. Generation -1: Is NameNode Scalable in Generation 1 Hadoop.
   b. Generation -2: HDFS Federation.

Module -4: Duration: 120 MIN

15) Hadoop Cluster Modes.
   a. Standalone

b. Psuedo-Distributed
c. Fully-Distributed Mode
d. Lab: Show configuration changes to run different cluster modes.

16) Core Configuration files
    a. Core-site.xml
    b. Hdfs-site.xml
    c. Mapred-site.xml
    d. Yarn-site.xml
    e. Hadoop-env.sh
    f. Masters
    g. Slaves.
17) Running teragen example.
18) Dump of MR Log
19) Hadoop copy commands.

Module-5 Duration: 120 MIN

20) Introduction to MapReduce Framework.
21) MR Framework beyond scenes
22) Traditional way of solving the problem, Lab: Show sample running of WordCount process.
23) MapReduce way of solving the problem, Lab: Show sample running of WordCount process.
24) Generation 1: Executing WordCount MR Job
    a. Job Tracker
    b. Task Tracker
25) Generation 2: Executing WordCount Application
26) How to debug the log files for MapReduce.
27) Differences between Gen1 and Gen2 Hadoop
28) Anatomy of MapReduce Job
    a. Output Collector
    b. Circular Memory
    c. Split files
29) Advantage of MapReduce
    a. Parallel Processing.
    b. Data Locality
30) What is speculative Execution?

Module-6: Duration: 120 MIN

31) Introduction to YARN
    a. Generation -2 Architecture
    b. Gen-2 Components

i. Client
ii. Resource Manager
    1. Scheduler
    2. Application Manager
iii. Node Manager
iv. Container
v. Application Master

c. MapReduce Application Phases in YARN
    i. Application Submission
    ii. Job Initialization
    iii. Task Assignment
    iv. Task Execution
    v. Status Update
    vi. Failure Recovery
        1. Container Failure
        2. Node Manager Failure
        3. AM Failure
        4. Resource Manager Failure
    vii. Lab: MR Program Execution on YARN
d. Moving beyond MapReduce in YARN
e. Introduction to Job Queues, Lab: How to define Queues?
f. Schedulers
    i. FIFO Scheduler
    ii. Fair Scheduler
    iii. Capacity Scheduler

Module 7 Duration: 120 MIN

32) Difference in MRv1 and MRv2 Java API chart
33) Writing MapReduce applications using MRv1 API
    a. Job : Configured, Tools, ToolRunner, **Run, Configuration etc… discussion**
    b. Mapper
    c. Reducer
34) Writing MapReduce applications using MRv2 API
    a. Job
    b. Mapper
    c. Reducer
35) Writing Weather Temperature Use Case.
36) DE Identification of Person Information
37) Fixed Width File to CSV file

38) JSON to CSV file
39) Combiner
40) Partitioner
41) Assignment: Secondary Sorting Use Case, Matrix Calculation use Case.

Module 8 Duration: 120 MIN

42) Joins in MapReduce
    a. MapSide Join
    b. Reduce Side Join
    c. BroadCast Join
43) Chain Mappers, Reducers
44) Custom InputFormat
45) OutputFormat
46) Data Types
    a. Pre-Defined Data Types
    b. Custom Data Types
    c. Writable Comparable.
47) MR Unit
48) Distributed Cache
49) Sequential File
50) AVRO File.
51) Lab: Writing and executing each use cases.

Module 8: Apache PIG

1) Why PIG?
2) Compare PIG Vs MR
3) Where to Use Pig?
4) Pig Execution Modes:
    a. Local
    b. MapReduce
    c. Tez_on_local
    d. Tez.
5) Pig Latin
6) Pig Data Types
    a. Primitive Data Types
    b. Scalar Data Types
        i. Bag
        ii. Tupple
        iii. Field
        iv. Map
    c. NULL

7) Pig Data Flow Language
8) Pig Operators
9) Load and Store Operators
    a. Load
    b. Store
    c. DUMP
10) Transform Operators
    a. FILTER
    b. FOREACH
    c. GROUP
    d. PARALLEL
    e. COGROUP
    f. INNER JOIN
    g. OUTER JOIN
    h. UNION
    i. SPLIT
11) DIAGNOSIS Operators
    a. DESCRIBE
    b. EXPLAIN
    c. ILLUSTRATE
12) Built in Functions in PIG
13) Pig Properties
14) UDFs
15) Pig HBASE storage Handler
    a. Load from HBASE
    b. Store into HBASE
16) Pig Schema
17) Synopsis: Pig Read Hive Table

Module: Apache Hive

1) History of Hive
2) Hive Architecture
3) Hive Metastore
    a. Embedded Metastore
    b. Local Metastore
    c. Remote Metastore
4) Hive Components
    a. Driver
    b. Shell
    c. Compiler
    d. Execution Engine

5) HiveQL
6) HiveQL data types
7) ACID Hive
    a. Hive Transactions
    b. Hive Updates
8) Partitioning
9) Bucketing
10) Hive Loads
11) Hive Tables
    a. Managed Table
    b. External Table
    c. Native Table
    d. Non-Native Tables
    e. Temporary Tables
12) Hive Views
13) Hive Diagnosis operators

Module : Advance Hive

14) Hive on HBASE
15) Join in Hive
    a. Inner Joins
    b. Outer Joins
16) Dynamic Partition
17) Hive SerDe: JSON Serdes
18) Hive UDF
19) Hive Parameters
20) Lab: Creating Hive tables, ORC Tables, Avro Tables, Jason tables

Module: HBASE

1) Introduction to Nosql Databases.
2) CAP Theorem
3) HBASE
4) History of HBASE
5) Three major HBASE Components
    a. HBASE Master
    b. Region Server
    c. Client Library
6) HBASE vs RDBMS
7) HBASE Versioning
8) HBASE shell
9) HBASE Column Family

10) Sparse Datastore
11) Horizontal Sharding
12) ROW KEY
13) HBASE Architecture
    a. Zookeeper
    b. WAL
    c. HFILE
    d. MEM store
14) REGION
15) HBASE Write
16) HBASE Read

Module: Advance HBASE

17) HBASE Loading Techniques
    a. HBASE SHELL
    b. HBASE Java Client
    c. PIG to HBASE
    d. SQOOP to HBASE
    e. Hive to HBASE
18) Coprocessors
19) Joins in HBASE
20) BLOOM FILTER

Module: Zookeeper

1) Introduction to Zookeeper
2) Introduction to ZNODE
3) Zookeeper ENSEMBLE
4) ZAP Protocol
5) Atomic Broad Cast
6) Zookeeper during failures

Module: OOZIE

7) Introduction to OOZIE
8) OOZIE Components
9) Coordinators
10) Workflow
11) Bundle
12) Creating and Running Workflow
13) Creating and Running Coordinator
14) OOZIE actions

15) OOZIE nodes
16) OOZIE WebUI
17) OOZIE Client

MODULE: FLUME

18) Introduction to FLUME
19) AGENTS
20) SOURCE,CHANNEL,SINK
21) DIFFERENT SOURCE
22) Working with Twitter Source
23) Dumping MR Logs into FLUME

MODULE: SQOOP

24) Introduction to SQOOP
25) SQOOP Working principle
26) Sqoop import
27) Sqoop Export
28) Sqoop import query
29) Sqoop Number of Mapper

MODULE: Project.

1) Discussion on Project discussion -1
    a. Implement the project
    b. OOZIE
    c. FLUME
    d. HADOOP COPY COMMANDS
    e. MAPREDUCE
    f. PIG
    g. HIVE
    h. HBASE
    i. MYSQL
    j. WebUI
2) Discussion on Project –II
3) Discussion on Project –III
4) Discussion on Project –IV
5) Discussion on Project -V