Hadoop

Course Duration: 25 days (60 hours duration).

Bigdata Fundamentals

Day1: (2hours)

1. Understanding BigData.
    a. What is Big Data?
    b. Big-Data characteristics.
    c. Challenges with the traditional Data Base Systems and Distributed Systems.

2. Hadoop Distributions:
    a. Hortonworks
    b. Cloudera
    c. Pivotal HD
    d. Greenplum.

Day2: (2hours)

3. Introduction to Apache Hadoop.
    a. Flavors of Hadoop: Big-Insights, Google Query etc..
4. Hadoop Eco-system components: Introduction
    a. MapReduce
    b. HDFS
    c. Apache Pig
    d. Apache Hive
    e. HBASE
    f. Apache Oozie
    g. FLUME
    h. SQOOP
    i. Spark.
    j. Kafka
    k. Crunch

Day3: (2hours)

5. Understanding Hadoop Cluster
6. Hadoop Core-Components.
    a. NameNode.
    b. JobTracker.
    c. TaskTracker.
    d. DataNode.
    e. SecondaryNameNode.
7. HDFS Architecture
    f. Why 64MB?
    g. Why Block?
    h. Why replication factor 3?

Hadoop

Course Duration: 25 days (60 hours duration).

Day4: (2hours)

8. Discuss NameNode and DataNode.
9. Discuss JobTracker and TaskTracker.
10. Typical workflow of Hadoop application
11. Rack Awareness.
    a. Network Topology.
    b. Assignment of Blocks to Racks and Nodes.
    c. Block Reports
    d. Heart Beat
    e. Block Management Service.

Day5: (4hours)

12. Anatomy of File Write.
13. Anatomy of File Read.
14. Heart Beats and Block Reports
15. Map Reduce Overview
16. Cluster Configuration
    a. Core-default.xml
    b. Hdfs-default.xml
    c. Mapred-default.xml
    d. Yarn-site.xml
    e. Hadoop-env.sh
    f. Slaves
    g. Masters
17. Map Reduce Framework
18. Why Map Reduce?
19. Use cases where Map Reduce is used.
20. YARN Architecture
21. Hadoop Classic vs YARN
22. YARN Demo

Day6: (2hours)

23. MR Practicals
    a. Setup environment for the programs.
    b. Possible ways of writing Map Reduce program with sample codes find the best code and discuss.
    c. Configured, Tool, GenericOptionParser and queues usage.

24. Limitations of traditional way of solving word count with large dataset.

Hadoop

Course Duration: 25 days (60 hours duration).

Day7: (2hours)

25. Map Reduce way of solving the problem.
26. Complete overview of MapReduce.
27. Unit testing of mapreduce programs using Junit, MRUnit frameworks.
28. Challenges in Hadoop Testing and options available
29. Manual testing of MapReduce programs

Day8: (2hours)

30. Split Size
31. Combiners
32. Multi Reducers
33. Parts of Map Reduce
34. Shuffle, Sort and Merge phases
35. Map Reduce Design Patterns

Day 9: (2hours)

1. Cloudera Distribution of Hadoop(CDH) – VM Setup
2. HDFS Practicals (HDFS Commands)
3. Map Reduce Anatomy
    a. Job Submission.
    b. Job Initialization.
    c. Task Assignments.
    d. Task Execution.

Day10: (4hours)

4. Schedulers
5. Map Reduce Failure Scenarios
6. Speculative Execution
7. Sequence File
8. Input File Formats
9. Output File Formats
10. Writable DataTypes
11. Custom Input Formats
12. Example List, show and run examples in map reduce.
13. Debugging Map Reduce Programs
14. Error Tracing of Map Reduce programs
15. Discussion on most common issues in MR.
16. Calculating the stats of MR Programs

Day11: (2hours)

Map Reduce Advance Concepts with usecases(Hands On):

17. Partitioning and Custom Partitioner
18. Joins

Hadoop

Course Duration: 25 days (60 hours duration).

19. Multi outputs
20. Counters
21. MR unit testcases
22. MR Design patterns
23. Distributed Cache
    a. Command line implementation
24. MapReduce API implementation

Day12: (2hours)

Sqoop:

1. Sqoop Theory
2. Demo for Sqoop and Practicals.
3. Sqoop Imports and Exports
4. Sqoop Tuning.

Day 13: (2hours)

Hive:

1. Hive Background.
2. What is Hive?
3. Where to Use Hive?
4. Hive Architecture
5. Metastore
6. Hive execution modes.
7. External, Managed, External tables.

Day 14: (2hours)

8. Hive Partitioning
9. Hive Bucketing
10. Hive Data Model
11. Hive Data Types
    a. Primitive
    b. Complex
12. Queries:
    c. Create Managed Table
    d. Load Data
    e. Insert overwrite table
    f. Insert into Local directory.
    g. CTAS.
    h. Insert Overwrite table select.
13. Joins
    a. Inner Joins
    b. Outer Joins

Hadoop

Course Duration: 25 days (60 hours duration).

c. Skew Joins

Day 15: (4hours)

14. Hive Sort By, Order By
15. Multi-table Inserts
16. Multiple files, directories, table inserts.
17. Serde
    a. RegexSerde
    b. AvroSerde
18. Storing in Sequence and ORC File Format
19. UDF
20. Hive through CLI, Batch and Hue
21. Hive Practical's and Usecases
22. Hive Configuration and Hive-site.xml
23. Optimizing hive queries
24. Best Practices in Hive
25. Debugging Hive Scripts and Error Tracing.
    a. Common Issues in Hive.
    b. Hive Optimization Techniques and Best Practices

Day 16: (2hours)

Pig:

1. Need of Pig?
2. Why Pig Created?
3. Introduction to skew Join.
4. Why go for Pig when Map Reduce is there?
5. Pig use cases.
6. Pig built in operators
7. Pig store schema.

Day 17: (2hours)

8. Operators:
    a. Load
    b. Store
    c. Dump
    d. Filter.
    e. Distinct
    f. Group
    g. CoGroup
    h. Join
    i. Stream
    j. Foreach Generate
    k. Parallel.
    l. Distinct

Hadoop

Course Duration: 25 days (60 hours duration).

m. Limit
n. ORDER
o. CROSS
p. UNION
q. SPLIT
r. Sampling

Day 18: (2hours)

9. Dump Vs Store
10. DataTypes
    a. Complex
        i. Bag
        ii. Tuple
        iii. Atom
        iv. Map
    b. Primitives.
        v. Integers
        vi. Float
        vii. Chararray
        viii. byteArray
        ix. Double

Day 19: (2hours)

11. Diagnostic Operators
    c. Describe
    d. Explain
    e. Illustrate
12. UDFs.
13. Physical and Logical Execution Plans
14. Storage Handlers.
15. Pig Practicals and Usecases.
16. Pig vs Hive
17. Testing suite's for Pig like Pig Lipstick, Pig Penny for debugging and error tracing.
18. Data loading using HCatalog Loader
19. Pig Debugging using Explain and Illustrate commands
20. Pig Stats

Day 20: (4hours)

Impala:

1. Impala Architecture
2. ImpalaD Deamon
3. Impala StateStore
4. Impala Catalog

Hadoop

Course Duration: 25 days (60 hours duration).

5. MPP Architecture
6. Impala Practicals
7. Adhoc Querying in Impala.
8. Impala integration with Hive

Day 21: (2hours)

Hadoop File Formats:

1. Sequence File
2. Avro File
3. ORC File Format
4. Parquet File Format
5. Storing Hive data in these File Formats
6. Comparing File Formats
7. Compression techniques like snappy, lzo, bgzip,etc
8. AVRO Shemas

Comparing BigData Execution Engines (Tez, MR, DAG, RDD, MPP)

Day 22: (2hours)

Introduction to NOSQL Databases:

1. Problem with RDBMS
2. Row Oriented vs Column Oriented
3. Introduction to NOSQL DB's
4. CAP Theorem

Day 23: (2hours)

HBase:

1. Introduction to NOSQL Databases.
2. NOSql Landscapes
3. Introduction to HBASE
4. HBASE vs RDBMS
5. Create Table on HBASE using HBASE shell
6. Where to use HBASE?
7. Where not to use HBASE?
8. Write Files to HBASE.
9. Major Components of HBASE.
    a. HBase Master.
    b. HRegionServer.
    c. HBase Client.
    d. Zookeeper.
    e. Region.
10. Compactions
11. HBase Practicals

Hadoop

Course Duration: 25 days (60 hours duration).

12. Bulk Loads
13. HBase Command Line
14. Using Map Reduce for HBase Operations
15. HBase Java Client Programming

Day 24: (2hours)

Flume:

1. Flume Architecture
2. Real time streaming in Flume
3. Defining and Deploying Flume Agents
4. Access data from multiple sources to collectors.
5. Different types of channels
6. Configuring Flume Agents
7. Running a Usecase

Day 25: (4hours)
Kafka:

1. Learn how to Develop Game-Changing Real Time Applications
2. Master kafka & its components
3. Understand architecture of kafka
4. Install kafka on single node as well as on multi-node cluster
5. Configure consumer, producer and brokers
6. Perform various Kafka Operations like adding and removing topics, modifying topics etc.

Oozie:

1. Oozie Architecture
2. Workflow designing in Oozie
3. Scheduling workflows in Oozie
4. Oozie practicals.
5. Automate the testing process using Oozie