HADOOP DEVELOPMENT TRAINING CURRICULUM - 60 HRS

1. Introduction

1.1 Big Data Introduction

- What is Big Data
- Data Analytics
- Bigdata Challenges
- Technologies supported by big data

1.2 Hadoop Introduction

- What is Hadoop?
- History of Hadoop
- Basic Concepts
- Future of Hadoop
- The Hadoop Distributed File System
- Anatomy of a Hadoop Cluster
- Breakthroughs of Hadoop
- Hadoop Distributions:
    - Apache Hadoop
    - Cloudera Hadoop
    - Horton Networks Hadoop
    - MapR Hadoop

2. Hadoop Daemon Processes

- **N**ame **N**ode
- **D**ata**N**ode
- **S**econdary **N**ame **N**ode/High Availability
- **J**ob **T**racker/Resource Manager
- **T**ask **T**racker/Node Manager

3. HDFS (Hadoop Distributed File System)

- Blocks and Input Splits
- Data Replication

- Hadoop Rack Awareness
- Cluster Architecture and Block Placement
- Accessing HDFS
  - JAVA Approach
  - CLI Approach

4. Hadoop Installation Modes and HDFS

- Local Mode
- Pseudo-distributed Mode
- Fully distributed mode
- Pseudo Mode installation and configurations
- HDFS basic file operations

5. Hadoop Developer Tasks

5.1 Writing a MapReduce Program

- Basic API Concepts
- The Driver Class
- The Mapper Class
- The Reducer Class
- The Combiner Class
- The Partitioner Class
- Examining a Sample MapReduce Program with several examples
- Hadoop's Streaming API
- Examining a Sample MapReduce Program with several examples
- Running your MapReduce program on Hadoop 1.0
- Running your MapReduce Program on Hadoop 2.0

5.2 Performing several hadoop jobs

- Sequence Files
- Record Reader
- Record Writer
- Role of Reporter
- Output Collector

- Processing XML files
- Counters
- Directly Accessing HDFS
- ToolRunner
- Using The Distributed Cache

5.3 Advanced MapReduce Programming

- A Recap of the MapReduce Flow
- The Secondary Sort
- Customized Input Formats and Output Formats
- Map-Side Joins
- Reduce-Side Joins

5.4 Practical Development Tips and Techniques

- Strategies for Debugging MapReduce Code
- Testing MapReduce Code Locally by Using LocalJobRunner
- Testing with MRUnit
- Writing and Viewing Log Files
- Retrieving Job Information with Counters
- Reusing Objects

5.5 Data Input and Output

- Creating Custom Writable and Writable-Comparable Implementations
- Saving Binary Data Using SequenceFile and Avro Data Files
- Issues to Consider When Using File Compression

5.6 Tuning for Performance in MapReduce

- Reducing network traffic with Combiner, Partitioner classes
- Reducing the amount of input data using compression
- Reusing the JVM
- Running with speculative execution
- Input Formatters
- Output Formatters
- Schedulers

    - FIFO schedulers
    - FAIR Schedulers
    - CAPACITY Schedulers

5.7 YARN

- What is YARN
- How YARN Works
- Advantages of YARN

6. Hadoop Ecosystems

6.1 PIG

- PIG concepts
- Install and configure PIG on a cluster
- PIG Vs MapReduce and SQL
- PIG Vs HIVE
- Write sample PIG Latin scripts
- Modes of running PIG
- Programming in Eclipse
- Running as Java program
- PIG UDFs
- PIG Macros
- Accessing Hive from PIG

6.2 HIVE

- Hive concepts
- Hive architecture
- Installing and configuring HIVE
- Managed tables and external tables
- Partitioned tables
- Bucketed tables
- Complex data types
- Joins in HIVE
- Multiple ways of inserting data in HIVE tables
- CTAS, views, alter tables

- User defined functions in HIVE
  - Hive UDF
  - Hive UDAF
  - Hive UDTF

6.3 SQOOP

- SQOOP concepts
- SQOOP architecture
- Install and configure SQOOP
- Connecting to RDBMS
- Internal mechanism of import/export
- Import data from Oracle/Mysql to HIVE
- Export data to Oracle/Mysql
- Other SQOOP commands

6.4 HBASE

- HBASE concepts
- ZOOKEEPER concepts
- HBASE and Region server architecture
- File storage architecture
- NoSQL vs SQL
- Defining Schema and basic operations
  - DDLs
  - DMLs
- HBASE use cases
- Access data stored in HBASE using clients like CLI, and Java
- Map Reduce client to access the HBASE data
- HBASE admin tasks

6.5 OOZIE

- OOZIE concepts
- OOZIE architecture
  - Workflow engine
  - Job coordinator
- Install and configuring OOZIE

- HPDL and XML for creating Workflows
- Nodes in OOZIE
  - Action nodes
  - Control nodes
- Accessing OOZIE jobs through CLI, and web console
- Develop sample workflows in OOZIE on various Hadoop distributions
  - Run HDFS file operations
  - Run MapReduce programs
  - Run PIG scripts
  - Run HIVE jobs
  - Run SQOOP Imports/Exports

6.6 FLUME

- FLUME Concepts
- FLUME architecture
- Installation and configurations
- Executing FLUME jobs

6.7 IMPALA

- What is Impala
- How Impala Works
- Imapla Vs Hive
- Impala's shortcomings
- Impala Hands on

6.8 ZOOKEEPER

- ZOOKEEPER Concepts
- Zookeeper as a service
- Zookeeper in production

7. Integrations

- Mapreduce and HIVE integration
- Mapreduce and HBASE integration

- Java and HIVE integration
- HIVE - HBASE Integration
- SAS – HADOOP

8. Spark

- Introduction to Scala
- Functional Programming in Scala
- Working with Spark RDDs

9. Hadoop Administrative Tasks:

Setup Hadoop cluster: Apache, Cloudera and VMware

- Install and configure Apache Hadoop on a multi node cluster
- Install and configure Cloudera Hadoop distribution in fully distributed mode
- Install and configure different ecosystems
- Basic Administrative tasks

10. Course Deliverables

- Workshop style coaching
- Interactive approach
- Course material
- Hands on practice exercises for each topic
- Quiz at the end of each major topic
- Tips and techniques on Cloudera Certification Examination
- Linux concepts and basic commands
- On Demand Services
  - Mock interviews for each individual will be conducted on need basis
  - SQL basics on need basis
  - Core Java concepts on need basis
  - Resume preparation and guidance
  - Interview questions