Hadoop Course Content

1. ***Introduction to Hadoop and its Ecosystem, Map Reduce and HDFS***

   - Big Data, Factors constituting Big Data
   - What is Hadoop?
   - Overview of Hadoop Ecosystem
   - Map Reduce -Concepts of Map, Reduce, Ordering, Concurrency, Shuffle, Reducing, Concurrency
   - Hadoop Distributed File System (HDFS) Concepts and its Importance
   - Deep Dive in Map Reduce – Execution Framework, Partitioner, Combiner, Data Types, Key pairs
   - HDFS Deep Dive – Architecture, Data Replication, Name Node, Data Node, Data Flow
   - Parallel Copying with DISTCP, Hadoop Archives

Assignment - 1

2. ***Hands on Exercises***
   - Installing Hadoop in Pseudo Distributed Mode, Understanding Important configuration files, their Properties and Demon Threads
   - Accessing HDFS from Command Line
   - Map Reduce – Basic Exercises
   - Understanding Hadoop Eco-system
   - Introduction to Sqoop, use cases .
   - Introduction to Hive, use cases.
   - Introduction to Pig, use cases.
   - Introduction to Oozie, use cases.
   - Introduction to Flume, use cases.
   - Introduction to Yarn

Assignment - 2 & 3

   - Mini Project – Importing Mysql Data using Sqoop and Querying it using Hive

3. ***Deep Dive in Map Reduce and Yarn***
   - How to develop Map Reduce Application, writing unit test
   - Best Practices for developing and writing, Debugging Map Reduce applications
   - Joining Data sets in Map Reduce
   - Hadoop API's

- Introduction to Hadoop Yarn
- Difference between Hadoop 1.0 and 2.0

- Project 1- Hands on exercise – end to end PoC using Yarn or Hadoop 2.
    a) Real World Transactions handling of Bank
    b) Moving data using Sqoop to HDFS
    c) Incremental update of data to HDFS
    d) Running Map Reduce Program
    e) Running Hive queries for data analytics
- Project 2- Hands on exercise – end to end PoC using Yarn or Hadoop 2.0
    a) Running Map Reduce Code for Movie Rating and finding their fans and average rating

Assignment -4 & 5

4. Deep Dive in Pig

- ***Introduction to Pig***
    - What Is Pig?
    - Pig's Features
    - Pig Use Cases
    - Interacting with Pig
- ***Basic Data Analysis with Pig***
    - Pig Latin Syntax
    - Loading Data
    - Simple Data Types
    - Field Definitions
    - Data Output
    - Viewing the Schema
    - Filtering and Sorting Data
    - Commonly-Used Functions
    - Hands-On Exercise: Using Pig for ETL Processing
- ***Processing Complex Data with Pig***
    - Complex/Nested Data Types
    - Grouping
    - Iterating Grouped Data
    - Hands-On Exercise: Analyzing Data with Pig

Assignment - 6

5. *Deep Dive in Hive*

- ***Introduction to Hive***
  - What Is Hive?
  - Hive Schema and Data Storage
  - Comparing Hive to Traditional Databases
  - Hive vs. Pig
  - Hive Use Cases
  - Interacting with Hive
- ***Relational Data Analysis with Hive***
  - Hive Databases and Tables
  - Basic HiveQL Syntax
  - Data Types
  - Joining Data Sets
  - Common Built-in Functions
  - Hands-On Exercise: Running Hive Queries on the Shell, Scripts, and Hue
- ***Hive Data Management***
  - Hive Data Formats
  - Creating Databases and Hive-Managed Tables
  - Loading Data into Hive
  - Altering Databases and Tables
  - Self-Managed Tables
  - Simplifying Queries with Views
  - Storing Query Results
  - Controlling Access to Data
  - Hands-On Exercise: Data Management with Hive
- ***Hive Optimization***
  - Understanding Query Performance
  - Partitioning
  - Bucketing
  - Indexing Data

Assignment - 7

6. *Introduction to Hbase architecture*

- What is Hbase
- Where does it fits

- What is NOSQL

Assignment -8

7. ***Hadoop Cluster Setup and Running Map Reduce Jobs***
    - ❖ Running Map Reduce Jobs on Cluster

Assignment - 9& 10

8. ***Advance Mapreduce***
    - ❖ Delving Deeper Into The Hadoop API
    - ❖ More Advanced Map Reduce Programming, Joining Data Sets in Map Reduce
    - ❖ Graph Manipulation in Hadoop

Assignment - 11&12

9. ***Job and certification support***
    - ❖ Major Project, Hadoop Development, cloudera Certification Tips and Guidance and Mock Interview Preparation, Practical Development Tips and Techniques, certification preparation

Project Work

I. ***Project:-*** Working with Map Reduce, Hive, Sqoop

Problem Statement:It describes that how to import mysql data using sqoop and querying it using hive and also describes that how to run the word count mapreduce job.

II. ***Project:-***Hadoop Yarn Project – End to End PoC

Problem Statement:It includes:-

- ❖ Import Movie data
- ❖ Append the data
- ❖ How to use sqoop commands to bring the data into the hdfs
- ❖ End to End flow of transaction data
- ❖ How to process the real word data or huge amount of data using map reduce program in terms of movie etc.

Hue:

- More Ecosystems
- HUE (Cloudera).
- Oozie
- Workflow (Action, Start, Action, End, Kill, Join and Fork), Schedulers, Coordinators and Bundles.
- Workflow to show how to schedule Sqoop Job, Hive, MR and PIG.
- Real world Use case which will find the top websites used by users of certain ages and will be scheduled to run for every one hour.
- Zoo Keeper

Impala:

- ***Objective***
- ***Impala Architecture***
  - Impala Daemon
  - Impala Statestore
  - Impala Catalog Service
- ***Impala Query Processing Interfaces***
  - Impala-shell
  - Hue interface
  - ODBC/JDBC drivers
- ***Impala Query Execution Procedure***
- ***Apache Kafka:***
  - Kafka producer
  - Kafka consumer
  - Kafa topics

Project:

- ***Kafka with spark streaming and mongodb.***