SIEVE ***SOFTWARE*** **Flat. No: 307A, 3rd floor, adhitya enclave, nilgiri block, ameerpet, hyd. 8019242423, 7386977110...**

Duration: 45 days

INTRODUCTION TO BIG DATA & HADOOP

- What is big data ?
- What are the challenges for processing big data ?
- What technologies support big data ?
- What is hadoop ?
- Why hadoop ?
- History of hadoop
- Use cases of hadoop
- Hadoop eco system
- HDFS
- Map reduce
- Statistics understanding the cluster
- Typical workflow
- Writing files to HDFS
- Reading files from HDFS
- Rack awareness
- 5 daemons

HDFS commands hands on

Installing the cluster (cdh4 pseudo Cluster & configuration)

- CDH4 Pseudo Cluster
- Configuration

- Hands –on exercises
    - Routine Admin and Monitoring Activities
- Name node High Availability

Hadoop Architecture and HDFS

- Hadoop 2.x Cluster Architecture
- Splits and blocks
- Input splits
- Hdfs splits
- Replication of data
- Awareness of Hadoop racking
- Hadoop cluster modes
- Common Hadoop Shell Commands
- Hadoop 2.x Configuration Files (xml files)
- Data loading techniques : Hadoop copy commands
- Hands-on Exercises

Map Reduce

- Before MapReduce
- MapReduce Overview
- Word count Flow and solution
- Word count Problem
- Mapreduce Flow
- Algorithms for simple problems
- Algorithms for complex problems

Developing the MapReduce Application

- Data types
- File formats
- Explain the driver,Mapper and reducer code

- Configuring development environment – Eclipse
- Writing unit test
- Running locally
- Running on cluster
- Hands on exercises

How map reduce works

- Map reduce Job run
- Job Submission
- Job initialization
- Task Assignment
- Job scheduling
- Job Failures
- Shuffle and sort
- Oozie workflows
- Hands on exercises

Map reduce Types and formats

- Input Formats – binary input, database input, text Input, Input splits & records etc..
- Output Formats – text Output, binary output, database output
- Hands –on Exercises

Map reduce Features

- Counters
- Joins – Map side and reduce side | Sorting
- MapReduce Combiner
- MapReduce Partitioner
- MapReduce Distributed Cache
- Hands- on Exercises

Hive

- What is hive?
- What Hive is not?
- Hive Architecture
- SQL vs Hive QL
- Data Types
- Managed Tables and external Tables
- Partitions | Buckets
- Storage Formats
- SerDes
- Importing data
- Joins
- UDFS
- Hands-on Exercises

Pig

- What is pig?
- Running pig?
- Data types
- Pig lation statements
- Schemas | validations
- Functions and macros
- UDFs
- When to use PIG and HIVE
- Hands-on Exercises

No SQL and HBASE

- Why No SQL?
- Problems with RDBMS
- CAP theorem

- HBASE Concepts
- Use cases for HBASE
- HBase data model | Hbase shell
- HBASE Architecture
- Minor and major compaction
- Bloom filter & Block filter
- Schema design
- Hands-on Exercises

Sqoop

- What is Sqoop
- Sqoop commands
- Importing data to (HBASE,HIVE,HDFS)
- Exporting Data
- Sqoop connectors
- Hands-on Exercises

Impala

- Impala Overview
- Impala Architecture
- Hands-on Exercises

Flume

- What is Flume
- Use cases
- Flume topology
- Hands-on Exercises

Introduction to Oozie

- Oozie Components
- Oozie Workflow
- Scheduling with Oozie
- Hands-on Exercises

Project – case study (Torrent data and Analyze with Hive)

Optional: (if required)

1. **SQL (oracle)**
2. **Core Java**