HANZOLABS

Course Contents

Introduction and Motivation of Hadoop
- What is Big Data
- Challenges in Big Data
- Challenges in Traditional Application
- New Requirements
- What is Hadoop
- Brief history of Hadoop
- Features of Hadoop
- Hadoop v/s RDBMS
- Hadoop Ecosystem's overview
- Overview of HDFS and MapReduce

Understanding Hadoop Distributed File System
- Understanding Configuration
- HDFS Concepts
  - Blocks
  - Replication
  - Version File
  - Safe mode
  - Namespace IDs
- Reading and Writing in HDFS
- Understanding NameNode
- Understanding Data Node
- Understanding Secondary NameNode
- Understanding Job Tracker
- Understanding Task Tracker

HDFS Shell Commands
- Hands On Exercise

Accessing HDFS using API
- Understanding HDFS Java classes and methods
- Hands On Exercise

Map Reduce Programming
- Understanding block and input splits
- Common Input and Output Formats - MapReduce Data types
- Understanding Writable and WritableComparable (Introduction)
  - Data Flow in MapReduce Application
- Understanding WordCount problem
- Writing MapReduce Application
  - Understanding Mapper function
  - Understanding Reducer Function
  - Understanding Driver
  - Understanding Tool Runner
  - Hands on Exercise

- Older and Newer API
- Solving common problems
  - Average word length
  - Inverted Index
  - Word Co-Occurrence
  - Searching
  - Sorting
  - Hands on exercise

MapReduce Continued
- Using Combiner
- Using Distributed Cache
- Passing the parameters to mapper and reducer
- Using Counter
- Hands On Exercise
- Writing Custom key values
- Writing Custom Partitioner
- Hands On Exercise
- Writing Custom Input Format
- Using Hadoop Archives

Secondary Sorting
- Motivation
- Understanding
- Hands On

Joins
- Map Side Join
- Reduce Side join
- Hands on

Practical development
- Calculating Number of Reducers
- Map Only Jobs
- Compression

Introduction to PIG
- Terminology
- Understanding Pig Program, structure and Execution
- Pig Data types
- Loading and Dumping Data
- Filtering
- Group and Co-Group
- Joins
  - Inner Join
  - Left Outer Join
  - Right Outer Join
  - Full Outer Join
- Schema Merging and Redefining Schema
- Pig Functions

- Writing Custom UDFs
- Hands on

Introduction to HIVE
- Motivation and Understanding Hive
- Using Hive Command line Interface
- Data types and File Formats
- Basic DDL operations
- Schema Design

Introduction to Pig –

- Motivation and Understanding Pig
- Using Pig Latin Scripts
- Data types and File Formats
- Basic DDL operations
- Schema Design

SQOOP
- Motivation and Understanding Sqoop
- Sqoop Configuration
- Sqoop and MySql Connection
- Import and Export Operator
- Troubleshooting

Flume
- Motivation and Understanding Flume
- Flume Configuration
- Flume Agents
- Weblog Crawling using Flume
- Troubleshooting

R
- Intro to R
- Descriptive Statistics
- Hypothesis Testing
- Regression
- Manipulating Data and Data Extraction

Machine Learning
- Regression Techniques
- Clustering Techniques
- Anomaly Detection
- Recommender Systems

Data Analytics
- Bayesian statistics
- Regression
- Hypothesis testing

- Data visualization
- Histograms