

# Predictive Analytics using R and Hadoop

## Table of Contents:

- 1. Hadoop Distributed File System (Hadoop 2.0)** **(Day 1)**
  - a. Origin of Bigdata, challenges of big data, attributes of big data
  - b. HDFS architecture 2.0, a solution to big data challenges
    - i. Introduction to Name node, data nodes and their functions
    - ii. Interactions between Name nodes and data nodes
    - iii. HDFS parameters including block size, replication factor & strategy, rack awareness
  - c. Hadoop file formats advantages and disadvantages
  - d. Schema design good practices
  - e. Uploading and downloading files to and from HDFS (commonly used commands)
  - f. (Optional - Install 4 node hadoop cluster on Amazon cloud)
- 2. Map Reduce programming paradigm (ver 2.0 / YARN)** **(Day 1)**
  - a. Map reduce framework with all the stages in a map reduce job
  - b. Introduction to mapper and reducer classes and functions in Java
  - c. Hello World of map reduce program – word count
  - d. **Lab 1** – Write mapreduce programs to generate simple reports on dummy data
  - e. Anatomy of map reduce work flow
- 3. Introduction to Hive and HQL based programming** **(Day 2)**
  - a. Overview
  - b. Hive internals (interaction between Hive, HDFS and Metastores services)
  - c. Creating Hive objects such as database, tables and partitions
  - d. Understanding Hive managed and external tables
  - e. Understanding partitioned tables
  - f. Introduction to Hive Query Language
  - g. **Lab 2** - Write queries to generate simple reports
- 4. Introduction to Flume . Capture streaming data of tweets from witter.** **(Day 3)**
  - a. Overview
  - b. Invoking Flume
  - c. Configuring Flume for twitter downloads
  - d. **Lab 3** -Capture live tweets (Internet should be accessible)

## 5. Introduction to Sqoop

(Day3)

- a. Overview
  - b. Invoking Sqoop
  - c. Configuring Sqoop to import/ export data from HDFS to mysql and vice versa
6. **Lab 4** - Develop a working prototype for analyzing tweets on pre-fixed topics. Define external tables in hive to store raw tweets. Extract user text from raw tweets. Using sqoop, move user text to mysql database, perform sentiment analysis on user text using Naive Bayesian algorithms in Python scripts

## 7. Introduction to R and R based Machine Learning

(Day 4 - 8)

- a. R Data Types
  - a. Installing R, RStudio
  - b. R Datatypes
  - c. Basic syntax
  - d. Variables
  - e. Vectors
  - f. Matrices
  - g. DataFrames
  - h. Lists
- b. Data Interfaces
  - a. CSV files
  - b. Excel files
  - c. Text files
  - d. Databases
  - e. Web links
- c. Charts and graphs
  - a. Pie chart
  - b. Line charts
  - c. Scatter plots
  - d. Histograms
- d. Basic statistical concepts
  - a. Descriptive statistics (Mean, Median, Mode, Spread measures)
  - b. Normal distribution and it's characteristics
  - c. Standard scores, confidence levels and intervals
- e. Machine Learning for Predictive Data Analytics
  - a. What is predictive Data Analytics
  - b. Predictive Data Analytics Project Lifecycle - CRISP-DM
  - c. What is machine learning models
- f. Data Exploration / Preparing data for analytics
  - a. Identifying data quality problems
  - b. Handling data quality issues
  - c. Tidy Data
  - d. Analysing attributes using tools such as SPLOM (Scatter Plot Matrix)
  - e. Dimensionality Reduction concept / method
- g. Supervised Learning Methods (Concepts, creating, evaluating and improvising)
  - a. Support Vector Machines
  - b. Neural Networks
  - c. Decision trees
  - d. Naive Bayesian classifiers

- e. K-NN Clustering
  - f. Linear Regression
- h. Unsupervised Learning Methods (Concepts, creating, evaluating and improvising)
  - a. K-Means Clustering
  - b. Association rules mining
- i. Important concepts of machine learning
  - a. Curse of dimensionality
  - b. Dimensionality reduction
  - c. Linearly separable data, Kernels and kernel tricks
  - d. Bias variance & model effectiveness
  - e. Model performance measures