BD5211 – BIG DATA QUERY LANGUAGES LABORATORY

OBJECTIVES:

- To understand the basic programming constructs of R and understand the use of R in Big Data analytics.
- To solve Big data problems using Map Reduce Technique in R, HADOOP.
- To develop Pig scripts for analyzing large un-structured and semi-structured data.
- To develop program for Query processing using Hive. • To perform analytics on Big data streams using Hadoop Streaming API.
- To learn to work on Sqoop.

LIST OF EXPERIMENTS:

1. Perform descriptive and predictive analytics using “R programming”
2. MapReduce application for word counting on R HADOOP after successful installation of three R packages(rhdfs, rmr, and rhbase)
3. Understand data pipeline using Pig Interactive Shell Commands after successful “Pig” installation
4. Develop Pig Scripts and call UDF’s to accomplish functionalities to meet the problem objectives
5. Embedding PIG Latin in Python
6. Log analysis using “Pig” on semi structured data
7. Perform query processing on data warehousing after successful installation of “Hive”
8. Perform adhoc query on HDFS data using Hive Query Language (HQL)
9. Accomplish MapReduce Job by using Hadoop Streaming API
10. Perform various HDFS commands
11. Loading data into HDFS using Sqoop

Exercise I (22.06.2022) – Explore R

1. **Download R and explore basic commands**
2. **Download the file**

   http://stats.ma.ic.ac.uk/das01/public_html/RCourse/hills.txt

   and save it in the directory. This file contains the Scottish hill races data set. Now, explore the following commands

   (i) Read the file, assigning the result to the object hills
   (ii) Examine the object. Note column and row names
   (iii) Construct a scatterplot matrix
   (iv) How many variables do you think such plots are suitable for?
   (v) Make the columns of the hills object available by name
   (vi) Construct a scatter plot. The function call in this way means the first argument is the horizontal axis
   (vii) Interact with the plot to label points - right click to finish
   (viii) Compute a linear regression of time against distance
   (ix) Obtain more information about the regression
   (x) Add the least squares regression line - note anonymous function call
   (xi) Obtain some diagnostics plots - note the different arguments to the plot function. Be aware of the prompt in the Console.
   (xii) there are many pre-defined system objects. Display the value of pi - note that this is a reserved word
   (xiii) List objects in current working space
   (xiv) Display the object ls
   (xv) Create a copy of the hills object
   (xvi) List objects in current working space
   (xvii) Delete the copy. Note that there is no undelete functionality.

3. **Vectors and Arithmetic**
   (i) Create a vector of coefficients for a quadratic equation, using the sample function. Here, we draw a sample of size 3 from $-20, -19, \ldots, 19, 20$ with replacement
   (ii) Determine the class of the object coeffs.
   (iii) Determine the length of the object coeffs
   (iv) Determine the names associated with the vector
   (v) Assign some names. Note the function call occurring on the right hand of the assignment operator.
   (vi) Prepare to plot the equation, by constructing a regularly spaced vector for the horizontal axis
   (vii) Evaluate the quadratic at each point in vector x
   (viii) Construct the plot
   (ix) Does the equation have real roots? Compute the discriminant
   (x) Create a vector of type character, and display the second element
4. Compute the real roots of the quadratic equation
   $x\,2 + x + 1 = 0$
5. Generate a regular grid between -50 and 50. Construct separate plots of log(x), exp(x), sin(x), sin(2x), sin(x)/cos(x). Examine the cumulative sum of the final function. Experiment with the argument type of the plot function.