Department: *Computer Science and Engineering*	***Course Type:*** *Programme Core*
Course Title*: Big Data Technologies Laboratory*	***Course Code:****18CSL66*
L-T-P:*0-0-2*	***Credits:****1*
Total Contact Hours:*36 Hours*	***Duration of SEE:****3Hours*
SEE Marks*: 50*	***CIE Marks:****50*

COURSE DESCRIPTION

In this course you will learn how to program in R and how to use R for effective data analysis and will learn scala and Spark to drive better business decisions and solve real-world problems.

PREREQUISITES

- Basics of Java (preferred), Python or another object-oriented language.

COURSE OBJECTIVES

At the end of the course students will be able to

1. Get a solid understanding of the fundamentals of the language, the tooling, and the development process.
2. Tackle data analysis problems involving Big Data, Scala and Spark.
3. Design and write efficient programs using R to perform routine and specialized data manipulation/management and analysis tasks.
4. Document, share, and collaborate on code development using a suite of Open Source standards and tools.
5. Develop a good application of more advanced features.

LAB EXERCISES

Part A

Implement the following exercises using R

1. **a.** Create three different variables, one that is numeric type and other two are vector of characters. Use these to create data frame of student.(USN, Name, Marks)
   b. Add a new numeric data column to the existing data frame (Age). Provide summary of the data
   c. Display the list of students whose Age is less than 20 and Marks greater than 25.
2. Write a program to create the csv file for storing Employee data, containing the fields (EmpID, EmpName, DOJ,Dept, Desig.)

a. Read the suitable number of employee details from the user.
b. Create a dataframe of Employee
c. Store the dataframe in the csv file
d. Read the data from csv and Display the contents
e. Append a new row into the csv file

3. Exploring Dataset
a. List the data set available in your system using suitable command
b. Select "mtcars" data set, find and display the number of rows and columns in that data set
c. Find are there more automatic (0) *or* manual (1) transmission-type cars in the dataset?
Hint: 9th column indicates the transmission type
d. Get a scatter plot of 'hp' vs 'weight'.
e. Change 'am', 'cyl' and 'vs' to *integer* and store the new dataset as 'newmtc'.
f. Extract the cases where cylinder is less than 5

4. Consider "Airquality" dataset
a. Display the dimension ofthe dataset
b. Display the class of each fields in the data set
c. Test the missing values
d. Recode the missing values, as mean ofthe column values
e. Exclude the missing values

Implement the following exercises using Scala

5. Write a program that reads words from a file. Use a mutable map to count how often each word appears.
6. Write a function minmax (values: Array[Int]) that returns a pair containing thesmallest and largest values in the array.
7. Write the menu driven program to implement quick sort algorithm using imperative style and functional style.
8. Write the program to illustrate the use of pattern matching in scala, for the following
Matching on case classes. Define two case classes as below:
abstractclassNotification
caseclassEmail(sender: **String**, title: **String**, body: **String**) **extendsNotification**
caseclassSMS(caller: **String**, message: **String**) **extendsNotification**
Define a function showNotification which takes as a parameter the abstract type Notification and matches on the type of Notification (i.e. it figures out whether it's an Email or SMS).
In the case it's an Email(email, title, _) return the string: s"You got an email from $email with title: $title"
In the case it's an SMS return the String: s"You got an SMS from $number! Message: $message"

Part B

Implement the following exercises using Spark

9. WordCount: Here the goal is to count how many times each word appears in a file and write out a list ofwords whose count is strictly greater than 4.

Use the file log.txt accompanying this assignment to count the words.Save the wordcounts in text form in the "wordcountsDir" using the saveAsTextFile RDD method.Examine the contents ofthe above directory, andthe contents ofthe files ofthe directory.

10. Tweet Mining: A dataset with the 8198 reduced tweets, reduced-tweets.json will be provided.The data contains reduced tweets as in the sample below:

{"id":"572692378957430785",
"user":"Srkian_nishu :)",
"text":"@always_nidhi @YouTube no idnt understand bti loved of this mve is rocking",
"place":"Orissa",
"country":"India"}

A function to parse the tweets into an RDD will be provided. The task is to print the top 10 tweeters.

Self Demonstartion of the below programs

1. IPLTossWinStats: You will be provided with a dataset from the Indian Premier League containing the following files:
   Ball_by_Ball.csv, Match.csv, Player.csv, Player_Match.csv, Season.csv, Team.csv.
   We want to find the percentage of game wins by teams which win the toss.Solets say N games have been played. Let us say there are M games where the team which has won the toss has also won the game. So we are looking for the percentage (M * 100 / N).Perform the task using SQL code only.
2. Streaming Rainfall Averages: Consider the scenario that there are three weather stations in Bangalore which report the rainfall at the respective locations once every 15 minutes. You have to write a Spark Streaming application which will gather the rainfall data from the three stations and print the average rainfall, also once every 15 minutes.
   You will be provided with a scala program, generate Events, which can simulate generation of the rainfall data from the three stations in JSON format as shown below to a folder:{"Creation_Time": 1.53633593969400013E18,"Station": "Bengaluru-1","Rainfall": 100.0} Write a Spark streaming application which reads the files written to the above folder and updates the average rainfall value every 15 minutes and prints the averages to the console.

ASSESSMET METHODS:

Parameters	Marks
Experiment Write up + Execution + Viva	15
Lab Record Writing	10
Lab Internals Test	15
Total	**50**
Final Exam will be conducted for 100 marks (SEE)	

COURSE OUTCOMES

At the end of the course student will be able to

COs	Description	Bloom's Level
CO 1	Apply the concept of R programming for cleansing, imputation, and computation of simple statistical measures on the data.	**L3**
CO 2	Understand the basics of Scala for data analysis.	**L2**
CO 3	Design a Spark code for basic data manipulation and aggregate analysis.	**L4**
CO 4	Implement real time application such as word count, data mining into a set of distributed computations and implement the same in Spark using Scala.	**L3**
CO 5	Implement analytics on higher level data objects like Tables using Spark SQL and analytics on Streaming datasets using Spark Streaming.	**L3**

Mapping of Course outcomes (COs) to Program outcomes (POs*)& PSO **

	Course Outcomes mapping to Program Outcomes												PSOs		
POs / COs	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**PSO1**	**PSO2**	**PSO3**
CO1	3	2											2		
CO2	3	3	1			2	2							2	
CO3	3	3	2		2									2	
CO4	3	3	2	3	2									2	
CO5	3	3	3		3	2									2

3: Strong, 2: Medium, 1: Weak ** H: Highly related S: Supportive