ADL332	BIG DATA ANALYTICS LAB	CATEGORY	L	T	P	CREDIT	YEAR OF INTRODUCTION
		PCC	0	0	3	3	2019

Preamble: The purpose of the course is to offer the students a hands-on experience on Big Data concepts using open source technologies such as Hadoop, Map Reduce, Hive, Pig and Apache Spark. The hands-on experience with R Programming language helps in statistical analysis and equip the students with data driven solutions for the next-generation data management. As data continues to grow it is known that via big data solutions, organizations generate insights and make well-informed decisions, discover trends, and improve productivity and the learner will be able to work on and solve data processing problems.

Prerequisite: Fundamental knowledge in Java programming, Statistics and Python and Big Data Analytics

Course Outcomes: At the end of the course, the student should be able to :

CO1	Illustrate the setting up of and Installing Hadoop in one of the three operating modes.**(Cognitive knowledge: Understand)**
CO2	Implement the file management tasks in Hadoop and explore the shell commands **(Cognitive knowledge: Apply)**
CO3	Implement different tasks using Hadoop Map Reduce programming model.**(Cognitive knowledge: Apply)**
CO4	Implement Pig Scripting operations and Spark Application functionalities.**(Cognitive knowledge: Apply)**
CO5	Implement data extraction from files and other sources and perform various data manipulation tasks on them using R Program.**(Cognitive knowledge: Apply)**
CO6	Illustrate the knowledge of R gained to data analytics for real life applications. **(Cognitive knowledge: Understand)**

Mapping of course outcomes with program outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	✓	✓			✓			✓		✓		✓
CO2	✓	✓	✓		✓			✓		✓		✓
CO3	✓	✓	✓		✓			✓		✓		✓
CO4	✓	✓	✓		✓			✓		✓		✓
CO5	✓	✓	✓		✓			✓		✓		✓
CO6	✓	✓	✓		✓			✓		✓		✓

Abstract POs defined by National Board of Accreditation			
PO#	**Broad PO**	**PO#**	**Broad PO**
PO1	Engineering Knowledge	PO7	Environment and Sustainability
PO2	Problem Analysis	PO8	Ethics
PO3	Design/Development of solutions	PO9	Individual and teamwork
PO4	Conduct investigations of complex problems	PO10	Communication
PO5	Modern tool usage	PO11	Project Management and Finance
PO6	The Engineer and Society	PO12	Lifelong learning

Assessment Pattern:

Bloom's Category	Continuous Assessment Test(Internal Exam) Marks in percentage	End Semester Examination Marks in percentage
Remember	20	20
Understand	20	20
Apply	60	60
Analyse		
Evaluate		
Create		

Mark Distribution

Total Marks	CIE Marks	ESE Marks	ESE Duration
150	**75**	**75**	**3 hours**

Continuous Internal Evaluation Pattern:

Attendance : 15 marks
Continuous Evaluation in Lab : 30 marks
Continuous Assessment Test : 15 marks
Viva Voce : 15 marks

Internal Examination Pattern: The marks will be distributed as Algorithm30 marks, Program 20 marks, Output 20 marks and Viva 30 marks. Total 100 marks which will be converted out of 15 while calculating Internal Evaluation marks.

End Semester Examination Pattern: The percentage of marks will be distributed as Algorithm 30 marks, Program 20 marks, Output 20 marks and Viva30 marks. Total 75 marks.

Operating System to Use in Lab : Linux
Compiler/Software to Use in Lab :
Programming Language to Use in Lab : Java, R, Python

Fair Lab Record:
All Students attending the Big Data Lab should have a Fair Record. The fair record should be produced in the University Lab Examination. Every experiment conducted in the lab should be noted in the fair record. For every experiment in the fair record, the right-hand page should contain Experiment Heading, Experiment Number, Date of experiment, Aim of the Experiment and the operations performed on them, Details of experiment including algorithm and result of Experiment. The left-hand page should contain a print out of the code used for experiment and sample output obtained for a set of input.

SYLLABUS

BIG DATA ANALYTICS LAB

* Mandatory

1. Perform setting up and Installing Hadoop in any of the three operating modes: Standalone, Pseudo distributed, Fully distributed.*
2. Explore the various shell commands in Hadoop.
3. Implement the following file management tasks in Hadoop:
   - Adding Files and Directories
   - Retrieving Files
   - Deleting Files
4. Implement a word count program using Map Reduce.
5. Write a R program to find the factorial and check for palindromes.*
6. Write a R program to solve linear regression and make predictions.*
7. Write a R program to solve logistic regression.*
8. Implement statistical operations using R.*
9. Implement a program to find variance, covariance and correlation between different types of attributes.*
10. Implement SVM/Decision tree Classifier.*
11. Implement clustering algorithm.*

12. To explore Hive with its basic commands
13. Write Pig Latin scripts to sort, group, join, project, and filter your data.
14. Install, Deploy and configure Apache Spark.

BIG DATA PROCESSING LAB - PRACTICE QUESTIONS

1. Write a MapReduce Program to retrieve data from documents.
2. Write word count program that only count the words starting with 'a'
3. Write a word count program that only counts the words whose length is longer than 10.
4. Using the structure of the Word Count program, write a Hadoop program that calculates the average word length of all words that start with each character.
5. Implement matrix multiplication with Hadoop Map Reduce
6. Write a Map Reduce program for removing stop words from the given text files.
7. Write a MapReduce Program to count the number of lines in a document.
8. Write Pig Latin script to count the number of occurrences of each word in an input text file.
9. Write a program to simulate Singular Value Decomposition
10. Write a program to simulate PCA.
11. Write a single Spark application that:
    a. Transposes the original Amazon food dataset, obtaining a Pair RDD of the type: user-id – list of the product-ids reviewed by user-id
    b. Counts the frequencies of all the pairs of products reviewed together;
    c. Writes on the output folder all the pairs of products that appear more than once and their frequencies.
    d. The pairs of products must be sorted by frequency..
12. Write a program to implement a stop word elimination problem. Input: A large textual file containing one sentence per line. A small file containing a set of Stop Words (One Stop Word per line) Output: A textual file containing the same sentences of the large input file without the words appearing in the small file
13. Implement matrix multiplication with Map Reduce.
14. Implement basic Pig Latin Scripts based on different scenarios.
15. Implement Frequent Item set algorithm

16. Implement Clustering algorithm
17. Implement Page Rank algorithm
18. Implement Bloom Filter
19. Write a R program to create a sequence of numbers from 20 to 50 and find the mean of numbers from 20 to 60 and sum of numbers from 51 to 91.
20. Write a R program to create a vector which contains 10 random integer values between -50 and +50.
21. Write a R program to find the maximum and the minimum value of a given vector.
22. Write a R program to get the unique elements of a given string and unique numbers of vectors.
23. Write a R program to create a list of random numbers in normal distribution and count occurrences of each value.
24. Write a R program to read the .csv file and display the content.
25. Write a R program to create an array, passing in a vector of values and a vector of dimensions. Also provide names for each dimension.
26. Write a R program to create a simple bar plot of five subjects' marks.
27. Write a R program to compute the sum, mean and product of a given vector element.
28. Write a R program to create a Data Frames which contain details of 5 employees and display the details.