Course Title	Big data Analytics Lab				Course Type		HC	
Course Code	M24TC0202	Credits	1		Class		II Semester	
Course Structure	TLP	Credits	Contact Hours	Work Load	Total Number of Classes Per Semester		Assessment in Weightage	
	Lecture	1	2	2				
	Tutorial	-	-	-	Theory	Practical	CIE	SEE
	Practical	-	-	-				
	Total	1	2	2	-	28	25	25

COURSE OVERVIEW:

This course is to familiarize the students with most important information technologies used in manipulating, storing, and analyzing big data. The basic tools for big data analysis: Python, Hadoop HDFS, Hadoop MapReduce, Pig, Hive and Flume are demonstrated in this course through the demonstration of real life examples.

COURSE OBJECTIVES:

1. Discuss the fundamentals of Hadoop distributed file system and Big Data Analytics.
2. Demonstrate Big Data Processing with MapReduce and Batch Analytics.
3. Describe the implementation of Real-Time Analytics with Apache Hadoop in real world Applications.
4. Illustrate the working of Pig, Hive and Stream Processing and also discuss the fundamentals of Flume.

COURSE OUTCOMES:

On successful completion of this course; the student will be able to:

CO#	Course Outcomes	POs	PSOs
CO1	Illustrate the fundamentals of Hadoop distributed file system and Big Data Analytics	1 to 5,9,10,11	1,2,3
CO2	Demonstrate Big Data Processing with MapReduce and Batch Analytics with Apache Hadoop to simple real world problems. .	1 to 5,9,10,11	1,2,3
CO3	Design Real-Time Analytics with Apache Pig and Hive for real world Applications.	1 to 5,9,10,11	1,2,3
CO4	Develop data and processing models using Hadoop eco-system for real world Big data Applications	1 to 5,9,10,11	1,2,3
CO5	Design Real-Time Analytics incorporating the structured data model using Apache Hive to solve real world Big Data Analytics Applications.	1 to 5,9,10,11	1,2,3
CO6	Develop data and processing models using Hadoop eco-system for real world Big data Applications	1 to 5,9,10,11	1,2,3

BLOOM'S LEVELOF THECOURSE OUTCOMES

CO#	Bloom's Level					
	Remember (L1)	Understand (L2)	Apply (L3)	Analyze (L4)	Evaluate (L5)	Create (L6)
CO1		√				
CO2			√			
CO3			√			
CO4			√			
CO5			√			
CO6			√			

COURSE ARTICULATIONMATRIX

Course Outcomes	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PSO1	PSO2	PSO3
CO1	3	2	3	3	3				1	1	1	3	3	3
CO2	3	3	3	3	3				2	1	2	3	3	3
CO3	3	3	3	3	3				2	1	2	3	3	3
CO4	3	3	3	3	3				3	3	2	3	3	3
CO5	3	3	3	3	3				2	1	2	3	3	3
CO6	3	3	3	3	3				3	3	2	3	3	3

PRACTICE:

SL. N.	Title of the Experiment	Tools and Techniques	Expected Skill/Ability
	PART-A Introduction: Installing PySpark on Colab		
1.	Create a Python script to test PySpark. a. Create a SparkContext object b. Create an RDD from for the input file "Sample.txt" c. Find the total count of the words in the RDD Print the string with highest occurrence.	Windows/Linux OS, IDE	Understanding the process of Installation of Hadoop in different modes
2.	Given the two RDDs: a. x created from the ordered pairs: ("spark", 1) and ("hadoop", 4) b. y created from the ordered pairs: ("spark", 2), ("hadoop", 5). c. Perform the join operation on the RDDs created above, and print the resulting RDD. Run the Usecases for Right Join, Left Join, Inner Join, Outer Join	Windows/Linux OS, IDE	Managing Files and performing operations on them on HDFS
3.	a.Create an RDD of set of numbers and perform the sum of these numbers using an *accumulator()* function in Spark context. b.Create an RDD from the existing file having CSV data, using *read()* and *load()* functions and display the top 5 rows of the data set. And also display the statistical results	Windows/Linux OS, IDE	Understanding the MapReduce Process

<table>
<tr><td></td><td>from the data frame (Note: It only works
for numerical values).</td><td></td><td></td></tr>
<tr><td>4.</td><td>Given the following tables, Perform Full outer join on dataframe.
<table><tr><th>Age</th><th>Name</th></tr><tr><td>2</td><td>Alice</td></tr><tr><td>5</td><td>Bob</td></tr></table>
<table><tr><th>Height</th><th>Name</th></tr><tr><td>80</td><td>Tom</td></tr><tr><td>85</td><td>Bob</td></tr></table>
<table><tr><th>Age</th><th>Name</th></tr><tr><td>2</td><td>Alice</td></tr><tr><td>5</td><td>Bob</td></tr></table>
<table><tr><th>Age</th><th>Height</th><th>Name</th></tr><tr><td>10</td><td>80</td><td>Alice</td></tr><tr><td>5</td><td>None</td><td>Bob</td></tr><tr><td>None</td><td>None</td><td>Tom</td></tr><tr><td>None</td><td>None</td><td>None</td></tr></table>
a. Write query and Perform Join of df1 and df2 to find the height
b. Write query and Perform outer join of df1 and df2 to find the heights
c. Write query and Perform outer join of df1 and df2 to find the age</td><td>Windows/Linux OS, IDE</td><td>Performing Big Data Analytics using MapReduce</td></tr>
<tr><td>5.</td><td>Given the following data
data = [("1", "john jones"), ("2", "tracey smith"), ("3", " amy sanders")], along with the following schema
of the data
columns = ["Seqno","Name"]
Perform the following using the afore mentioned data.
i. create an RDD from the above data using its schema
ii. create the PySparkdataframe from the RDD created.
iii. Write python functions to convert the first letter of every string into upper case.
iv. Use the above python function as udf in pySpark to convert the data in the dataframe and display
the result.</td><td>Windows/Linux OS, IDE</td><td>Understanding the MapReduce Process</td></tr>
<tr><td>6.</td><td>Given the following data.
Data = [("James","Sales","NY",90000,34,10000),
("Michael","Sales","NV",86000,56,20000),
("Robert","Sales","CA",81000,30,23000),
("Maria","Finance","CA",90000,24,23000),
("Raman","Finance","DE",99000,40,24000),
("Scott","Finance","NY",83000,36,19000),
("Jen","Finance","NY",79000,53,15000),
("Jeff","Marketing","NV",80000,25,18000),
("Kumar","Marketing","NJ",91000,50,21000)
], with the following schema schema =
["employee_name","department","state","salary","age","bonus"]
Perform the following using the aforementioned data.
i. create an RDD from the above data using its schema
ii. create the PySparkdataframe from the RDD created.
iii. Using groupBy() function, display the salaries of
the employees state-wise.</td><td>Windows/Linux OS, IDE, Pig Tool</td><td>Performing Big Data Analytics using Pig Scripts</td></tr>
</table>

	iv. Display the state-wise salaries that are greater than 1 lakh V. Display the state-wise salaries in descending order.		
7	Create a dataframe and then perform the following operations i. Check the lifestage of each person into Adult, Child and Teenager ii. Write query to Display entries of teenager and adult only iii. Write query to average age iv. Write query to group by entries by life_stage v. Insert a record "Frank,4, Child) into new data frame vi. Write a query to display teenage entries	Windows/Linux OS, IDE, Hive Tables	Performing Big Data Analytics using Pig Scripts
8	Write a Word Count Map Reduce program to understand Map Reduce Paradigm.	Windows/Linux OS, IDE, Hive Tables	Performing Big Data Analytics using Pig Scripts
	PART-B		
1	Implement and demonstrate any real life big data problem using any of the publicly available big data sets.	Windows/Linux OS, IDE, Hadoop-eco system	Literature Surveying, Project Implementation, Seminars, IPR Filing, Paper Publication

TEXT BOOKS:

Sridhar Alla, "Big Data Analytics with Hadoop 3", Packt Publishing Ltd, 2018
Gates, Alan, and Daniel Dai. Programming pig: Dataflow scripting with hadoop. " O'Reilly Media, Inc.", 2016.
Capriolo, Edward, Dean Wampler, and Jason Rutherglen. Programming Hive: Data warehouse and query language for Hadoop. " O'Reilly Media, Inc.", 2012.

REFERENCE BOOKS:

Michael Minelli, Michele chambers, AmbigaDhiraj,"Big data, big analytics", Wiley,2013
P. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining", Addison-Wesley, 2005.
J. Han, M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed. Morgan Kaufmann 2005.

JOURNALS/MAGAZINES

IEEE,Introduction to the IEEE Transactions on Big Data.
Elsevier,Big data research journal Elsevier.
Springer, Journal on Big Data Springer.