BIG DATA ANALYTICS LAB

III B. TECH. – I SEMESTER								
Course Code	Category	Hours / Week			Credits	Maximum Marks		
A6DS07	PCC	L	T	P	C	CIE	SEE	Total
		-	-	3	1.5	40	60	100

COURSE OBJECTIVES

1. Get familiar with Hadoop distributions, configuring Hadoop and performing File management tasks.
2. Understand the storage of data in a distributed file system.
3. Introduce the basics required to develop map reduce programs.
4. Introduce programming tools PIG & HIVE in the Hadoop ecosystem.
5. Demonstrate the usage of PySpark to implement ML concepts.

COURSE OUTCOMES

At the end of the course, students will be able to:

1. Connect to Hadoop cluster, experiment with various Linux and HDFS commands to store data.
2. Apply the knowledge of MapReduce programming to process the stored data in HDFS.
3. Write scripts using Pig and retrieve data from Hadoop using HiveQL.
4. Create data processing pipelines with Spark
5. Build and tune machine learning models with Spark ML

LIST OF EXPERIMENTS:

WEEK 1

i) Perform setting up and installing Vmware for Hadoop and Linux.

ii) Basic Linux Commands

iii) Run basic HDFS shell commands

WEEK 2

Implement the following file management tasks in Hadoop:

i) Adding files and directories

ii) Retrieving files

iii) Deleting files and directories.

WEEK 3

i) Develop a MapReduce program to calculate the frequency of a given word in a given file.

ii) Develop a MapReduce program to find the maximum temperature in each year.

WEEK 4

Design MapReduce algorithms to take a very large file of integers and produce as output:

i) The largest integer

ii) The average of all the integers.

iii) The same set of integers, but with each integer appearing only once. *

iv) The count of the number of distinct integers in the input.*

WEEK 5

Implement **Matrix** Multiplication on **Hadoop** Using **Map Reduce**.

WEEK 6

i) Run Pig and perform basic PIG commands.

ii) Write Pig Latin scripts to sort, group, join, project, and filter your data.

WEEK 7

i) Practice Basic HiveQL Commands, read data from various File Formats and create Data Definition Statements and Data Manipulation Statements.

ii) Write Queries using select.

WEEK 8

i) Interactive Analysis with the Spark Shell

ii) Writing and running Spark program

WEEK 9

Implement the following algorithms for classification using PySpark.

i) Logistic Regression

ii) Decision Tree Classifier

iii) Naïve Bayes

WEEK 10

Implement the following algorithms for clustering using PySpark.

i) K-Means

ii) Latent Dirichlet Allocation (LDA)
iii) Gaussian Mixture Model (GMM)

WEEK 11

Implement collaborative filtering using spark ML library.

WEEK 12

Implement FP-Growth using Spark ML Library.

TEXT BOOKS

1. Hadoop: The Definitive Guide, 4th Edition – O'Reilly Media
2. Singh, Pramod. *Machine Learning with PySpark: With Natural Language Processing and Recommender Systems*. Apress, 2018.
3. Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015.

INDEPENDENTSTUDY/MOOC'S

IIIB.TECH-IISEMESTER

CourseCode	Category	Hours/Week			Credits	MaximumMarks		
		L	T	P	C	CIA	SEE	Total
A5DS12	PWC	-	-	-	1	-	100	100