COURSE TITLE	**BIG DATA TOOLS AND ANALYTICS**
COURSE CODE	**05MD0202**
COURSE CREDITS	**2**

Course Outcomes: After completion of this course, student will be able to:

1 Define the concept of Big Data and basic Big Data terminologies.
2 Implement HDFS and Map Reduce in Hadoop.
3 Implement Hive queries and create Pig script.
4 Discuss the concept of spark and differentiate Hadoop and Spark.
5 Apply the machine learning concepts in Spark using MLib.

Pre-requisite of course: Knowledge of Java Programming, Python Programming and Database Management System

Teaching and Examination Scheme

Theory Hours	Tutorial Hours	Practical Hours	ESE	IA	CSE	Viva	Term Work
0	0	4	0	0	0	25	25

Contents : Unit	Topics	Contact Hours
Total Hours		

Suggested List of Experiments:

Contents : Unit	Topics	Contact Hours
1	**Understanding Big Data** Introduction of syllabus, Introduction of data, dataset, big data, challenges of big data, big data characteristics, Why Big Data?, Traditional BI v/s Big Data, Clusters, File systems and Distributed File Systems, Sharding, Replication, Sharding and Replication, CAP Theorem, On – Disk Storage Devices, In – memory storage devices, Scaling – Horizontal and Vertical	8

Suggested List of Experiments:

Contents : Unit	Topics	Contact Hours
2	**Hadoop :** Introduction, Distributed Computing Challenges, History of Hadoop, Overview of Hadoop and Hadoop Ecosystems, Features and key advantages of Hadoop, Versions of Hadoop, Hadoop distributions, RDBMS versus Hadoop, Hadoop vs SQL, Hadoop Core Components, HDFS : HDFS features, HDFS components (Daemons), HDFS Read/Write anatomy, HDFS Commands, 1. Hadoop installation steps. 2. Which path you need to set in hadoop-env? 3. Properties setting in : a. Core-site.xml b. Hdfs-site.xml c. Mapred.xml d. Yarn-site.xml 4. How to start and stop hadoop server? 5. Which command is used to list the server names? Start all the servers and list the server names., MapReduce : Introduction, Steps of Mapper and Reducer, MapReduce programming in Python example, Hadoop streaming, mrjob (installation, wordcount example), Create a hello.txt file in local system and copy it into HDFS. Word count program – MapReduce (List all the steps with code), Vowel count program using map reduce., Matrix multiplication using map reduce.	14
3	**Hive and Pig** Hive : Introduction (What is HIVE?, HIVE Architecture, HIVE data Types, HIVE File Formats, Hive : Question : Write steps to start Hive. Ex – 1 : a. Create database Payroll b. Create Table Employee, Department c. Run DDL and DML commands d. Create table via loading data from files e. Create table form existing schema f. Run Data Retrieval queries , Joins, HIVEQL, word count using Hive, static and dynamic partition with practicals, Convert emp.xml file into structure format using Hive., Introduction(What is Pig? The anatomy of Pig, Pig on Hadoop, Pig philosophy, Use Case for Pig-ETL Processing, Pig Latin overview, Datatypes in Pig, running Pig, Execution modes of Pig, HDFS commands, Ex – 1 : Create Movies.csv file a. Running Pig program in Local and MapReduce Mode b. Working with Pig Operators (FOREACH, ASSERT, FILTER, GROUP, ORDER BY, DISTINCT, JOIN, LIMIT, SAMPLE, SPLIT) c. Working with Pig functions d. Error handling in Pig e. Debugging in Pig, Pig script for word count, split relational operator with practical	10
4	**Introduction to Spark, Resilient Distributed Dataset and Data Frames** Introduction of Spark : What is Spark and what is its purpose?, Components of the Spark unified stack, Resilient Distributed Dataset (RDD), Spark installation steps, Word Count using Spark. (Python/Scala), Understand how to create parallelized collections and external datasets, Work with Resilient Distributed Dataset (RDD) operations, Utilize shared variables and key-value pairs.	8

Suggested List of Experiments:

Contents : Unit	Topics	Contact Hours
5	**Machine Learning with MLlib (Spark Tool)** Machine Learning with MLlib (Spark Tool): Overview, System Requirements, Machine Learning Basics,Example: Spam Classification, Data Types, Working with Vectors, 1. Spam classifier, 2. Feature extraction, 3. Linear Regression, practical test - 1, practical test - 2	10
Total Hours		**50**

Textbook :

1 Big Data and Analytics, Seema Acharya, Subhashini Chellappan , Wiley, 2015

2 Hadoop with Python, Zachary Radtka and Donald Miner, O'Reilly Media, 2016

3 Learning Spark, Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau, O'Reilly Media, 2015

Suggested Theory Distribution:

The suggested theory distribution as per Bloom's taxonomy is as follows. This distribution serves as guidelines for teachers and students to achieve effective teaching-learning process

Distribution of Theory for course delivery and evaluation					
Remember / Knowledge	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Higher order Thinking**

Supplementary Resources:

1 https://www.tutorialspoint.com/hbase

2 https://cognitiveclass.ai