BIG DATA ANALYTICS		Semester	Vil
Course Code	BCS714D	CIE Marks	50
Teachin Hours Week		SEE Marks	50
Total Hours of Peda o	40 hours Theo + 8-10 Lab slots	Total Marks	100
Credits	04	Exam Hours	3
Examination nature SEE	Theo ractical		

Course objectives:

1. To implement MapReduce programs for processing big data.
2. To realize storage and processing of big data using MongoDB, Pig, Hive and Spark
3. To analyze big data using machine learning techniques.

Teaching-Learning Process (General Instructions)

These are sample Strategies: that teachers can use to accelerate the attainment of the various course outcomes. 1. Lecturer method (L) needs not to be only a traditional lecture method, but alternative effective teaching methods could be adopted to attain the outcomes.

2. Use of Video/Animation to explain functioning of various concepts.

g. Encourage collaborative (Group Learning) Learning in the class.

4. Ask at least three HOT (Higher order Thinking) questions in the class, which promotes critical thinking.
5. Discuss how every concept can be applied to the real world - and when that's possible, it helps improve the students' understanding.
6. Use any of these methods: Chalk and board, Active Learning, Case Studies.

MODULE-I

Classification of data, Characteristics, Evolution and definition of Big data, What is Big data, Why Big data, Traditional Business Intelligence Vs Big Data,Typical data warehouse and Hadoop environment.

Big Data Analytics: What is Big data Analytics, Classification of Analytics, Importance of Big Data Analytics, Technologies used in Big data Environments, Few Top Analytical Tools , NoSQL, Hadoop.

TBI: Ch 1: 1.1, Ch2: 2.1-2.5,2.7,2.9-2.11, Ch3: 3.2,3.5,3.8,3.12,Ch4:

MODULE-Z

Introduction to Hadoop: Introducing hadoop, Why hadoop, Why not RDBMS, RDBMS vs Hadoop, History of Hadoop, Hadoop overview, Use ease of Hadoop, HDFS (Hadoop Distributed File System),Processing data with Hadoop, Managing resources and applications with Hadoop YARN(Yet Another Resource Negotiator).

Introduction to Map Reduce Programming: Introduction, Mapper, Reducer, Combiner, Partitioner, Searching, Sorting, Compression.

TBI: Ch . 5.10-5.12, Ch 8: 8.1 -8.8

MODULE-3

Introduction to MongoDB: What is MongoDB, Why MongoDB, Terms used in RDBMS and MongoDB, Data Types in MongoDB, MongoDB Query Language.

TBI: Ch 6: 6.1-6.5

MODULE-4

Introduction to Hive: What is Hive, Hive Architecture, Hive data types, Hive file formats, Hive Query Language (HQL), RC File implementation, User Defined Function (UDF).

Introduction to Pig: What is Pig, Anatomy of Pig, Pig on Hadoop, Pig Philosophy, Use case for Pig, Pig Latin Overview, Data types in Pig, Running Pig, Execution Modes of Pig, HDFS Commands, Relational Operators, Eval Function, Complex Data Types, Piggy Bank, User Defined Function, Pig Vs Hive.

TBI: Ch 9: 9.1-9.6,9.8, Ch 10: 10.1 - 10.15, 10.22

MODULE-5
Spark and Big Data Analytics: Spark, Introduction to Data Analysis with Spark. Text, Web Content and Link Analytics: Introduction, Text Mining, Web Mining, Web Content and Web
Usage Analytics, Page Rank, Structure of Web and Analyzing a Web Graph. TB2: Ch5: 525.3, Ch 9: 9.1-9.4

PRACTICAL COMPONENT OF IPCC

SI.NO	Ex eriments ava P on R
	Install Hadoop and Implement the following file management tasks in Hadoop: Adding files and directories Retrieving files Deleting files and directories. Hint: A typical Hadoop workflow creates data files (such as log files) elsewhere and copies them into HDFS using one of the above command line utilities.
2	Develop a MapReduce program to implement Matrix Multiplication
3	Develop a Map Reduce program that mines weather data and displays appropriate messages indicating the weather conditions of the day.
4	Develop a MapReduce program to find the tags associated with each movie by analyzing movie lens data.
5	Implement Functions: Count — Sort — Limit — Skip — Aggregate using MongoDB
6	Write Pig Latin scripts to sort, group, Join, project, and filter the data.
7	Use Hive to create, alter, and drop databases, tables, views, functions, and indexes.
8	Implement a word count program in Hadoop and Spark.
9	Use CDH (Cloudera Distribution for Hadoop) and HUE (Hadoop User Interface) to analyze data and generate reports for sample datasets

Course outcomes (Course Skill Set):

At the end ofthe course, the student will be able to:

- Identify and list various Big Data concepts, tools and applications.
- Develop programs using HADOOP framework.
- Use Hadoop Cluster to deploy Map Reduce jobs, PIG,HIVE and Spark programs.
  Analyze the given data set and identify deep insights from the data set.

Assessment Details (both CIE and SEE)

The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50% The minimum passing mark for the CIE is 40% of the maximum marks (20 marks out of 50) and for the SEE minimum passing mark is 35% of the maximum marks (18 out of 50 marks). A student shall be deemed to have satisfied the academic requirements and earned the credits allotted to each subject/ course if the student secures a minimum of 40% (40 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together.

CIE for the theory component of the IPCC (maximum marks 50)

IPCC means practical portion integrated with the theory of the course.

CIE marks for the theory component are 25 marks and that for the practical component is 25 marks.

25 marks for the theory component are split into 15 marks for two Internal Assessment Tests (Two Tests, each of 15 Marks with 01-hour duration, are to be conducted) and 10 marks for other assessment methods mentioned in 220B4.2. The first test at the end of 40-50% coverage of the syllabus and the second test after covering 85-90% of the syllabus.

Scaled-down marks of the sum of two tests and other assessment methods will be CIE marks for the theory component of IPCC (that is for 25 marks).

The student has to secure 40% of 25 marks to qualify in the CIE of the theory component of IPCC CIE for the practical component of the IPCC

15 marks for the conduction of the experiment and preparation of laboratory record, and 10 marks for the test to be conducted after the completion of all the laboratory sessions.

On completion of every experiment/program in the laboratory, the students shall be evaluated including viva-voce and marks shall be awarded on the same day.

The CIE marks awarded in the case of the Practical component shall be based on the continuous evaluation of the laboratory report. Each experiment report can be evaluated for 10 marks. Marks of all experiments' write-ups are added and scaled down to 15 marks.

The laboratory test (duration 02/03 hours) after completion of all the experiments shall be conducted for 50 marks and scaled down to 10 marks.

Scaled-down marks of write-up evaluations and tests added will be CIE marks for the laboratory component of IPCC for 25 marks.

The student has to secure 40% of 25 marks to qualify in the CIE of the practical component of the
IPCC.

SEE for IPCC

Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the course (duration 03 hours)

1. The question paper will have ten questions. Each question is set for 20 marks.
2. There will be 2 questions from each module. Each of the two questions under a module (with a maximum of 3 sub-questions), should have a mix oftopics under that module.
3. The students have to answer 5 full questions, selecting one full question from each module.
4. Marks scored by the student shall be proportionally scaled down to 50 Marks

The theory portion of the IPCC shall be for both CIE and SEE, whereas the practical portion will have a CIE component only. Questions mentioned in the SEE paper may include questions from the practical component.

Suggested Learning Resources:

Books:

1. Seema Acharya and Subhashini Chellappan "Big data and Analytics" Wiley India Publishers, 2nd Edition, 2019.
2. Rajkamal and Preeti Saxena, "Big Data Analytics, Introduction to Hadoop, Spark and Machine Learning", McGraw Hill Publication, 2019.

Reference Books:

1. Adam Shook and Donald Mine, "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems" O'Reilly 2012
2. Tom White, "Hadoop: The Definitive Guide" 4th Edition, O'reilly Media, 2015.
3. Thomas Erl, Wajid Khattak, and Paul Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, Pearson India Education Service Pvt. Ltd., 1st Edition, 2016
4. John D. Kelleher, Brian Mac Namee, Aoife D'Arcy -Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, MIT Press 2020, 2nd Edition

Web links and Video Lectures (e-Resources):

- https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset
- https://www.youtube.com/watch?v=bAyrOb17TYE&list=PLEiEAq2VkUUJqp1k-g5W1m037urJQOdCZ
- https://www.youtube.com/watchftAJm00QgPCbZY&list=PLEiEAq2VkUUJqp1kg5W1m037urJQOdCZ&index=4
- https://www.youtube.com/watch?v=GG-VRm6XnNk https://www.youtube.com/watch?v=Jg102Nv_92A

Activity Based Learning (Suggested Activities in Class)/ Practical Based learning

1. Implement MongoDB based application to store big data for data processing and analyzing the results [10 marks]