FACULTY OF ENGINEERING & TECHNOLOGY

Effective from Academic Batch: 2022-23

Programme: **Bachelor of Technology (Information Technology)**

Semester: **VII**

Course Code: **202047803**

Course Title: **Big Data Analytics**

Course Group: **Professional Elective Course - III**

Course Objectives: This course gives an overview of Big Data, the characteristics of Big Data and its applications in Big Data Analytics. In addition, it also focuses on the tools and algorithms that covers a wide range of analytics platforms and databases, including Hadoop, Sqoop, Hive, Pig, HBase and Spark.

Teaching & Examination Scheme:

Contact hours per week			Course Credits	Examination Marks (Maximum / Passing)				
				Theory		J/V/P*		Total
Lecture	Tutorial	Practical		Internal	External	Internal	External	
3	0	2	4	50/18	50/17	25/9	25/9	150/53

* **J**: Jury; **V**: Viva; **P**: Practical

Detailed Syllabus:

Sr.	Contents	Hours
1	**Introduction to Big Data:** Classification of Digital Data, Structured Data, Semi- Structured data, Unstructured Data, Characteristic of Data, Evolution of Big Data, Definition of Big Data, 4Vs of Data- Volume, Velocity, Variety and Veracity, Big Data requirement, Traditional Business intelligence versus Big Data, Introduction to Big Data Analytics.	05
2	**NoSQL:** What is it? Where It is Used, Types of NoSQL databases, Why NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL, NewSQL.	05
3	**Introduction to Hadoop:** Features of Hadoop, Key Advantages of Hadoop, Versions of Hadoop, Hadoop Ecosystems, Hadoop Vs SQL, Hadoop Components, Use case of Hadoop, Processing data with Hadoop, YARN Components, YARN Architecture, YARN MapReduce Application, Execution Flow, YARN Workflow, Anatomy of MapReduce Program, Input Splits, Relation between Input Splits and HDFS Blocks.	10

4	**HDFS, SQOOP, HIVE, PIG AND HBASE:** HDFS: Daemons, Anatomy of File Read, Anatomy of File Write, Replica Placement Strategy, Working with HDFS Commands Sqoop: Introduction, import and export command Hive: Hive Architecture and Installation, Comparison with Traditional Database, HiveQL Querying Data, Sorting and Aggregating, Map Reduce Scripts, Joins & Sub queries PIG: PIG Architecture & Data types, Shell and Utility components, PIG Latin Relational Operators, PIG Latin: File Loaders and UDF, Programming structure in UDF, PIG Jars Import, limitations of PIG. HBase: HBase concepts, Advanced Usage, Schema Design, Advance Indexing Zookeeper: How it helps in monitoring a cluster, HBase uses Zookeeper and how to Build Applications with Zookeeper.	**12**
5	**SPARK:** Introduction to Data Analysis with Spark, Features of Apache Spark, Components of Spark, Downloading Spark and Getting Started, RDD Transformations, RDD Actions, Programming with RDDs, Machine Learning with MLlib.	**08**
	Total	**40**

List of Practicals / Tutorials:

1	Configure Hadoop cluster in pseudo distributed mode. Try Hadoop basic commands.
2	Write Map Reduce code for following: a. Count frequency of words from a large file. b. Find year wise maximum temperature using the weather data set which consists of year, month, and temperature. c. Patent data files consist of patent id and sub patent id. One patent is associated with multiple sub patents. Write a map reduce code to find out the total sub patent associated with the patent.
3	Write a word count program using partitioner and combiner.
4	Configure multimode Hadoop Cluster.
5	Configure Sqoop. Try sqoop import and export command.
6	Configure Hive and try basic Hive query.

7	Write Hive Query for the following task for movie dataset. Movie dataset consists of movie id, movie name, release year, rating, and runtime in seconds. A sample of the dataset is as follows: a. The Nightmare Before Christmas,1993,3.9,4568 b. The Mummy,1932,3.5,4388 c. Orphans of the Storm,1921,3.2,9062 d. The Object of Beauty,1991,2.8,6150 e. Night Tide,1963,2.8,5126 Write a hive query for the following a. Load the data b. List the movies that are having a rating greater than 4 c. Store the result of previous query into file d. List the movies that were released between 1950 and 1960 e. List the movies that have duration greater than 2 hours f. List the movies that have rating between 3 and 4 g. List the movie names and its duration in minutes h. List all the movies in the ascending order of year. i. List all the movies in the descending order of year. j. list the distinct records. k. Use the LIMIT keyword to get only a limited number for results from relation. l. Use the sample keyword to get a sample set from your data. m. M. View the step-by-step execution of a sequence of statements using ILLUSTRATE command.
8	Configure Pig and try different Pig commands.
9	Configure HBase and try different HBase commands.
10	Write a java program to insert, update and delete records from HBase.
11	Install Apache Spark and try basic commands.
12	Write a scala program to process CSV, JSON and TXT File.
13	Write a scala program a. To get the character at the given index within a given String. Also print the length of the string b. To compare two strings lexicographically c. To concatenate a given string to the end of another string d. To exchange the first and last characters in a given string and return the new string e. To exchange the first and last characters in a given string and return the new string
14	Capstone project.

Reference Books:

1	BIG Data and Analytics, Sima Acharya, Subhashini Chhellappan, Willey
2	DT Editorial Services, "Black Book- Big Data (Covers Hadoop 2, MapReduce, Hive, Yarn, PIG, R, Data visualization)", Dream tech Press edition 2016.
3	Learning Spark: Lightning-Fast Big Data Analysis Paperback by Holden Karau
4	Chris Eaton, Dirk Derooset al., "Understanding Big data", McGraw Hill, 2012.
5	Tom White, "HADOOP: The Definitive Guide", O Reilly 2012.
6	Vignesh Prajapati, "Big Data Analytics with R and Hadoop", Packet Publishing 2013.

7	Learning Spark: Lightning-Fast Big Data Analysis Paperback by Holden Kara
8	http://www.bigdatauniversity.com/

Supplementary learning Material:	
1	NPTEL - Swayam Course: Big Data Computing- https://nptel.ac.in/courses/106104189
2	Coursera -Introduction to Big Data with Spark and Hadoop - https://www.coursera.org/learn/introduction-to-big-data-with-spark-hadoop#syllabus

Pedagogy:

- Direct classroom teaching
- Audio Visual presentations/demonstrations
- Assignments/Quiz
- Continuous assessment
- Interactive methods
- Seminar/Poster Presentation
- Industrial/ Field visits
- Course Projects

Suggested Specification table with Marks (Theory) (Revised Bloom's Taxonomy):

Distribution of Theory Marks in %						**R**: Remembering; **U**: Understanding; **A**: Applying. **N**: Analyzing; **E**: Evaluating; **C**: Creating
R	**U**	**A**	**N**	**E**	**C**	
15%	25%	25%	15%	20%	---	

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

Course Outcomes (CO):

Sr.	Course Outcome Statements	%weightage
CO-1	Understand Big Data and its analytics in the real world.	**15**
CO-2	Analyze the Big Data frameworks like Hadoop and NOSQL to efficiently store and process Big Data for analytics.	**25**
CO-3	Design algorithms to solve Data Intensive problems using the Map Reduce paradigm.	**20**
CO-4	To solve data intensive problems and generate analytics using Pig, Spark, Hive and Sqoop.	**40**

Curriculum Revision:	
Version:	2.0
Drafted on (Month-Year):	June -2022
Last Reviewed on (Month-Year):	-
Next Review on (Month-Year):	June-2025