GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Bachelor of Engineering
Level: UG
Subject Code : 3174807
Subject Name :Big Data Engineering and Analytical Methods

w. e. f. Academic Year:	A.Y. 2025-26
Semester:	VII
Category of the Course:	PEC

Prerequisite :	NA
Rationale:	Today's world is a data-driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of data. Big data analytics helps us to examine these data to uncover hidden patterns, correlations, and other insights. It is a fast-growing field and skills in the area are some of the most in-demand today.

Course Outcome:

After Completion of the Course, Student will able to:

No	Course Outcomes	RBT Level
1	Recognize key application domains where big data solutions can be applied effectively.	U
2	Utilize appropriate big data frameworks and tools for data processing and management.	A
3	Apply relevant analytical methods and modeling techniques to extract insights from data.	An
4	Design a workflow integrating big data tools and technologies to address real-world problems.	E
5	Evaluate the effectiveness of big data solutions in solving complex data-driven challenges.	C

**Revised Bloom's Taxonomy (RBT)*

Teaching and Examination Scheme:

Teaching Scheme (in Hours)			Total Credits L+T+ (PR/2)	Assessment Pattern and Marks				Total Marks
				Theory		Tutorial / Practical		
L	T	PR	C	ESE (E)	PA / CA (M)	PA/CA (I)	ESE (V)	
3	0	2	4	70	30	20	30	150

GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Bachelor of Engineering
Level: UG
Subject Code : 3174807
Subject Name :Big Data Engineering and Analytical Methods

Course Content:

Unit No.	Content	No. of Hours
1.	**Introduction to Big Data:** Introduction to Big Data, Big Data characteristics, Challenges of Conventional System, Types of Big Data, Intelligent data analysis, Traditional vs. Big Data business approach, Case Study of Big Data Solutions.	04
2.	**Hadoop:** Introduction to Hadoop and HDFS, including node architecture and key components. Overview of data processing using MapReduce, its working, job execution flow, error handling, and scheduling. Discussion on HDFS design, Hadoop Streaming, Java HDFS APIs, and cluster setup. Basics of cluster configuration, security, monitoring, administration, and Hadoop's use in cloud environments and performance benchmarking.	12
3	**NoSQL:** NoSQL refers to non-relational databases designed to handle large volumes of diverse and fast-changing data. It emerged due to business needs for scalable, flexible data storage. Common types include key-value, graph, column-family, and document stores. NoSQL systems use varied architectures to manage big data, often relying on shared-nothing models. They solve challenges using distribution methods like master-slave and peer-to-peer, each optimized for different scalability and performance needs	07
4	**Mining Data Stream:** Introduction to Streams Concepts, Stream Data Model and Architecture, Stream Computing, Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Estimating moments, Counting oneness in a Window, Decaying Window, Real time Analytics Platform (RTAP) applications, Case Studies, Real Time Sentiment Analysis, Stock Market Predictions. Using Graph Analytics for Big Data: Graph Analytics	10
5.	**Frameworks:** Applications on Big Data Using Pig and Hive, Data processing operators in Pig, Hive services, HiveQL, Querying Data in Hive, fundamentals of HBase and ZooKeeper, IBM InfoSphere BigInsights and Streams.	08
6.	**Spark:** Introduction to Data Analysis with Spark, In-Memory Computing with Spark, Spark Basics, Interactive Spark with PySpark, Writing Spark Applications	07

Suggested Specification Table with Marks (Theory):

Distribution of Theory Marks (in %)					
R Level	**U Level**	**A Level**	**N Level**	**E Level**	**C Level**
–	**15**	**30**	**25**	**15**	**05**

Where R: Remember; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create (as per Revised Bloom's Taxonomy)

References/Suggested Learning Resources:

(a) Books:

1. Michael Berthold, David J. Hand, "Intelligent Data Analysis", Springer, 2007
2. Bill Franks , "Taming The Big Data Tidal Wave: Finding Opportunities In Huge Data Streams With Advanced Analytics", Wiley
3. Anand Rajaraman and Jeff Ullman "Mining of Massive Datasets", Cambridge University Press,
4. Michael Minelli, Michele Chambers, Ambiga Dhiraj, "Big Data Big Analytics: Emerging Business Intelligence And Analytic Trends For Today's Businesses", Wiley India
5. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley.
6. Chris Eaton, Dirk derooset al., "Understanding Big data", McGraw Hill, 2012.
7. BIG Data and Analytics , Seema Acharya, Subhashini Chhellappan, Willey
8. MongoDB in Action, Kyle Banker, Piter Bakkum , Shaun Verch, Dream tech Press
9. Tom White, "HADOOP: The Definitive Guide", O Reilly 2012.
10. Vignesh Prajapati, "Big Data Analytics with R and Hadoop", Packet Publishing 2013.
11. Learning Spark: Lightning-Fast Big Data Analysis Paperback by Holden Karau

List of Open Source Software/learning website:

1. http://in.reuters.com/tools/rss

2. http://www.altova.com/xmlspy.html

List of Experiments and Design based Problems (DP)/Open Ended Problem:

1. **Data Ingestion and Cleaning:** Load and preprocess a large dataset using Apache Spark and perform basic cleaning operations (e.g., missing value handling, formatting).
2. **Distributed File System Operations:** Perform file operations (create, read, write) using HDFS command line and APIs.
3. **MapReduce Fundamentals**: Write a MapReduce job to perform log parsing or session summarization.
4. **Batch vs. Stream Comparison**: Implement a word count problem using both Hadoop MapReduce and Apache Spark Streaming.
5. **NoSQL with MongoDB or Cassandra**: Insert, query, and update large volumes of records using a NoSQL system

6. **Graph Analytics on Social Network Data**: Use Apache Spark GraphX or Neo4j to perform centrality and path analysis on a small network dataset.
7. **Real-Time Data Analysis**: Ingest live Twitter data using Apache Kafka and Spark Streaming, and perform sentiment classification.
8. **HiveQL Queries**: Create databases and tables in Hive, and run aggregation and join operations on structured datasets.
9. **Performance Benchmarking**: Compare job execution time between Hadoop and Spark for a data-intensive task.
10. **Capstone Mini Project**: Develop an end-to-end mini project—such as COVID-19 case trend analysis, retail sales prediction, or financial fraud detection—integrating data ingestion, processing, analysis, and visualization.

****