GUJARAT TECHNOLOGICAL UNIVERSITY

BACHELOR OF ENGINEERING SYLLABUS

Subject Code: 3164604

Subject Name: Big Data Analytics

WEF Academic Year :	AY 2022-23
Semester :	6
Category of the Course :	Professional Elective III

Prerequisite: Programming skills

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 Scheme:

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

Course Content:

Sr. No.	Course Content	No. of Hours	% of Weightage
1	**Introduction to Big Data:** Introduction to Data, Types of Data, Structured Data, Unstructured Data, Semi- Structured Data, Meta Data, The Emergence of 'New Data', Comparison of New and Traditional Data Data Analytics, Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Perceptive Analytics, analysis of problem space and data needs Introduction to Big Data, Why and Where, Big Data characteristics, Types of Big Data, Transition to Big Data, Definition of Big Data, The V's, Sources of Big Data Challenges of Conventional System, Intelligent data analysis, Traditional vs. Big Data business approach, Common Application and Case Study of Big Data Solutions. Getting values out of Big Data	06	15%
2	**Hadoop:** History of Hadoop, Hadoop Distributed File System: Physical organization of Compute Nodes, Components of Hadoop, Analyzing the Data with Hadoop, Scaling Out, Hadoop Streaming, Design of HDFS, Java interfaces to HDFS Basics, Developing a Map Reduce Application, How Map Reduce Works, Anatomy of a Map Reduce Job run, Failures, Job Scheduling,	08	20%

	Shuffle and Sort, Task execution, Map Reduce Types and Formats, Map Reduce Features, Hadoop environment. Hadoop Hybrids, Cluster, Setting up a Hadoop Cluster, Cluster specification, Cluster Setup and Installation, Hadoop Configuration, Security in Hadoop, Administering Hadoop,		
3	**NoSQL:** What is NoSQL? NoSQL business drivers; NoSQL Databases in the Light of CAP Theorem, NoSQL case studies; NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, NoSQL Database: Cassandra, Variations of NoSQL architectural patterns; Using NoSQL to manage big data: What is a big data NoSQL solution? Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models: master-slave versus peer-to-peer; Four ways that NoSQL systems handle big data problems	08	20%
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,	08	15%
5	**Frameworks:** Applications on Big Data Using Hive and Pig, Hive Architecture, Data Flow and Data Types in Hive, Different Types of Tables in Hive, Partitioning and Bucketing in Hive, Why Apache Pig, Features of Apache Pig, Pig vs Mapreduce, Pig Architecture, Data processing operators in Pig, Hive services, HiveQL, Querying Data in Hive, fundamentals of HBase and ZooKeeper, IBM InfoSphere BigInsights and Streams.	08	20%
6	**Spark:** Introduction to Data Analysis with Spark, In-Memory Computing with Spark, Spark Basics, Interactive Spark with PySpark, Writing Spark Applications,	06	10%

Reference Book:

1) Sourabh Mukherjee , Amit Kumar Das and Sayan Goswami “Big Data Simplified”, Pearson
2) Anand Rajaraman and Jeff Ullman “Mining of Massive Datasets”, Cambridge University Press
3) Thomas Eri, Wajid Khattak and Paul Buhler, “Big Data Fundamentals: Concepts, Drivers & Techniques", Pearson
4) Big Data and Analytics , Seema Acharya, Subhashini Chhellappan, Willey
5) Chris Eaton, Dirk Deroos, Tom Deutsch, George Lapis and Paul Zikopoulcs “Understanding Big data”, McGraw Hill, 2012
6) Tom White, “HADOOP: The Definitive Guide”, O Reilly 2012
7) Learning Spark: Lightning-Fast Big Data Analysis by Holden Karau

Course Outcome:

After Completion of the Course, Student will able to:

No	Course Outcomes	RBT Level*
01	Identify big data application areas	UN
02	Use big data framework	AP
03	Model and analyze data by applying selected techniques	AN
04	Demonstrate an integrated approach to big data	AP

*RM: Remember, UN: Understand, AP: Apply, AN: Analyze, EL: Evaluate, CR: Create

Suggested Course Practical List:

Design based Problems (DP)/Open Ended Problem:

Case Study:

Stage 1:
Selection of case study topics and formation of small working groups of 2/3 students per group. Students engage with the cases, read through background material provided in the session and work through an initial set of questions to deepen the understanding of the case. Sample applications and data will be provided to help students familiarize themselves with the cases and available (big) data.

Stage 2:
The groups are given a specific task relevant to the case in question and are expected to develop a corresponding big data concept using the knowledge gained in the course and the parameters set by the case study scenario. A set of questions that help guide through the scenarios will be provided.

Stage 3:
Each group prepares a short 2 –5 page report on their results and a 10 min oral presentation of their big data concept.

Apart from case study, students can perform following programming exercises:

1. Implement following using Map- Reduce
   a. Matrix multiplication
   b. Sorting
   c. Indexing
2. Distributed Cache & Map Side Join, Reduce side Join Building and Running a Spark Application Word count in Hadoop and Spark Manipulating RDD.
3. Implementation of Matrix algorithms in Spark Sql programming.
4. Implementing K-Means Clustering algorithm using Map-Reduce.
5. Implementing any one Frequent Item set algorithm using Map-Reduce.
6. Create A Data Pipeline Based On Messaging Using PySpark And Hive - Covid-19 Analysis

List of Open Source Software/learning website:

1. http://in.reuters.com/tools/rss
2. http://www.altova.com/xmlspy.html
3. https://www.w3.org/RDF/
4. https://www.oreilly.com/library/view/data-analytics-with/9781491913734/ch04.html
5. https://data-flair.training/blogs/spark-in-memory-computing/

* * * * * * *