GUJARAT TECHNOLOGICAL UNIVERSITY

Bachelor of Engineering
Subject Code: 3170722
BIG DATA ANALYTICS
B.E. 7th Semester

Type of course: Elective

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.

Teaching and Examination Scheme:

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

Syllabus:

Sr. No.	Content	Total Hrs
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**: History of Hadoop, Hadoop Distributed File System: Physical organization of Compte 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, Shuffle and Sort, Task execution, Map Reduce Types and Formats, Map Reduce Features, Hadoop environment. Setting up a Hadoop Cluster, Cluster specification, Cluster Setup and Installation, Hadoop Configuration, ecurity in Hadoop, Administering Hadoop, Monitoring-Maintenance, Hadoop benchmarks, Hadoop in the cloud	12
3	**NoSQL:** What is NoSQL? NoSQL business drivers; NoSQL case studies; NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, 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	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					
R Level	U Level	A Level	N Level	E Level	C Level
15	15	30	20	15	5

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom's Taxonomy)

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.

Reference 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

Course Outcome:

After learning the course, the students should be able to:

Sr. No.	**CO Statement**	**Marks % Weightage**
1	identify big data application areas	15%
2	use big data framework	30%
3	model and analyze data by applying selected techniques	25%
4	demonstrate an integrated approach to big data	30%

List of Experiments and 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 student students will perform at the 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/