GUJARAT TECHNOLOGICAL UNIVERSITY
Syllabus for Integrated MSc, 9th Semester
Branch: Information Technology
Subject Name: Big Data
Subject Code: 1390501

Teaching and Examination Scheme

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

Content:

Sr. No.	Content	Teaching Hours	Module Weightage (%)
1.	**Essentials of Big Data and Analytics:** Types of Digital Data: Classification of Data (Structured, Semistructured and Unstructured), Characteristics of Data, Evolution of Big Data, Definition of Big Data, Challenges of Big Data, Characteristics of Big Data (Volume, Velocity, Variety), Other characteristics of Big Data which are not Definitional Traits of Big Data, Why Big Data?, Are we Information Consumer or Producer? Traditional BI vs Big Data, Typical Data Warehouse Environment, Typical Hadoop Environment.	06	15%
2.	**Introduction to NoSQL and Hadoop** **NoSQL:** Introduction: Where is it used? What is it? Types of NoSQL databases, Why NoSQL? Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL. **Hadoop:** Introducing Hadoop, Need of Hadoop, limitations of RDBMS, RDBMS versus Hadoop, Distributed computing challenges, History of Hadoop , Hadoop overview, Use case of Hadoop, Hadoop distributors, HDFS (Hadoop Distributed File System) , Processing data with Hadoop, Managing resources and applications with Hadoop YARN (Yet another Resource Negotiator), Interacting with Hadoop Ecosystem.	12	25%
3.	**Introduction to MongoDB and MapReduce** **MongoDB:** Introduction: What is MongoDB? Why MongoDb? (using JSON, Creating or generating a unique key, Support for Dynamic Queries, Storing Binary Data, Replication, Sharding, Updating information in – place), Terms used in RDBMS and MongoDb, Data types in MongoDb, MongoDB Query Language. **MapReduce:** Data Flow, Map, Shuffle, Sort, Reduce, Hadoop	08	25%

	Streaming, mrjob, Installation, word count in mrjob, Executing mrjob.		
4.	**Introduction to HIVE and Pig** **HIVE:** Introduction to Hive, Hive architecture, Hive data types, Hive file format, Hive Query Language (HQL), User-Defined Function (UDF) in Hive. **Pig:** The anatomy of Pig , Pig on Hadoop, Pig Philosophy, Use case for Pig; ETL Processing , Pig Latin overview , Data types in Pig , Running Pig , Execution modes of Pig, HDFS commands, Relational operators, Piggy Bank , Word count example using Pig.	10	25%
5.	**Overview of SPARK** Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs.	04	10%

Reference Books:

1) Seema Acharya, Subhashini Chellappan, " Big Data and Analytics", Wiley India Pvt.Ltd.,2015
2) Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau ,"Learning Spark",O'Reilly Media,2015
3) Zachary Radtka and Donald Miner,"Hadoop with Python",O'Reilly Media,2016 (Free ebook is available on the following link)(As on 12-10-2018)
4) Shashank Tiwari, "Professional NoSQL", Wiley India Pvt. Ltd.,2011
5) Kyle Banker,Peter Bakkum,Shaun Verch,Douglas Garrett,Tim Hawkins,"MongoDB in Action", DreamTech Press, 2nd Edition ,2016
6) Chris Eaton,Paul Zikopoulos,Tom Deutsch,George Lapis,Dirk Deroos,"Understanding Big Data : Analytics for Enterprise Class Hadoop and Streaming Data", Mcgraw Hill Education (India)Pvt.Ltd.,2012

Course Outcome:

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

No.	CO statement
CO-1	Understand Big Data and its analytics in the real world.
CO-2	Analyze the Big Data framework like Hadoop and NOSQL to efficiently store and process Big Data to generate analytics.
CO-3	Design of Algorithms to solve Data Intensive Problems using Map Reduce Paradigm.
CO-4	Design and Implementation of Big Data Analytics using pig and implement Big Data Activities using Hive.
CO-5	Implement Spark to solve data intensive problems and to generate analytics.