GUJARAT TECHNOLOGICAL UNIVERSITY
Syllabus for Integrated MSc, 9th Semester
Branch: Computer Science
Subject Name: Big Data Analytics
Subject Code: 1390307

Teaching and Examination Scheme:

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

Content:

Sr. No.	Content	Teaching Hours	Module Weightage (%)
1	**DATA COLLECTION AND VISUALIZATION:** Concept of measurement, details of measurement, design of data collection format with illustrations, data qualities and issues with data collection systems with examples from business, cleaning and treatment of missing data. Principles of visualization and different methods involved.	6	20
2	**STATISTICS AND CONTNGENCY TABLES:** Frequency table, histogram, measure of location, skewness, kurtosis, correlation and simple linear regression, probability distribution as statistical model, fitting probability distribution.2-way contingency tables, testing of dependence	6	20
3	**BIG DATA**: Concepts and terminology, Big Data Characteristics, Different types of Data, Identifying Data Characteristics - Big Data Architecture - Big Data Storage: File system and Distributed File System, No SQL, Sharding, Replication, Sharding and Replication, ACID and BASE Properties.	8	20
4	**HADOOP:** Hadoop Architecture - Hadoop Distributed File System (HDFS) –YARN – Hadoop I/O – Map Reduce: Developing a map-reduce application – Map-reduce working procedure – Types and Formats - Features of Map reduce: sorting and joins- Pipelining Map Reduce jobs. Hadoop Technologies: Introduction, Parallel processing using Pig, Pig Architecture, Grunt, Pig Data Model-scalar and complex types. Pig Latin- Input and output, Relational operators, User defined functions - Working with scripts. Hadoop Operations.	10	20
5	**HIVE AND SPARK:** Introduction-Hive modules, Data types and file formats, Hive QL-Data Definition and Data Manipulation-Hive QL queries, Hive QL views- reduce query complexity. Hive scripts. Hive QL Indexes- Aggregate functions- Bucketing vs Partitioning. Spark: Overview of Spark – Hadoop Overview of Spark – Hadoop vs. Spark – Cluster Design – Cluster Management – performance, Application	10	20

	Programming interface (API): Spark Context, Resilient Distributed Datasets, Creating RDD, RDD Operations, and Saving RDD - Lazy Operation-SPARK JOBS.		

Reference Books:

1. Thomas Erl, Wajid Khattak, and Paul Buhler, Big Data Fundamentals: Concepts, Drivers &Techniques, Pearson India Education Service Pvt. Ltd., First Edition, 2016.
2. Tom White, Hadoop: The Definitive Guide, O'Reilly Media, Inc., Fourth Edition, 2015.
3. Alan Gates, Programming Pig Dataflow Scripting with Hadoop, O'Reilly Media, Inc, 2011.
4. Jason Rutherglen, Dean Wampler, Edward Caprialo, Programming Hive, O'Reilly Media Inc,2012
5. Mike Frampton, "Mastering Apache Spark", Packt Publishing, 2015.
6. Mohammed Gulle , "Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis" – Apress 2015

Course Outcome:

Upon completion of the course the student will be able to:

1. Understand the need of new frame work to deal with huge amounts of Data.
2. Demonstrate the Hadoop framework Hadoop Distributed File System and Map Reduce.
3. Demonstrate the Pig architecture and evaluation of pig scripts.
4. Describe the Hive architecture and execute SQL queries on sample data sets.
5. Demonstrate spark programming with different programming languages and graph algorithms.

List of Experiments:

1. Installing and configuring the Hadoop frame work. HDFS Commands.
2. Map Reduce Program to show the need of combiner
3. Map Reduce I/O Formats – Text, Key – Value
4. Map Reduce I/O Formats – N Line – Multiline
5. Installing and Configuring Apache PIG and HIVE
6. Sequence File Input / Output Formats
7. Distributed Cache & Map side Join, Reduce Side Join
8. Building and Running Spark Application
9. Word count in Hadoop and Spark
10. Manipulation RDD 11. Spark Implementation of Matrix algorithms in Spark Spark Sql programming, Building Spark Streaming application

List of open Source software/learning Websites:

- https://hadoop.apache.org/
- https://spark.apache.org/
- https://www.kaggle.com/