



# GUJARAT TECHNOLOGICAL UNIVERSITY

Syllabus for Integrated Master of Computer Applications, 8<sup>th</sup> Semester

Subject Name: Big Data Analytics

Subject Code: 2688601

With effective  
from academic  
year 2020-21

Prerequisite: NA

Rationale: NA.

Content:

Sr. No.	Content	Total Hrs	Weightage %
1	<b>INTRODUCTION TO BIG DATA</b> Introduction– distributed file system–Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications. Algorithms using map reduce	06	13
2	<b>INTRODUCTION TO HADOOP AND HADOOP ARCHITECTURE</b> Big Data – Apache Hadoop & Hadoop Eco System, Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce -, Data Serialization.	12	25
3	<b>HDFS, HIVE AND HIVEQL, HBASE</b> HDFS-Overview, Installation and Shell, Java API; Hive Architecture and Installation, Comparison with Traditional Database, HiveQL Querying Data, Sorting And Aggregating, Map Reduce Scripts, Joins & Sub queries, HBase concepts, Advanced Usage, Schema Design, Advance Indexing, PIG, Zookeeper , how it helps in monitoring a cluster, HBase uses Zookeeper and how to Build Applications with Zookeeper.	08	15
4	<b>SPARK</b> Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs, Machine	12	20
5	<b>NoSQL</b> What is it?, Where It is Used Types of NoSQL databases, Why NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL, NewSQL	05	12
6	<b>Data Base for the Modern Web</b> Introduction to MongoDB key features, Core Server tools, MongoDB through the JavaScript's Shell, Creating and Querying through Indexes, Document-Oriented, principles of schema design, Constructing queries on Databases, collections and Documents, MongoDB Query Language.	08	15



# GUJARAT TECHNOLOGICAL UNIVERSITY

Syllabus for Integrated Master of Computer Applications, 8<sup>th</sup> Semester

Subject Name: Big Data Analytics

Subject Code: 2688601

With effective  
from academic  
year 2020-21

## Suggested Specification table with Marks (Theory):

Distribution of Theory Marks					
R Level	U Level	A Level	N Level	E Level	C Level
10	20	25	28	16	0

**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:

- Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.
- Chris Eaton, Dirk deRoos et al., "Understanding Big data", McGraw Hill, 2012.
- BIG Data and Analytics, Sima Acharya, Subhashini Chellappan, Wiley
- MongoDB in Action, Kyle Banker, Peter Bakum, Shaun Verch, Dream tech Press
- Tom White, "HADOOP: The definitive Guide", O Reilly 2012.
- Vignesh Prajapati, "Big Data Analytics with R and Hadoop", Packet Publishing 2013.
- <http://www.bigdatauniversity.com/>
- Learning Spark: Lightning-Fast Big Data Analysis Paperback by Holden Karau

## Course Outcome:

Upon completion of this course, students will be able to do the following:

- CO1. Students will be able to build and maintain reliable, scalable, distributed systems with Apache Hadoop.
- CO2. Students will be able to write Map-Reduce based Applications
- CO3. Students will be able to design and build MongoDB based Big data Applications and learn MongoDB query language
- CO4. Students will learn difference between conventional SQL query language and NoSQL basic concepts
- CO5. Students will learn tips and tricks for Big Data use cases and solutions.



## List of Experiments:

1. To understand the overall programming architecture using Map Reduce API
2. Store the basic information about students such as roll no, name, date of birth , and address of student using various collection types such as List, Set and Map
3. Basic CRUD operations in MongoDB
4. Retrieve various types of documents from students collection
5. To find documents from Students collection
6. Develop Map Reduce Work Application
7. Creating the HDFS tables and loading them in Hive and learn joining of tables in Hive

## Design based Problems (DP)/Open Ended Problem:

1. Create a system which can use of Web search, web crawlers and web information retrieval.
2. Analyze and implement a system with Web graph mining.
3. Implement and Subscribe RSS News feeds to get latest news in India.

## Major Equipment:

XMLSpy, RSS Feed, RSS Reader.

## List of Open Source Software/learning website:

- <http://in.reuters.com/tools/rss>
- <http://www.altova.com/xmlspy.html>
- <https://www.w3.org/RDF/>

## ACTIVE LEARNING ASSIGNMENTS:

Preparation of power-point slides, which include videos, animations, pictures, graphics for better understanding theory and practical work – The faculty will allocate chapters/ parts of chapters to groups of students so that the entire syllabus to be covered. The power-point slides should be put up on the web-site of the College/ Institute, along with the names of the students of the group, the name of the faculty, Department and College on the first slide. The best three works should submit to GTU.