



GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Post Graduate Diploma In Data Science

Level: PG

Branch: Data Science

Subject Code: DS02080031

Subject Name: Big Data Analytics

w. e. f. Academic Year:	2024-25
Semester:	2
Category of the Course:	Core Course

Prerequisite:	Nil
Rationale:	Nil

Objectives

- To provide knowledge about scalable, reliable and distributed storage environment for huge amount of data.
- To provide insights of data analysis in distributed environment.
- To optimize business decisions and create competitive advantage with Big Data analytics
- To introduce programming tools PIG & HIVE in Hadoop ecosystem
- To able to realistically assess the application of big data analytics technologies for different usage scenarios

Course Outcomes:

Sr. No.	CO statement
1	analyze the HADOOP and Map Reduce technologies associated with big data analytics.
2	work with big data platform and explore the big data analytics techniques in business applications.
3	Design and build MongoDB based Big data Applications and learn MongoDB query language.
4	Differentiate between conventional SQL query language and NoSQL basic concepts.
5	Build a complete business data analytics solution.



GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Post Graduate Diploma In Data Science

Level: PG

Branch: Data Science

Subject Code: DS02080031

Subject Name: Big Data Analytics

Teaching and Examination Scheme:

Teaching / Learning Scheme (in Hours per semester)			Total Credits	Assessment Pattern and Marks				Total Marks
L	T	PR		Theory		Tutorial / Practical		
				ESE (E)	PA / CA (M)	PA/CA (I)	ESE (V)	
4	0	2	5	70	30	0	50	150

Content:

Sr. No.	Content	Total Hrs
1	Introduction to Big Data 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	02
2	Introduction to Hadoop and its Architecture Big Data – Apache Hadoop & Hadoop EcoSystem, Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce -Data Serialization	06
3	HDFS, HIVE AND HIVEQL, HBASE 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	10
4	SPARK Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs, Machine Learning with MLlib	10
5	NoSQL 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	06
6.	Database for the Modern Web 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	06
TOTAL		



GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Post Graduate Diploma In Data Science

Level: PG

Branch: Data Science

Subject Code: DS02080031

Subject Name: Big Data Analytics

Reference Books:

1. Professional Hadoop Solutions by Boris Iubinsky, Kevin T. Smith, Alexey Yakubovich Wiley, ISBN-13: 978-1118611937, ISBN-10: 1118611934, 2015
2. Understanding Big data by Chris Eaton and Paul C. Zikopoulos McGraw Hill, 2012
3. Big Data and Analytics by Seema Acharya, Subhashini Chhappan Willey, Second edition
4. MongoDB in Action, by Kyle Banker, Peter Bakum, Shaun Verch DreamTech, ISBN 9781935182870, Second edition
5. HADOOP: The definitive Guide by Tom White O'Reilly, 2012
6. Big Data Analytics with R and Hadoop by Vignesh Prajapati Packet Publishing, 2013

- **Suggested list of Practical (at least 10 practical are to be performed by students. These practical should cover majority of all topics of syllabus.)**
- **This is the suggested list of practical but it may not be limited only to this list.**

1. To draw and explain Hadoop Architecture and Ecosystem with the help of a case study using WordCount example. To define and install Hadoop.
2. To understand the overall programming architecture using Map Reduce API.
3. To implement the following file management tasks in Hadoop System (HDFS): Adding files and directories, Retrieving files, Deleting files.
4. 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.
5. Basic CRUD operations in MongoDB.
6. To perform NoSQL database using MongoDB to create, update and insert.
7. Retrieve various types of documents from students collection.
8. To find documents from Students collection.
9. To run a basic Word Count MapReduce program to understand MapReduce Paradigm: To count words in a given file, To view the output file, and To calculate execution time.
10. Creating the HDFS tables and loading them in Hive and learn joining of tables in Hive.
