GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Master of Computer Applications

Level: Post Graduate

Course / Subject Code: MC03094081

Course / Subject Name: Big Data Tools

w. e. f. Academic Year:	2025-26
Semester:	3
Category of the Course:	Elective Group-1

Prerequisite:	Basic understanding of databases, data structures, and programming concepts. Familiarity with SQL and any programming language (preferably Python or Java) is recommended.
Rationale:	• To build a strong foundation in big data concepts, architectures, and tools used for handling large-scale data. • To provide hands-on experience with leading big data technologies such as Hadoop, NoSQL, MongoDB, Hive, Pig, and Spark. • To understand data storage, processing, and analysis in distributed computing environments. • To develop practical skills for managing, querying, and analyzing structured and unstructured data. • To enable students to apply big data tools and techniques in solving real-world problems and to pursue careers in fields leveraging large-scale data technologies.

Course Outcome:

After completion of the course, the student will be able to:

No	Course Outcomes	RBT Level
01	Discuss the core concepts of Big Data, including its characteristics, challenges, and evolving ecosystem.	UN
02	Apply the concepts of NoSQL to develop the non relational databases using MongoDB	AP
03	Implement the various concepts of Hadoop, Hadoop ecosystem, Hadoop Components, HDFS and Map Reduce in managing large-scale data environments.	CR
04	Design and analyze data processing workflows using HIVE and Pig	AP
05	Explain the basic concepts of SPARK	UN

**Revised Bloom's Taxonomy (RBT)*

Teaching and Examination Scheme:

Teaching Scheme (in Hours)			Total Credits L+T+ (PR/2)	Assessment Pattern and Marks				Total Marks
				Theory		Practical		
L	T	PR	C	ESE (E)	PA / CA (M)	PA(I)	ESE (V)	
3	0	2	4	70	30	20	30	150

Course Content:

Unit No.	Content	No. of Hours	% of Weightage
1.	**Introduction to Big Data** 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, what is Changing in Realms of Big Data? Terminologies used in Big Data Environments	5	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, NewSQL **Hadoop:** Introduction, Distributed Computing Challenges, History of Hadoop, Overview of Hadoop and Hadoop Ecosystems, Features and key advantages of Hadoop, Versions of Hadoop, Hadoop distributions, RDBMS versus Hadoop, Hadoop vs SQL, Integrated Hadoop Systems offered by leading market vendors, Cloud based Hadoop solutions, HDFS, Processing data with Hadoop, Managing Resources and applications with Hadoop YARN, Interacting with Hadoop Ecosystem	13	25%
3.	**Introduction to MongoDB and Map Reduce** **MongoDB:** Introduction: What is MongoDB? Why Mongo DB? (using JSON, Creating or generating a unique key, Support for Dynamic Queries, Storing Binary Data, Replication, Shading, Updating information in place), Terms used in RDBMS and Mongo DB, Data types in Mongo DB, MongoDB Query Language **Map Reduce:** Data Flow, Map, Shuffle, Sort, Reduce, Hadoop Streaming, mrjob, Installation, word count in mrjob, Executing mrjob	12	25%
4.	**Introduction to HIVE and Pig** **HIVE:** Introduction: What is HIVE? HIVE Architecture, HIVE data Types, HIVE File Formats, HIVE Query Language, RCFile implementation, SerDe, User-Defined Functions (UDF). **Pig:** Introduction: What is 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, Eval function, Complex Data Types, Piggy Bank,	10	25%

	User-defined Functions, Parameter substitution, Diagnostic Operator, Word Count Example using Pig, when to use and not use Pig?, Pig at Yahoo, Pig vs HIVE.		
5.	Overview of SPARK Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs.	5	10%
	Total	**45**	**100**

Suggested Specification Table with Marks (Theory):

Distribution of Theory Marks (in %)					
R Level	**U Level**	**A Level**	**N Level**	**E Level**	**C Level**
10	25	30	25	5	5

Where R: Remember; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create (as per Revised Bloom's Taxonomy)

References/Suggested Learning Resources:

(a) 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. Shashank Tiwari, "Professional NoSQL", Wiley India Pvt. Ltd.,2011
4. Kyle Banker,Peter Bakkum,Shaun Verch,Douglas Garrett,Tim Hawkins,"MongoDB in Action", DreamTech Press, 2nd Edition ,2016
5. 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
6. Tom White,"Hadoop: The Definitive Guide",O'Reilly Media,4th Edition,2015
7. Vignesh Prajapati,"Big Data Analytics With R and Hadoop", Packt Pub Ltd ,2013

(b) Web Resources:

1. http://www.bigdatauniversity.com
2. http://www.mongodb.com
3. http://hadoop.apache.org/

(b) Open-source software and website:

1. Development Environment & IDEs
   - Apache Zeppelin – https://zeppelin.apache.org/
   - Jupyter Notebook – https://jupyter.org/
   - Visual Studio Code – https://code.visualstudio.com/
2. Big Data Frameworks
   - Apache Hadoop – https://hadoop.apache.org/
   - Apache Spark – https://spark.apache.org/
   - Apache Hive – https://hive.apache.org/
   - Apache Pig – https://pig.apache.org/

3. NoSQL & Databases
   - MongoDB (Community Edition) – https://www.mongodb.com/try/download/community
   - Apache HBase – https://hbase.apache.org/
   - CouchDB – https://couchdb.apache.org/
   - Cassandra – https://cassandra.apache.org/
4. Data Processing & Querying Tools
   - Apache Flink – https://flink.apache.org/
   - Apache Drill – https://drill.apache.org/
5. Cluster & Resource Management
   - Apache YARN - https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
   - Apache Mesos – https://mesos.apache.org/
   - Kubernetes – https://kubernetes.io/
6. Cloud & Distributed Systems
   - Google Cloud Big Data Tools – https://cloud.google.com/products/big-data
   - AWS Big Data Solutions – https://aws.amazon.com/big-data/
   - Microsoft Azure HDInsight – https://azure.microsoft.com/en-in/products/hdinsight/

SUGGESTED PROJECT LIST:

Note: Hadoop programs can be implemented using either Java or Python

MongoDB

1. **Create a Student Master Database**
   Create a database named StudentDB and a collection called Student containing documents with the following fields: StudentRollNo, StudentName, Grade, Hobbies.
   Perform the following operations:
   a) Insert at least 5 student records.
   b) Update the grade of a student.
   c) Delete a student record.
   d) Retrieve all students whose grade is above 'B'.
2. **Employee Management System**
   Create a database CompanyDB with a collection Employees. Each document should have: EmpID, EmpName, Department, JoiningDate, Salary.
   Perform queries to:
   a) Find employees from a specific department.
   b) Sort employees by salary in descending order.
   c) Find employees who joined after 2022.
3. **Use of Aggregation Framework**
   Use any existing collection (e.g., Employees from above) and perform the following tasks:
   a) $group to calculate total salary by department.
   b) $match to filter records by condition.
   c) $project to include/exclude specific fields.
   d) $sort based on a numeric field.
4. **Online Store Inventory System**

Create a collection Products with fields: ProductID, ProductName, Category, Price, StockAvailable.
Tasks:

a) Insert 10 product documents.
b) Find products in the 'Electronics' category.
c) Update stock for a particular product.
d) Delete out-of-stock products.

5. **Library Database**
Create a Library collection with fields: BookID, Title, Author, Genre, AvailableCopies.
Perform:
a) Insert multiple book records.
b) Find all books by a specific author.
c) Increment the number of copies for a book when new stock arrives.
d) Use $in operator to find books in 'Fiction' or 'Science' genres.
6. **Customer Feedback System**
Create a collection Feedback with: FeedbackID, CustomerName, Rating, Comments, Date.
Perform the following:
a) Insert 5 feedback entries.
b) Find feedback with rating >= 4.
c) Use projection to retrieve only CustomerName and Rating.
d) Count total feedback entries.
7. **Working with Date and Time**
Create a collection Events with fields: EventID, EventName, EventDate, Location.
Performed the following tasks:
a) Insert 5 events with different dates.
b) Retrieve upcoming events (EventDate > current date).
c) Format the date using MongoDB operators.
d) Sort events based on date.

Hadoop/HDFS

8. **Install and Set Up Hadoop Environment**
a) Install Java Development Kit (JDK)
b) Download and extract the Hadoop package
c) Configure environment variables for Java and Hadoop
d) Verify the installation of Java and Hadoop using version commands
9. **HDFS – Basic File Operations**
Perform the following tasks using Hadoop HDFS commands:
a) Create a file in the local file system and add some sample content.
b) Create a directory in HDFS.
c) Upload the local file to the HDFS directory using different commands (-put, -copyFromLocal).
d) List the contents of the HDFS directory (-ls, -ls -R).
e) Display the contents of the file in HDFS using the -cat command.
f) Copy the file within HDFS using the -cp command.
g) Download the file back to the local system using -get or -copyToLocal.

h) Delete the file and directory from HDFS using -rm -r.

10. Word Count using mrjob (Python-based MapReduce)

Using mrjob library in Python:

a) Create a Python script to count the frequency of words in a text file
b) Run the job locally and on Hadoop (if configured)

HIVE/PIG

11. Create a Hive table and load data

a) Create a Hive table with sample fields.
b) Load data from a local or HDFS file.
c) Display all records using a SELECT query.

12. Perform basic SQL operations in Hive

a) Use SELECT, WHERE, GROUP BY, and ORDER BY clauses.
b) Run queries to filter and sort data.

13. Load data in Pig and view structure

a) Load a data file using Pig Latin.
b) Use DUMP and DESCRIBE to view content and schema.

14. Filter and group data using Pig

a) Apply FILTER and GROUP operations.
b) Count grouped results using FOREACH and COUNT.

15. Use functions and expressions in Pig

a) Use eval functions like TOKENIZE, UPPER, or SUBSTRING.
b) Apply transformations on loaded data.
c) View results using DUMP.

16. Perform word count using Pig Latin

a) Load a text file.
b) Split lines into words.
c) Count frequency of each word and store results.

CO- PO Mapping:

Semester 3	**Big Data Tools**							
	POs							
Course Outcomes	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8
CO1	3	-	-	-	-	-	-	-
CO2	3	2	2	3	-	-	-	-
CO3	3	2	2	3	-	-	-	-
CO4	3	3	3	3	-	-	-	-
CO5	3	3	-	3	-	-	-	-

*Legend: '**3**' for high, '**2**' for medium, '**1**' for low and '**-**' for no correlation of each CO with PO.*

Note: The CO-PO mapping is indicative; the institute/faculty member can change as required.

* * * * * * *