Reg. No.

B.E./ B. TECH. DEGREE EXAMINATIONS, MAY 2024
Sixth Semester
AD18602 – BIG DATA ANALYTICS AND VISUALIZATION
(Artificial Intelligence and Data Science)
(Regulation 2018 / 2018A)

TIME:3 HOURS   **MAX. MARKS: 100**

COURSE OUTCOMES	STATEMENT	RBT LEVEL
CO 1	Students will be able to apply analytics to big data applications.	**3**
CO 2	Students will be able to understand the hadoop frameworks.	**2**
CO 3	Students will be able to use MongoDB and Cassandra for big data storage and retrieval.	**3**
CO 4	Students will be able to work with hadoop ecosystem tools such as MapReduce, Hive, Pig.	**3**
CO 5	Students will be able to visualize data to transform into information for further analysis.	**3**

PART- A(10x2=20Marks)
(Answer all Questions)

		CO	RBT LEVEL
1.	Identify the various kinds of technologies that are required to face the challenges of big data.	**1**	**2**
2.	Summarize the advantages of shared nothing architecture.	**1**	**2**
3.	Differentiate the characteristics of RDBMS and Hadoop.	**2**	**2**
4.	Enumerate the key components of Master Node and its uses in Hadoop environment.	**2**	**2**
5.	Write a query in MongoDB to import the data from **'student.csv'** to the **'student'** collection present in **'school'** database.	**3**	**3**
6.	Write a query in Cassandra to drop a column **'st_name'** from the column family **'student'**.	**3**	**3**
7.	Enumerate the types of partition in HIVE and its effects on handling huge degree of I/O.	**4**	**3**

8.	Write the equivalent commands for the two modes of execution of a PIG program.	**4**	**3**
9.	Analyse the reason behind transforming the data into information to infer valuable insights.	**5**	**3**
10.	Write the perception behind Gestalt principles for data visualization.	**5**	**3**

PART- B (5x 14=70Marks)

			Marks	**CO**	**RBT LEVEL**
11. (a)	**(i)**	Demonstrate how the digital data is being classified in the real world scenario.	**(7)**	**1**	**3**
	(ii)	Interpret the skills required for a data scientist in implementing a data science project.	**(7)**	**1**	**3**

(OR)

(b)		Illustrate the following key terminologies used in big data environment	**(14)**	**1**	**3**
	i.	In-Memory Analytics			
	ii.	In-Database Processing			
	iii.	Symmetric Multiprocessor System			
	iv.	Massively Parallel Processing			
	v.	CAP Theorem			

12. (a)	**(i)**	Explain the responsibilities of name node and data node in Hadoop Distributed File System.	**(7)**	**2**	**2**
	(ii)	Discuss the steps involved in YARN Architecture.	**(7)**	**2**	**2**

(OR)

(b)	**(i)**	Explain how the jobtracker and tasktracker handle job scheduling in a	**(7)**	**2**	**2**

mapreduce framework.

(ii) Consider the text file 'test.txt' whose contents are as follows: (7) 2 2

An apple is a round, edible fruit produced by an apple tree. Apple trees are cultivated worldwide and are the most widely grown species in the genus Malus. The apple tree comes from southern Kazakhstan, Kyrgyzstan, Uzbekistan, and northwestern part of China. Apples have been grown for thousands of years in Asia and Europe.

Perform a word count on the file 'test.txt' using MapReduce to find the unique words and the number of occurrences of those unique words. Construct the various stages of output in MapReduce Process. (Consider there are 6 mappers initially)

13. (a) Write the equivalent query for the following commands in MongoDB. (14) 3 3

Match_ID	Venue	Place
M1	Salt Lake Stadium	Kolkatta
M2	Jawaharlal Nehru Stadium	Delhi
M3	Greenfield International Stadium	Thiruvananthapuram
M4	EMS Stadium	Kozhikode
M5	KD Singh Babu Stadium	Lucknow

a. Create a database **'sport'**.

b. Create a collection **'football'** with Match_ID as primary key and insert the documents given in the table.

c. Display the Venue alone from the collection.

d. Display the documents whose place starts with 'K'.

e. Change the Venue of Match_ID 'M1' to 'SL Stadium'.

f. Display the last 3 documents in the collection.

g. Export the collection to 'football.csv'.

(OR)

(b) Write the equivalent query for the following commands in Cassandra. **(14)** **3** **3**

flower_id	flower_name	flower_color	flower_family
101	hollyhock	pink	Malvaceae
102	coreopsis	yellow	Asteraceae
103	calendula	yellow	Asteraceae
104	geranium	red	Geraniaceae
105	star jasmine	white	Apocynaceae

Create a keyspace **'flower'** with replication factor 3.

a. Create a column family **'flowers_tbl'** with flower_id as primary key and insert the records given in the table.

b. Display the record whose flower_id is 104.

c. Update the flower_color of flower_id '103' as 'lemon yellow'.

d. Display the record whose flower_name is 'star jasmine'.

e. Sort the data in the descending order of flower_family.

f. Export the column family to 'flower.csv'.

14. (a) **(i)** Demonstrate the use of various components in the HIVE architecture. **(7)** **4** **3**

(ii) Write down the DDL commands used in HIVE for creation and deletion of database, table, view and index. **(7)** **4** **3**

(OR)

(b)	**(i)**	Apply any six relational operators in PIG to a table containing fields such as RNO (INT), NAME (CHARACTER ARRAY) and MARKS (INT).	**(7)**	**4**	**3**
	(ii)	Write a PIG program to generate the word count for the given text.	**(7)**	**4**	**3**
15. (a)	**(i)**	Examine the different tools in python used for data visualization.	**(7)**	**5**	**3**
	(ii)	Interpret the use of each data visualization plot in data analytics.	**(7)**	**5**	**3**

(OR)

(b)	Demonstrate in detail the best practices used for data visualization.	**(14)**	**5**	**3**

PART- C (1x 10=10Marks)

(Q.No.16 is compulsory)

		Marks	CO	RBT LEVEL
16.	Write a mapreduce program to perform word search on the text. Generate the output of each step involved in performing word search.	**(10)**	**4**	**5**

****