Q. Code:330324

Reg. No. | | | | | | | | | | | | | |

B.E./ B. TECH.DEGREE EXAMINATIONS, MAY 2024

Seventh Semester

IT18702 – BIG DATA ANALYTICS

(Information Technology)

(Regulation 2018 /2018A)

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

COURSE OUTCOMES	STATEMENT	RBT LEVEL
CO 1	Identify the characteristics of datasets and compare the trivial data and big data for various applications.	**3**
CO 2	Interpret business models and scientific computing paradigms, and apply software tools for big data analytics.	**4**
CO 3	Apply scaling up machine learning techniques and associated computing techniques and technologies.	**3**
CO 4	Integrate machine learning libraries and mathematical and statistical tools with modern technologies like Hadoop and MapReduce.	**4**
CO 5	Investigate how Big Data is managed.	**5**

PART- A(10x2=20Marks)

(Answer all Questions)

		CO	RBT LEVEL
1.	Categorize the types of data analytics with examples.	**1**	**2**
2.	Differentiate probabilistic and non-probabilistic sampling techniques.	**1**	**2**
3.	State real-time applications of the stream data model.	**2**	**3**
4.	Identify widely used Real-Time Analytics Platforms (RTAPs).	**2**	**3**
5.	Why data visualization is necessary for big data analytics?	**3**	**3**
6.	Point out the challenges involved in text analysis.	**3**	**3**
7.	How the replication technique is used in HDFS to provide fault tolerance?	**4**	**4**
8.	How does MapReduce organize the execution of a job?	**4**	**4**
9.	Differentiate PIG and HIVE.	**5**	**4**
10.	Determine the different types of client applications for performing queries on the Hive.	**5**	**5**

PART- B (5x 14=70Marks)

			Marks	CO	RBT LEVEL
11. (a)	**(i)**	Identify how the different phases of data analytic lifecycle is used to address the Big data problems.	**(7)**	**1**	**3**
	(ii)	Classify the various tests of significance and write down the steps in testing a hypothesis.	**(7)**	**1**	**3**

(OR)

(b)	**(i)**	A personal fitness trainer claimed that all the new boys are above average weight. A random sample of thirty boys weight have a mean score of 112.5 kg and the population mean weight is 100 kg and the standard deviation is 15. Is there a sufficient evidence to support the claim of the trainer?(Critical table value is 1.645)	**(7)**	**1**	**3**
	(ii)	The marks of boys and girls are given: Boys: 12, 14, 10, 8, 16, 5, 3, 9, and 11 Girls: 21, 18, 14, 20, 11, 19, 8, 12, 13, and 15 Is there any significant difference between the marks of males and females i.e. population means are different? (Critical table value is 2.11)	**(7)**	**1**	**3**
12. (a)		Examine the working of the Flajolet Martin algorithm for a sample data stream to count the number of distinct elements.	**(14)**	**2**	**4**

(OR)

(b)		Inspect sentiment analysis techniques used in social media monitoring and discuss their effectiveness in understanding public opinion and trends.	**(14)**	**2**	**4**
13. (a)		Construct decision tree for the following example data to make predictions based on the age attribute.	**(14)**	**3**	**3**

Age	Likes dogs	Likes gravity	Going to be an astronaut
24	0	0	0
30	1	1	1
36	0	1	1
36	0	0	0
42	0	0	0
44	1	1	1
46	1	0	0
47	1	1	1
47	0	1	0
51	1	1	1

(OR)

(b)	x	0	1	2	3	4	**(14)**	**3**	**3**
	y	2	3	5	4	6			

a) Find the least square regression line y= ax+b

b) Estimate the value of y when x=10

c) Plot the given points and regression line in the same rectangular system of axes.

		Marks	CO	RBT LEVEL
14. (a)	Examine the functionalities of four different entities in the anatomy of MapReduce Job run.	**(14)**	**4**	**4**
	(OR)			
(b)	Explore the different types of Java Interfaces and APIs to perform file operations in HDFS.	**(14)**	**4**	**4**
15. (a)	Interpret the functionality of Apache PIG components with data types and operators to perform operations on the data.	**(14)**	**5**	**5**
	(OR)			
(b)	Determine the principles, processes, and tools used in the design and development of data visualization system.	**(14)**	**5**	**5**

PART- C (1x 10=10Marks)

(Q.No.16 is compulsory)

		Marks	CO	RBT LEVEL
16.	Evaluate the performance of IBM-specific tools and technologies used in the IBM Infosphere BigInsights platform for processing stream data.	**(10)**	**5**	**5**

****