Q. Code: 662685

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

B. E / B. TECH.DEGREE EXAMINATION, MAY 2023

Seventh Semester

IT18702–BIG DATA ANALYTICS

(Information Technology)

(Regulation2018)

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

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

PART- A(10x2=20Marks)

(Answer all Questions)

		CO	RBT LEVEL
1	List the main characteristics of Big Data.	**1**	**1**
2	Why does one choose analytical system over conventional system?	**1**	**3**
3	What is the need of sampling and list out the characteristics of good sampling?	**2**	**2**
4	Give one real time example for overfitting and underfitting in model.	**2**	**2**
5	In a corpus of N documents, one document is randomly picked. The document contains a total of T terms and the term "data" appears K times. Write the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term "data" appears in approximately one-third of the total documents.	**3**	**3**
6	What is the importance of the F-test in a linear model?	**3**	**2**
7	What is the need of Rack Awareness algorithm?	**4**	**2**
8	Compare and contrast the scheduling algorithms used in MapReduce?	**4**	**3**
9	How the relational operators are used in Pig Scripts?	**5**	**2**
10	Is Scaling possible in Hadoop? If yes, List the types of scaling.	**5**	**2**

Q. Code: 662685

PART- B (5x 14=70Marks)

(Restrict to a maximum of TWO subdivisions)

			Marks	CO	RBT LEVEL
11(a)	Draw the conceptual architecture of Big Data Analytics and explain the tools.		**(14)**	**1**	**3**
	(OR)				
11(b)	**(i)**	In a random sample of 100 men taken from a village A, 60 are found to be consuming alcohol. In another sample of 200 men taken from village B, 100 were found to be consuming alcohol. Do the two villages differ significantly in respect of their consuming alcohol?	**(7)**	**1**	**3**
	(ii)	Consider a controlled clinical trial in which 90 of 100 patients received treatment A got cured compared with 105 of 150 who received Treatment. Test the hypothesis that Treatment A is more effective than Treatment B?	**(7)**	**1**	**3**

	Response to treatment		
	Cured	Not cured	Total (Rj)
TREATMENT A	90 (a11)	10 (a12)	100 (R1)
TREATMENT B	105 (a21)	45 (a22)	150 (R2)
Total (Cj)	195 (C1)	55 (C2)	250

12(a)	Consider a following stream of data 101011 000 10111 0 11 00 101 10 where 1 represents the movie viewed by the customer. Explain the algorithm to count the number of ones (after k=15 timestamp) after the following data stream 0111		**(14)**	**2**	**3**
	(OR)				
12(b)	Explain how the stream processing model manages the huge data and propose a method to increase the efficiency of blooms?		**(14)**	**2**	**3**
13(a)	A database has five transactions. Let min sup = 60% and min conf=80% TID ITEMS T100 Milk, Onion, Nuts, Kiwi, Egg, Yoghurt T200 Dhal, Onion, Nuts, Kiwi, Egg, Yoghurt T300 Milk, Apple, Kiwi, Egg T400 Milk, Curd, Kiwi,Yoghurt T500 Curd, Onion, Kiwi, Ice cream,Egg Find all frequent item sets using Apriori method and write the association rules. Write appropriate R code.		**(14)**	**3**	**3**

Q. Code: 662685

		(OR)			
13(b)	**(i)**	Explain Decision tree with an example to predict whether the customers will buy a product.	**(10)**	**3**	**3**
	(ii)	Differentiate Linear regression and Logistic regression.	**(4)**	**3**	**3**
14(a)		Consider a collection of literature survey made by a researcher in the form of a text document with respect to cloud and big data analytics. Using Hadoop and Map Reduce, write a program to count the occurrence of pre dominant key words.	**(14)**	**4**	**3**
		(OR)			
14(b)		Summarize the significances of MapReduce and discuss about Hadoop distributed file system architecture with neat diagram.	**14**	**4**	**3**
15(a)		Evaluate how the distributed synchronization is achieved using the key components of HBase and Zookeeper architectures.	**14**	**5**	**5**
		OR			
15(b)		Assess how the design goals of IBM Infosphere Big-Insights and streams are achieved with example code snippets.	**14**	**5**	**5**

PART- C (1x 10=10Marks)

(Q.No.16 is compulsory)

		Marks	**CO**	**RBT LEVEL**
16	Taking sentiment analysis as a case study, elaborate on the Real-time Sentiment Analysis Platform with an example.	**10**	**2**	**4**

****