Paper / Subject Code: 42172 / BIG DATA ANALYTICS

BE Sem VII (R-2019, C Scheme) "Computer" Dec'2023

28/12/2023

Time: 03 Hours **Marks: 80**

Note: 1. Question 1 is compulsory
2. Answer any three out of the remaining five questions.
3. Assume any suitable data wherever required and justify the same.

Q1 a) What is the basic difference between traditional RDBMS and Hadoop? [5]

b) What are the 3 V's of big data? Give two big data case studies indicating respective V's with justification. [5]

c) Explain how node failure is handled in Hadoop. [5]

d) List down all six constraints that must be satisfied for representing a stream by buckets using DGIM algorithm with examples. [5]

Q2 a) Describe the four ways by which big data problems are handled by NoSQL. [10]

b) Write a map reduce pseudo code to multiply two matrices. Apply map reduce working to perform following matrix multiplication. [10]

$$M = \begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{matrix} \quad X \quad V = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$$

Q3 a) Suppose the stream is S = {4, 2, 5 ,9, 1, 6, 3, 7}. Let hash functions h(x) = x + 6 mod 32 for some a and b, treat result as a 5-bit binary integer. Show how the Flajolet- Martin algorithm will estimate the number of distinct elements in this stream. [10]

b) i. Create a data frame from the following 4 vectors and demonstrate the output: [10]

```
emp_id = c (1:5)
emp_name = c("Rick","Dan","Michelle","Ryan","Gary")
start_date = c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11", "2015-
        03-27")
salary = c(60000, 45000, 75000, 84000, 20000)
```

ii. Display structure and summary of the above data frame.

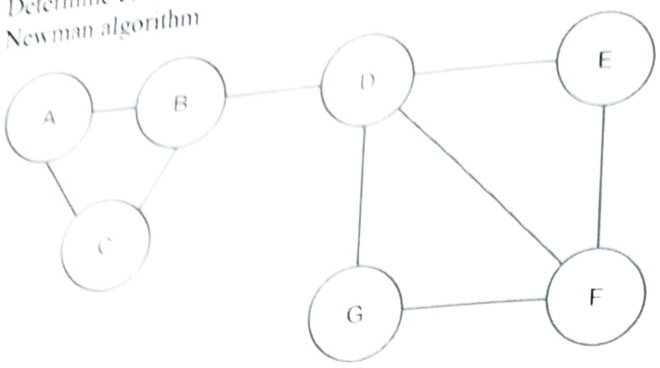
iii. Extract the emp_name and salary columns from the above data frame.

iv. Extract the employee details whose salary is less than or equal to 60000.

Q4 a) Explain Map Reduce execution pipeline with suitable example [10]

b) Explain DGIM algorithm for counting ones in a stream with example. [10]

Q5 a) Determine communities for the given social network graph using Girvan-Newman algorithm [10]



b) List and explain various functions that allow users to handle data in R workspace with appropriate examples. [10]

Q6 a) i. What are the advantages of using functions over scripts? [10]

ii. Suppose you have two datasets A and B.
Dataset A has the following data: 6 7 8 9.
Dataset B has the following data: 1 2 4 5.
Which function is used to combine the data from both datasets into dataset C.
Demonstrate the function with the input values and write the output.

b) How recommendation is done based on properties of the product? Explain with the help of an example. [10]

****