**ST JOSEPH’S UNIVERSITY, BENGALURU - 27**

Registration Number:

Date & Session:

**M.Sc. COMPUTER SCIENCE – III SEMESTER**

**SEMESTER EXAMINATION: OCTOBER 2023**

**(Examination conducted in November/December 2023)**

**CS 9122: DATA ANALYTICS with HADOOP**

**(For current batch students only)**

**Time: 2 Hours Max Marks: 50**

**This paper contains TWO printed pages and THREE parts.**

**NOTE: Students are allowed to use a scientific calculator**

**PART A**

Answer **ALL** the following questions.                   (2 x 5 = 10)

1. Differentiate between variability and variety in the characteristics of big data.
2. What is the goal of Support Vector Machine Algorithm?
3. Write the stopping criteria for merging or splitting the clusters.
4. Mention the functionality of the hive services given below.

 a. hwi b. jar

1. What does Hadoop YARN do?

**PART B**

Answer any **FIVE** of the following question.          (4 x 5 = 20)

1. Construct a MapReduce word count process to demonstrate the steps involved for the input

{Dog, Cat, Mouse, Dog, Dog, cat, Dog, Cat, Duck}.

1. Distinguish between Traditional Analytic Architecture with Modern In-Database Architecture.
2. Implement Apriori algorithm for the dataset given below and generate the association rules.

|  |  |
| --- | --- |
| Transactions  | Items |
| T1 | A,B,C |
| T2 | B,D,E |
| T3 | A,B,D |
| T4 | A,B,C,D |
| T5 | D,E |

1. Mention the object specifications associated with BFR algorithm. Explain.
2. Mention the two types of execution mode in Pig environment. Explain.
3. HIVE is called schema on read. Justify in comparison with schema on write approach.
4. a. What does the node and arc signify in the Bayesian Network? (2m)

 b. Discuss the types of time-series data. (2m)

**PART C**

Answer any **TWO** of the following:         (10 x 2 = 20)

1. a. How is memory mapping done in Park Chen Yu Algorithm? (4m)

b. Apply PCY algorithm on the following transaction to find the candidate sets. with minimum support = 2 (6m)

|  |  |
| --- | --- |
|  T\_ID | ITEMS |
| 1 | 1,3,4,5 |
| 2 | 2,3,4 |
| 3 | 1,2,3,5 |
| 4 | 2,5 |

1. a. What is linear regression? Why linear regression is important? (4m)

b. What is an analytic sandbox? Compare and contrast between internal and external sandbox. (6m)

1. a. What is Hadoop? Explain the core components of HDFS. (4m)

b. With a SEQUENCE DIAGRAM depict how job processing requests is performed in HDFS. (6m)