**ST. JOSEPH’S COLLEGE (AUTONOMOUS), BENGALURU -27**

Registration Number:

Date 2-12-2022 (1PM)

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

**SEMESTER EXAMINATION: OCTOBER 2022**

**(Examination conducted in December 2022)**

**CS 9122: DATA ANALYTICS with HADOOP**

**Time: 2.5 Hours Max Marks: 70**

**This paper contains FOUR printed pages and THREE parts**

**PART A**

Answer **ALL** the following question.                   (15 x 1 = 15)

1. Which of the following statements are true about using visualizations to display a

dataset?

I. Visualizations are visually appealing, but do not help the viewer understand relationships

that exist in the data

II. Visualizations like graphs, charts, or visualizations with pictures are useful for conveying

information, while tables just filled with text are not useful.

III. Patterns that exist in the data can be found more easily by using a visualization

a) I AND II

b) II AND III

c) I AND III

d) ONLY III

1. Which of the following is the simplest NoSQL database format for storing unstructured

data?

a) Key-value

b) Document

c) Wide column

d) All of the above

1. When the values of two variables move in the opposite directions, correlation is said to

be \_\_\_\_\_\_\_\_\_\_\_\_\_

a) Linear

b) Non-linear

c) Positive

d) Negative

1. What is perceptron?

a) a single layer feed-forward neural network with pre-processing

b) an auto-associative neural network

c) a double layer auto-associative neural network

d) a neural network that contains feedback

1. Identify the large scale clustering algorithm which uses a combination of partition based

and hierarchical algorithms

a) PCY Algorithm

b) SON Algorithm

c) CURE Algorithm

d) FM Algorithm

1. The algorithm that is also called as partition algorithm, which divides the database into

non-overlapping subsets, such that each partition fits into the main memory using two scans

of the database is

a) SON Algorithm

b) FM Algorithm

c) PCY Algorithm

d) Simple Randomized Algorithm

1. Consider an example of customers and their scores. Using one technique, the customers may be scored poorly but very well using another technique. By combining the intelligence from multiple models, a scoring algorithm becomes better in aggregate. Such a model is called as \_\_\_\_\_\_\_\_\_\_

a) Ensemble model

b) Commodity Model

c) Aggregate Model

d) Statistical Model

1. The characteristic of big data that specifies that the same data, which keeps changing

constantly and focuses on understanding and interpreting the correct meanings of raw data,

is \_\_\_\_\_\_\_\_\_\_

a) Variability

b) Variety

c) Value

d) Visualization

1. The results of a hive query can be stored as

a) Local File

b) HDFS file

c) Both the above

d) Cannot be stored

1. Which kind of keys(CONSTRAINTS) Hive can have?
2. Primary Keys
3. Foreign Keys
4. Unique Keys
5. None of the above
6. If the schema of the table does not match with the data types present in the file containing the table then Hive
7. Automatically drops the file
8. Automatically corrects the data
9. Reports Null values for mismatched data
10. Does not allow any query to run on the table
11. \_\_\_\_\_\_\_\_\_ is a framework for collecting and storing script-level statistics for Pig Latin.

a) Pig Stats

b) Statistics

c) Pig Statistics

d) None of the mentioned

1. Which of the following is true about MapReduce?

a) It provides the resource management

b) An open-source data warehouse system for querying and analyzing large datasets stored in Hadoop files

c) Data processing layer of Hadoop

1. Which of the following phases occur simultaneously in MapReduce.
2. Shuffle and Map
3. Shuffle and Sort
4. Reduce and Sort
5. Which of the following is a column-oriented database that runs on top of HDFS:
6. Hive
7. Sqoop
8. HBase
9. Flume

**PART B**

Answer the following questions.         (5 x 5 = 25)

1. With a diagrammatic representation, explain Model and Score Management.

OR

Differentiate traditional analytic architecture with modern in-database architecture.

1. What is the aim of Bayesian network? What does the node and arc signify in the network? Explain it with an example.

OR

Abbreviate SVM. Explain the two types of SVM with suitable scenarios.

1. Implement Apriori algorithm for the dataset given below and generate the association rules. Given the Minimum support is 2 and minimum confidence is 50%

|  |  |
| --- | --- |
| **TID** | **Itemset** |
| **T1** | **A, B** |
| **T2** | **B, D** |
| **T3** | **B, C** |
| **T4** | **A, B, D** |
| **T5** | **A, C** |
| **T6** | **B, C** |
| **T7** | **A, C** |
| **T8** | **A, B, C, E** |
| **T9** | **A, B, C** |

OR

Explain Park Chen Yu Algorithm? How is memory mapping done in PCY?

1. Explain Map task and reduce task using the mapper input given below.

Dear, Bear, River, Car, Car, River, Deer, Car, Bear

 OR

Compare Hadoop with RDBMS and Grid Computing

1. Why is HIVE used? Explain the architecture of HIVE.

OR

Write the structure of Pig Latin. Explain the four major categories of Pig functions.

**PART C**

Answer any **THREE** of the following questions.       (3 x 10 = 30)

1. What is regression? Explain simple linear regression with its cost function.
2. Explain K-means clustering algorithm for larger-dataset. Mention the object specifications associated with BFR algorithm.
3. a. Explain EDA. Mention the various methods of performing EDA. (6m)

b. Compare HiveQL with SQL. (4m)

1. a. What is Hadoop? Explain the core components of HDFS. (4m)
2. With a SEQUENCE DIAGRAM depict how JOB PROCESSING REQUESTS is performed. (6m)