Jahangirnagar University
Department of Computer Science and Engineering
Fourth Year Second Semester B.Sc. (Hons.) Final Examination -2022

Course Title: **Data Mining and Big Data Analysis**
Full Marks: 60

Course No: CSE-451
Time: 3 Hrs.

[Answer each of the following questions. Each question carries equal marks. Figures in the right margin indicate marks.]

1. Answer all questions:
   a) Define the terms: i) Data mining ii) sequence. 2
   b) Describe the architecture of typical data mining system with diagram. 2
   c) List down the objectives of data warehouse. 2
   d) Write down the primitives for defining data mining tasks. 2
   e) Describe the application of data mining methods in market analysis and management. 2
   f) Draw the high level architecture of Hadoop. Define the functions of the Master Node. 2

2. Answer Any Three out of Four questions:
   a) i) Explain Decision tree. 4
      ii) Describe Hunt's algorithm for classification
   b) Differentiate between clustering and classification. Organize 5 clusters by applying clustering technique to the following sample data set considering 5 initial random centroids using k-Means algorithm: 4

      {130, 50, 870000, 345, 7500, 8000, 205, 150000, 34, -55, 45, -150, 200}

   c) Explain 3 V's of Big Data. What is Big Data analytics? 4
   d) i. Illustrate the main features of NoSQL. Explain that NoSQL database can easily scale up and down. 4
      ii. Describe different types of NoSQL database with examples.

3. Answer Any Three out of Four questions:
   a) Draw the typical system architecture of a data mining system, and define the functions of its components. 4
   b) Sketch the three-tier data warehouse architecture and explain the functions of its different components. 4
   c) Explain the subject-oriented and non-volatile features of data warehouse. 4
   d) Show the frequent item set generation for the following data set given in Table 1 using a diagram based on the Apriori algorithm. Consider *minsup* (e.g., case 1: 70% and case 2: 85%) and *minconf* (e.g., 80% in both cases) such that *at least two items* are determined as frequent items. Also construct the association rules using these frequent items determined by computing the support and confidence of the items. You can update the values of support and confidence if at least two frequent items cannot be generated. 4

i) sharding ii) Dynamic schema

Table 1.

Sl. No.	*Item1*	*Item2*	*Item3*	*Item4*
1	p	q	r	
2	q	r	s	
3	q	r	s	
4	p	q	r	s
5	q	r	s	

4. **Answer Any Three out of Four questions:**

a) Explain Naïve Bayesian classification algorithm with an appropriate example. 4

b) In relation to data mining classification problems, 4

i. Illustrate the general approach for building a classification model.

ii. Criticize constructively on why we use Bayesian classifier instead of decision-tree based classification for some data mining problems with example.

c) i. Differentiate between OLTP and OLAP. 4

ii. Illustrate that "Can we extract all and only interesting patterns using data mining technique?"

d) Illustrate the File Write anatomy used in Hadoop Distributed File System with diagram. 4

5. **Answer Any Two out of Three questions:**

a) Write down the DMQL query for the following data mining problem. 6

As a marketing manager of an electronics shop, you need to characterize the *buying habits* of customers who purchase items priced at no less than TK 2000, with respect to the customer's age, the type of item purchased, the place in which the item was made, and the percentage of customers having that characteristic. In particular, you are only interested in purchases in Dhaka, and paid for with a "SC Bank" credit card for the current year. You need to define the attributes of the tables as mentioned below. The resulting descriptions can be viewed in the form of a table.

Tables: *item, customer, purchases*, and *items_sold*

Relevant attributes: *name* and *price* from the ***item*** table, and *income* and *age* from the ***customer*** table

b) Evaluate the Hadoop Distributed File System Architecture in designing Big-Data processing applications for large organizations using HDFS architecture. Also, describe the functions of each of its components. 6

c) Consider a student's fee payment system. The system has to keep records about students personal information, course registration information, total credit registered and semester payment information. 6

Answer the following questions.

i. Design the dimension tables and a fact table to create a data warehouse schema for student's fee payment system with defining their attributes using the *Star* model.

ii. Identify the *measures* required for the fact table and define the measures using the aggregate functions.

Jahangirnagar University

Department of Computer Science and Engineering

Tutorial Examination, 4th Year 2nd Semester B.Sc. (Hons.) 2021-2022

CSE-451: Data Mining and Big Data Analysis Marks: 30

1. Explain Classification. Draw a decision tree from a sample mammal classification data set
2. Describe Hunt's algorithm for classification. Explain its termination.
3. Define the functions used in the Decision tree induction algorithm.
4. Illustrate Rule-based classifier.
5. What is association analysis? Define support and confidence of an association rule.
6. What is frequent item? Show the candidate items and frequent item set generation process for the following data set. Consider *minsup* and *minconf* such that ***at least two items*** are determined as frequent items. Also define a rule using these frequent items.

Sl. No.	Item1	Item2	Item3	Item4
1	a	b	c	
2	b	c	d	
3	c	d		
4	a	b	c	d
5	b	c	d	

Jahangirnagar University

Department of Computer Science and Engineering

Tutorial Examination, 4th Year 2nd Semester B.Sc. (Hons.) 2021-2022

CSE-451: Data Mining and Big Data Analysis Marks: 30

1. Explain 3 V's of Big Data. What is Big Data analytics?
2. Illustrate the features of NoSQL. Explain in-database analytics.
3. What is Hadoop? Write down the features of Hadoop.
4. Describe the data processing framework of Hadoop.
5. Explain different types of NoSQL database with example.
6. What is HDFS? Explain HDFS organization and functioning.

Jahangirnagar University

Department of Computer Science and Engineering

Tutorial Examination, 4th Year 2nd Semester B.Sc. (Hons.) 2021-2022

CSE-451: Data Mining and Big Data Analysis Marks: $5 \times 6 = 30$

1. Explain the solution to the data explosion problem. What is meant by we are drowning in data but starving for knowledge.
2. Explain how KDD process generates knowledge for decision making using a diagram.
3. Can we extract all and only interesting patterns? Explain the solutions in brief.
4. Define the terms: Classification, decision tree, cluster analysis.
5. Differentiate between OLTP and OLAP. Explain any one OLAP operation.
6. Suppose as a marketing manager of ABC Super shop, you would like to characterize the buying habits of customers who purchase items priced at no less than TK 500 with respect to the customers age, the type of item purchased, and the place in which the item was made. For each characteristic discovered, you would like to know the percentage of customers having that characteristic. In particular, you are only interested in purchases in Dhaka, and paid for with an "GL Bank" credit card. You would like to view the resulting descriptions in the form of a table.

   Define the database tables and write down the DMQL expression for the above data mining problem.