University of Ruhuna - Faculty of Technology

Bachelor of Information & Communication Technology Honours Degree

Level 3 (Semester II) Examination – October/November-2022

Academic Year 2020/2021

Time: 2.5 hours

Course Unit: ICT3222 Data Mining and Business Intelligence

Answer all **four (04)** questions

1)

a) A data warehouse system is the main solution for a complete data set which enables fast and accurate data analysis. State the main problem of data analysis that is solved by data warehouse systems? Write **two (02)** reasons for that problem. (20 Marks)

b) *"A DW is a subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making."* Explain what is meant by Time Variant by using a suitable example. (20 Marks)

c) Assume that, you are working as a senior Business Analyst in a reputed software company. Your organization has received a requirement to develop a data warehouse for a multinational company which is maintaining palm tree plantations and palm oil business. They want to keep records on their harvest which is collected from various states. They are facing some practical issues when collecting harvest from various locations. In some cases, they do not receive the palm fruit harvest on the same day. Organization owns various mills in different areas in country and oil is taken from these mills. Client wants to keep their mill related information and their trading related information in addition to harvest related information within the Data warehouse for their decision-making purposes. Large amount of data is exchanging within a day related to their operations on harvesting, milling and trading. Data warehouse must be available for multiple users, who are involved in decision making process and should be accessible 24*7 hours. Client expects to have ad-hoc reports by getting the use of Data warehouse.

i) What is the suitable approach (Query Driven/Update Driven) for develop above mentioned Data warehouse? Justify your answer by providing **two (02)** reasons. (20 Marks)

ii) Write **five (05)** Online Analytical Processing (OLAP) rules that can be considered in designing Data warehouse according to the nature of the multinational company. (20 Marks)

iii) Development team has decided to follow the Multidimensional Online Analytical Processing (MOLAP) Model on Data warehouse. Illustrate it using a suitable diagram. (20 Marks)

2) You are working as a data design specialist on the data warehouse project team for a Grocery Shop which is functioning as a supermarket. Organization maintains a wide range of products and they have number of stores in the same location to store their products. Supermarket owns a large customer base and qualified staff is there to perform their functions within the outlet. They keep each transaction records on sales daily basis to analyze their sales information. Unit Price, Quantity Sold and Amount Sold will be stored for future analysis. Top management is going to make necessary actions based on data which is going to store within the newly designed Data warehouse.

a) What are the **three (03)** basic architectures that can be used to construct a Data warehouse? Explain one of them using a suitable diagram. (20 Marks)

b) Classify facts, dimensions and measures that can be included in designed Data warehouse according to the given requirements. (10 Marks)

c) Development team have been selected star schema instead of fact constellation schema to design the conceptual model of proposed Data warehouse. Do you agree with this statement? Justify your answer. (20 Marks)

d) Design a star schema/ fact constellation schema for the proposed Data warehouse. (30 Marks)

e) Development team has to pay their attention on security issues when they design the physical design for the Data warehouse. Explain the importance of security in Data warehouse using a suitable example. (20 Marks)

3)

a) Explain the importance of data cleaning in data warehousing using an example. (20 Marks)

b) A reputed super market maintains their customer information within their data warehouse. Table Q3b shows some collected data records. (30 Marks)

Customer_ID	City	Profession
C1	Matara	Doctor
C2	Galle	Technical Officer
C3	Kandy	Nurse
C4	Colombo	Engineer
C5	Kandy	Doctor
C6	Matara	Teacher
C7	Matara	Engineer

Table Q3b

i) Construct a bitmap index on City and Profession. (20 Marks)

ii) Write **two (02)** issues that can be faced when apply bitmap index in high cardinality. (10 Marks)

c) Yearly crop harvest of three provinces of Sri Lanka is represented by the below ordered set of data. The identified dimensions of the data cube are crop type, year and province. The measure is the weight in Metric tons. (30 Marks)

Rice, 2018, Western, 530
Tea, 2020, Western, 240
Coconut, 2021, Western, 430
Coconut, 2019, Western, 420
Rubber, 2018, Western, 510
Rice, 2019, Western, 580
Tea, 2020, Central, 650
Coconut, 2018, Central, 480
Rubber, 2019, Central, 320
Rice, 2021, Central, 430
Rice, 2019, Central, 340
Tea, 2021, Central, 720
Rice, 2019, North Central, 780,
Tea, 2020, North Central, 280
Rice, 2020, North Central, 690
Rice, 2021, North Central, 540
Coconut, 2021, North Central, 220
Rubber, 2018, North Central, 340
Rubber, 2019, North Central, 280
Tea, 2018, North Central, 260

i) Calculate all the necessary aggregated cuboids and clearly show the data cube representation of above information.

ii) Calculate the following aggregate cuboid values.

(a) (*, *, Central)
(b) (Rice, *, North Central)
(c) (*, 2019, *)
(d) (*, *, *)

d) Sketch the diagram of knowledge discovery process (KDD) and briefly describe the stages of the process. (20 Marks)

4)

a) Consider the following transactions show in the following Table Q4a.

T_ID	Transaction_Date	Item_List
T01	15/09/2022	c, a, f, g, b, k
T02	15/09/2022	a, b
T03	15/09/2022	a, c, f, g
T04	15/09/2022	b, c, f, a
T05	15/09/2022	a, k
T06	15/09/2022	a, c, f, i
T07	15/09/2022	f, k
T08	15/09/2022	f, c, b, g, a
T09	15/09/2022	a, b
T10	15/09/2022	a, b, c

Table Q4a

i) "Those who purchase item **a** and item **b** also purchase item **f.**" Show this association as a rule. (5 Marks)

ii) Find the support and confidence for the associations given below. (15 Marks)

(a) a, b → f

(b) a, b, f → c

(c) a → b

iii) Assuming a minimum level of support min_sup = 5 and a minimum level of confidence min_conf = 80%: Find the most frequent itemset using the Apriori algorithm. For each iteration show the candidate and acceptable frequent itemsets. (25 marks)

b) You are going to analyze the kidney failure cases from a randomly selected set of patients. Table Q4b shows the data collected from the individual patients. (45 Marks)

Age	fbs	hbp	Heart Disease	Kidney Failure (Class Label)
<=20	High	Yes	No	No
21…40	Low	Yes	No	No
>40	Medium	No	Yes	Yes
21…40	Medium	Yes	No	No
<=20	Low	No	Yes	Yes
>40	High	Yes	Yes	Yes
<=20	High	No	Yes	No
>40	Medium	No	No	Yes
21…40	Low	Yes	Yes	Yes
>40	High	Yes	No	Yes
21…40	High	Yes	Yes	Yes
>40	Low	No	Yes	No
21…40	Medium	Yes	No	Yes
>40	High	No	No	Yes

Table Q4b

Details of the attributes of the Table Q4b are given below,

fbs: Fasting Blood Sugar
hbp: High Blood Pressure
Heart Disease: Suffering from Heart disease or Not

i) Calculate the overall Entropy before splitting

ii) Calculate the overall Entropy after splitting each attribute

iii) At which attribute should the decision tree split first? Explain Why?

c) Compare and contrast the differences between **Supervised Learning** and **Unsupervised learning.** (10 Marks)

Log Table

$\log_2(x)$										
x	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0	inf.	-6.64	-5.64	-5.06	-4.64	-4.32	-4.06	-3.84	-3.64	-3.47
0.1	-3.32	-3.18	-3.06	-2.94	-2.84	-2.74	-2.64	-2.56	-2.47	-2.4
0.2	-2.32	-2.25	-2.18	-2.12	-2.06	-2	-1.94	-1.89	-1.84	-1.79
0.3	-1.74	-1.69	-1.64	-1.6	-1.56	-1.51	-1.47	-1.43	-1.4	-1.36
0.4	-1.32	-1.29	-1.25	-1.22	-1.18	-1.15	-1.12	-1.09	-1.06	-1.03
0.5	-1	-0.97	-0.94	-0.92	-0.89	-0.86	-0.84	-0.81	-0.79	-0.76
0.6	-0.74	-0.71	-0.69	-0.67	-0.64	-0.62	-0.6	-0.58	-0.56	-0.54
0.7	-0.51	-0.49	-0.47	-0.45	-0.43	-0.42	-0.4	-0.38	-0.36	-0.34
0.8	-0.32	-0.3	-0.29	-0.27	-0.25	-0.23	-0.22	-0.2	-0.18	-0.17
0.9	-0.15	-0.14	-0.12	-0.1	-0.09	-0.07	-0.06	-0.04	-0.03	-0.01

.... **End of the Paper**............