EIE3112 Test 2 (2021/22 sem 2)

Name: Sze Kin Sang

Student ID: 19062606D

# Instructions:

* You may write your answer on a piece of paper or use text editor to type your answer below. (You may need to consider that there may be technical problem encountered using computer or text editor. This cannot be the reason for exemption from late submission penalty.)
* Then upload your answer to Blackboard before the submission deadline.
* You have approximately 85 minutes to work on the answer and another 5 minutes for uploading the answer. Please take your time.
* Late submission will be subject to mark deduction. No score will be given to the submission uploaded 10 minutes after the submission deadline.
* If you encounter any problem during the test, please contact Dr Pauli Lai at +852-27666217 or pauli.lai@polyu.edu.hk.
* Please pay attention to the Blackboard announcement or email for any updated information on the question (if any).
* WARNING: If you are found sharing or copying the question and answer with or from your classmate, both you and your classmate will score 0 mark due to plagiarism.

# Part 1: Data Warehousing (40 marks)

Consider the following relational database for Grand Travel Airlines.

Grand Travel Airlines has to keep track of its flight and airplane history. A flight is uniquely identified by *the combination of* a flight number and a departure date. Every passenger who has flown on Grand Travel has a unique passenger number. For a particular passenger who has taken a particular flight, the company wants to keep track of the fare that he/she paid for it and the date that he/she made the reservation for it. Clearly, a passenger may have taken many flights (he/she must have taken at least one to be in the database) and every flight has had many passengers on it.

A pilot is identified by a unique pilot number. A flight on a particular date has exactly one pilot. Each pilot has typically flown many flights, but a pilot may be new to the company, is in training, and has not flown any flights, yet. Each airplane has a unique serial number. A flight on a particular date used one airplane. Each airplane has flown on many flights and dates, but a new airplane may not have been used at all, yet.

|  |  |  |  |
| --- | --- | --- | --- |
| Pilot  Number | Pilot  Name | Date of  Birth | Date of Hire |

PILOT Relation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Flight  Number | Departure  Date | Departure Time | Arrival  Time | Pilot  Number | Airplane  Number |

FLIGHT Relation

|  |  |  |  |
| --- | --- | --- | --- |
| Passenger  Number | Passenger  Name | Address | Telephone  Number |

PASSENGER Relation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Flight  Number | Departure  Date | Passenger  Number | Fare | Reservation  Date |

RESERVATION Relation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Airplane  Number | Model | Passenger  Capacity | Year  Built | Manufacturer |

AIRPLANE Relation

(a) Draw a star schema for a data warehouse for the Grand Travel Airlines business environment. The fact table will be RESERVATION, which represents a particular passenger reservation on a particular flight. Be sure to keep track of the fare that the passenger paid for the flight and the date of the reservation. The dimension tables are PASSENGER, FLIGHT, and TIME PERIOD. The Reservation Date will be represented by the Time Period dimension. The Departure Time in the FLIGHT dimension table will show the actual dates of the flights. You should include attributes for all tables, indicate primary key attributes using "PK" and foreign key attributes using "FK".

[11 marks]

|  |
| --- |
|  |

(b) Describe one OLAP use of this data warehouse. (Question that the management will raise for business analysis.)

[3 marks]

|  |
| --- |
|  |

(c) Describe one data mining use of this data warehouse. (To discover some hidden patterns/associations from the data.)

[3 marks]

|  |
| --- |
| Find out if there is a relationship between the fare and reservation date, so discount can be made before travelling season. |

(d) Draw the OLAP cube for the query below.

[10 marks]

SELECT Time, Region, Department, sum(Profit) AS Profit

FROM sales

GROUP BY Time, Region, Department;

一張含有 桌 的圖片

自動產生的描述

|  |
| --- |
| West  East  Central  VideoSales  VideoRental  2001  2000 |

(e) Circle any one cube from the above OLAP cube and give the values of Time, Region, Department and Profit for your encircled cube.

[4 marks]

|  |
| --- |
| Time 2000  Region West  Department VideoSales  Profit = 86000 |

(f) The following query lists the number of employees for each combination of gender-city, total number of employees per gender, and total number of employees in the entire table.

**SELECT** GENDER, CITY, COUNT(\*)

**FROM** EMPLOYEES

GROUP BY GENDER, CITY WITH ROLLUP

Please fill in the blanks in the following result table:

[9 marks]

一張含有 桌 的圖片

自動產生的描述

F

M

NULL

NULL

NULL

NULL

14

9

5

# Part 2: Data Mining (17 marks)

Fig. 1 shows a data set of driving or not based on the weather data.

一張含有 桌 的圖片

自動產生的描述

Fig. 1

(a) Given that Drive is an output class, provide one example of classification rule.

[3 marks]

|  |
| --- |
| If the outlook is Sunny and purpose is Leisure then no drive. |

(b) Given that Drive is an output class, provide one example of association rule among attributes.

[3 marks]

|  |
| --- |
|  |

(c) Can linear regression be applied to predict the class for this data set? Why?

[3 marks]

|  |
| --- |
| No it is because rather than values in different attributes, we have different classes. Therefore, no linear regression can be done. |

(d) Assume that each instance in Fig. 1 is a transaction. Calculate the support and confidence of the following 2 Associative-Rules:

[4 marks]

(i) {Destination=Near, Work=Leisure } -> {Drive=Yes}

|  |
| --- |
| #{Destination=Near, Work=Leisure, Drive=Yes}/#{Destination=Near, Work=Leisure }  = 2/9 / 4/9 = 50% |

(ii){Destination=Far} -> {Drive=Yes}

|  |
| --- |
| #{Destination=Far, Drive=Yes}/#{Destination=Far}  = 3/9 / 4/9 = 75% |

(e) What is the difference between classification and clustering? Give one machine learning algorithm for classification and one algorithm for clustering.

[4 marks]

|  |
| --- |
| Classification: linear support vector machines  Clustering: K-means  Classification is done by analyzing two column while Clustering is done by analyzing one column. |

# Part 3: Big Data (25 marks)

(a) Give an example (not definition) to illustrate each V of the 5 V's of Big Data.

[5 marks]

|  |
| --- |
| Velocity: online game update the current database in real time  Volume: big data may have huge TB size database  Variety: video and audio can be also included in big data  Veracity: Backup are required to ensure no data loss due to accident  Value: big data may contain personal information which is worthy |

(b) What is the difference between MapReduce and Hadoop?

[5 marks]

|  |
| --- |
|  |

(c) Describe how MapReduce provide automatic parallelization and distribution.

[5 marks]

|  |
| --- |
| MapReduce will first split and map the big file into multiple smaller file, after finding the aggregate values that according to the keys, it can compare and reduce the results. |

(d) Describe the data locality characteristic of MapReduce and its benefits.

[5 marks]

|  |
| --- |
| Data locality means MapReduce allow to do the computation near the data location , rather than send all the data along the node. As a result, the calculation can be done quickly. |

(e) Describe how Hadoop provides fault tolerance.

[5 marks]

|  |
| --- |
| Fault tolerance allow Hadoop to restart a failed task only, but not the entire job. |

# Part 4: Database Security (18 marks)

Fig. 2 shows the HTML code of a web application’s login form.

一張含有 文字 的圖片

自動產生的描述

Fig. 2

A dynamic SQL statement is generated at run time using parameters email and password from the web form. Suppose the statement at the backend for checking user ID is as follows:  
SELECT \* FROM users WHERE email = '$email' AND password = md5('$password');

The $email will be replaced by the text input to the email field while the $password will be replaced by the text input to the password field. The function md5() is a hash function to encrypt the password.

(a) Give an example of using SQL injection technique in the email field so that any password can be used to log in to the web application.

[5 marks]

|  |
| --- |
| By using `# on the email input, the password input can be skipped, so any password can be used. |

(b) Give an example of using SQL injection technique in the password field so that any email address can be used to log in to the web application.

[5 marks]

|  |
| --- |
| By using 1=1 on the password input, any email can be used. |

(c) Discuss the methods of preventing SQL injection and what are their limitations.

[8 marks]

|  |
| --- |
| Allow-list Input Validation |