# THE HONG KONG POLYTECHNIC UNIVERSITY

## DEPARTMENT OF COMPUTING

## EXAMINATION

---

Course : BSc Scheme-61031, BSc G-34014

Subject : COMP417 Data Warehousing & Data Mining Tech. In Business & Commerce

Group : 1011, 1111

Session : 2010 / 2011 Semester I

Date : 11 December 2010 Time : 14:00-17:00

Time Allowed: 3 Hours Subject Lecturer: Keith Chan

---

This question paper has 11 pages (cover included).

---

### Instructions to Candidates:

Answer **ALL** questions.

Note:

1. This is an OPEN-BOOK examination.
2. Students are allowed to use a standard non-programmable calculator.

**Do not turn this page until you are told to do so!**

**Section A (Answer ALL questions):**

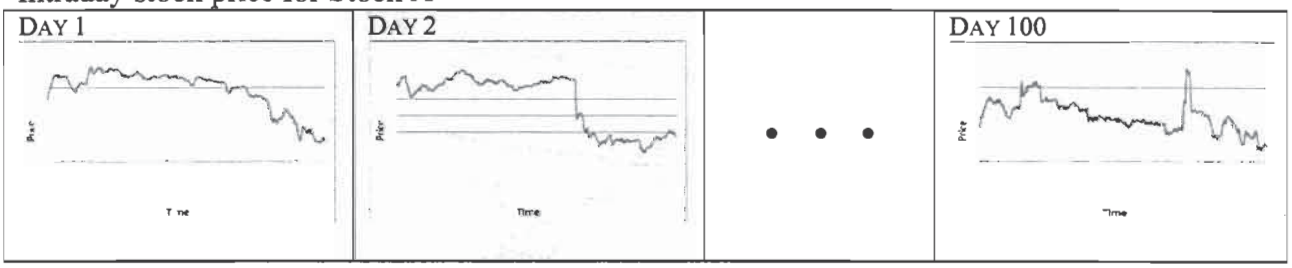
***Instructions (please read):*** *Answer questions in this section based on the following case descriptions. Marks will only be given to answers that provide sufficient details to allow markers to understand how data mining algorithms are used. Please state clearly all assumptions that you have made in answering any question.*

***Case Descriptions:***

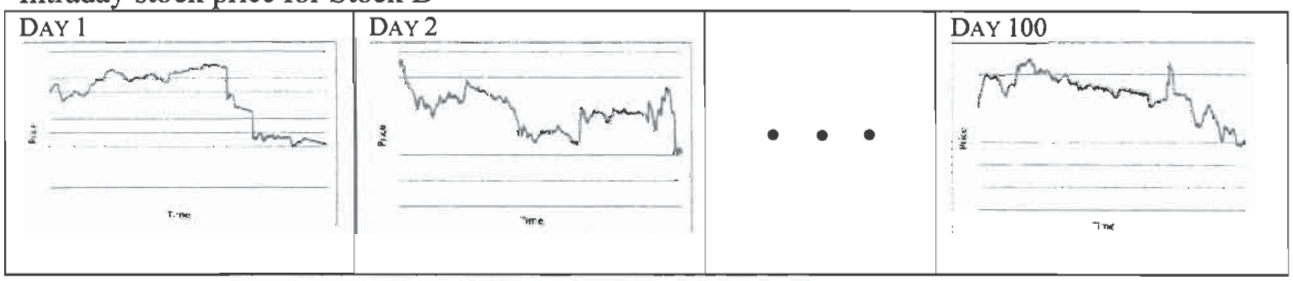
As a management trainee in an investment bank, John is given the responsibility to manage a portfolio of 20 stocks to maximize *ROI* (*Return On Investment*). Having taken COMP417, John believes that he can use the data mining techniques he learnt to help him make better investment decisions.

For the purpose of data mining, John has obtained the intra-day stock price data for the last 100 trading days for all the 20 stocks in his portfolio. For example, for Stock A and B, he has obtained 100 data sets corresponding to the 100 trading days as follows:

Intraday stock price for Stock A




Intraday stock price for Stock B




***Questions:***

1. After studying these charts for some time, John is under the impression that some of the stocks in his portfolio have very similar price movements (12 Marks).
   (a) Explain how John could use data mining techniques to find out whether or not he is correct and if he is correct, explain how he can use these techniques to identify stocks that have similar price movements. In answering this question, please be sure to:
      i. Explain your choice of data mining algorithm(s) for the task. If there are more than one alternative, discuss why you prefer one data mining technique to the other. If you have to use more than one algorithm, please describe the sequence of applications of these algorithms. Please give examples whenever necessary.
      ii. For each data mining algorithm you choose for i. above, give details as to how you would prepare input data for the algorithm. Describe all necessary data

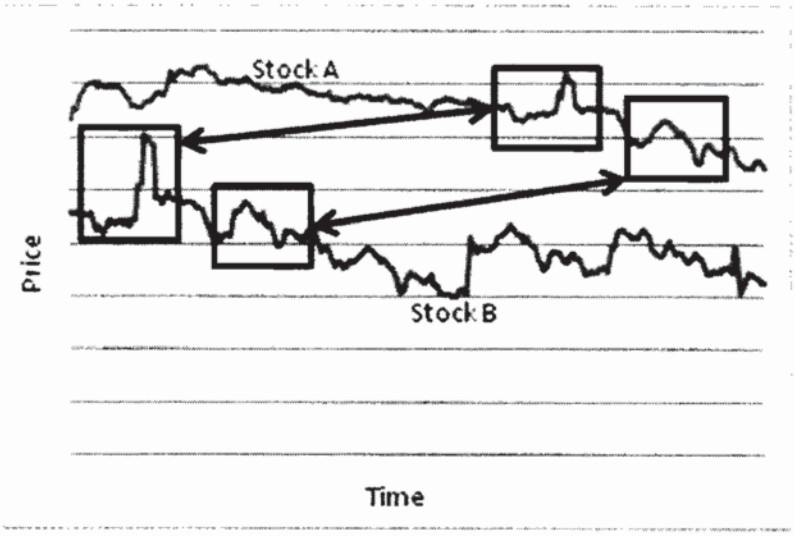
pre-processing or transformation that you need to perform before the data mining algorithms are used.

iii. Give examples of the output rules or patterns that you expect to obtain using the data mining algorithm(s) you propose in i. and the data in ii.

(b) How should John find out if the patterns or rules discovered by the data mining algorithms are useful?

(c) How may John's investment strategies be affected if stocks that exhibit similar price movements can indeed be identified?

2. By looking at the intra-day stock price charts, John discovered that the price of some stocks seem to follow that of some others at a certain time-lag. He discovered, for example, that the price movements of Stock B seem to follow that of Stock A except that they only move 2 to 3 hours (i.e. a time-lag of 2 to 3 hours) later than A (See example below) (12 Marks).




(a) Given John's observation, can you suggest to him what data mining techniques he can use to identify stocks whose stock price follows that of the others at a certain time-lag. In answering this question, please be sure to:

i. Explain your choice of data mining algorithm(s) for the task. If there are more than one alternative, discuss why you prefer one data mining technique to the other. If you have to use more than one algorithm, please describe the sequence of applications of these algorithms. Please give examples whenever necessary.

ii. For each data mining algorithm you choose for i. above, give details as to how you would prepare input data for the algorithm. Describe all necessary data pre-processing or transformation that you need to perform before the data mining algorithms are used.

iii. Give examples of the output rules or patterns that you expect to obtain using the data mining algorithm(s) you propose in i. and the data in ii.

(b) How should John find out if the patterns or rules discovered by the data mining algorithms are useful?

(c) How may this affect John's investment strategies if stocks whose prices move with each other at different time-lags can be identified?

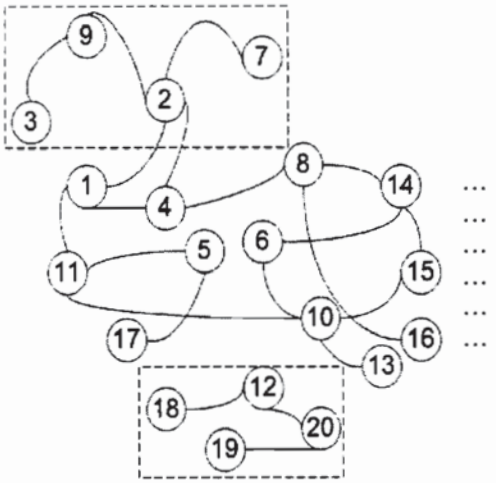
3. Among the 20 stocks in John's portfolio, 7 of them are bank stocks, 3 of them are insurance company stocks, 3 are telecom company stocks, 3 are utility company stocks, 2 are transportation company stocks and 2 are petroleum company stocks. Since both banks and insurance companies are *financial institutions*, John expects the price movements of bank and insurance company stocks to be similar to each other and different from the rest of the others (12 Marks).
   (a) Explain how John can use data mining techniques to find out if this is indeed the case. In answering this question, please be sure to:
      i. Explain your choice of data mining algorithm(s) for the task. If there are more than one alternative, discuss why you prefer one data mining technique to the other. If you have to use more than one algorithm, please describe the sequence of applications of these algorithms. Please give examples whenever necessary.
      ii. For each data mining algorithm you choose for i. above, give details as to how you would prepare input data for the algorithm. Describe all necessary data pre-processing or transformation that you need to perform before the data mining algorithms are used.
      iii. Give examples of the output rules or patterns that you expect to obtain using the data mining algorithm(s) you propose in i. and the data in ii.
   (b) How should John find out if the patterns or rules discovered are useful?
   (c) How may this affect John's investment strategies if bank stocks and insurance company stocks are found to have similar price movements and are different from those of the others?

4. A very experienced stock trader told John that the price movements of some stocks in his portfolio were *significantly correlated* in the sense that they usually move together in the same (or opposite) directions. If two correlated stocks are represented as two vertices in a graph, then they can be connected with an edge. If all significantly correlated stocks in John's portfolio are identified, a *correlation pattern* graph that can capture all such relationship can be drawn for each month of a year. According to the stock trader, the correlation patterns for the summer months of June, July, August, are different from that of the winter months of November, December and January (see graphs below) (16 Marks).
   (a) To find out if the stock trader's observations are correct, John obtained the intra-day stock price data for every trading day for the last 10 years for each of the 20 stocks in his portfolio. Explain how John can use data mining algorithm(s) to construct the correlation pattern graphs for each month from June, 2001 to December, 2010. In answering this question, please be sure to:
      i. Explain your choice of data mining algorithm(s) for the task. If there are more than one alternative, discuss why you prefer one data mining technique to the other. If you have to use more than one algorithm, please describe the

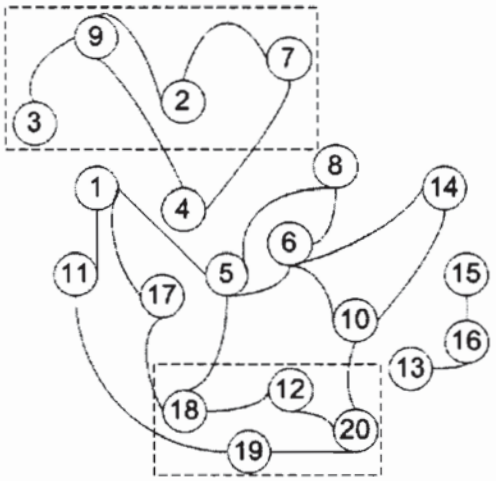
sequence of applications of these algorithms. Please give examples whenever necessary.

ii. For each data mining algorithm you choose for i. above, give details as to how you would prepare input data for the algorithm. Describe all necessary data pre-processing or transformation that you need to perform before the data mining algorithms are used.

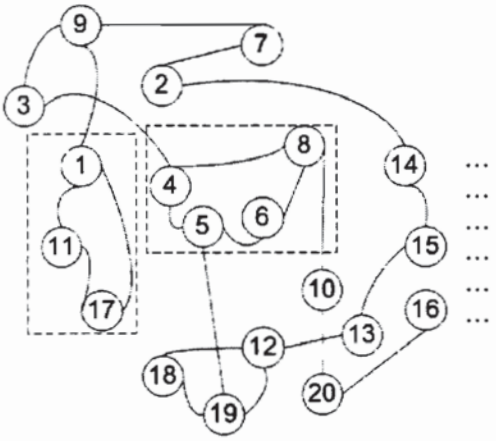
iii. Give examples of the output rules or patterns that you expect to obtain using the data mining algorithm(s) you propose in i. and the data in ii.



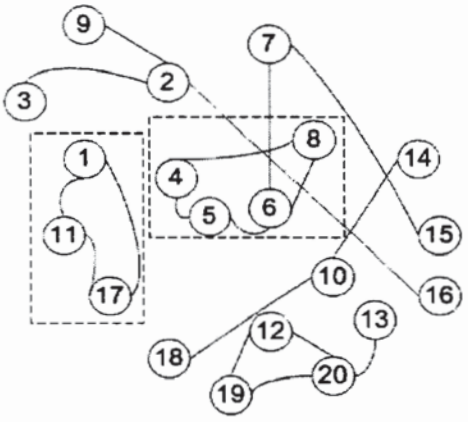
CORRELATION PATTERNS IN JUNE 2001



CORRELATION PATTERNS IN JUNE 2010



CORRELATION PATTERNS IN DECEMBER 2001



CORRELATION PATTERNS IN NOVEMBER 2010

(b) After the construction of the correlation pattern graphs, explain how John should use data mining to discover if the correlation patterns in the summer months are indeed different from that of the winter months. In answering this question, please be sure to:

i. Explain your choice of data mining algorithm(s) for the task. If there are more than one alternative, discuss why you prefer one data mining technique to the other. If you have to use more than one algorithm, please describe the sequence of applications of these algorithms. Please give examples whenever necessary.

ii. For each data mining algorithm you choose for i. above, give details as to how you would prepare input data for the algorithm. Describe all necessary data pre-processing or transformation that you need to perform before the data mining algorithms are used.

iii. Give examples of the output rules or patterns that you expect to obtain using the data mining algorithm(s) you propose in i. and the data in ii.

(c) How should John find out if the patterns or rules discovered are useful?

(d) How may this affect John's investment strategies if such correlation patterns can be identified and if there are indeed differences between the correlation patterns of the winter and summer months?

5. Of the 20 stocks in John's portfolio, 3 of them, companies A, B and C, are telecom companies. As they all belong to the same industry, John expected their stock price movements to be inter-related. By looking carefully at the intraday data, John is under the impression that the daily closing price of the stock of company B is dependent not only on its previous daily closing price but also on that of the other two companies, A and C. In order to find out if it is indeed the case, John obtained the daily closing data of 9 consecutive trading days for each of these companies as follows. (18 Marks)

| Day | Company A | Company B | Company C |
|---|---|---|---|
| 1 | $10.2 | $5.1 | $2.8 |
| 2 | $10.3 | $5.6 | $2.4 |
| 3 | $10.8 | $5.7 | $2.3 |
| 4 | $10.1 | $5.8 | $2.2 |
| 5 | $10.7 | $5.7 | $2.1 |
| 6 | $10.9 | $6.0 | $2.7 |
| 7 | $11.0 | $5.5 | $2.4 |
| 8 | $10.4 | $5.6 | $2.8 |
| 9 | $10.5 | $5.2 | $2.7 |

He would like to know if it is possible for the price movement of B to be predicted by that of companies A, B and C the day before.

(a) To help John determine if it is indeed the case, construct a decision tree that can allow you to predict the price movement of B. Please show your work in details.

(b) Can you use the Apriori Algorithm to predict the price movement of B? If so, please show your work in details.

(c) John believes that the price movements of B or any other stock can be determined more accurately if he uses the concepts of fuzzy set in data mining. Do you agree?

(d) In order to save time coming up with a fuzzy membership function for each stock in his portfolio, John is thinking of using only one single membership function for all the stocks in his portfolio. Is this possible? If so, can you come up with such a membership function?

**Section B (Answer ALL questions):**

6. You are given a set of customer data which had been classified according to the size of the loan (LOAN LEVEL) that they were previously approved (see table below). (16 Marks)

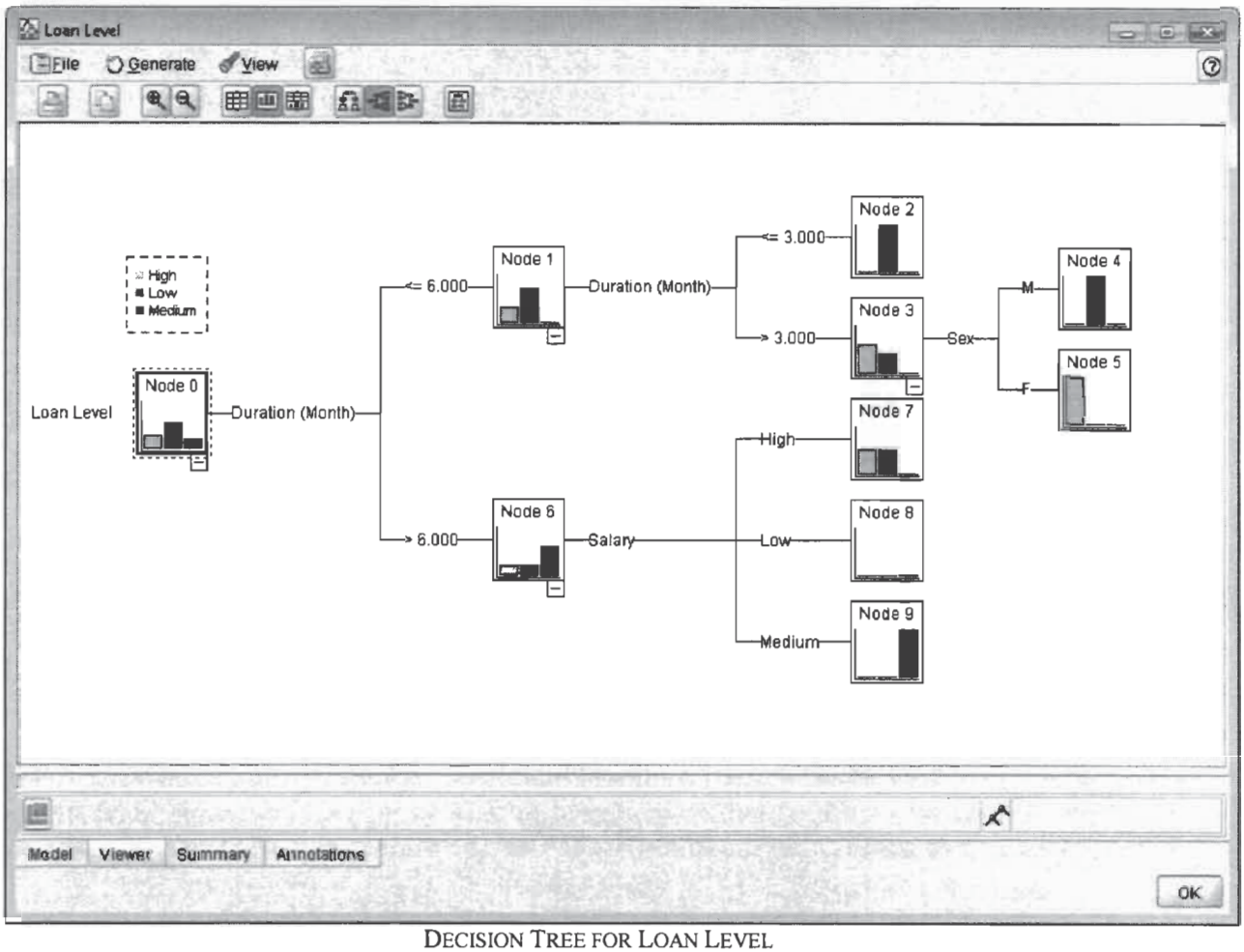
| CUSTOMER | SALARY | AGE | DURATION (MONTH) | SEX | LOAN LEVEL |
|---|---|---|---|---|---|
| c01 | LOW | OLD | 3 | ? | LOW |
| c02 | ? | YOUNG | 12 | M | MEDIUM |
| c03 | MEDIUM | OLD | 6 | M | LOW |
| c04 | MEDIUM | YOUNG | 6 | F | ? |
| c05 | LOW | OLD | 3 | M | LOW |
| c06 | MEDIUM | YOUNG | 12 | ? | MEDIUM |
| c07 | HIGH | ? | 12 | F | LOW |
| c08 | MEDIUM | YOUNG | ? | F | LOW |
| c09 | MEDIUM | OLD | 6 | F | HIGH |
| c10 | LOW | OLD | 3 | F | ? |
| c11 | HIGH | YOUNG | ? | M | HIGH |
| c12 | MEDIUM | OLD | 3 | F | LOW |
| c13 | ? | OLD | 6 | F | HIGH |
| c14 | LOW | OLD | 6 | M | LOW |
| c15 | MEDIUM | ? | 12 | M | MEDIUM |

LOAN RECORDS

(a) In order to determine if there is any differences in the customers who are approved different **LOAN LEVELS**, IBM's PASW was used to construct a decision tree for this purpose and this decision tree is given in the next figure. Show the data streams that you need to prepare to generate the decision tree in that figure. To help you prepare for the data streams, a list of nodes that you may use is given in the next page.

(b) Discuss the different approaches that you can take to handle missing values (shown in the above table as "?") when using ID3 or other decision tree construction algorithm.

(c) The missing values in the table above can be determined using data mining algorithms. Explain how they can be used to determine, say the **SALARY** of customer, c13. Please show the data streams that you need to prepare for this purpose. Again, to prepare for the data streams, please use the nodes listed in the next page.

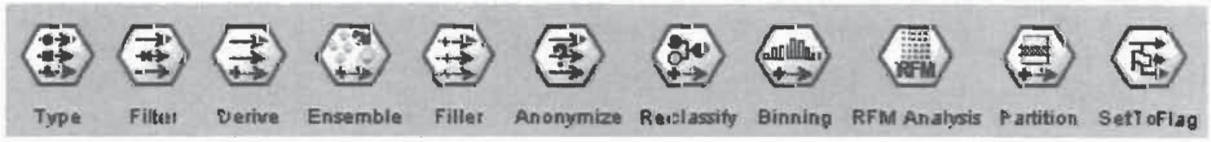



DECISION TREE FOR LOAN LEVEL

SOURCES


OUTPUT

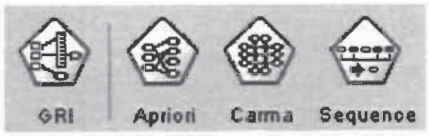

FIELD OPS




RECORD OPS


GRAPHS


MODELING – ASSOCIATION

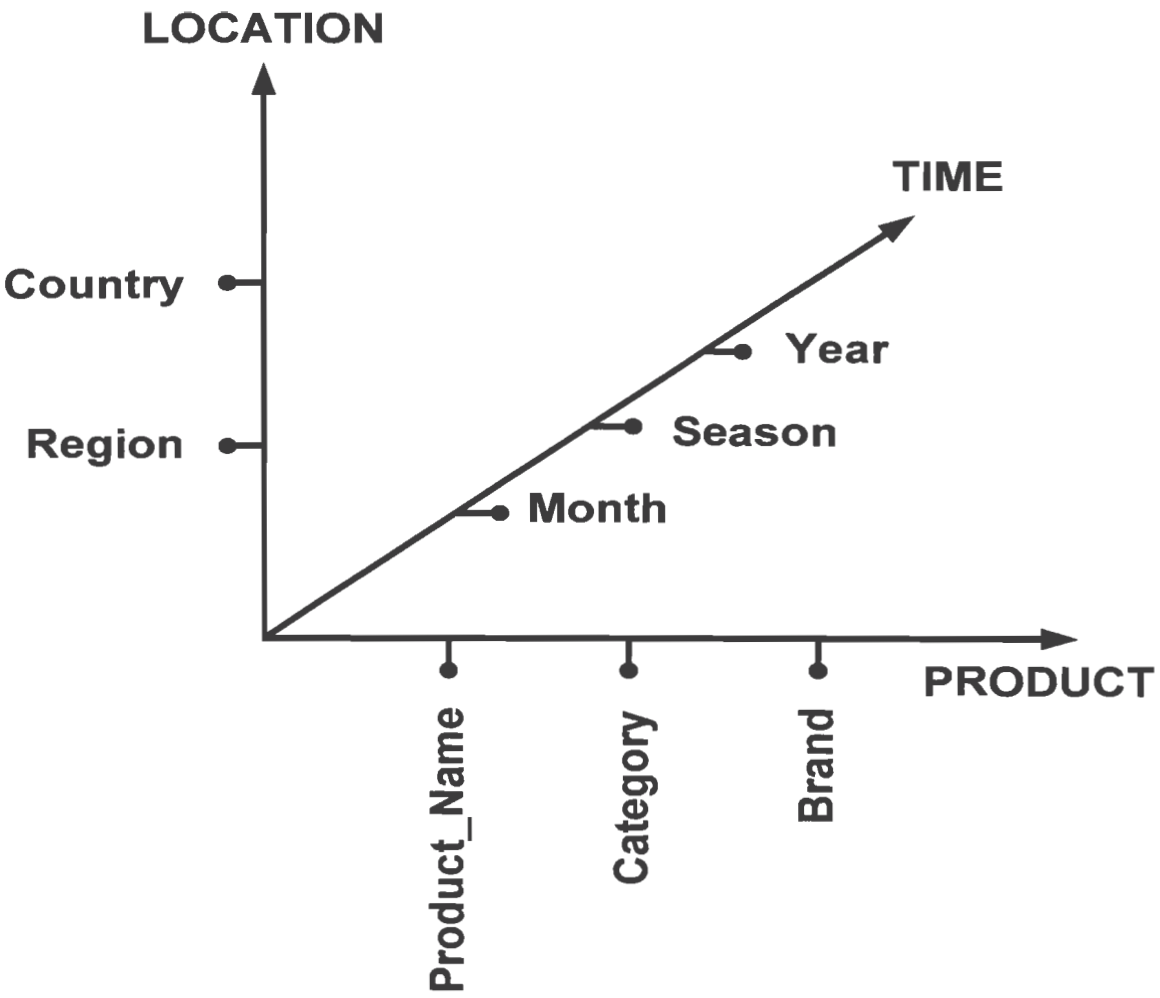



MODELING – CLASSIFICATION


MODELING – SEGMENTATION


7. A company has designed a data cube consisting of three dimensions, LOCATION, TIME, and PRODUCT to support executive decision making. Some sample data stored under such a data cube is given in the following table. Based on this design and the sample data shown in the table, show results after the following OLAP operations are performed: (7 Marks)
   - ROLL UP FROM CITY TO COUNTRY ON LOCATION [PRODUCT_NAME, COUNTRY, *]
   - DICE ON LOCATION AT COUNTRY = "US" AND ON PRODUCT AT BRAND = "SUNY"
   - ROLL UP TO ALL ON LOCATION
   - ROLL UP TO ALL ON TIME [PRODUCT_NAME, COUNTRY, *]
   - DRILL DOWN FROM BRAND TO CATEGORY ON PRODUCT
   - ROLL UP FROM PRODUCT_NAME TO BRAND [BRAND, COUNTRY, *]
   - SLICE ON LOCATION AT COUNTRY = "US"
   - SLICE ON PRODUCT AT BRAND = "SUNY"
   - PIVOT

8. What sequence of OLAP operators do you need to perform in order to obtain the Sum of Sales in US's REGIONS for each category of PENASONIC. Show all intermediate results that you can obtain step-by-step by filling in the attribute values in each dimension and the corresponding Sum of Sales. (7 Marks)




**Dimensions of a Data Ware House**

| YEAR | SEASON | BRAND | CATEGORY | P_NAME | COUNTRY | REGION | SALES |
|---|---|---|---|---|---|---|---|
| 2008 | SPRING | PENASONIC | CAMERA | LAMIX | CHINA | HK | 40 |
| 2008 | SPRING | PENASONIC | CAMERA | LAMIX | US | LA | 70 |
| 2008 | SPRING | PENASONIC | NOTEBOOK | W4 | CHINA | HK | 30 |
| 2008 | SPRING | PENASONIC | NOTEBOOK | W4 | US | LA | 60 |
| 2008 | SPRING | PENASONIC | PHONE | X60 | CHINA | HK | 60 |
| 2008 | SPRING | PENASONIC | PHONE | X60 | US | NY | 30 |
| 2008 | WINTER | PENASONIC | CAMERA | LAMIX | CHINA | HK | 40 |
| 2008 | WINTER | PENASONIC | CAMERA | LAMIX | US | LA | 70 |
| 2008 | WINTER | PENASONIC | NOTEBOOK | W4 | CHINA | HK | 30 |
| 2008 | WINTER | PENASONIC | NOTEBOOK | W4 | US | LA | 60 |
| 2008 | WINTER | PENASONIC | PHONE | X60 | CHINA | HK | 60 |
| 2008 | WINTER | PENASONIC | PHONE | X60 | US | NY | 30 |
| 2009 | SPRING | PENASONIC | CAMERA | LAMIX | US | NY | 10 |
| 2009 | SPRING | SUNY | CAMERA | DSC-T30 | US | LA | 100 |
| 2009 | SPRING | SUNY | CAMERA | DSC-T31 | CHINA | HK | 10 |
| 2009 | SPRING | SUNY | CAMERA | DSC-T32 | CHINA | HK | 70 |

| 2009 | SPRING | SUNY | GAME | PXP | CHINA | HK | 50 |
|---|---|---|---|---|---|---|---|
| 2009 | SPRING | SUNY | GAME | PXP | US | NY | 50 |
| 2009 | SPRING | SUNY | NOTEBOOK | VIO ZZ188 | US | LA | 110 |
| 2009 | WINTER | PENASONIC | CAMERA | LAMIX | US | NY | 10 |
| 2009 | WINTER | SUNY | CAMERA | DSC-T30 | US | LA | 100 |
| 2009 | WINTER | SUNY | CAMERA | DSC-T31 | CHINA | HK | 10 |
| 2009 | WINTER | SUNY | CAMERA | DSC-T32 | CHINA | HK | 70 |
| 2009 | WINTER | SUNY | GAME | PXP | CHINA | HK | 50 |
| 2009 | WINTER | SUNY | GAME | PXP | US | NY | 50 |
| 2009 | WINTER | SUNY | NOTEBOOK | VIO ZZ188 | US | LA | 110 |

********** END ********

***Please put your name and student ID on the question paper and return it together with your answer book.***