Student Number: .....

Question 1 [10 marks] Data mining and data warehousing concepts

Suppose you are the data analyst of the road transportation authority (RTA) in a medium-sized city. RTA wants to optimize its physical asset maintenance process. This process is responsible for ensuring the normal operation of all transportation assets, such as roadwork, signs, equipment and facilities. Your mission is to help on this goal by examining the following three databases.

1. Physical assets. These include transportation assets such as traffic lights, streetlights, road signs, traffic cameras, road/sidewalk/bikepath segments, roundabouts; operational assets such as trucks, as well as repairing tools.
2. Crew. These include information about each worker: name, title, address, contact number, responsibility, skill, hourly pay rate, reporting structure, shift and availability.
3. Maintenance. These include work orders being executed every day: the type of order (e.g. repair, repaint street lines, install street signs, pave new roads), its priority (critical/urgent/high/medium/low), location, day/time of different status updates (reported/dispatched/completed), which crew performed each order, which asset(s) was used and/or worked on, ...

Your first task is to build a data warehouse from these three databases to analyze maintenance needs and cost.

(a) Which one of the above is the transaction database? In one sentence give one example scenario where entries are **added** to this database daily? Give another example scenario where entries are **updated** in this database daily?

[2 marks]

Student Number: .....

Question 1 (continued)

Your design starts from the three-tier data warehouse diagram below.

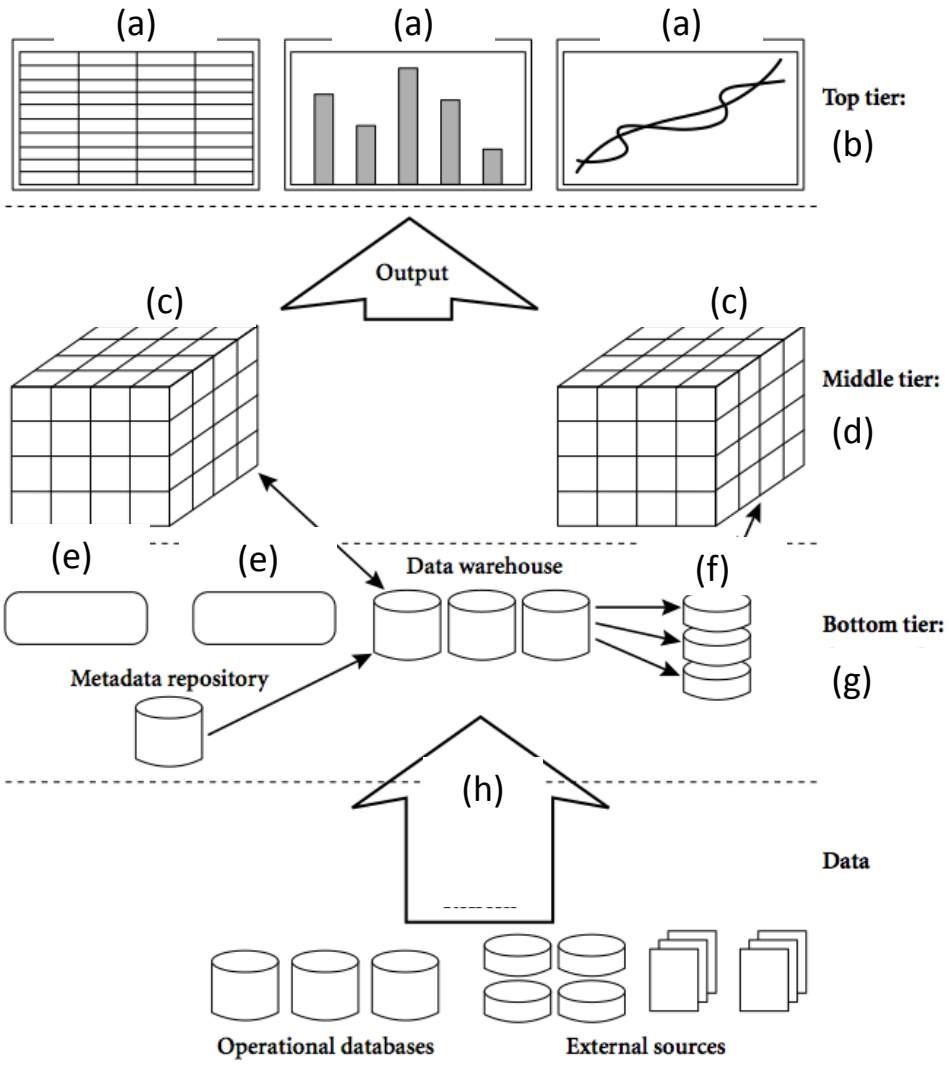


Figure 1: A three-tier data warehouse architecture.

(b) Name the three different layers in Figure :

Top tier (b):	Middle tier (d):	Bottom tier (g):

[2 marks]

Question 1 (continued)

(c) Then label the different components in Figure 1. You can choose from the following components: Administration, Monitoring, OLAP Server, Querying, Reporting, Data Marts, Visualization, Data Mining.
Note: there may be more than one labels that is appropriate for each letter, you need to get at least one name right for each label for the mark.

(a) in the top tier:	(c) in the middle tier:
(e):	(f):

[2 marks]

(d) There are several operations being performed at the arrow labeled as (h) in Figure 1. Name at least two of them.

[2 marks]

(e) What are two examples of **external** data sources you may use for the RTA data warehouse in Figure 1. Given that RTA staff often perform work outdoors in a complex urban environment, that other city agencies (such as water, electric and telecom suppliers) may also need to work on road segments.

[2 marks]

Student Number: .....

Question 2 [10 marks] Data pre-processing

Your next task is to pre-process the RTA data in Question 1 to construct the data warehouse.

(a) There are a number of different forms of data pre-processing that you can perform, including data cleaning and data transformation. Can you name two more forms?

[2 marks]

(b) In the RTA scenario, work-orders are either called in by citizens or recorded by RTA crew. The location of the work-orders can be in the form of an address, an intersection, or a (latitude, longitude) tuple from GPS devices carried by the crew. Give one example in which this process will generate noise in the Maintenance database, and write down your proposed method to clean such noise

[2 marks]

(c) You visualize the Maintenance database using a series of plotting tools. Name two of such plots. Specify the input of each plot, and its purpose. Your answers can be from both within the textbook and from real-world practices.

[2 marks]

Question 2 (continued)

(d) You examine a sample of the Maintenance database. It has a column recording the number of crew members needed for each job (`num_crew`), that takes integer values such as 1, 2, ... Over 1000 different work-orders, there were:

200 work-orders with `num_crew`=1;
300 work-orders with `num_crew`=2;
220 work-orders with `num_crew`=3;
200 work-orders with `num_crew`=5;
70 work-orders with `num_crew`=10;
8 work-orders with `num_crew`=15;
2 work-orders with `num_crew`=25;

If you were asked to convert numeric values of `num_crew` to four different types: *single-crew*, *small-team*, *large-team* and *multiple-teams*.
Write down you mapping from each of the observed `num_crew` value to one of the four target types. Briefly state why.

[2 marks]

(e) Compute the **mean** and **median** of `num_crew` among the 1000 work-orders in the previous part.

[2 marks]

Student Number: .....

Question 3 [10 marks] Data cubes and OLAP operations

You selected four dimensions from the RTA databases in Question 1 to build a basecube. There dimensions are: `date, work-order type, priority, location`. There are three measures over these dimensions: `count, time_taken`, and `cost`. Here `time_taken` is computed as the number of hours elapsed between job dispatch time and job completion time; and `cost` is computed as the sum of salary rates for all crew members involved, plus truck milage and consumables (e.g. gas).

(a) Draw a star schema for this data warehouse, with the three measures and four dimensions specified above.

[2 marks]

Question 3 (continued)

(b) List four popular OLAP operations.

[1 mark]

(c) Specify how you can use all or a subset of these OLAP operations to get the number of road-repair work-orders that were urgent or critical during the first quarter of 2011.

[1 mark]

(d) In this base cuboid, `date` has three levels: day, month and quarter; `work-order type` has two levels; `priority` and `location` each has one level. How many cuboids are there in the full data cube?

[1 mark]

Question 3 (continued)

(e) Assume that you are testing this datacube with a small data sample, containing all `date`s (365) in 2010, 20 different `work-order type`, 5 different `priority` values and the entire city partitioned into 10 distinct `location` zones based on postcode.
There are two work-orders in this data sample:

work-order-id	date	work-order type	priority	location
00123	2011-06-09	traffic_light_repair	urgent	4217
00234	2011-06-09	traffic_sign_install	medium	4222

What is the fraction of empty (zero) cells in the cuboid with dimensions `work-order type, priority, location` ?

[1 mark]

(f) Take the data cube in part (e) of this question. How many non-zero cells will you need to compute for the entire data cube? This includes the base cubiod, the apex cuboid, and everything in between.

[2 marks]

Question 3 (continued)

(g) Give two different examples about how you could use the available data in the RTA databases to help improve transportation service quality or save operating cost. For each example specify the following: which database(s) to take data from, what are the dimensions and measures of the data cube, what kind of mining or analysis you will do on the data cube, and what is the possible action(s) that can be taken from the analysis results.
Note: you can (but do not have to) use the data cube given in this problem in one of your examples.

First Example:

[1 mark]

Second Example:

[1 mark]

Student Number: .....

Question 4 [7 marks] Association mining

(a) Given the following small data set consisting of 7 transactions. Each contains item sets made of items **A** to **E**.

TID	Item set
1	**B**, **C**, **E**
2	**B**, **D**, **E**
3	**A**, **B**, **C**, **D**
4	**B**, **C**, **D**
5	**A**, **D**
6	**B**, **C**
7	**A**, **B**, **D**

(i) Following the *Apriori* algorithm, give all candidate item sets and all frequent item sets of lengths 1, 2 and 3 with a minimum support of 2 transactions.

(i)

[3 marks]

(ii) For all frequent item sets of length three from part (i) of this question (previous page), generate all rules with two items on the left-hand side and one item on the right-hand side (such as $\{A,B\} \rightarrow C$), and calculate their support and confidence (as ratios or percentage numbers).

(ii)

[2 marks]

(b) Explain the *Apriori* principle, and how it is incorporated in the *Apriori* algorithm to make the algorithm efficient and scalable to large data sets.

[1 mark]

(c) Explain how association rule mining can be applied to numerical data.

[1 mark]

Student Number: .....

Question 5 [7 marks] Cluster analysis

(a) What is the objective of cluster analysis?

[1 mark]

(b) Describe two application domains where cluster analysis can be very useful.

[1 mark]

(c) Describe five clustering requirements that are important to data mining.

[1 mark]

(d) What is the main advantage of k-means clustering over other clustering approaches?

[1 mark]

(e) (i) Explain how the *single link* and the *average link* measurements work.

(i)

[1 mark]

(ii) Which of these two measurements would you use to cluster a large real-world data set? Explain why.

(ii)

[1 mark]

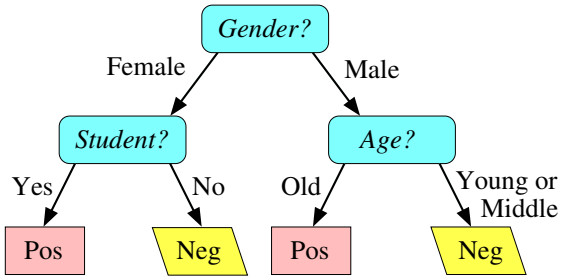
(f) Describe one advantage and one disadvantage of hierarchical clustering.

[1 mark]

Student Number: .....

Question 6 [7 marks] Classification and prediction

(a) Below you can see a decision tree that has been constructed using a training data set (unknown to us). Also shown is a *test* data set with ten test records. The attribute (variable) *Test* is the class (target) attribute with classes *Pos* and *Neg*.



Student	Age	Gender	Test
No	Young	Female	Pos
Yes	Young	Male	Neg
No	Old	Female	Pos
No	Young	Male	Neg
No	Middle	Female	Neg
No	Middle	Male	Neg
Yes	Old	Female	Pos
Yes	Old	Male	Pos
Yes	Young	Female	Neg
Yes	Middle	Female	Pos

(i) Write down the rules that can be generated from the above decision tree.

(i)

[1 mark]

(ii) For each of the ten test records, determine if it is a *true positive*, a *true negative*, a *false positive* or a *false negative*. Then write down the resulting error matrix with the absolute counts (i.e. number of records) in each cell of the error matrix.

(ii)

[1 mark]

(iii) What are the accuracy and the misclassification rate (both as percentage values or ratios) of the above decision tree on the above test data set?

(iii)

[1 mark]

(b) (i) Explain why decision trees are popular classifiers in data mining.

(i)

[1 mark]

(ii) Explain why it is important to prune decision trees.

(ii)

[1 mark]

(c) Describe a major advantage that lazy learners have over eager learners.

[1 mark]

(d) Explain how 10-fold cross-validation works.

[1 mark]

Student Number: .....

Question 7 [7 marks] Further data mining topics

(a) identify and describe four major differences between *data stream management systems* and *relational database management systems*.

[1 mark]

(b) Describe three methods of how a *trend curve* can be estimated.

[1 mark]

(c) Describe an application domain where using privacy-preserving data mining can be of great benefit.

[1 mark]

Question 7 (continued)

(d) Explain how the similarity between two text documents can be measured.

[1 mark]

(e) Describe the major steps involved in keyword based association rules mining.

[1 mark]

(f) What are the major characteristics of the Web as a data source?

[1 mark]

(g) Describe an example application of multimedia mining of Web sites.

[1 mark]