THE AUSTRALIAN NATIONAL UNIVERSITY

First Semester Examination – June 2012

COMP3420

Advanced Databases and Data Mining

Study Period: 15 minutes

Time Allowed: 3 hours

Permitted Materials: One A4 page with handwritten notes on both sides, non-programmable calculator is permitted.

Questions are NOT equally weighted.

This examination will be marked out of 60, and it will be worth 60% of your final course mark (1% per mark).

The questions are followed by labelled, framed blank panels into which your answers are to be written. Additional answer panels are provided (at the end of this examination paper) should you wish to use more space for an answer than is provided in the associated labelled panels. If you use an additional panel, be sure to indicate clearly the question and part to which it is linked.

The marking scheme will put a high value on clarity. As a general guide, it is therefore better to give fewer answers in a clear manner than to outline a greater number in a sketchy, half-answered fashion.

Please note that the questions and each of their parts are not sequenced in the order from easy to hard, please schedule your time accordingly.

Please write clearly – if we cannot read your writing you might lose marks!

Student Number:

The following are for use by the examiners.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Total

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

Suppose you are a data analyst of a major telecommunications operator, A-Tel.

A-Tel wants to look into its call-center operations and improve customer satisfaction. The call-centers of A-Tel are centralised offices that takes incoming customer phone calls to the service number 13 45 67. Each call-center consultant administers certain types of customer requests, such as product technical support, account services, or information inquiries. Your mission is to help on this goal by examining the following four databases.

1. Accounts. These include: information about each customer such as their name, title, address, contact info, and the date they had the first A-Tel account; a list of accounts that each customer has with A-Tel, including the account type (telephone/mobile phone/home internet/wireless data/cable TV/. . .), the open and close date of account, contract term, account limit, etc.
2. Consultants. These include information about each call-center consultant: name, title, address, contact number, work location, skill level, specialty area, hourly pay rate, reporting structure, shift, and so on.
3. Call records. This database contain information about each customer request: customer id, the type of request (e.g. repair, account update, technical support, etc), its priority (urgent/high/medium/low), location, day/time of different status updates (reported/dispatched/completed), additional notes about the request (its cause/resolution), and so on.
4. Surveys. This contains two types of surveys: general customer satisfaction about A-Tel products or services, and customer satisfaction about their call-center requests. They are collected by calling the customer landline or mobile telephone at random, or at the end of each call-center phone call.

Your overall task is to build a data warehouse from these databases to analyze customer needs and the way call-center operates.

(a) A customer calls A-Tel to change her billing address, and asked for instructions to setup her internet connection at the new address. Which databases (among 1–4 above) are changed during her call? And if changed, are entries added or updated? Please indicate your answer by circling the appropriate option for each DB below. **[2 marks]**

DB(1) Accounts: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(2) Consultants: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(3) Call records: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(4) Surveys: **Changed / Unchanged**; (if changed) **Added / Updated**

Question 1 (continued)

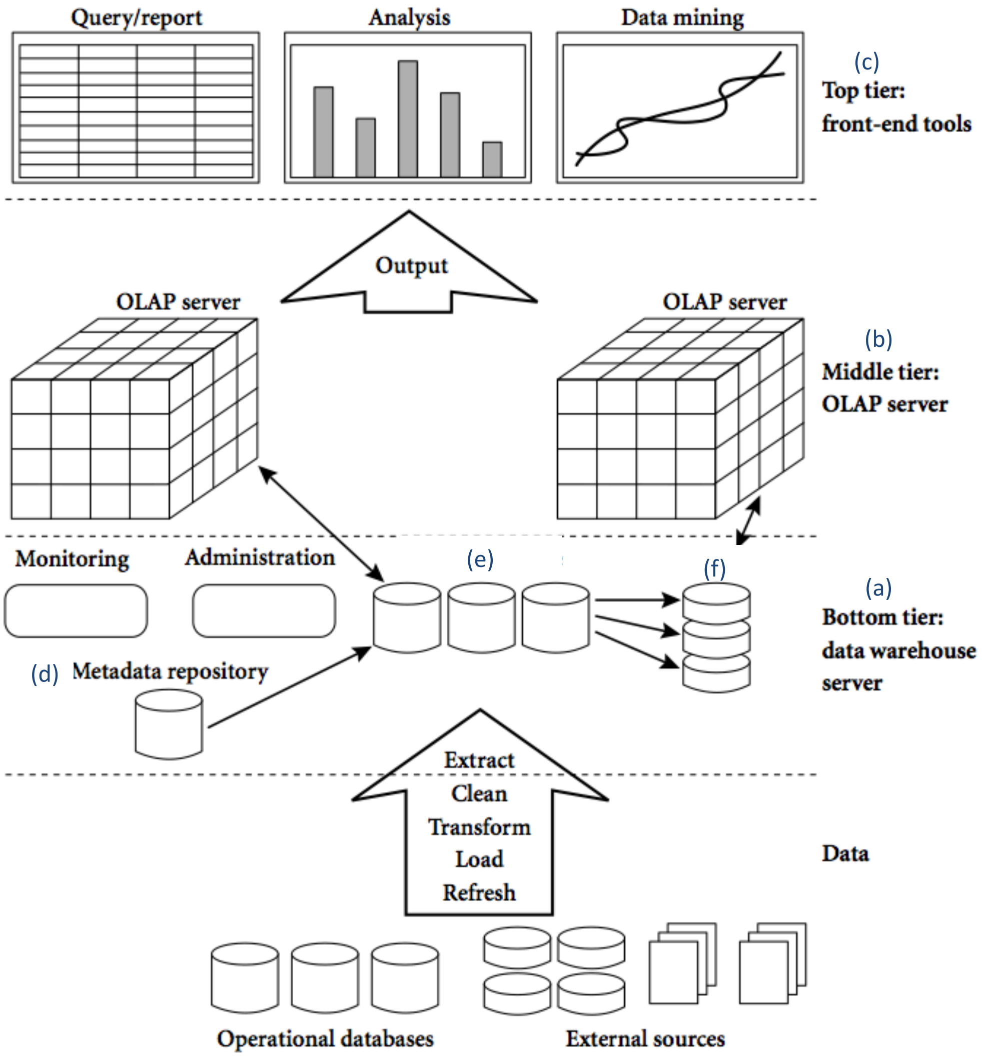

Figure 1: A three-tier data warehouse architecture.

(b) Consider a general data warehouse architecture in Figure 1. Please name the two components labeled as (e) and (f). **[2 marks]**

(e)	(f)

Student Number: ....

Question 1 (continued)

(c) What is metadata (box d in Figure 1)? Please provide a brief description. Please also give two examples of metadata about the Consultants database described on page 2. **[2 marks]**

(d) You are to find out the level of general customer satisfaction about different types of internet connections across different cities. Which database(s) among 1–4 (on page 2) do you need to pull data from? Briefly state what data you will use from each source. **[2 marks]**

Question 1 (continued)

(e) One of your intermediate analyses yields a table containing the number of satisfied customers in the general service survey, tabulated by their internet connection type (as shown below). Can you conclude that customers tend to be more satisfied with their ADSL connections than with other connections, why, or why not?

connection-type	# satisfied
Cable	2217
ADSL	6854
DSL	1065
Optical Fiber	534
Wireless 3G	1320

[2 marks]

Question 2 [10 marks] Data pre-processing

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

(a) Consider the following five attributes and their example values in the Call Records database.

1. `Request-type`: repair, technical support, bill correction.
2. `Priority`: urgent/high/medium/low.
3. `Location`: in the format of a street address, provided by the customer.
4. `Request-date/time`: in the format of `yyyy-mm-dd HH:MM:SS` (e.g. `2012-05-05 14:04:04`).
5. `Request-notes`: up to 300 characters documenting details of the request.

Which of the above attributes (if any) are of the numeric, nominal, ordinal types, or others? Please sort the attributes 1–5 above according to the appropriate types below, write `None` if no attribute belongs to the specified type. **[2 marks]**

Numeric attribute(s):

Nominal attribute(s):

Ordinal attribute(s):

Other types:
Sort the attribute here if it does not belong to any of numeric/nominal/ordinal types. Can each of the attributes here be converted to one of numeric/nominal/ordinal attributes (Yes/No)? If yes, to which type, and briefly describe how.

Question 2 (continued)

(b) Given two example call records, how do you compute the **distance** between each of their attributes? For each pair of example attribute below, please first name the distance function being used, and then outline the key steps for computing the distance (including any transformation you apply to the data, necessary assumptions you make, and data units being used). Please also write down the final distance value if you can.

You will get full marks for this part once 4 out of the 5 distances are computed correctly. [**2 marks**]

`Request-type`: R_1 =“line repair”, R_2=“bill correction”

`Priority`: P_1 =“urgent”, P_2=“low”

`Location`: L_1 =“2 Hutton St, Acton, ACT”, L_2=“7 London Circuit, Canberra, ACT”

`Request-date/time`: T_1 =“2012-05-03, 13:00”, T_2=“2012-05-02, 12:00”

`Request-notes`: N_1 =“Telephone line not working since April.”, N_2=“Telephone over-billed in April.”

Question 2 (continued)

(c) Take a sample of the Call Records database, an assume an additional column `hours-taken` has been computed as the difference between the `request-date/time` when the customer first called to lodge the request, and the `closed-date/time` when the request was marked as resolved, rounded to the closest number of hours.
200 requests with `hours-taken` = 0;
400 requests with `hours-taken` = 1;
100 requests with `hours-taken` = 2;
200 requests with `hours-taken` = 5;
50 requests with `hours-taken` = 12;
50 requests with `hours-taken` $\geq$ 24;

How many requests does this data sample contain? **[1 mark]**

Compute the **median** of `hours-taken` among these requests. **[1 mark]**

(d) Define *inlier mean* to be the average of values that lie between the 5-th and 95-th percentile of the data sample. Compute the *inlier mean* for `hours-taken` in the previous part. **[2 marks]**

Question 2 (continued)

(e) Detailed analysis is required to see why some requests are taking longer than others to resolve. We take 100 calls about "technical support", examine the `hours-taken` to resolve them ($>$ 1 hour or $\leq$ 1 hour) with the `experience-level` of the consultant answering the call ($<$ 1 year, $\geq$ 1 year). Resulting in the 2-by-2 table below.

	$<$ 1 year	$\geq$ 1 year
$>$ 1 hour	60	15
$\leq$ 1 hour	20	5

Compute the χ^2 correlation between `hours-taken` and `experience-level`. From your result, does experienced consultants seem to take longer/shorter to resolve customer inquiries? **[2 marks]**

Question 3 [10 marks] Data cubes and OLAP operations

You selected six dimensions from the A-Tel databases in Question 1 to build a datacube. The dimensions and the number of possible values they can take are:

1. `Request-type`, split into five sub-types.
   (a) Repairs, can take one out of 3 different values
   (b) Telephone technical support, can take one out of 2 different values
   (c) Internet technical support, can take one out of 5 different values
   (d) Billing questions, can take one out of 5 different values
   (e) Others, can take one out of 5 different values
2. `Hours-taken`, six different values.
3. `Experience-level`, two values, < 1 year or ≥ 1 year.
4. `Consultant-specialty`, five values.
5. `Customer-location`, ten values each corresponding to a geographic area.
6. `Call-center-location`, three values.

(a) How many cuboids are there in the entire data cube, if `request-type` can be materialized either along the five sub-types or all different values that each sub-type takes? **[1 mark]**

(b) How many cube cells are there in the base cuboid (i.e., the cuboid with the most number of cells)? **[1 mark]**

Question 3 (continued)

(c) Customer ratings after each call are measured on a scale of 1–5, 1 as very dissatisfied, 5 as very satisfied. The numeric average of these rating from different calls are then computed to create a measure called `average-customer-rating`. This is currently the only measure in this data cube.

You need to compute the `average-customer-rating` for all *repair* requests for each of the 10 customer locations. Is the current data cube sufficient for completing this task?

If Yes, describe which OLAP operation(s) you will use to get the required `average-customer-rating` from the base cuboid.

If No, which other measure(s) do you need? Add those measure(s) to your data cube, and then describe which OLAP operation(s) you will use to get the required `average-customer-rating` with the additional measure(s).

[2 marks]

(d) Suppose each of the five `consultant-specialty` values correspond to one sub-type in the dimension `request-type`, for example, consultants specializing in “billing” will only handle `request-type` of “billing question”. How many cells in the base cuboid will certainly be **empty** (i.e. having a value of zero)?

[1 mark]

Question 3 (continued)

(e) Suppose, *in addition to* the specification in part (d) above, each of the three `call-center-locations` values correspond to one or two sub-type in `request-type`, i.e., the *Sydney* center handles *repairs*; the *Melbourne* center handles *telephone tech support* and *internet tech support*; and the *Adelaide* center handles *billing* and *other* requests. At most how many cells in the base cuboid can be **non-empty**?

[1 mark]

(f) Which cube computation algorithm would you use, if you are to materialize the full data cube, for cells with `average-customer-rating`$\geq$ 5.0? Please name your choice among *Multi-dimensional aggregation*, *BUC (bottom-up-computation)* and *Star-cubing*, and very briefly state why.

[2 marks]

Question 3 (continued)

(g) Assume that you are testing this datacube by examining one of the 3-dimensional cuboids, which only has three non-empty cells, denoted as (a_1, a_2, a_3):

cell id	`hours-taken`	`experience-level`	`consultant-specialty`	`average-customer-rating`
a_1	0	> 1 year	billing	5.0
a_2	1	≤ 1 year	billing	4.8
a_3	1	≤ 1 year	repairs	5.0

What is the total number of ancestor cells of a_1, a_2 and a_3 (i.e., cells that can be derived from one or more roll-up operations using this 3-dimensional cuboid) that have `average-customer-rating`$\geq$ 5.0.

[2 marks]

Question 4 [7 marks] Association mining

(a) Given the following small data set consisting of 9 transactions. Each contains an item set made of some of the items **A** to **F**.

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

(i) Following the *Apriori* algorithm, give all candidate item sets and all large item sets of length 1, 2, 3 and 4 with a minimum support of 3 transactions.

(i)

[3 marks]

(ii) For the **alphabetically sorted** first two large item sets of length three (3) 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) In one sentence each, describe the two major bottlenecks of the *Apriori* algorithm.

[1 mark]

(c) In one or two sentences, describe an example application where multi-dimensional association mining can be useful.

[1 mark]

Question 5 [7 marks] Cluster analysis

(a) Describe what the characteristics of a good clustering are.

[1 mark]

(b) Two requirements for clustering in data mining are: (1) being able to discover clusters of arbitrary shapes; and (2) being insensitive to the order of input records. In one sentence each, explain why these two requirements are important.

[1 mark]

(c) Assume you have two numerical attributes, one containing age values (in the range 0 to 100), the other containing salary values (in the range 0 to 1,000,000). Is any data pre-processing required (or not) with such values before they can be used for clustering? Explain in one or two sentences.

[1 mark]

Question 5 (continued)

(d) Describe the difference between the single link and complete link methods, and explain one drawback these two methods have in common.

[2 marks]

(e) Explain how you would use a dendrogram to find 10 clusters in a data set.

[1 mark]

(f) Describe a situation where the use of density based clustering would be of advantage over the use of hierarchical or partitioning based clustering.

[1 mark]

Question 6 [8 marks] Classification and prediction

(a) The following table contains a small training data set with 10 records, three attributes with two or three different values each, and the class label attribute with two classes **1** (positive class) and **0** (negative class).

RecID	Attr1	Attr2	Attr3	Class
1	M	A	S	0
2	F	B	T	1
3	M	A	U	1
4	F	A	T	0
5	F	A	U	1
6	M	A	T	0
7	F	B	S	0
8	F	A	T	0
9	M	B	S	0
10	M	A	U	1

(i) Build a decision tree based on this training data set. At each step of building the tree, use the attribute which results in the purest splitting of the data (i.e. results in sub-sets of data with all – or most – of the records being in one class). Show your workings.

(i)

[3 marks]

(ii) For the following four testing tuples, decide if they are a true positive (TP), true negative (TN), false positive (FP), or a false negative (FN), using the tree you built in part (i) of this question.

(1) [Attr1=M, Att2=B, Attr3=U, Class=1]
(2) [Attr1=F, Att2=A, Attr3=T, Class=1]
(3) [Attr1=M, Att2=A, Attr3=T, Class=0]
(4) [Attr1=M, Att2=B, Attr3=T, Class=0]

(ii)

[1 mark]

Student Number: ....

Question 6 (continued)

(b) Describe the process of X-fold cross validation. What is the aim of this process?

[1 mark]

(c) Describe one advantage and one disadvantage of artificial neural network classifiers.

[1 mark]

(d) Describe one advantage and one disadvantage of using a small k (for example $k = 1$ or $k = 3$) in the k-nearest neighbour classifier (assuming a binary classification problem).

[1 mark]

(e) Describe the difference between classification and prediction, and give one example application each where you would use one or the other.

[1 mark]

Question 7 [8 marks] Further data mining topics

(a) Describe the difference between outlier and novelty detection.

[1 mark]

(b) Describe the challenge of class-imbalance in supervised outlier detection, and how it can be overcome.

[1 mark]

(c) Describe the difference between ‘relevant’ and ‘retrieved’ documents in a text retrieval system.

[1 mark]

(d) The Web is a highly dynamic source of information. How does this affect the way Web data mining is conducted?

[1 mark]

Question 7 (continued)

(e) Why is it often not enough to remove identifying values (such as names, dates-of-birth, or telephone numbers) from records to ensure the privacy of such data?

[1 mark]

(f) Besides maintaining privacy, what is a major challenge of privacy-preserving data linkage?

[1 mark]

(g) Histograms can be used for data stream processing. Describe how this method works, and what a major drawback of this method is.

[1 mark]

(h) The following sequence of ten values is assumed to be a time series. Calculate the moving average of this series of order 4.

15 1 4 0 3 5 8 4 3 1

[1 mark]

Student Number: ....

Continuation of answer to Question □ Part □

Continuation of answer to Question □ Part □

Continuation of answer to Question □ Part □

Student Number: ....

Continuation of answer to Question □ Part □

Continuation of answer to Question □ Part □

Student Number: ....

Continuation of answer to Question □ Part □

Continuation of answer to Question □ Part □