School of Electrical Engineering and Computer Science

CSI4142 Introduction to Data Science

Midterm 2018

Memorandum

Duration: 80 minutes

Total: 40 marks

Instructions:

1. This is a closed book examination, *subject to the following*.
2. You are allowed to bring along one letter-side note page (so-called cheat sheet), written or printed on both sides.
3. You are not allowed to use any magnifying glasses or any other magnifying devices.
4. No calculators are allowed, or required.
5. Answer all questions in the spaces provided.
6. This Midterm contains 7 pages, including this first page.

For the grader:

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	TOTAL
2	4	10	2	4	2	4	6	2	2	2	40

OSG Case Study

Consider the Ottawa Sporting Gear (OSG) cooperation, which offers customers a wide range of outdoor gear, clothing, equipment and accessories for sale. The company sells men's, women's and children's clothing and footwear. Other products range from equipment for activities such as bicycling, hiking and climbing, to outdoor gear such as tents and sunglasses. The company also rents out equipment (for pickup from, and drop off at, their stores). In addition, they offer bicycle repairs and perform ski equipment servicing during the winter.

All customers need to be members of the OSG cooperation and join by paying a one-time fee of $10.00. When registering, members are required to provide details such as their addresses and phone numbers, dates of birth, gender and occupations.

Customers may shop in one of three modes:

i) In-store purchases, where their purchases are scanned at the point of sale (POS) terminal;

ii) Online purchases (at www.osg.ca) with pick-up at the store nearest to their homes; or

iii) Online purchases (at www.osg.ca) with Canada Post Expedient Parcel home delivery.

Currently, there are 15 stores, located in British Columbia, Ontario, Quebec, Nova Scotia and Alberta. The headquarters is located in Ottawa.

A purchase (invoice) may contains many line items, each corresponding to a particular item or stock keeping unit (SKU). For instance, one may purchase a sleeping bag, a tent and gloves at the same time. One may also rent multiple equipment at the same time. For instance, a canoe, paddles, life jackets and a bear barrel could be rented together, when planning to go canoe camping.

Answer the following questions.

OGS currently maintains a transactional relational database than contains all the details of the daily in-store transactions at all their stores. This database also records equipment rentals, bicycle repairs and ski equipment servicing that are treated as "purchase" transactions. For instance, one would pay $200 to rent a canoe for a weekend.

OGS further maintain a separate relational database for online shopping.

Suppose that the CEO of OGS asked you to create a single OGS data mart in order to track all customer transactions, in-store as well as online, over the last five years.

1. Declare the grain of your data mart. (2)

The grain of the data mart is a single item (product) purchased, rented or returned by a customer, when shopping either in a store or online.

Note: returned items have a negative $sold (-$amount)

2. Explain the steps you would follow to integrate the two databases, with reference to the high-level data staging plan. (4)

Refer to slides on Data Staging.

Specifically, from slide 7:

1. **Create a very high-level, one-page schematic of the source-to-target flow**
2. **Identify starting and ending points**
3. **Label known data sources**
4. **Include placeholders for sources yet to be determined**
5. **Label targets**
6. **Include notes about known problems**

Some notes:

In this case, we have two relational databases. It will be easier if they are both of the same type, e.g. PostgreSQL, so this needs to be assessed.

Next, we evaluate the database table schemas, in order to map the attributes. Referential integrity rules between the two databases need to be noted and enforced.

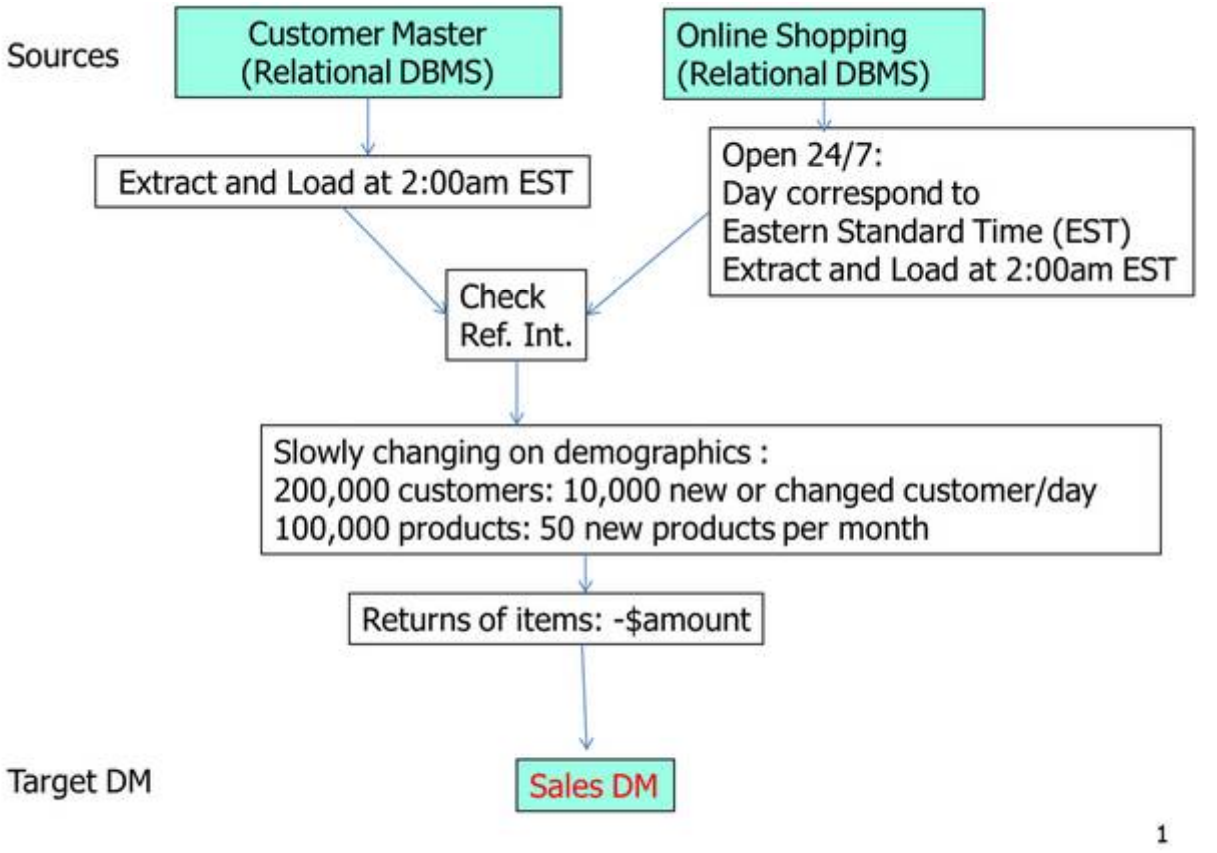
This is a data fusion (or integration) exercise. If needed we will transform different naming conventions, units of measure, time zones, and so on to standard units.

We next identify slowly changing dimensions.

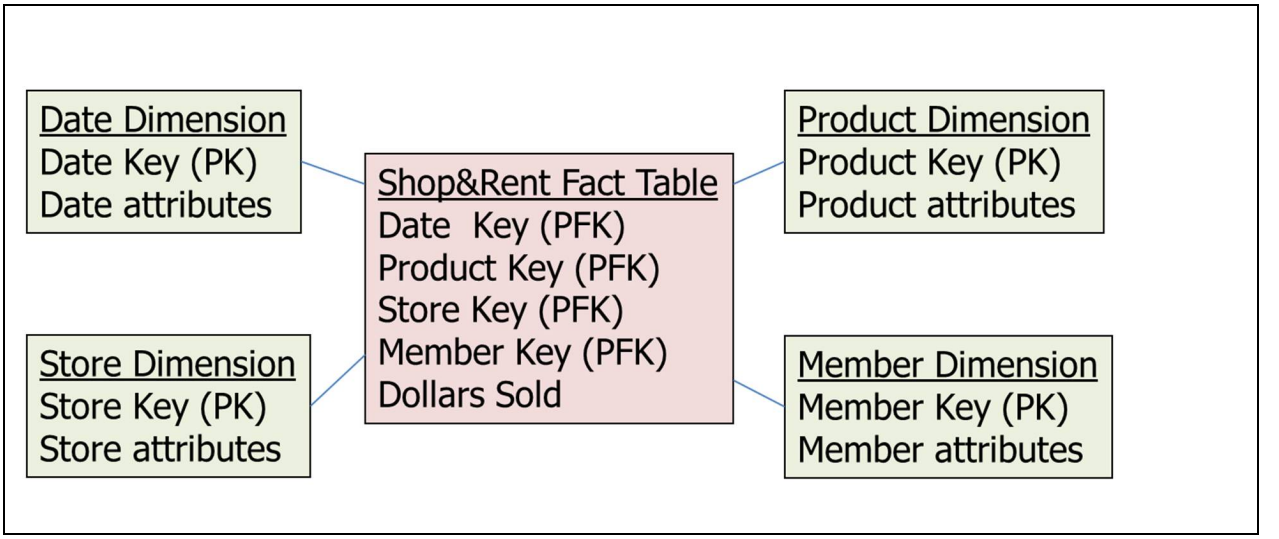
We note how to handle returns of merchandise.

Further, the online database is available 24/7, while the stores have specific hours of operation. The time of extraction is thus recorded.

Students could also have drawn a high level plan, such as the following.



3. Draw the dimensional model of your data mart. Your design should clearly show the following:
    a. The detailed fact table (2)
    b. At least one measure (or fact) (2)
    c. The relevant dimensions, including a Customer Dimension (6)



This data mart is very close to the "classical" Sales data mart. However, the requirement that we need to integrate online and in-store shopping complicates matters.

1. **The dollars sold ($amount) corresponds to a single item/product (e.g. clothes or equipment) that was sold to a Customer (or so-called Member).**
2. **The $amount for items that were returned need to be specified, since we monitor all customer transactions. We can do this by recording a negative amount (-).**
3. **The Product dimension includes items that are for sale, as well as items that are for rent. We therefore include a Rent(y/n) attribute in the Product dimension.**
4. **Online Sales are modelled as two rows in the Store dimension.**
    a. **One row for Online Shopping, with at Home delivery.**
    b. **One row Online Shopping, with in Store pickup (with the store name as an attribute).**
5. **Importantly, since we may need to ship items, care needs to be taken to keep the Customer's address up to date.**
6. **We could also have two dates that are role playing, namely the Purchase Date and the Return Date.**
    a. **For Products that are sold, the Return Date is (hopefully) usually empty.**
    b. **For rentals, the Purchase Date refers to the date of rental, and the Return Date refers to the date the equipment is returned.**
7. **One may add an attribute online (y/n) in the Fact table, in order to facilitate OLAP queries.**

Suppose the following is a list of queries (a) to (e) that are executed against the OGS data mart.

a) How does the total sales of *a product* during the current month (e.g. February 2018) compare to the sales during the Summer of 2017?
For instance, one may want to compare the sales of women's Keen Revel III hiking boots.

b) How does the total sales of *a product* during this month (e.g. February 2018) compare to the sales at the same time (e.g. February 2017) last year?
For instance, one may want to compare the sales of women's Keen Revel III hiking boots.

c) What are the names and brands of the six tents that had the *lowest* volume of sales, in Ontario, during the week starting on 25 June 2017? That is, we want to determine which tents did not sell well during the so-called "peak" summer selling season.
For instance, one may Determine that the Spark 1 tent by MEC was only sold 5 times, and so on.

d) What was the most popular colour for women's hats, as sold during 2017? Here, we use the term "popular" to refer to the volume of sales and we further assume that we only have hats with colours as contained in the set {red, black, lilac, green}. Strangely, we did not have any hats in stock of any other colour.

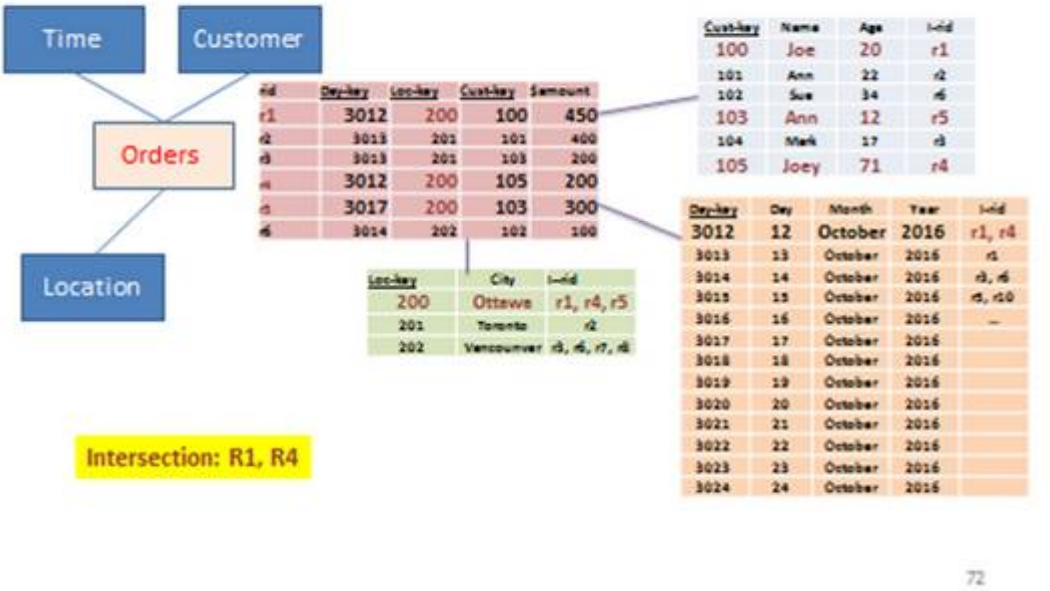
e) Name the five (5) bicycles are the most popular in the Ottawa store, in terms of the total number of sales in 2017, when compared to the sales of bicycles throughout Canada.
For instance, the Ghost Trekking 5 bicycle was the most popular seller in the Ottawa store, during 2017, However, it was ranked 3rd popular in Canada during the same period of time.

4. Give an example of one aggregate (or cube) that you would build to speed up both query (a) and query (b). (2)

An aggregate by Month would speed both queries up.
Here, we use the Year → Month → Day concept hierarchy.

5. Show how you would optimize the star join operation in order to speed up query (c). (4)

Students should refer to Slide 27 of the Physical Design deck.



72

We proceed as follows.

A. In the Product dimension: we select all the tents. These are linked with the Fact table, using a semi-join, in order to obtain the Fact table row-ids.

B. In the Date dimension: we select the rows for that week. These are linked with the Fact table, using a semi-join, in order to obtain the Fact table row-ids.

The Fact table rows that are of importance are those in the <u>intersection</u> of A and B. For each product-key, we calculate the count. Next, we sort the results and we list the names and brands of the 6 with the lowest counts.

This query is based on the idea of selectivity. Intuitively, we will have fewer tents. Also, assume we have the data of 10 years. In this case, retrieving the data of one week will reduce the workload considerably, especially if our data are partitioned (on disk or in the cloud) by date.

(An alternative is to use the method used by Informix, namely hash joins with pipelining, see Slide 74.)

6. Show how a bitmap index may be used to answer query (d). (2)

We maintain a bitmap for each one of the colours and link this to the FACT table. (We assume that all hats only have one colour.) This is done internally, by the OLAP engine.

For example, assume the following bitmap is maintained for the first 12 records in the FACT table.

Red	**100000001000**
Black	**011011000111**
Lilac	**000100000000**
Green	**000000110000**

In this case, the BLACK bitmap has the most 1s (7), so it will be chosen.

---

7. Provide the SQL statement to answer query (e). (4)

This is an example of using Iceberg Queries.

The following would get the results for Ottawa:

```
SELECT P.pname, S.dollars
FROM  Fact S, Products P, Date D, Customer C
WHERE S.pid = P.pid AND S.did = D.did AND S.cid = C.cid
AND   C.city = `Ottawa`
AND   D.year = 2017
AND   P.type = `bike`
ORDER BY S.dollars DESC
OPTIMIZE FOR 5 ROWS
```

The following would get the results for all of Canada:

```
SELECT P.pname, S.dollars
FROM  Fact S, Products P, Date D, Customer C
WHERE S.pid = P.pid AND S.did = D.did AND S.cid = C.cid
AND   D.year = 2017
AND    P.type = `bike`
ORDER BY S.dollars DESC
OPTIMIZE FOR 5 ROWS
```

8. Suppose that your OGS data mart contains a Customer Dimension that is slowly changing on marital status. Explain how you would handle such changes during data staging and motivate the answer. (6)

The answer will depend on whether we want to maintain a link to prior marital status(es). Refer to the Data Staging slides 54 to 60.

Type 1: overwrite the old value by the new one, but this is not recommended!

Type 2: we add a new record, for the new marital status. All records from the date of marital status change will refer to the newer record, while the older record will link to past records. In addition (Type 2b), we can add a "flag" that indicates whether the row is current. For Type 2c, we can add an "effective date" attribute.

Type 3: we add a new attribute, where we have "old marital status" as well as "new marital status", with an effective date

Type 4: we add a new "History" dimension (see example below):

- Customer dimension has current data:

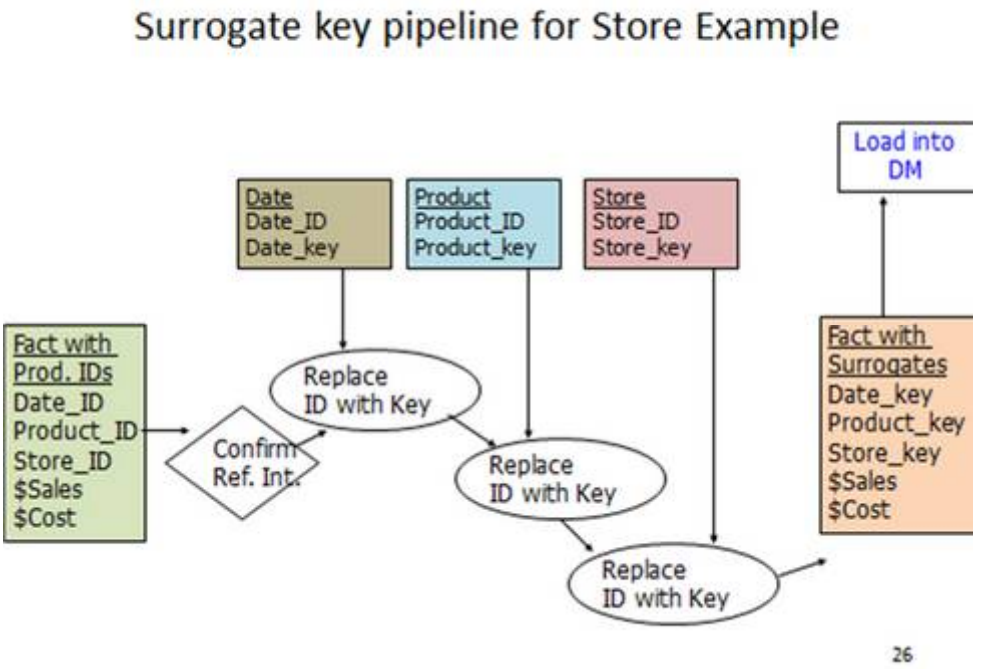
Cust-key	Name	Age	City	Marital-Status
2347	ANN	20	Montreal	Married

- "Customer-History" dimension keeps history:

Cust-key	Name	Age	City	Marital-Status	Effective-date
122	ANN	20	Ottawa	Single	13/2/2002
2346	ANN	20	Montreal	Single	1/1/2018
2347	ANN	20	Montreal	Married	14/2/2018

9. Explain what the surrogate key pipeline is. (2)

The surrogate key pipeline refers to the data staging steps we take to convert the Fact table from the transactional input data format to the dimensional model. Refer to slide 26 of the Data Staging nodes, as well as the textbook by Kimball et. al.



10. Explain how you would determine whether there is a correlation between the time of the day and the number of online sales. (2)

There are a number of ways to do this.
One could use a statistical test such as Pearson's Coefficient.
Alternatively, one could draw a scatter plot and see if there are any strong correlations between the two dimensions.

Correlated:

Uncorrelated:

11. Explain what attribute banding is and give one example in the OGS data mart where attribute banding may be used. (2)

Attribute banding refers to the process of discretizing numeric data into groups that are then useful for decision support. This is also sometimes done to avoid so-called monster dimensions.

Examples in this data mart could be to convert the dates of birth of members into age ranges, the exact time of purchase to hourly bands, the exact prices of products into price bands, and so on.

(You only had to list <u>one</u>.)