FIT3003 Business Intelligence and Data Warehousing

Sample Exam Paper 2 [SOLUTION]

Question 1

Employee Dimension is a *slowly changing dimension* (*SCD*). For simplicity, the dimension contains only these attributes: EmployeeID, Name and Salary. SCD has the six types: Types 0, 1, 2, 3, 4 and 6. Draw sample tables for Employee Dimension using each of the SCD types. Assume that the Salary attribute is the *Temporal attribute*, which may change from time to time, due to promotion, increase of salary, etc. Add more attributes to the table, whenever required by the SCD. For each type, you need to have at least two employee records, and these are the two employees:

- The first employee is *Adam* with an EmployeeID: A1. He started his job in 1 March 2016 with a salary of $3900. After his probation ended in 31 May 2016, his salary was increased to $4300. At the beginning of 2017, he received a pay increase, and his salary became $5000. From 1 July 2017 (until now), his salary has been $5500.
- The second employee is *Ben* with an EmployeeID: B2. He started his job in 1 February 2017 with an initial salary of $4000. After his probation ended in 31 May 2017, his salary became $4750 which is his current salary.

Solution:

Table 1.1 Employee Dimension Table (SCD Type 0)

EmployeeID	Name	Initial Salary
A1	Adam	$3900
B2	Ben	$4400
...	...	...

Table 1.2 Employee Dimension Table (SCD Type 1)

EmployeeID	Name	Latest Salary
A1	Adam	$5500
B2	Ben	$4750
...	...	...

Table 1.3 Employee Salary Dimension Table (SCD Type 2)

EmployeeID	Name	Start Date	End Date	Salary	Remarks	Current Flag
A1	Adam	1-3-2016	31-5-2016	$3900	...	
A1	Adam	1-6-2016	31-12-2016	$4300	...	
A1	Adam	1-1-2017	30-6-2017	$5000	...	
A1	Adam	1-7-2017	31-12-9999	$5500	...	Y
B1	Ben	1-2-2017	31-5-2017	$4000	...	
B1	Ben	1-6-2017	31-12-9999	$4750	...	Y
...	...	...	...	...	...	

Table 1.4 Employee Dimension Table (SCD Type 3)

EmployeeID	Name	Current Salary	Previous Salary
A1	Adam	$5500	$5000
B2	Ben	$4750	$4000
...	...	...	...

Table 1.5 Employee Dimension Table

EmployeeID	Name
A1	Adam
B2	Ben
...	...

Table 1.6 Employee Salary Dimension Table (SCD Type 4)

EmployeeID	Start Date	End Date	Salary	Remarks
A1	1-3-2016	31-5-2016	$3900	...
A1	1-6-2016	31-12-2016	$4300	...
A1	1-1-2017	30-6-2017	$5000	...
A1	1-7-2017	31-12-9999	$5500	...
B1	1-2-2017	31-5-2017	$4000	...
B1	1-6-2017	31-12-9999	$4750	...
...	...	...	...	...

Question 2

Bento Garden is a retail chain that specializes in selling Korean and Japanese bento sets. Customers can choose to dine-in or order online for home delivery. For continuous improvement, customers are encouraged to provide feedback where vouchers will be given for subsequent purchases. Each feedback is provided with a star rating, ranging from 1 (poor) to 5 (excellent).

The operational database of *Bento Garden* includes the following tables, with primary keys (PK) being underlined and foreign keys (FK) in italic:

Customer (PhoneNumber, FirstName, LastName, Street, Suburb, ZipCode, State, Points)

ProductCategory (CategoryID, Description)

Item (ProductCode, Price, Description, *CategoryID*)

Branch (BrandID, Street, Suburb, ZipCode, State, Phone)

Sales (SalesID, Date, Total, *PhoneNumber*)

Sales_Item (*SalesID*, *ProductCode*, Quantity)

Feedback (FeedbackID, StarRating, Feedback, Date, *PhoneNumber*, *ProductCode*)

After being in business for more than 10 years, the management would like to analyse the sales and feedback of their customers.

a) The management of *Bento Garden* would like to build a data warehouse that is able to answer the following questions. Total sales for each sale is calculated by summing up the unit price of each item, multiplied by the quantity in *Sales_Item*.

i) What is the best-selling product category by total sales in each season of the year?

ii) What is the most popular product category to customers in different suburbs, based on sales quantity?

iii) Which product category has the highest total sales?

iv) How many 5-star reviews are there for each product category?

Draw a star schema based on the requirements.

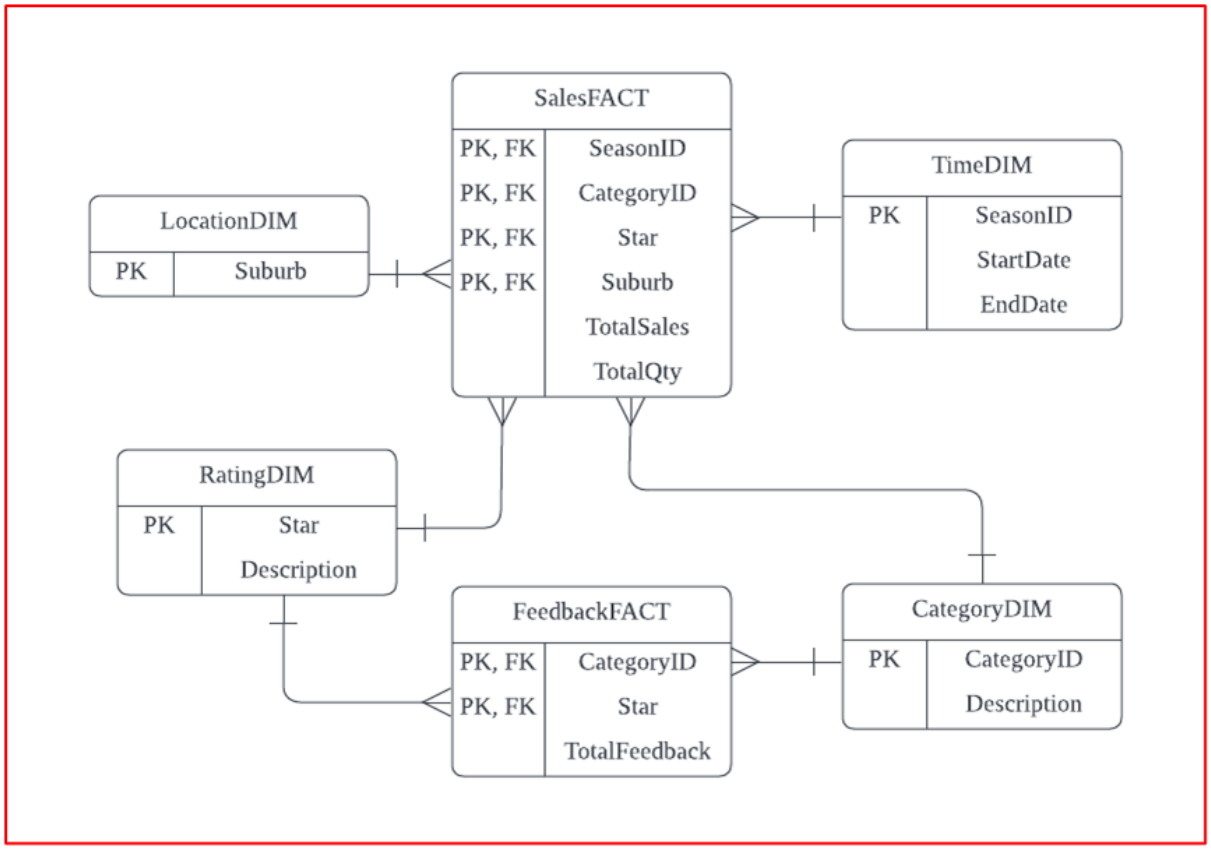
b) Assume that due to the volume of the data stored in the data warehouse, the management has decided to split the data warehouse based on the suburbs of the

customers. Between horizontal and vertical fact slicing, which one should be used? Explain your answer.

c) Assume that the suburbs are Carlton, East Melbourne, Parkville and Southbank. Write a sample of the SQL statement to implement fact slicing based on the suburb.

Solution:

a)



b) Horizontal fact slicing should be used.

Reasons: The fact table should be sliced based on the values of customer suburbs. In this case, the number of attributes in the fact table in all the slices should be the same, but each slice carries data for a specific suburb.

c) Implementation

CREATE TABLE SALESFACT_SOUTHBANK AS SELECT * FROM SALESFACT WHERE SUBURB = 'SOUTHBANK'';

CREATE TABLE SALESFACT_SOUTHBANK AS SELECT * FROM SALESFACT WHERE SUBURB = 'PARKVILLE';

CREATE TABLE SALESFACT_SOUTHBANK AS SELECT * FROM SALESFACT WHERE SUBURB = 'EAST MELBOURNE'';

CREATE TABLE SALESFACT_SOUTHBANK AS SELECT * FROM SALESFACT WHERE SUBURB = 'CARLTON';

Question 3

A medical clinic employs four general practitioners (doctors): Dr Adele, Dr Ben, Dr Kate and Dr Chris. Some of these doctors do not practise every day. For example, Dr Adele practises on Mondays and Wednesdays only, whereas Dr Ben is there only on Thursdays. When a patient comes to the clinic and has a consultation with a doctor, the patient pays a certain consultation fee, depending on the type of consultation the patient had. For example, a general consultation fee (code 113) is $37.50. Because of the nature of medical practice, there are more than 100 different codes for different types of consultations.

The clinic maintains an operational database that records every payment for each consultation by every doctor. A data warehouse is needed for reporting purposes. There are two versions of the star schema for this clinic.

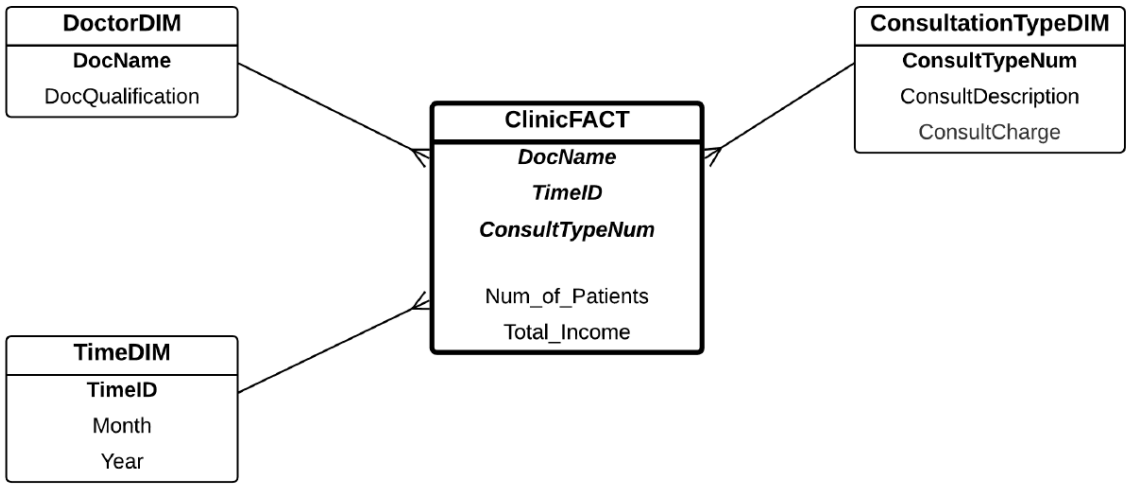


Fig. 1.1 Star Schema Version 1

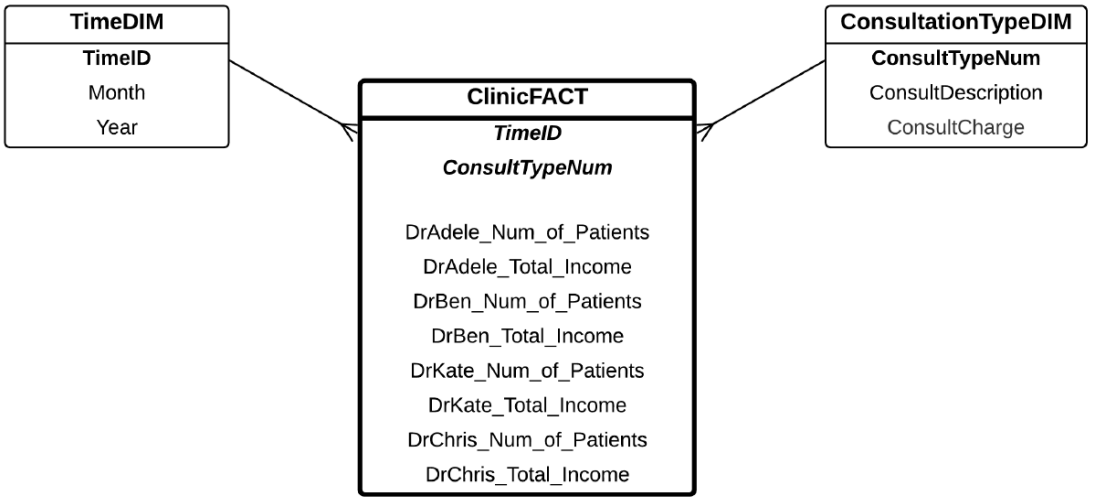


Fig. 1.2 Star Schema Version 2

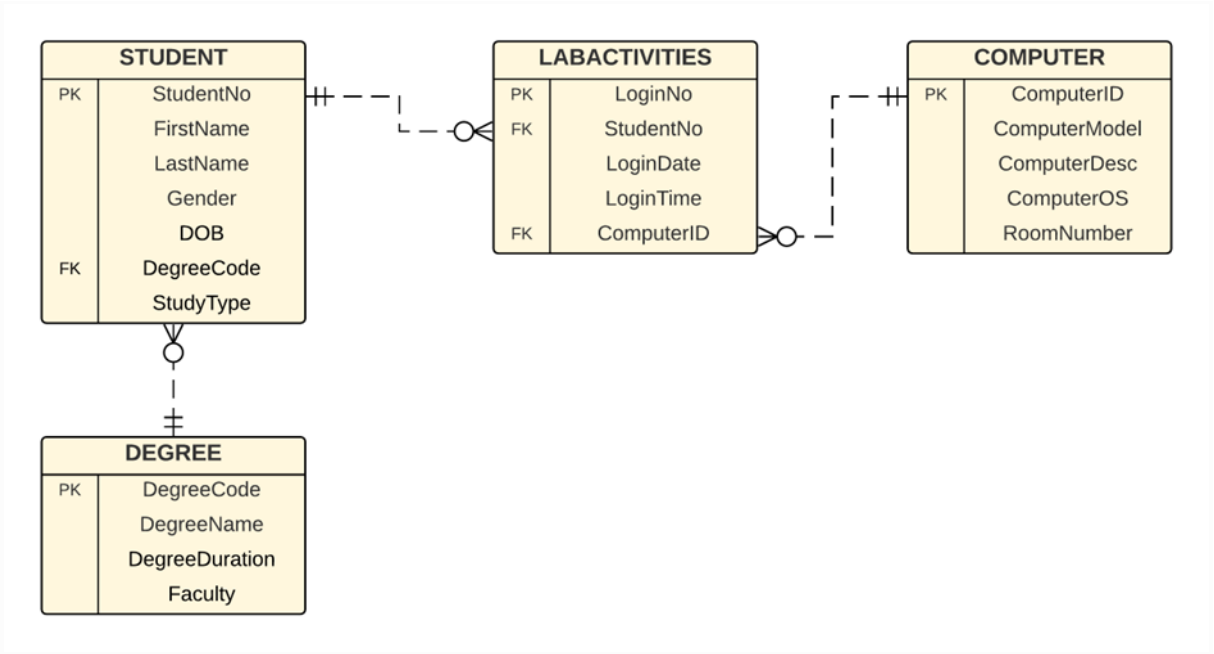
Question: Which schema (Star Schema Version 1 or Star Schema Version 2) has a higher level of aggregation, or are they of the same level of aggregation? State your answer, and explain your reasons as well.

Solution:

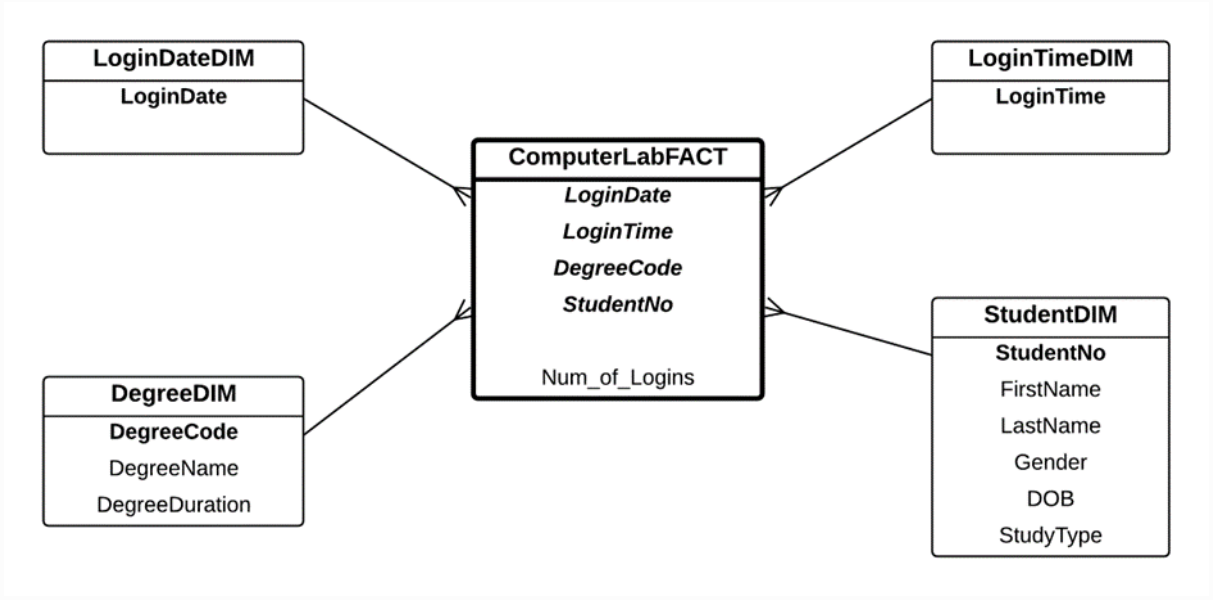
Both versions have the same level of aggregation, because one is not a more detail version of the other. Both have the same degree of detail, hence the same level of aggregation.

Question 4

Given the following E/R diagram, a Level-0 star schema has been created. The E/R diagram:



Level-0 star schema:



Your task is to create an *Active Data Warehouse*.

Questions:

(a) Explain what an Active Data Warehouse is.

(b) Assuming that the tables from the operational database are on the same environment as the data warehouse, and the operational database tables are accessible, explain two ways to create the dimension table for Level-0 star schema (Hints: use one dimension as an example, e.g. StudentDIM). In explaining the two ways to create the dimension table (e.g. StudentDIM table), you need to write the respective SQL commands.

(c) Discuss why Primary Key-Foreign Key (PK-FK) constraint does not add any values in the Traditional Data Warehousing.

(d) Explain why PK-FK constraint is necessary in Active Data Warehousing. Use an example to support your arguments.

Solution:

(a) An active data warehousing is where the data warehouse is immediately updated when the operational database is updated.

(b) Method 1: use create view

Method 2: use create table, but must be complimented by a database trigger

(c) In Passive DW, once the data warehouse is built, there will never be an update. Hence, data anomaly due to insert, update, and delete of records will never happen. Hence, PK-FK is unnecessary in Passive DW.

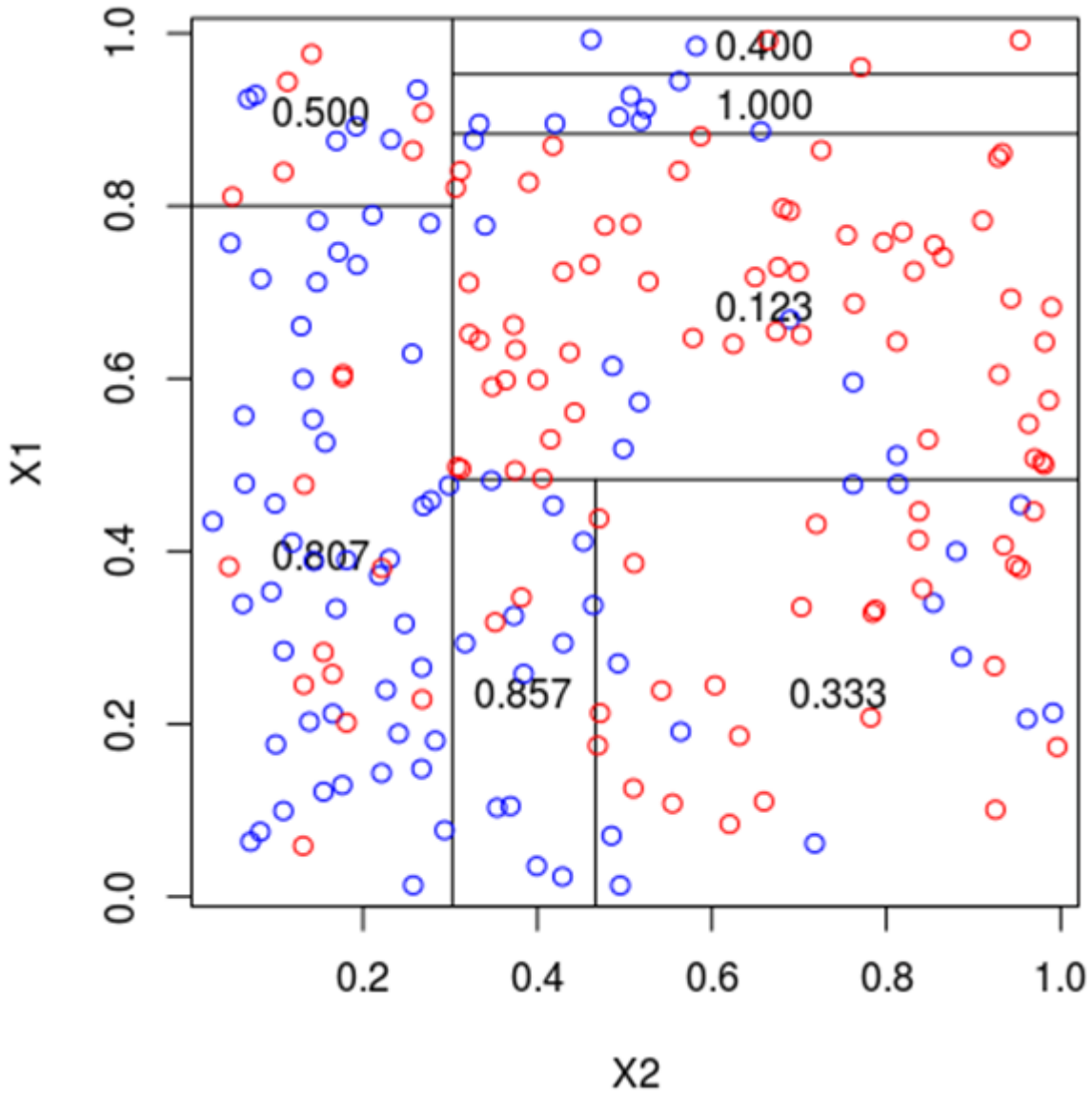
(d) In relational databases, PK-FK constraints are very important as they prevent anomalies (i.e. insert anomaly, update anomaly, delete anomaly).

In the traditional data warehousing, once the data warehouse is built, the data in the data warehouse will never be changed (e.g. there won't be any new insert, update or delete to the records in the data warehouse). Hence, the role of PK-FK in the traditional data warehousing is minimal.

However, in Active Data Warehousing, once the data warehouse is created, new data will be inserted, old data will be updated and deleted, depending on the changes happening with the underlying operational databases. Therefore, PK-FK in Active DW is very important, in order to guarantee that the data in the Active DW is always consistent, which can only be achieved if PK-FK constraints are maintained correctly.

Question 5

The following figure shows a map partitioning of the two-dimensional map: x1 and x2 axes. Each dot on the map represents a data point, whereas the number inside each grid (e.g. 0.500 in the top left-hand corner) represents the average value of the data points in that category.



Questions:

(a) (**2 marks**) How many regression trees can we draw for this map partitioning?

(b) (**8 marks**) Draw two regression trees based on this map partitioning.

Solution:

a) There are 3 possible regression trees:

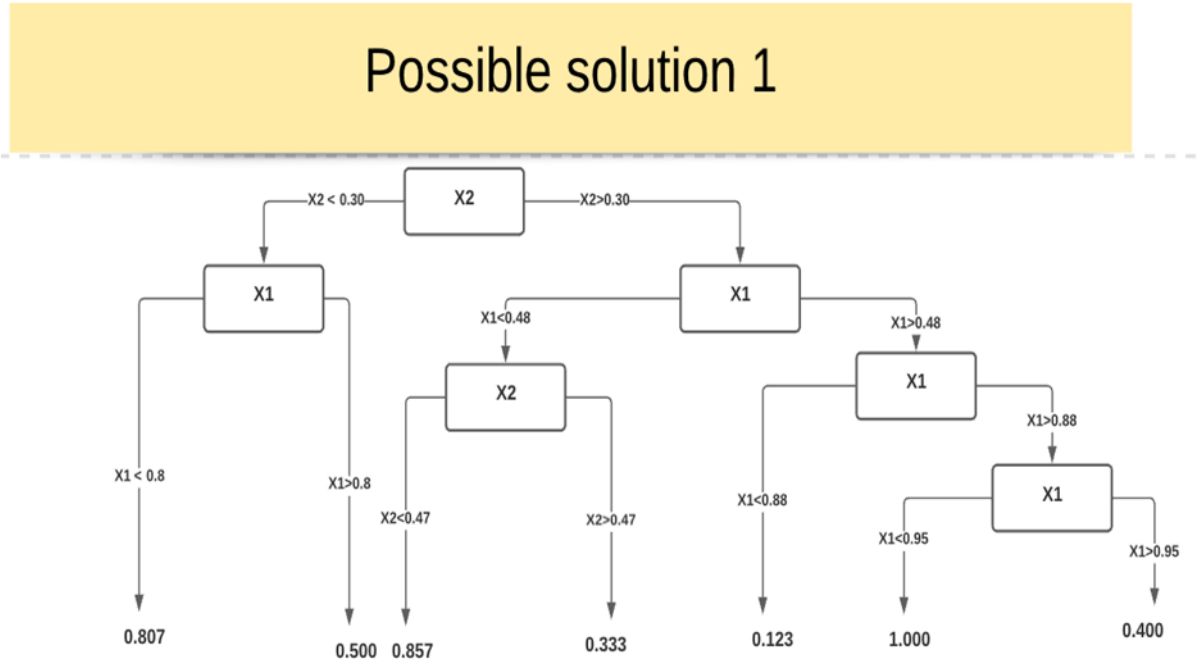
0.3, 0.95

0.3, 0.88

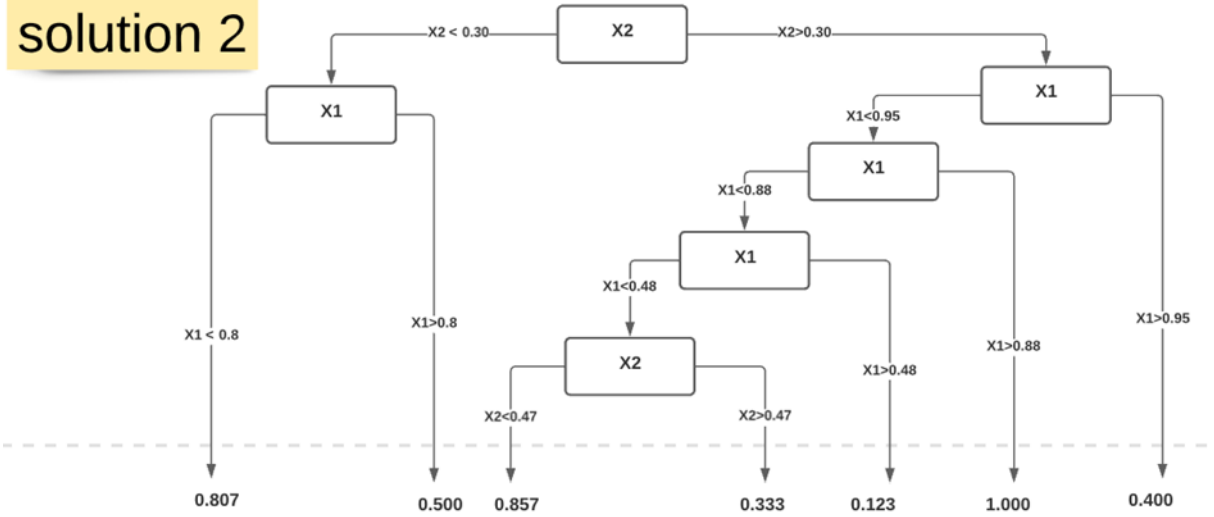
0.3, 0.48

b) Two possible regression trees are: (4 marks each, total 8 marks)

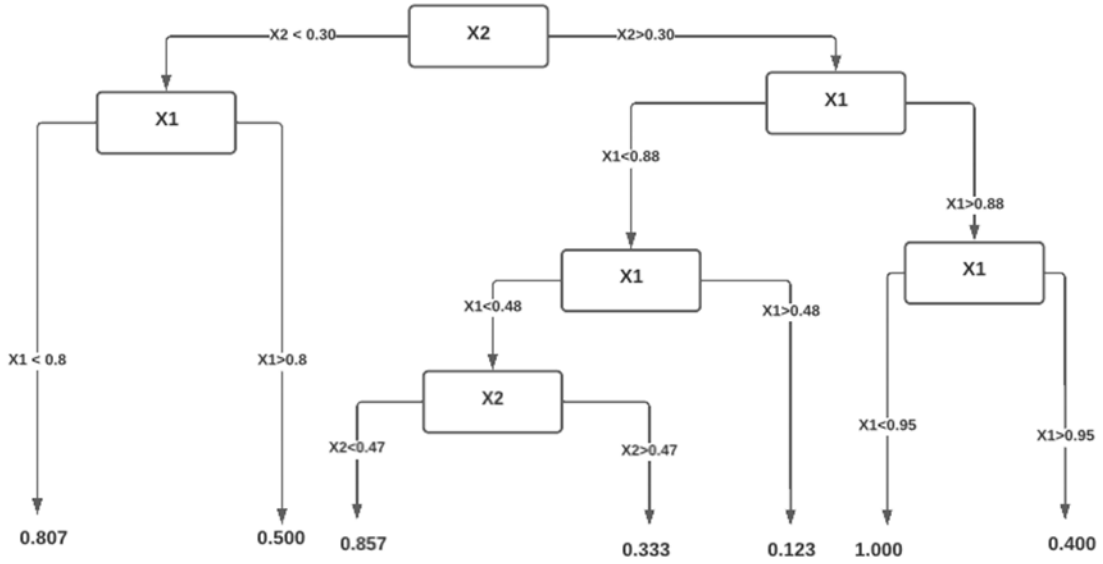
Note: the = sign could exist either left or right. Missing equal deduct 1 mark per tree.

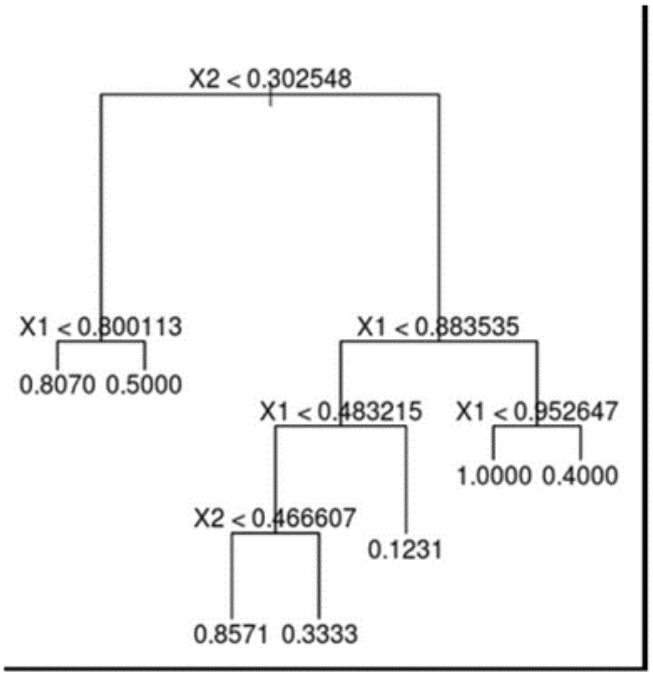


Possible solution 2



Possible solution 3

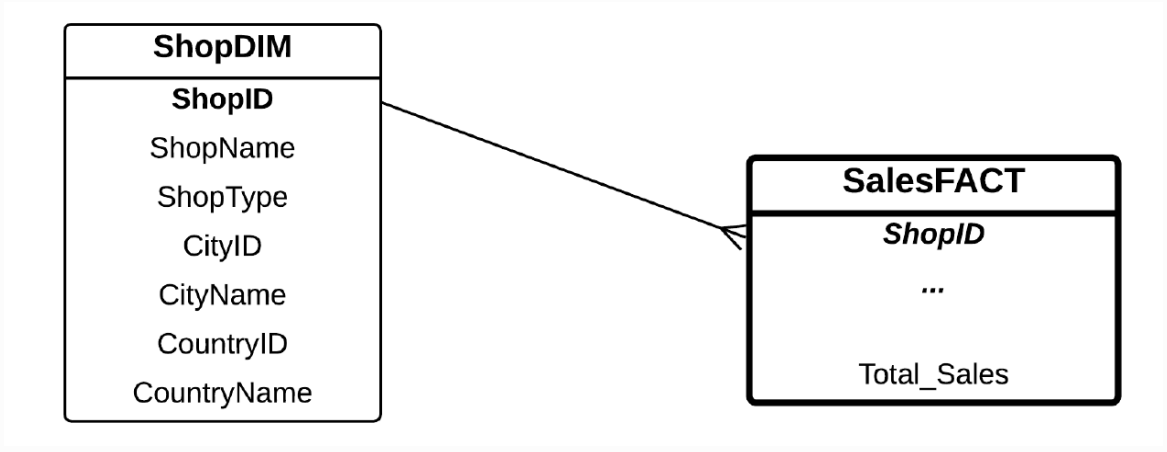




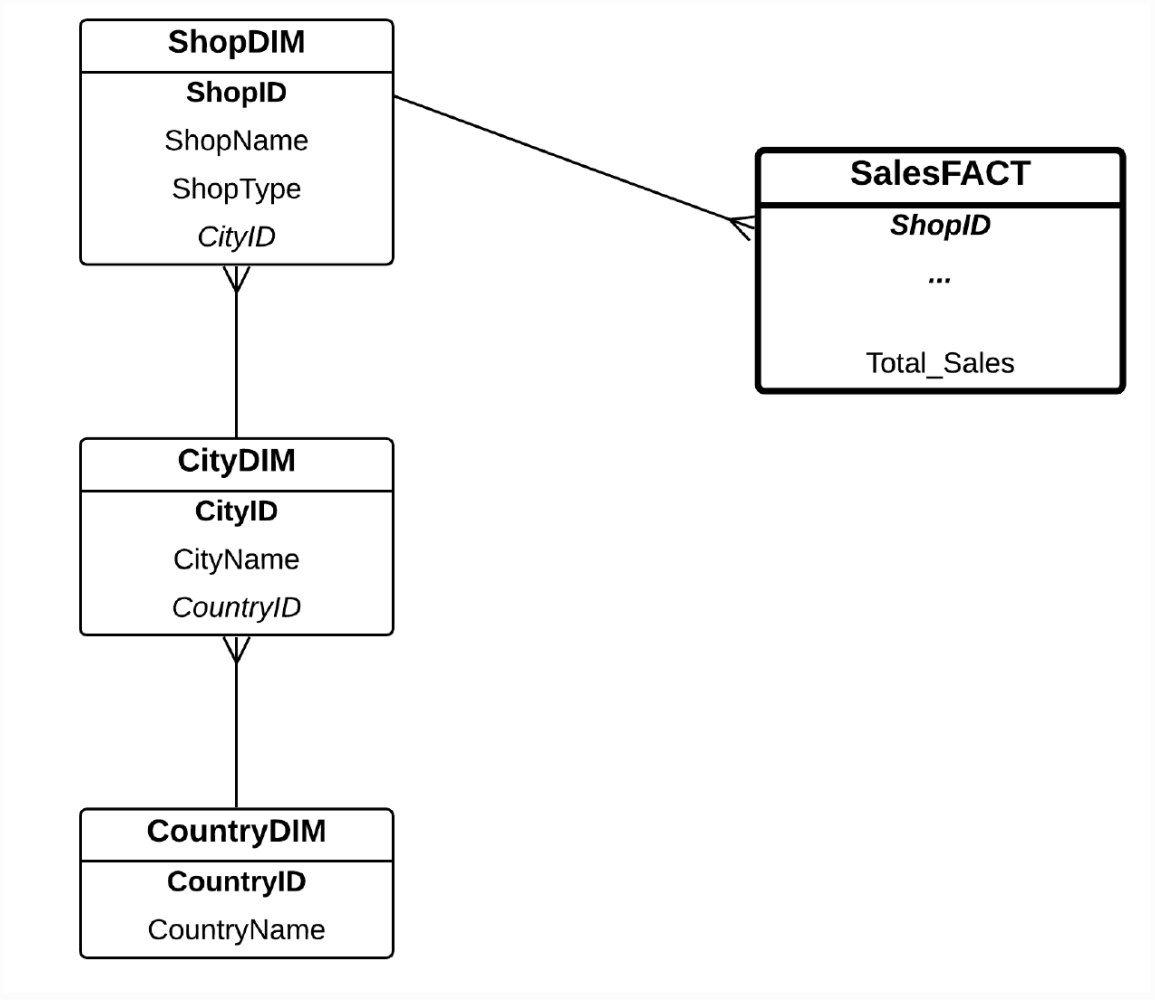
Question 6

Consider the following star schemas. The first star schema contains a hierarchy in the dimension, whereas the second star schema collapses the hierarchy into one dimension.

Star schema without Hierarchy



Star schema with Hierarchy



Questions:

a) Draw sample table contents of the fact and dimension tables of the two star schemas
b) Compare and contrast the two star schemas using the sample tables in question (a) above. Explain the pros and cons of each star schema.

Solution:

The Star Schema with Hierarchy has four tables: Shop Fact, Shop Dimension, City Dimension, and Country Dimension. They are shown in Tables 1.1 - 1.4.

Table 1.1 Star Schema with Hierarchy – Shop Fact Table

ShopID			Total Sales
1	...	...	$1,500,000
2	...	...	$2,750,000
3	...	...	$1,800,000

Table 1.2 Star Schema with Hierarchy – Shop Dimension Table

ShopID	Shop Name	Shop Type	CityID
1	Myer	Department Store	C
2	Coles	Supermarket	C
3	Big W	Department Store	M

Table 1.3 Star Schema with Hierarchy – City Dimension Table

CityID	City Name	CountryID
C	Canberra	AU
M	Melbourne	AU

The Star Schema without Hierarchy has only two tables: Shop Fact, and Shop Dimension. They are shown in Tables 1.5 - 1.6.

Table 1.4 Star Schema with Hierarchy – Country Dimension Table

CountryID	Country Name
AU	Australia

Table 1.5 Star Schema without Hierarchy – Shop Fact Table

ShopID			Total Sales
1	...	...	$1,500,000
2	...	...	$2,750,000
3	...	...	$1,800,000

Table 1.6 Star Schema without Hierarchy – Shop Dimension Table

ShopID	Shop Name	Shop Type	CityID	City Name	CountryID	Country Name
1	Myer	Department Store	C	Canberra	AU	Australia
2	Coles	Supermarket	C	Canberra	AU	Australia
3	Big W	Department Store	M	Melbourne	AU	Australia

b) Star Schema with Hierarchy:

- Pros: Normalized in 3NF, and minimized data duplication
- Cons: When producing a report, we need to join the fact table with three dimension tables: Shop, City, and Country Dimensions. Hence, we need three join operations.

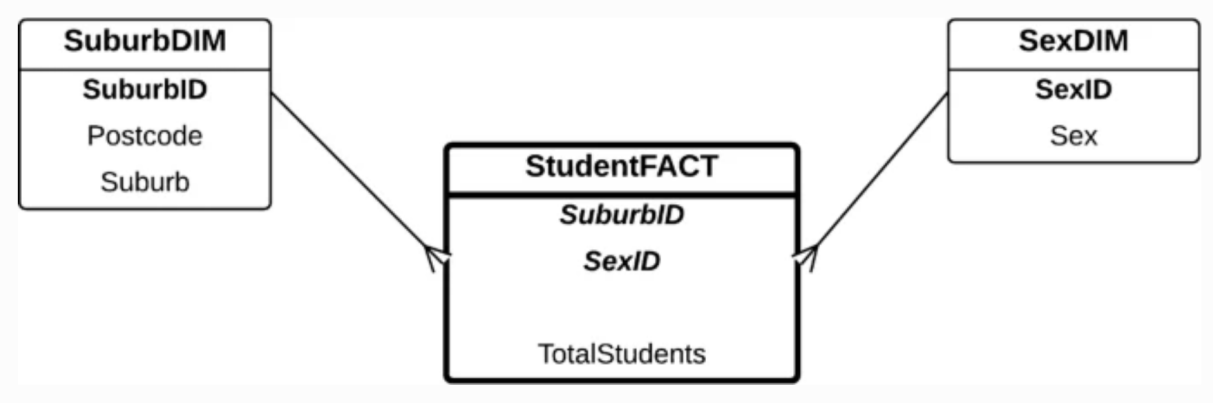
Star Schema without Hierarchy:

- Cons: Not in 3NF, and has more data duplication
- Pros: When producing a report, we need to join the Fact table with Shop Dimension only. Hence, we need one join operation only.

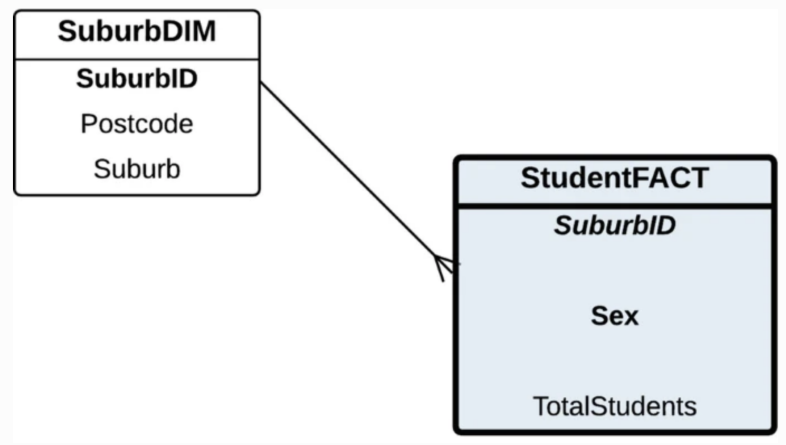
Question 7

Write an SQL statement to display all records from the Fact Table. Do this in two versions: version 1 is the star schema with two dimensions (Suburb Dimension and Sex Dimension) and version 2 is the star schema with a dimension-less key of Sex. Compare and contrast these two versions to retrieve the same information.

Star schema version 1 (with two dimensions)



Star schema version 2 (with dimension-less key of Sex)



Solution:

-- with two dimensions

CREATE TABLE SUBURBDIM AS SELECT DISTINCT SUBURB, POSTCODE FROM STUDENT;

CREATE TABLE SEXDIM AS SELECT DISTINCT SEX FROM STUDENT;

CREATE TABLE FACT AS SELECT SUBURB, SEX FROM STUDENT GROUP BY SUBURB, SEX;

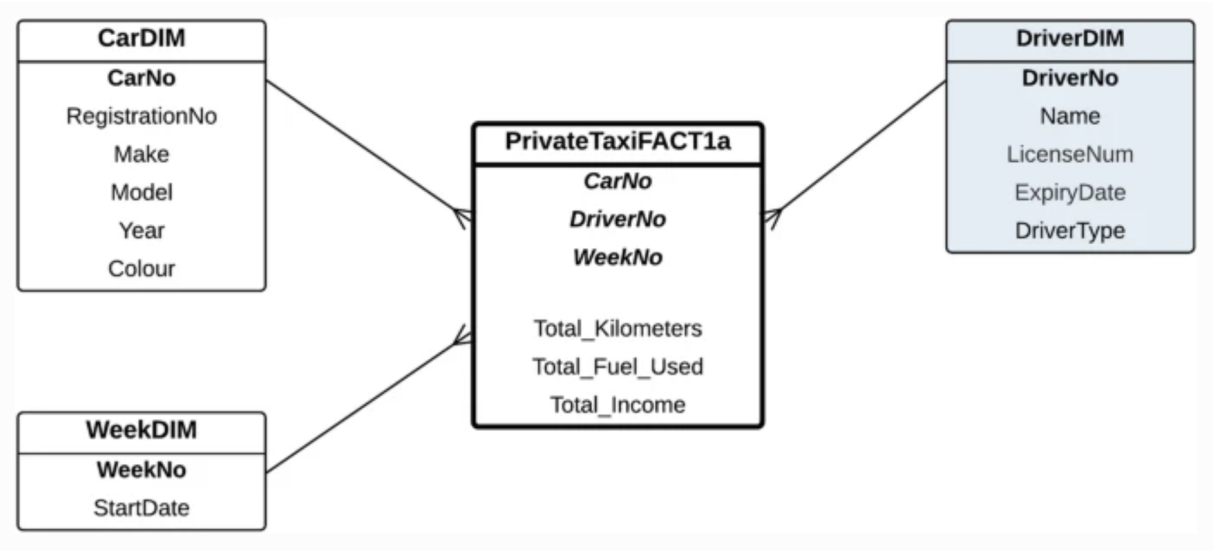
-- with dimension-less key of Sex

CREATE TABLE SUBURBDIM AS SELECT DISTINCT SUBURB, POSTCODE FROM STUDENT;

CREATE TABLE FACT AS SELECT DISTINCT SEX, SUBURB FROM STUDENT GROUP BY SUBURB, SEX;

Question 8

A Taxi company also employs two kinds of drivers, Full-Time drivers and Part-Time drivers. The star schema below shows three dimensions: Car, Week and Driver Dimensions. The Driver Dimension also includes the Driver Type attribute to indicate if the driver is a Full-Time or a Part-Time driver.



Question: Your task is to slice the fact by creating two new star schemas, one focusing on Full-Time drivers and the other on Part-Time drivers, as indicated by the Driver Type attribute in the Driver Dimension.

Solution:

This is a Horizontal Slice based on Driver Type attribute in the Driver Dimension. Two new star schemas are created from this horizontal slice.