National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Data Warehousing and Data Mining**	**Course Code:**	**CS409**
Program:	**BS(Computer Science)**	**Semester:**	**Fall 2018**
Out Date:	**Thu 13-Sep-2018**	**Total Marks:**	
Due Date:	**Tue 25-Sep-2018 *(start of class)***	**Weight:**	
Section	**CS**	**Page(s):**	**10**
Assignment Solution:	**1 (Dimensional Modeling & OLAP)**		

Question 1. Why is the entity-relationship modeling technique not suitable for the data warehouse? How is the dimensional modeling different?

Answer: Refer Page 209-210 of text book (Data warehousing fundamentals by Paulraj Ponniah)

Question 2. Differentiate between fully additive and semi additive measures. Provide Examples.

Answer: Refer Page 215 of text book (Data warehousing fundamentals by Paulraj Ponniah)

Question 3. Discuss the design impact of choosing between star schema and snow flake schema with respect to following factors:

a) Performance of loading data from source to data warehouse (ETL).
b) Performance of querying/analyzing the data.

Answer: https://www.computerweekly.com/answer/Star-schema-vs-snowflake-schema-Which-is-better

Question 4. What will be your design decision?

a) If two departments of the same business use different attributes of the same dimension i.e. both the departments have different definition of the dimension.

Answer: We can create two tables for the same dimension, each table conforming to the requirements of the respective department i.e. each table will have attributes that are relevant to the respective department.

b) If a dimension is used only by one fact table, will you store it directly into the fact table or create a separate dimension table for it?

Answer: We will not store the dimension table directly into the fact table:

(i) if the dimension has a lot of descriptive attributes or
(ii) if the fact table is too large
(iii) Otherwise, we can store the dimension table directly into the fact table to eliminate the cost of joining at run time.

Note: One of the above reasons or any other valid reason will result in full credit.

Question 5. Suppose there are 4000 products sold by the store, 5 brands and each brand has 800 products each, there are 10 store locations in the country, also assume there are at least one sale per product per store per week. Estimate the number of rows of fact table retrieved and summarized for following types of queries:

	Product	Store	Time	# of Rows retrieved
Query 1	1 product	1 store	1 week	
Query 2	1 product	All stores	1 week	
Query 3	1 brand	1 store	1 week	
Query 4	1 brand	All stores	1 year	
Query 5	All brands	All stores	1 year	

For which of the above queries Aggregate fact tables should be used and why?

Also draw the appropriate dimensional model showing aggregate fact tables.

Suppose you created an aggregate fact table for the third query… Then how many rows you need to retrieve for Queries 3, 4 and 5?

Answer:

# of Rows retrieved = number of rows read after applying the given filter e.g. in Query 2, following ten rows will be retrieved after applying the given filter.

Product 1	Store 1	Week1
Product 1	Store 2	Week1
Product 1	Store 3	Week1
Product 1	Store 4	Week1
Product 1	Store 5	Week1
Product 1	Store 6	Week1
Product 1	Store 7	Week1
Product 1	Store 8	Week1
Product 1	Store 9	Week1
Product 1	Store 10	Week1

These rows will be aggregated to give product1's sales across all the ten stores in 1st week. So, the number of rows retrieved in this case is 10.

	Product	Store	Time	# of Rows retrieved
Query 1	1 product	1 store	1 week	1*1*1=1
Query 2	1 product	All stores	1 week	1*10*1=10
Query 3	1 brand	1 store	1 week	800*1*1=800
Query 4	1 brand	All stores	1 year	800*10*52(Weeks per Year)=416,000
Query 5	All brands	All stores	1 year	**Solution: Products =4000, Brands=5** 4,000*10*52= 2,080,000

For which of the above queries Aggregate fact tables should be used and why?

Answer: Aggregate fact tables should be used for queries 3, 4 and 5.Otherwise, we need to retrieve a lot of rows and aggregate them at run time.

Also draw the appropriate dimensional model showing aggregate fact tables.

Will be discussed in lab

Suppose you created an aggregate fact table for the third query… Then how many rows you need to retrieve for Queries 3, 4 and 5?

Query 3	1 brand	1 store	1 week	1*1*1=1
Query 4	1 brand	All stores	1 year	1*10*52=520
Query 5	All brands	All stores	1 year	**Number of brands * Number of Stores * Number of weeks per year =5 * 10 * 52 =2080**

Question 6. You are required to design a Dimensional Model in the way that it fulfills the requirement for the following system.

Grocery System (POS)
The following queries shall be generated through your design:

1. Total sales of a particular product from all stores in the last quarter
2. Total sales by product by store by month
3. Yearly profit generated by stores in the north region
4. How customer deviates from store to store with particular products
5. When I promote one thing how does it affect the other
6. Check if more Products are sold on 1st 10 days and 20th to 25th date of the month than the whole month.
7. Average daily sales (in dollars) of product categories.
8. The total number of customers purchasing a particular product.
9. The total number of customers visiting a particular store in a month.
10. Count how many people buy with coupon.

Design Requirements
Here is the eight points of the complete dimensional modeling design:

1. The processes, and hence the identity of the fact tables
2. The grain of each fact table.
3. The dimensions of each fact table
4. The facts, including pre-calculated facts
5. The dimension attributes with complete descriptions and proper terminology
6. How to track slowly changing dimensions
7. The historical duration of the database
8. The urgency with which the data is extracted and loaded into the data warehouse.

Grocery System (POS):

1. **The processes, and hence the identity of the fact tables**

Following are the major processes in the Grocery System Data ware:

Sales, inventories, Cost, revenue, Buyer (Customer) etc

Identities of Fact Tables:

Base Fact Table:

i) time_key
ii) Product_key

iii) Store_key

Quarterly Agg Fact Table:

i) Product_key
ii) Quarter_key
iii) Store_key

Monthly Aggregate Fact Table:

i) Month_key
ii) Product_key
iii) Store_key

Yearly Agg Fact Table:

i) Product Key
ii) Store Key
iii) Yearly_key
iv) Region_key

Customer Fact Table:

i) Customer key
ii) Product key
iii) Store key
iv) Promo key
v) Time key

2. The grain of each fact table:

The grain of a fact table is the least level of each dimension.

Base Fact Table:

The grain level is the fact measurements by Day by Store and By Product Wise

Quarterly Agg Fact Table:

The grain level is the sales by product by quarter by All store wise.

Monthly Aggregate Fact Table:

The grain of this table is Total sales by product by store by month

Yearly Agg Fact Table:

The gain level of this fact table is sales by year by product by Region wise.

Customer Fact Table:

The grain level of this fact table is by day ,by store by customer by promotion wise.

3. The dimensions of each fact table:

Base Fact Table:

Time, Product, Store

Quarterly Agg Fact Table:

Quarter, Product, Store

Yearly Agg Fact Table:

Year, Product, Region

Customer Fact Table:

Time, Product, Store, Promotion, Customer

4. The facts, including pre-calculated facts:

The major facts and pre calculated facts are

i) Quantity_sold
ii) Dollar profit
iii) Cost
iv) Average_sales
v) Quartely_qauntiy_sold
vi) Quartely_cost
vii) Yearly_profit
viii) Customer Count
ix) Product Count
x) Product Quantity_per Customer

5. The dimension attributes with complete descriptions and proper terminology:

i) Time Dimension

Time_key It is the primary key of the time Dimension

Day Is keeps the day info

Week_key It stores the primary key of the week dimension

Day_one_to_ten It keeps dates from 1st to 10th

Day_twenty_to_twentyfive It keeps dates from 20 to 25

ii) Product Dimension

Product_key it is the primary key of the product dimension

SKU_number It is the SKU number of the production

SKU_Desc It is the Description of the SKU

Package_size It keeps the size of package of the product.

Brand It stores the Brand of the product

Subcategory IT stores the subcategory of the product.

Category

Department

Diet It keeps the diet info of the product.

Weight_type It is the weight of the product

Cases_per_pallet It keeps cases per pallet.

iii) Store Dimension

Store_number it stores the store number of the store

Store_key It is the primary key of the store dimension

Store_address it keeps the address of the Store

City

Country

State

Zip

Manager

Phone

Fax

Floor_plan_type It stores the current floor plan type of the store

Region_key It stores the Region key of the Region level dimension

iv) Customer Dimension:

Customer_key It is the primary key of the customer dimension

Customer_SSN It keeps the SSN no of the customer for tracking

Customer_name It keeps the name of the customer

Customer_phone_no

Customer_region

v) Promotions Dimension

Promotion_key It is the primary key of the promotion Dimension

Promotion_name It is the name of the promotion

Price_reduction_type

Ad_type It stores the advertisement type of the promotion

Display_type It stores the display type of the promotion

Cupon_type It stores the coupon types offered in the promotion

Ad_media_name It stores the media name of the advertisement

Display_provider It stores the provider of the displayer.

Prmo_cost It keeps the cost of the promotion

Promo_begin_date It keeps the start date of promotion

Promo_end_date It keeps the end date of promotion

vi) Quarter Dimension:

Quarter_key it is the key of the Quarter Dimension

Quarter

Year_key

vii) Year Dimension

Year_key

Year

viii) Region Dimension

Region_key

Region

ix) Week Dimension

Week_key

Week

Month_key

x) Month Dimension

Month_key

Month

Quarter_key

6. How to track slowly changing dimensions:

We will track slowly changing Dimensions by Type TWO approach where we generate a new account record every time a meaniningful account attribute changes.

7. The historical duration of the database:

The historical duration of database is 7 to 8 years approximately however it varies according to the type of Database and under certain requirements and constraints.

8. The urgency with which the data is extracted and loaded into the data warehouse:

The urgency with which the data is extracted and loaded depends upon your assumptions, your answer will be marked correct given your reason is right.

Question 7: Give brief answers to the following questions.

1. What is the difference between OLAP and OLTP systems? Give examples.

Answer: Refer Page 348 of text book (Data warehousing fundamentals by Paulraj Ponniah)

2. Briefly explain multidimensional analysis.

Answer: Refer Page 362 of text book (Data warehousing fundamentals by Paulraj Ponniah)

3. What are hypercubes? How do they apply in an OLAP system?

Answer: Refer Page 375 of text book (Data warehousing fundamentals by Paulraj Ponniah)

4. Briefly explain OLAP models.

Answer: Refer Page 381 of text book (Data warehousing fundamentals by Paulraj Ponniah)

5. What does it mean by horizontal and vertical partitioning of a cube? Be clear and concise.

Answer: Refer Page 465 (Paragraph 2) of text book (Data warehousing fundamentals by Paulraj Ponniah)

6. Explain the two specific goals for horizontal partitioning.

Answer: Refer Page 465 of text book (Data warehousing fundamentals by Paulraj Ponniah)

7. Difference between Drill down and roll-up.

Answer: Refer Page 378 of text book (Data warehousing fundamentals by Paulraj Ponniah)

8. What is Slice and dice?

Answer: Refer Page 380 of text book (Data warehousing fundamentals by Paulraj Ponniah)

Question 8: Case Study (Be clear and concise)

Zunair, a manager in the I.T. department in his company, is trying to explain to Ahmad, the head of marketing, why the company has two major information systems, an ERP OLTP system and a data warehouse OLAP system. Write an explanation for Zunair that he can give to Ahmad that discusses why these two systems are necessary and then explain to Ahmad how his department should use each of the systems. It would be helpful to Ahmad to cite marketing examples of the use to each of these systems.

Answer: OLAP vs OLTP. Benefits of each of them.