DATAWAREHOUSE ASSESSEENT

[Pavithra Nagaral]

1)

a. How many dimensions and Facts are present?

Ans: Dimensions - 6

Fact -1

b. Please identify the cardinality between each table?

Ans:

Tables	Cardinality
Sales:Customer	N:1
Sales:Store	N:1
Sales:Products	N:1
Sales:Time	N:1
Time:Month	N:1
Year:Month	1:N

c. How to create a Sales_Aggr fact using the following structure (SQL Statement):

Ans: Create table Sales_Aggr(Year_ID int, Customer_key int, Store_key, Product_key, dollar,

foreign key(year_ID) references Year(Yearkey),

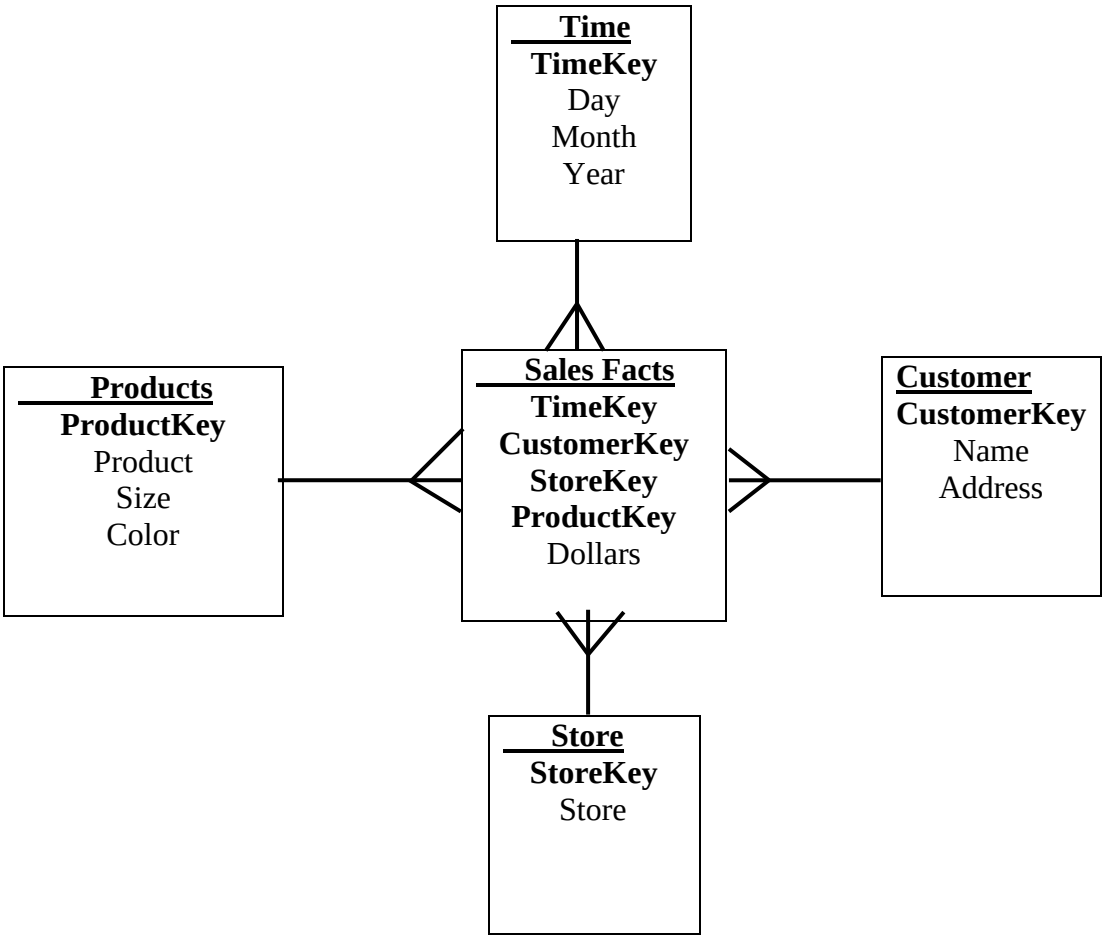
foreign key (Customer_key) references Customer(CustomerKey),

foreign key (Store_key) references Store(Storekey),

foreignkey(Product_key) references Products(Productkey));

d. Can you Please Modify the above snowflake schema to Star schema and draw the dimension model, showing all the cardinality?

Ans:



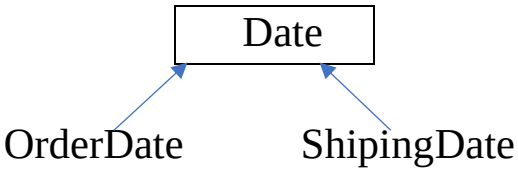
2) For the following dimension Model can you please give an example of Circular Join and how to avoid it:

Ans: Circular join occurs when attribute in one table referring to the two attributes Of another table. That is inter connection of contents of different tables.

In the given case, we have date table and sales table. In sales table there are Order date and Shipping date referring to the same date of date table.

Date table:

Date	Month	MonthNumber
1/12/2019	Dec	12
5/12/2019	Dec	12



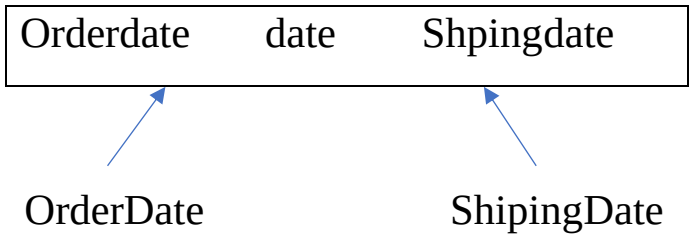
Sales table:

OrderDate	ShipingDate	SalesAmount
1/12/2019	5/12/2019	10000

To avoid this circular join , we can use alias name, that is in date table

We can differentiate order date and shipping date by giving alias name from

Which we can separate the both tables.



Query to remove circular join by using alias,

```
Select s.OrderDate , s.ShippingDate
from date as “order_date”, date as “shiping_date”
where order_date.date = s.OrderDate AND  shiping_date.date= s.ShippingDate
```

3) For the given Dimension Model, can you please generate a sql to get the total divergence between Quantity sold and Quantity Forecast for the current month for all the stores:

Query:

```
SELECT  SUM (F.QUANTITY_FORECAST ) – SUM(S.QUANTITY_SOLD)
AS DIVERGENCE
FROM  DAILY_FORECAST AS F, DAILY_SALES AS S
WHERE  F.PERKEY = (SELECT PERKEY FROM PERIOD
                        WHERE MONTH= TO_CHAR(SYSDATE,'MM')
                        AND
                         (S.PERKEY=  (SELECT PERKEY FROM PERIOD
                        WHERE MONTH= TO_CHAR(SYSDATE,'MM') );
```

4) For the mentioned dimension model, please identify the conformed and nonconformed dimensions. Additionally, identify the measure types?

Ans:

Conformed dimensions - store, period, product.

[Conformed dimensions means two fact tables sharing common contents. In the given table, store period and product dimensions are common between daily_sales and daily_fact tables]

Non conformed dimensions – customer and promotion.

[Non conformed dimensions are not related any two fact tables. In the given table, customer and promotion are not common between between daily_sales and daily_fact tables]

Measure types:

Measures	Measure Types
QUANTITY_SOLD	Additive
EXTENDED_PRICE	Semi Additive
EXTENDED_COST	Additive
QUANTITY_SOLD_ETENDED	Additive
EXTENDED_PRICE_FORECAST	Semi Additive
EXTENDED_COST_FORECAST	Additive

5) Make a list of differences between DW and OLTP based on Size, Usage, Processing and Data Models.

Ans: DW – DataWare house.

Technique for collecting and managing data from various sources , which helps in management decision making process.

OLTP – Online Transaction Process.

Original source of data which control and runs fundamental business tasks.

	DW	**OLTP**
Size	Large [100GB- 2TB] Due to aggregation structures and historical data	Small [10MB – 10GB] Due to the absence of historical data
Usage	Used in planning, problem Solving and support to decision making process.	Used to support business transaction that is to control and run fundamental business tasks.
Processing	Depends on the amount of data involved. Transaction processing.	Typically very fast. Query processing.
Data Models	Dimension modeling.	E-R modelling

Part B

1) a) Category of a product may change over a period of time. Historical category information (current category as well as all old categories) has to be stored. Which SCD type will be sutiable to implement this requirement? What kind of structure changes are required in a dimenstion table to implement SCD type 2 and type 3.

Ans:

SCD type 2 should be used.

Because in this complete history is maintained.

Consider an example of 2 employee, e1 and e2 and their job is SE and SSE respectively. And after 1 day if e1 job is changed to TL then after 2 day again job of e1 changed to MGR .

In SCD type 2 complete history is maintained, that is it will store the

All changes of e1.

Full load SCD type 2 :

Emp no	Emp name	Job	St Date	End Date
01	E1	SE	D1	Null
02	E2	SSE	D1	Null

SCD Type 2:

Emp no	Emp name	Job	St Date	End Date
01	E1	SE	D1	D2
02	E2	SSE	D1	Null
01	E1	TL	D2	D3
01	E1	MGR	D3	Null

In SCD type 3 immediate history is maintained, that is e1 has changed job twice but it will store only the previous job details.

Full load SCD type 3:

Emp no	Emp name	Job	Previous job
01	E1	SE	Null
02	E2	SSE	Null

SCD Type 3:

Emp no	Emp name	Job	St Date
01	E1	TL	D2
02	E2	SSE	D1
01	E1	MGR	Null

2) What is surrogate key? Why it is required?

Ans: Surrogate Key is sequential system generated unique number attached with each and every record in a table in any Data Warehouse. It is basically used to avoid the ambiguity in selecting the particular row.

If we take the example of above SCD type 2, the e1 job is changed twice but If we want to select row of job TL it will create confusion because emp no will be 1 for both the cases, so we use surrogate key.

Sk	Emp no	Emp name	Job	St Date	End Date
01	01	E1	SE	D1	D2
02	02	E2	SSE	D1	Null
03	01	E1	TL	D2	D3
04	01	E1	MGR	D3	Null

- Surrogate key is used to handle Slowly changing dimensions(SCD).
- Joins will be faster as it is number to number to join.
- Key is independent of data contained in the table so does not effect anywhere.
- Instead of using huge natural keys , storing of concise SKs would result in less amount of space needed.

3) Stores are grouped in to multiple clusters. A store can be part of one or more clusters. Design tables to store this store-cluster mapping information.

Ans: Cluster is a process of grouping similar kind of objects. Here stores are grouped into clusters, but a store can be a part of another cluster also.

Let us consider two location l1 and l2 , Based on the those locations We are making two clusters c1 and c2.

Cluster_01:

Store_id	Store_name	Loc_id	Loc_name
01	S1	01	L1
02	S2	01	L1

Cluster_02:

Store_id	Store_name	Loc_id	Loc_name
03	S3	02	L2
02	S2	02	L2

These two clusters are mapped into the single table as follows,

Cl_id	Cl_name	Store_id	Store_name	Loc_id	Loc_name
01	C1	01	S1	01	L1
01	C1	02	S2	01	L1
02	C2	03	S3	02	L2
02	C2	02	S2	02	L2

4) What is a semi-additive measure? Give an example.

Ans: Semi Additive measures are values that you can summarise across any related dimension except time. That is it can not be summed up based on the time.

Examles: stocks

Date	Stock name	Stocks
01/12	NASD:AMZN	100
02/12	NASD:AMZN	50

Here in this example, if you had 100 in stock yesterday, and 50 in stock today, you're total stock is 50, not 150. It doesn't make sense to add up the measures over time, you need to find the most recent value.

So it is called semi additive measure where one can not summed up the values over the time.