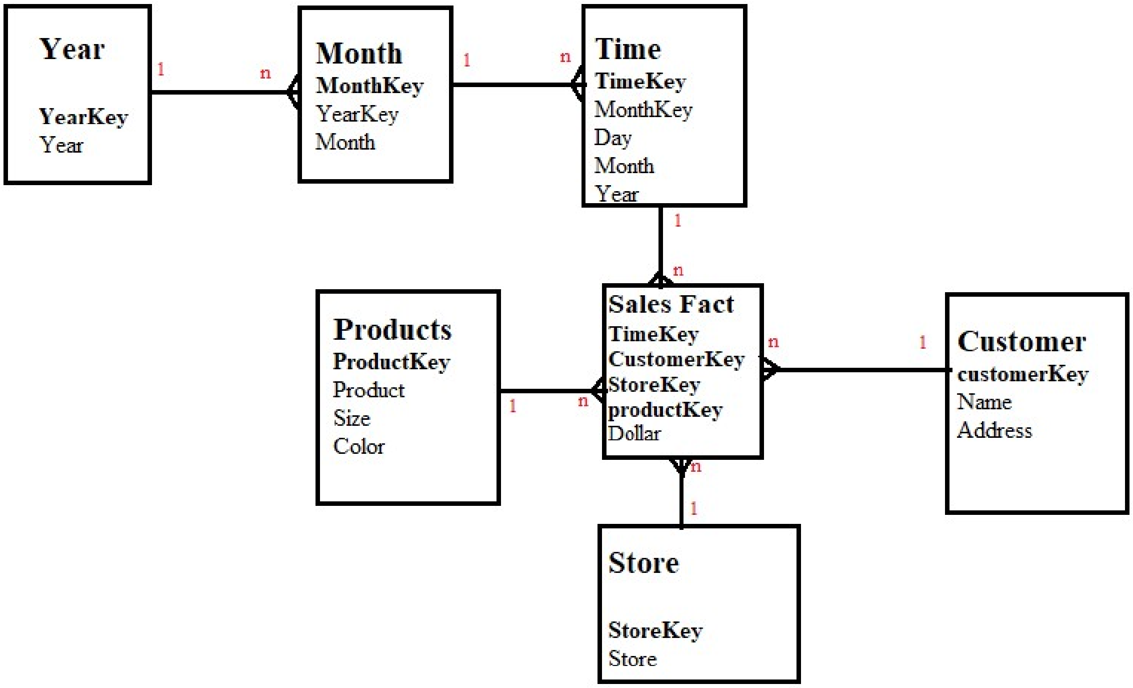
ASSESSMENT PART 1

1. For the given Dimensional Modeling, please identify the following:

- How many dimensions and Facts are present?

Above given schema is called snowflakes schema. We have **sales Facts** as a fact table in this schema and **six dimension tables** where two tables (year and Month) are normalized dimension tables.

- Please identify the cardinality between each table?



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

CREATE TABLE sales_aggr (

YearKey as YearID int(10) FOREIGN KEY REFERENCES Year(YearKey),

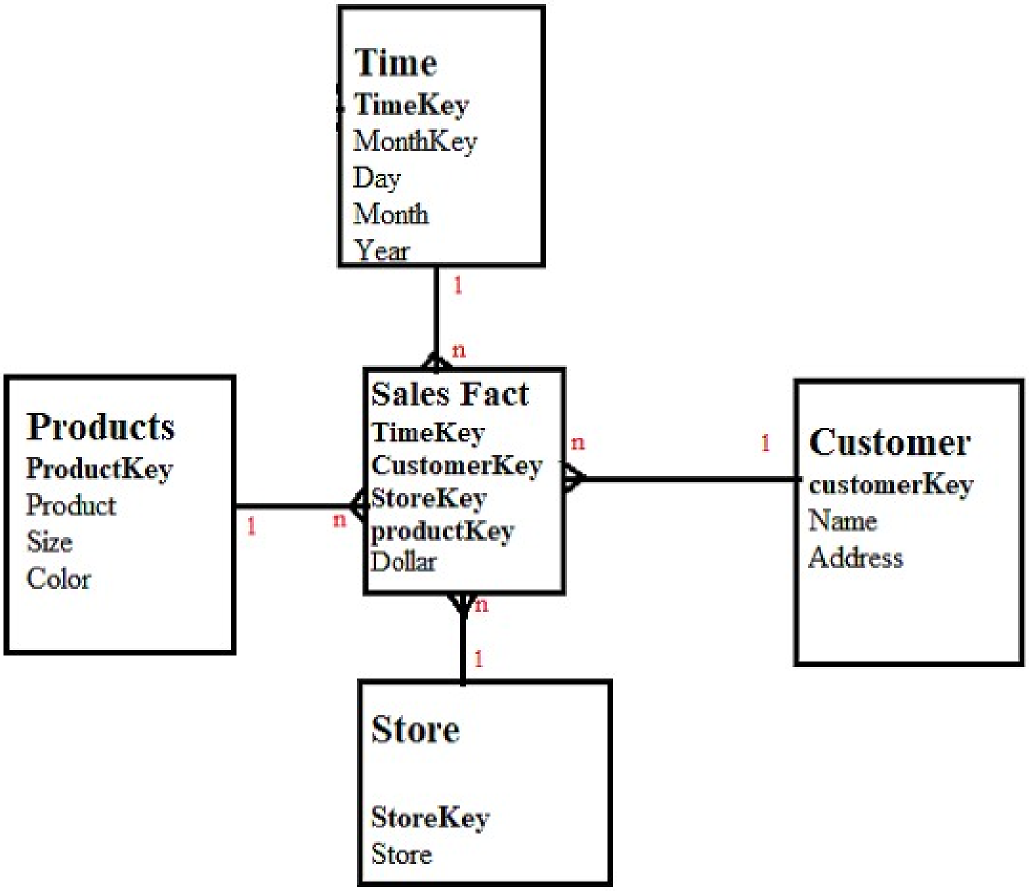
Customer_Key int(10) FOREIGN KEY REFERENCES Customer(CustomerKey),

Store_key int(10) FOREIGN KEY REFERENCES Store(Storekey),

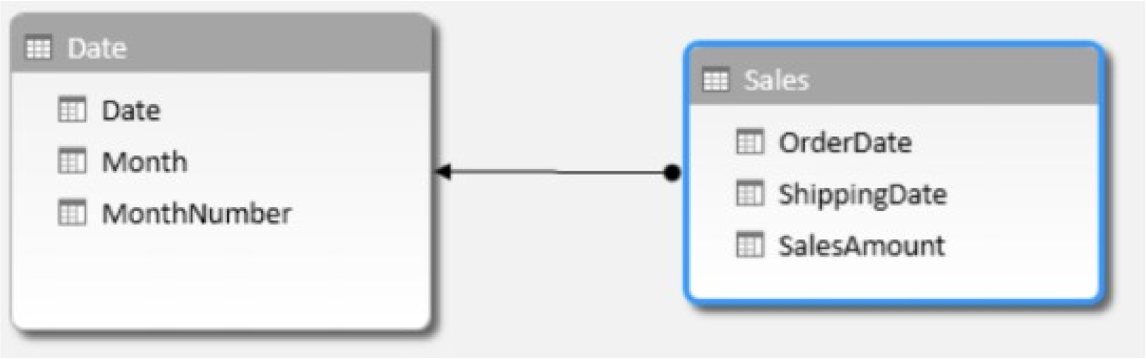
Product_key int(10) FOREIGN KEY REFERENCES Product(ProductKey), Dollar double,

PRIMARY KEY (YearID, Customer_Key, Store_key, Product_key)

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



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



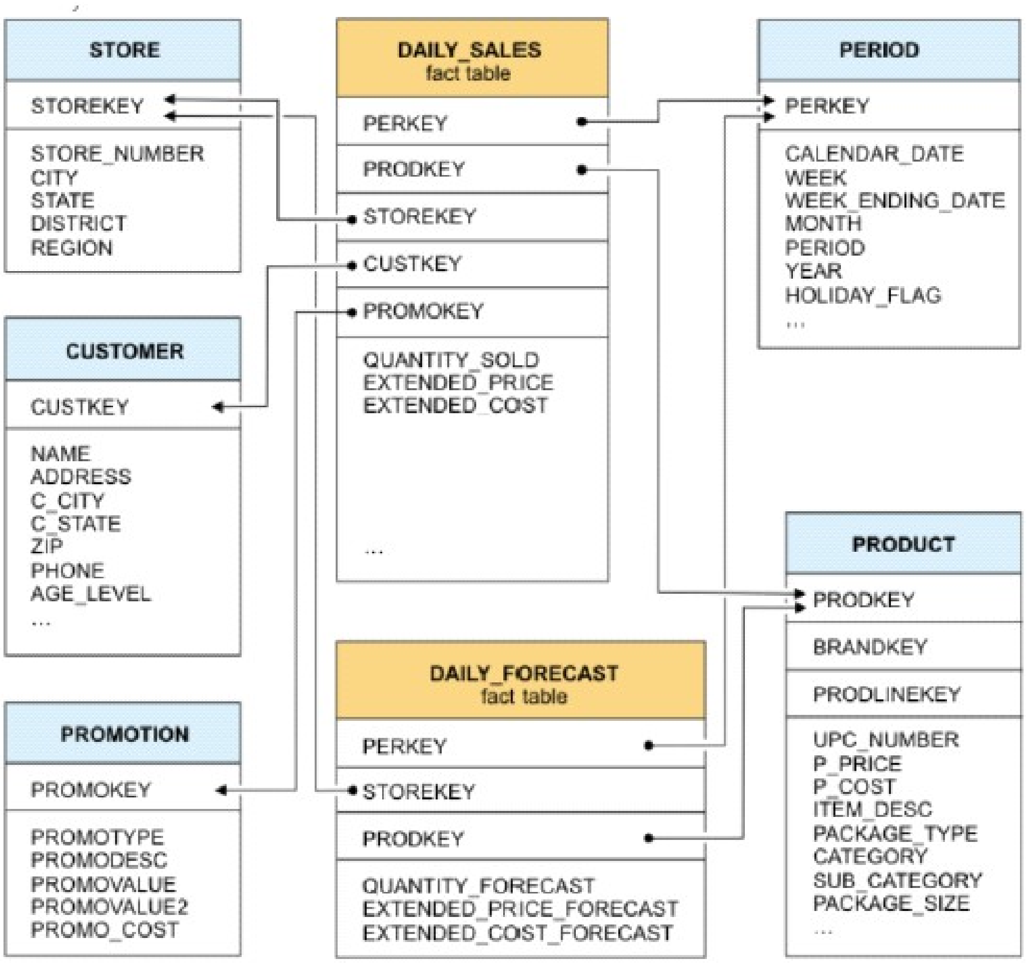
The following query will create circular join

```
SELECT S.order_date, s.shipping_date
FROM Date d, Sales s WHERE d.date=s.OrderDate AND d.date=s.ShippingDate;
```

We can avoid this by giving two alias name to date attribute as

```
SELECT  S.order_date, S.shipping_date
FROM Date AS order_date,
Date AS shipping_date,
Sales AS S
WHERE
order_date.date=S.OrderDate AND
shipping_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:



```
SELECT (sum(s.quantity_sold)-sum(f.quantity_forecast)) AS divergent

FROM store st, daily_Forecast f INNER JOIN daily_sales s ON s.perkey=f.perkey

INNER JOIN period p ON s.perkey=p.perkey

WHERE p.month=to_char(sysdate, 'mm');
```

4. For the above-mentioned dimension model, please identify the conformed and non-conformed dimensions. Additionally, identify the measure types?

We have three confirmed dimensions as product dimensions, period dimension and store dimensions.

Non-confirm dimensions are promotion dimension table and Customer dimension table

extended_price_forecast, extended_cost_forecast, extended_price, extended_cost have semi-additive type of measures.

quantity_forecast, quantity_sold are additive type of measures.

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

	DATAWAREHOUSE	**OLTP**
Size:	Large amount of data is stored here	Comparatively less amount of data store
Usage:	Help in business analysis, and runs fundamental business task	Helps in fast transaction, maintains data integrity in multiple environment.
Processing:	Depends on complex queries and as data get refreshed every interval so complex query may take little time.	Typically very fast
Data Models:	De-normalized with few tables creating star and/or snowflakes schema.	Normalized with many tables

Assessment Part 2

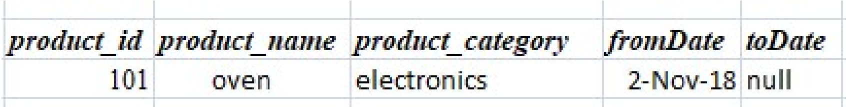
- Category of a product may change over a period of time. Historical category information (current category as well as all old categories) has to be stored. Which SCD type will be suitable to implement this requirement? What kind of structure changes are required in a dimension table to implement SCD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

We can either use SCD type 2 or SCD type 3 based on storing limited history of data or full history of data. If we choose to use SCD type 2, that means we can store the full history of old categories of products. If SCD type 3 is used, only limited history say last 3 changes in categories of product has been stored as history.

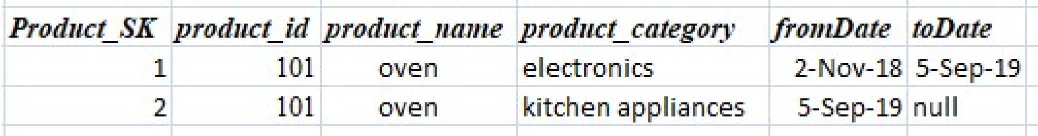
In SCD type 2, we can add columns like “fromDate”, “toDate” and surrogate key .

fromDate and toDate can specify the time period when specific product had particular category.

Once the category is modified, in previous row with older category will be updated with date modified value in toDate column and a new row is generated with new category with fromDate (as when category was changed to new category) and toDate is assigned as null until next change in category.

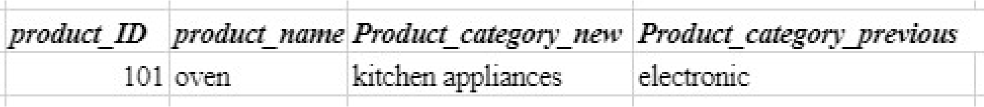


product_id	product_name	product_category	fromDate	toDate
101	oven	electronics	2-Nov-18	null



Product_SK	product_id	product_name	product_category	fromDate	toDate
1	101	oven	electronics	2-Nov-18	5-Sep-19
2	101	oven	kitchen appliances	5-Sep-19	null

In SCD type 3, we can add columns like "previous_category" and "new_category" in dimension table, which will have only the immediate history as previous_category.



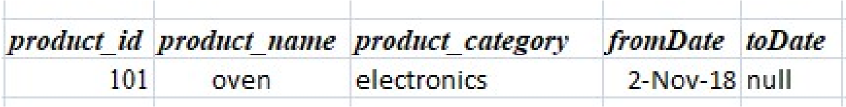
product_ID	product_name	Product_category_new	Product_category_previous
101	oven	kitchen appliances	electronic

- What is surrogate key? Why it is required?

Surrogate key is sequentially generated key (can also be auto generated) attached with each and every record. They don't have major meaning behind (since it doesn't carry any business meaning regarding records its attached to)

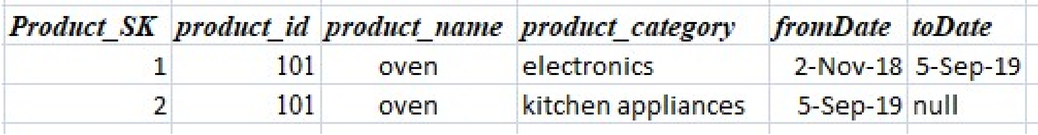
It's an artificial key used as substitute for natural key. For example, in above question given scenario, considering product_id as a primary key, every time there is a change in a category of specific product_id, a new row is generated with new category, implies in a single table we have duplicate product_id so we use surrogate key as substitute of product_id to determine the change in category.

Initially



product_id	product_name	product_category	fromDate	toDate
101	oven	electronics	2-Nov-18	null

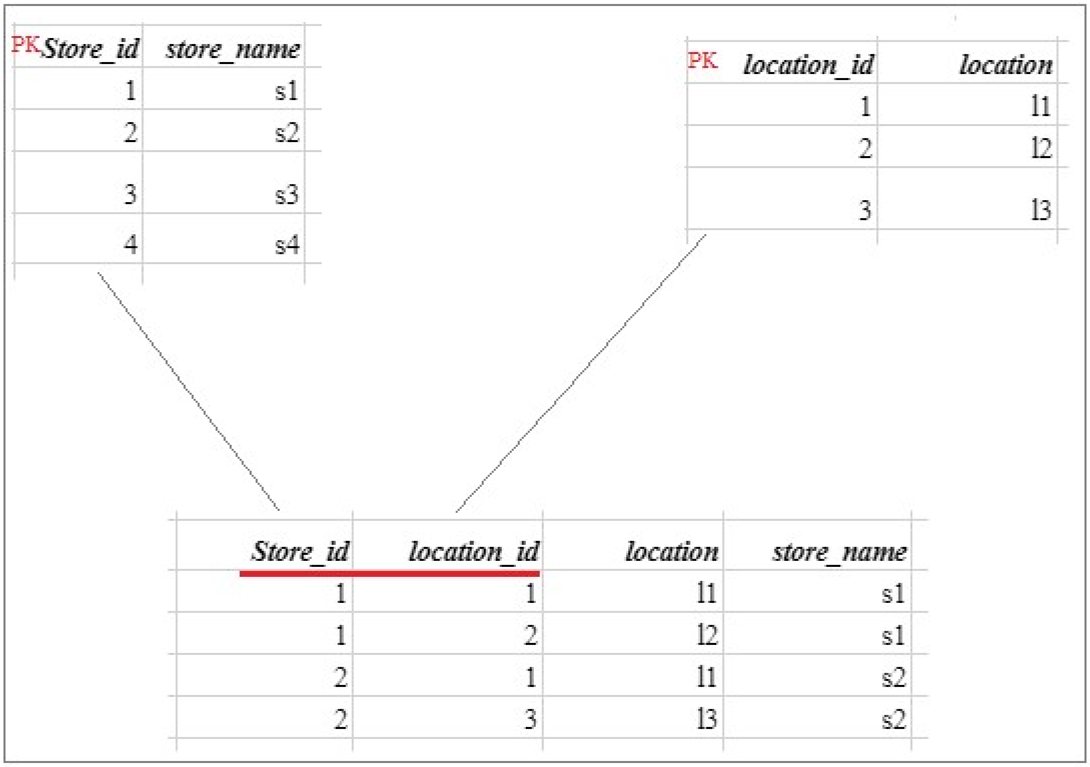
After category changed



Product_SK	product_id	product_name	product_category	fromDate	toDate
1	101	oven	electronics	2-Nov-18	5-Sep-19
2	101	oven	kitchen appliances	5-Sep-19	null

Here product_SK is taken as surrogate key

- 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.



PK Store_id	store_name
1	s1
2	s2
3	s3
4	s4

PK location_id	location
1	l1
2	l2
3	l3

Store_id	location_id	location	store_name
1	1	l1	s1
1	2	l2	s1
2	1	l1	s2
2	3	l3	s2

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

Semi-additive measures values are those which can be summarized across any related dimention except time variant.

For example measures sales yesterday might be 50 and today sales is 100, wwe can say total sales is 150. So this is Additive measures. Stock yesterday was 50 and stock today is 100, we can say stock is 100 not 150 because addong stock results makes no-sense, so we can call this as semi-additive.