1) 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.

SCD type 2 will be suitable for storing data about current category as well as historical data. Because SCD type 2 retains the full history of values. When the value of a chosen attribute changes, the current record is closed. A new record is created with the changed data values and this new record becomes the current record. Each record contains the effective time and expiration time to identify the time period between which the record was active.

- Structural changes required in a dimension table to implement SCD type 2:

Dimension

STATUS_ID	NAME	AADHAAR_NO
1	Rachel	4685 5434 4566
2	Monica	4388 0326 0793

SCD Type 2:

STATUS_ ID	NAME	AADHAAR_NO	START_DATE	END_DATE
1	Rachel	4685 5434 4566	01-01-1987	15-06-2019
2	Monica	4388 0326 0793	01-01-2019	
3	Rachel Green	4685 5434 4566	16-06-2019	

In the above example candidate name was Rachel from 01-01-1987 to 15-06-2019 then name got changed to Rachel green from 16-06-2019 all this data is stored in the SCD type 2 table.

- Structural changes required in a dimension table to implement SCD type 3:
  SCD type 3

A Type 3 SCD stores two versions of values for certain selected level attributes. Each record stores the previous value and the current value of the selected attribute. When the value of any of the selected attributes changes, the current value is stored as the old value and the new value becomes the current value.

STATUS_ ID	NAME	UPDATED_NAME	AADHAAR_NO
1	Rachel	Rachel Green	4685 5434 4566
2	Monica		4388 0326 0793

In the above example only limited data is stored like only the candidate's updated name is stored other details like date is not stored.

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

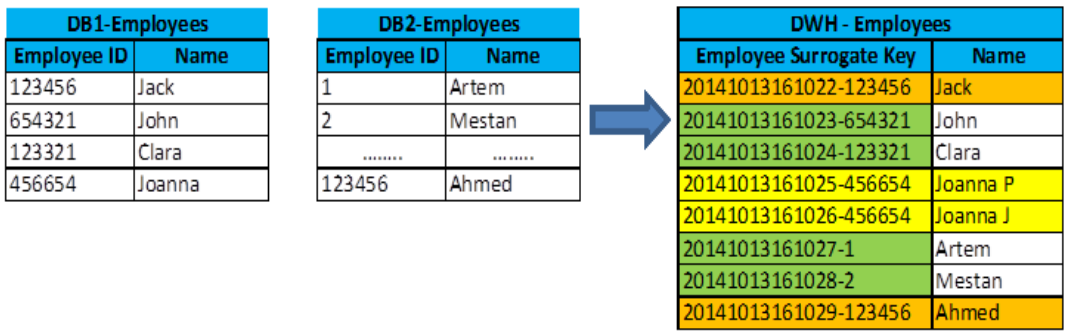
A surrogate key is a key which does not have any contextual or business meaning. It is manufactured "artificially" and only for the purposes of data analysis. The most frequently used version of a surrogate key is an increasing sequential integer or "counter" value (i.e. 1, 2, 3). Surrogate keys can also include the current system date/time stamp, or a random alphanumeric string.

- **Surrogate keys are unique -** surrogate keys are system-generated, it is impossible for the system to create and store a duplicate value.

- **Surrogate keys apply uniform rules to all records**-The surrogate key value is the result of a program, which creates the system-generated value. Any key created as a result of a program will apply uniform rules for each record.

- **Surrogate keys allow for unlimited values-**Sequential, timestamp, and random keys have no practical limits to unique combinations.

- **Surrogate keys stand the test of time-** Because surrogate keys lack any context or business meaning, there will be no need to change the key in the future.



DB1-Employees	
Employee ID	Name
123456	Jack
654321	John
123321	Clara
456654	Joanna

DB2-Employees	
Employee ID	Name
1	Artem
2	Mestan
……	……
123456	Ahmed

DWH - Employees	
Employee Surrogate Key	Name
20141013161022-123456	Jack
20141013161023-654321	John
20141013161024-123321	Clara
20141013161025-456654	Joanna P
20141013161026-456654	Joanna J
20141013161027-1	Artem
20141013161028-2	Mestan
20141013161029-123456	Ahmed

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.

Store_Id	Store_Name
1	Max
2	Trends
3	W
4	Ajio

Cluster_Id	Cluster_Name
1	A
2	B
3	C
4	D

Store_Id	Store_Name	Cluster_Id	Cluster_Name
1	Max	1	A
2	Trends	2	B
3	W	3	C
4	Ajio	4	D
4	Ajio	3	C

Store_Id	Store_Name
1	Max
2	Trends
3	W
4	Ajio

Cluster_Id	Cluster_Name
1	A
2	B
3	C
4	D

Store_ID	Cluster_ID
1	1
2	2
3	3
4	3
4	4

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

Semi Additive measures are values that you can summarise across any related dimension except time.

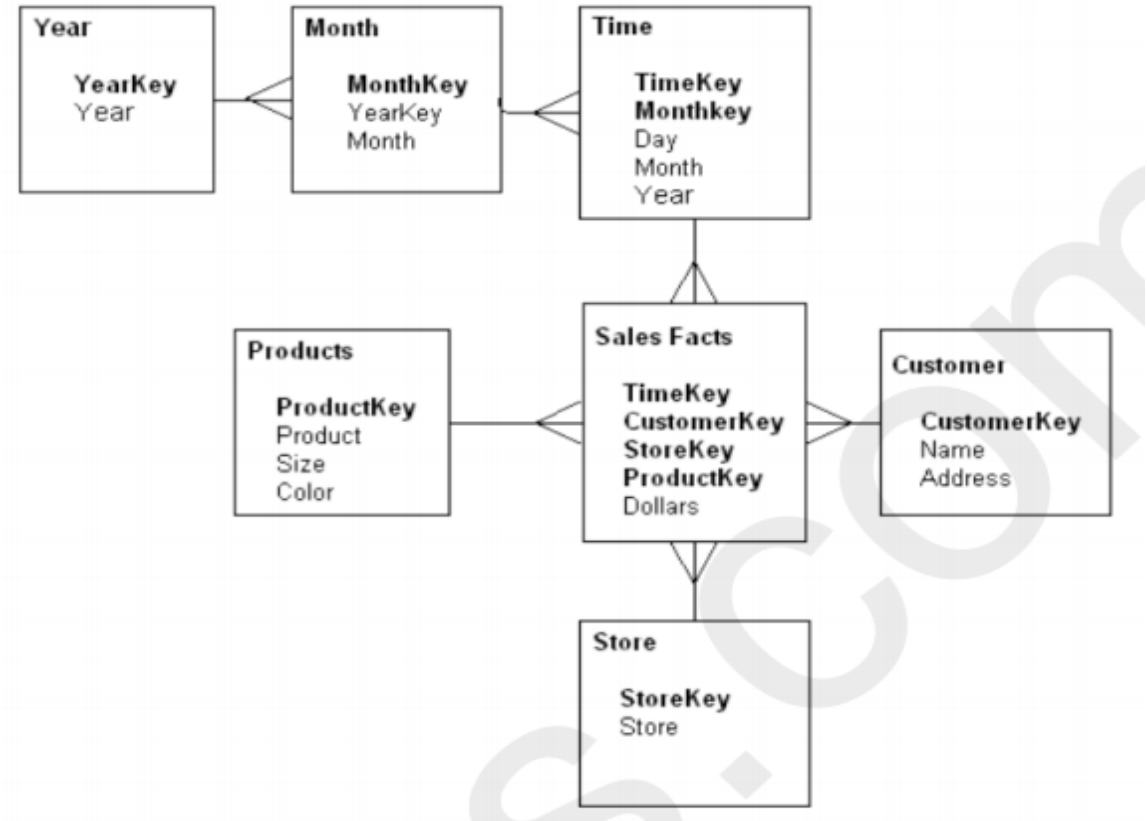
For example, Sales and costs are fully additive; if you sell 100 products and 50 products then you've sold 150 in total. You can add them up over time.

Stock levels however are semi additive; if you had 1000 products in stock yesterday, and 500 in stock today, you're total stock is 500, not 1500. It doesn't make sense to add up the measures over time, you need to find the most recent value.

Semi-additive measures

DateKey	ProductKey	StockCount
20120101	25	23
20120101	99	118
20120102	25	22

For the given Dimensional Modelling, please identify the following:



1a) How many dimensions and F

acts are present?

1 Fact Table -> Sales Facts.

6 Dimension Tables -> Time, Customer, Products, Store, Time, Month, Year.

1b) Please identify the cardinality between each table?

Between Store and Sales Facts -> one-to-many

Between Customer and Sales Facts -> one-to-many

Between Products and Sales Facts -> one-to-many

Between Time and Sales Facts -> one-to-many

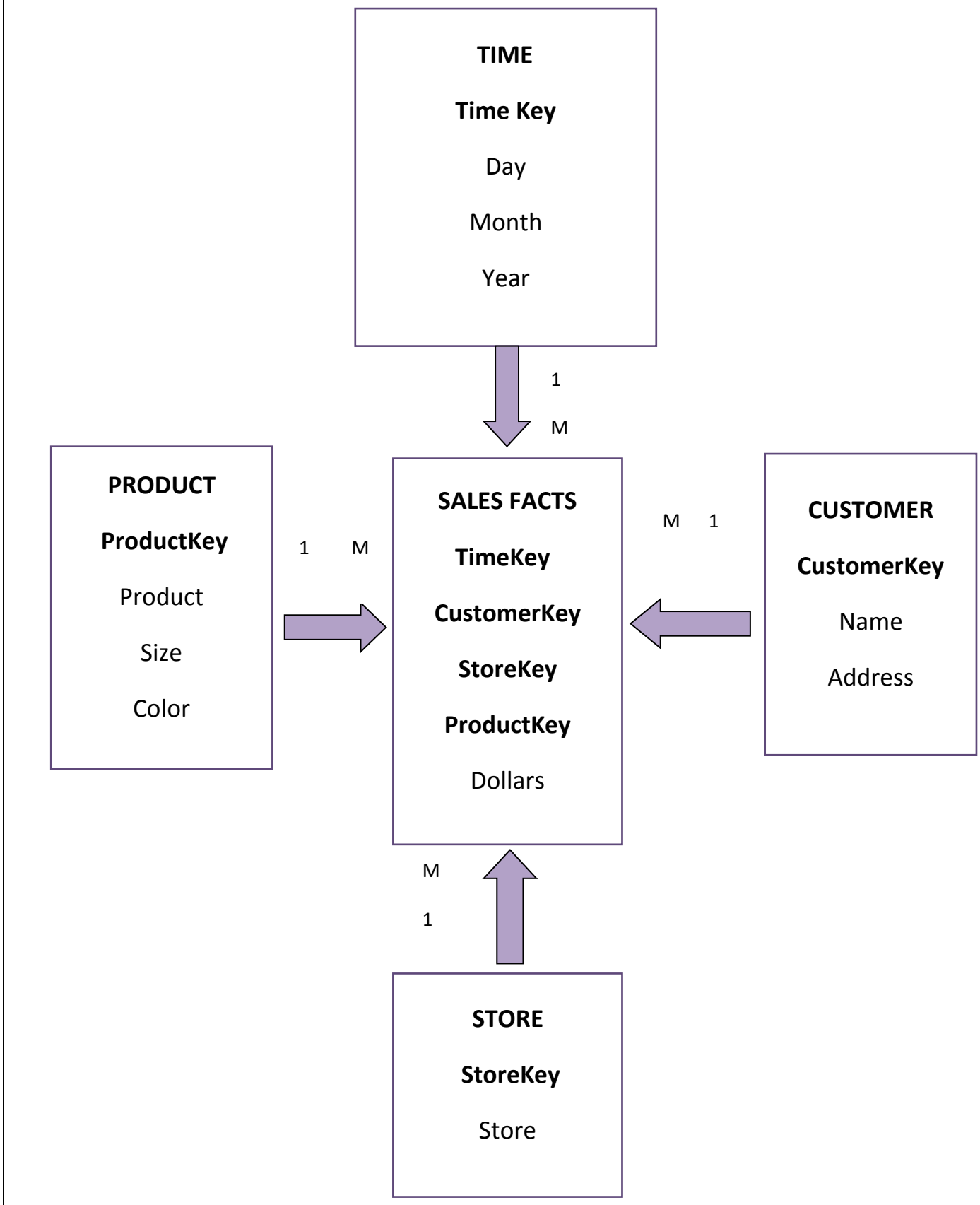
Between Month and Time -> one-to-many

Between Year and Month -> one-to-many

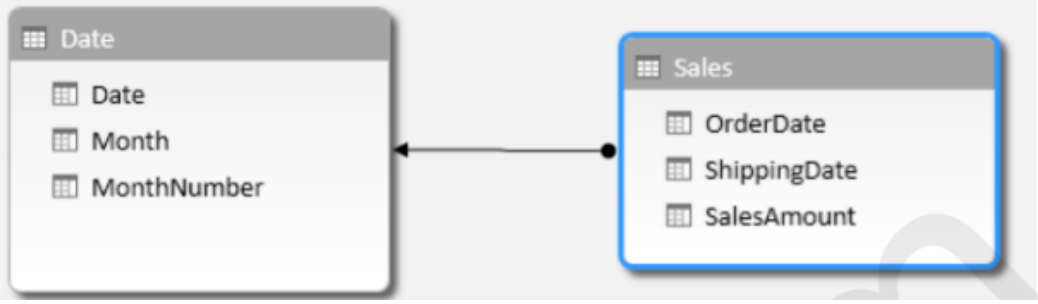
1c) How to create a Sales_Aggr fact using the following structure (SQL Statement):

```
Create table Sales_Aggr

(Year_ID  year  not  null,

Customer_key  Int not null,

Store_Key Int  not null,

Product_key Int not null,

Dollars Double,

Foreign Key (Year_ID) REFERENCES      Year(Year_ID),

Foreign Key (Customer_key) REFERENCES Customer(Customer_Key),

Foreign Key (Store_Key) REFERENCES  Store(Store_Key),

Foreign Key (Product_Key) REFERENCES Product(Product_Key));
```

1d)Can you Please Modify the above snowflake schema to Star schema and draw the dimension model, showing all the cardinality?


TIME
Time Key
Day
Month
Year
1
M
PRODUCT
ProductKey
Product
Size
Color
1
M
SALES FACTS
TimeKey
CustomerKey
StoreKey
ProductKey
Dollars
M
1
CUSTOMER
CustomerKey
Name
Address
M
1
STORE
StoreKey
Store

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



Circular Join:

Circular Joins occur when a table A is joined to table B and in turn joined to table A. So the loops should be avoided.

Select Avg(SalesAmount) from Sales, Date

From Sales , Date

Where Sales.OrderDate = Date.Date,

Sales.ShippingDate = Date.Date;

To avoid circular join, we can make use of alias name.

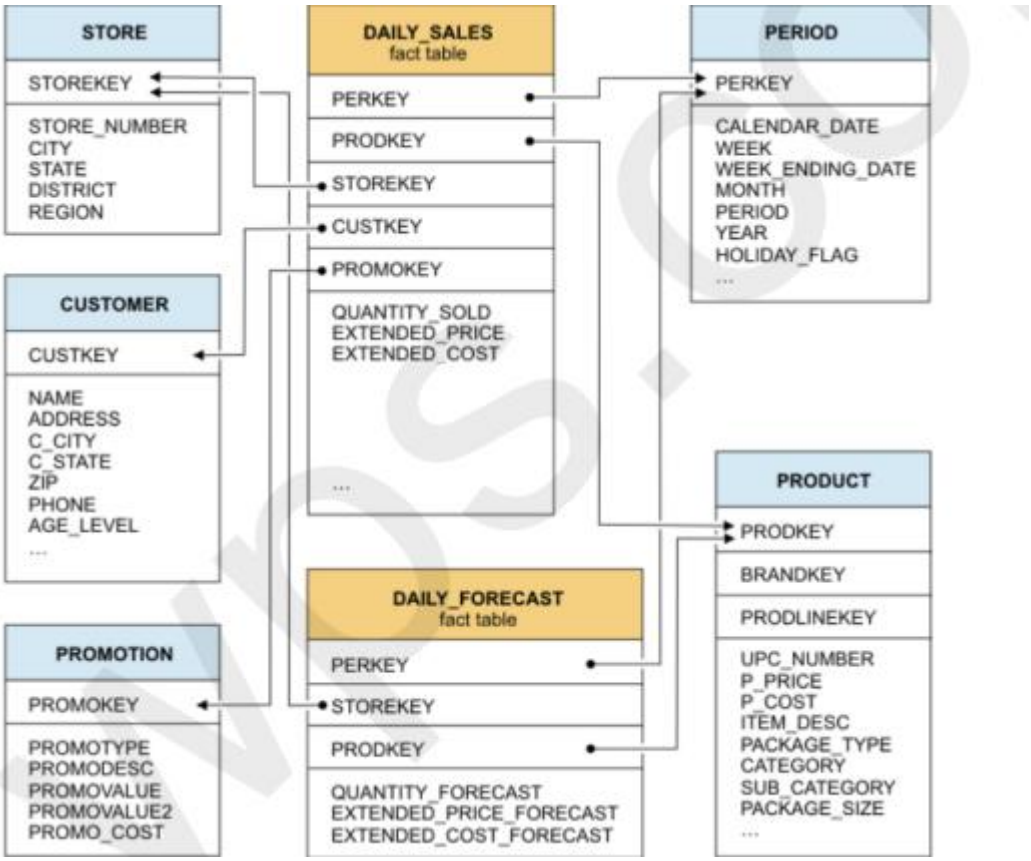
Select Avg(SalesAmount) from Sales S, Date D1, Date D2

From Sales,Date

Where S.OrderDate = D1.Date,

S.ShippingDate = D2.Date;

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 ((select sum(QUANTITY_SOLD) from DAILY_SALES, PERIOD
where PERIOD.MONTH = tochar(sysdate,'MM'))

- (select sum(QUANTITY_FORECAST) from DAILY_FORECAST, PERIOD

where PERIOD.MONTH = tochar(sysdate, 'MM'));

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

Conformed Dimensions: STORE

PERIOD

PRODUCT

Non-Conformed Dimensions: CUSTOMER

PROMOTION

Measures: QUANTITY_SOLD -> Additive Measure

EXTENDED_PRICE -> Semi-Additive Measure

EXTENDED_COST -> Semi-Additive Measure

QUANTITY_FORECAST -> Additive Measure

EXTENDED_PRICE_FORECAST -> Semi-Additive Measure

EXTENDED_COST_FORECAST -> Semi-Additive Measure

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

	OLTP	DW
Size	Size ranges from 100MB-GB	Size Ranges from 100GB-TB
Usage	Used for every business tasks	Used for Analysis
Processing	Databases which are OTLP require sub-second response time.	Analytical processing may require several minutes to run.
Data Models	E-R modeling	Star Schema or Snowflake Schema