DATA WAREHOUSE ASSESSMENT

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.

Ans:- As mentioned in the above question we need to store all the old categories i.e. storing all the historical data for that the suitable SCD type is SCD2 because in SCD2 we keep all previously updated value.

- As in SCD2 we do full load for the very first time and after every update we do incremental load.

The structure changes required in dimensional table to implement SCD2 and SCD3 is:-

SCD2

SCD2 stores entire history of the dimensional table.

In SCD2 we can store data in three different ways:-

1. Versioning
2. Flagging
3. Effective date

Status_Id	Customer_Name	Customer_City
1	Steve	California
2	Bill	New York
3	Jeff	Georgia

4. Costumer Dimension

Costumer Dimension(SCD2) Effective date

Status_Id	Customer_Name	Customer_City	Start_Date	End_Date
1	Steve	California	28/Nov/19	29/Nov/19
2	Bill	New York	28/Nov/19	
3	Jeff	Georgia	28/Nov/19	
1	Steve	Miami	29/Nov/19	

SCD3

SCD3 does not stores entire history of the dimensional table it stores the exact previous history of the table.

Costumer Dimension

Status_Id	Customer_Name	Customer_City
1	Steve	California
2	Bill	New York
3	Jeff	Georgia

Costumer Dimension(SCD3)

Status_Id	Customer_Name	Customer_City	Previous_City
1	Steve	Miami	California
2	Bill	New York	
3	Jeff	Georgia	

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

Ans:-Surrogate keys is a key which does not have any business meaning in a schema. It is mostly a sequentially increasing integer or it can also be current date/time-stamp and random generated numbers.

- It is required when the dimensions does not contain only the current state but also the historical data. If we use business key after every change we need to update the key and that would affect the whole schema. So, to overcome of that we use surrogate key(fact-less key) to store the updated record .
- We can identify the updated data by using surrogate keys because it is system generated unique keys.
- Sequential, time-stamp, and random keys have no practical limits to unique combinations.
- It increases the performance because it is less expensive to join.

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:- Store-Cluster Schema is:-

Store_Id	Store_Name
1	Big bazar
2	Nike
3	Adidas

Cluster_Id	Cluster_name
1	C1
2	C2

Loc_Id	Location
10	New York
20	Berlin
30	Vatican

SK	Store_Id	Cluster_Id	Loc_Id
1	1	1	10
2	1	2	20
3	2	2	30
4	3	1	10

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

Ans:- Semi-additive measures are the attributes that you can summarize across any related dimension accept time.

- Ex-In a sales table total sales quantity and costs are fully additive but stocks are semi additive because stocks can be added up for a particular time but it may varies like yesterday we had 100 cars stock but today we have 50 if we add up we'll end up with wrong measures. So, these measures are called as semi-additive measures.

- Some common measures are stocks, balance left in account,etc.

- Ex:-

Sales Fact
Prod_Id
Cost(Additive measures)
Quantity(Additive measures)
Stock_Left(Semi-additive measures)
Discount(Non additive measures)