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 kinds 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

2. 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.
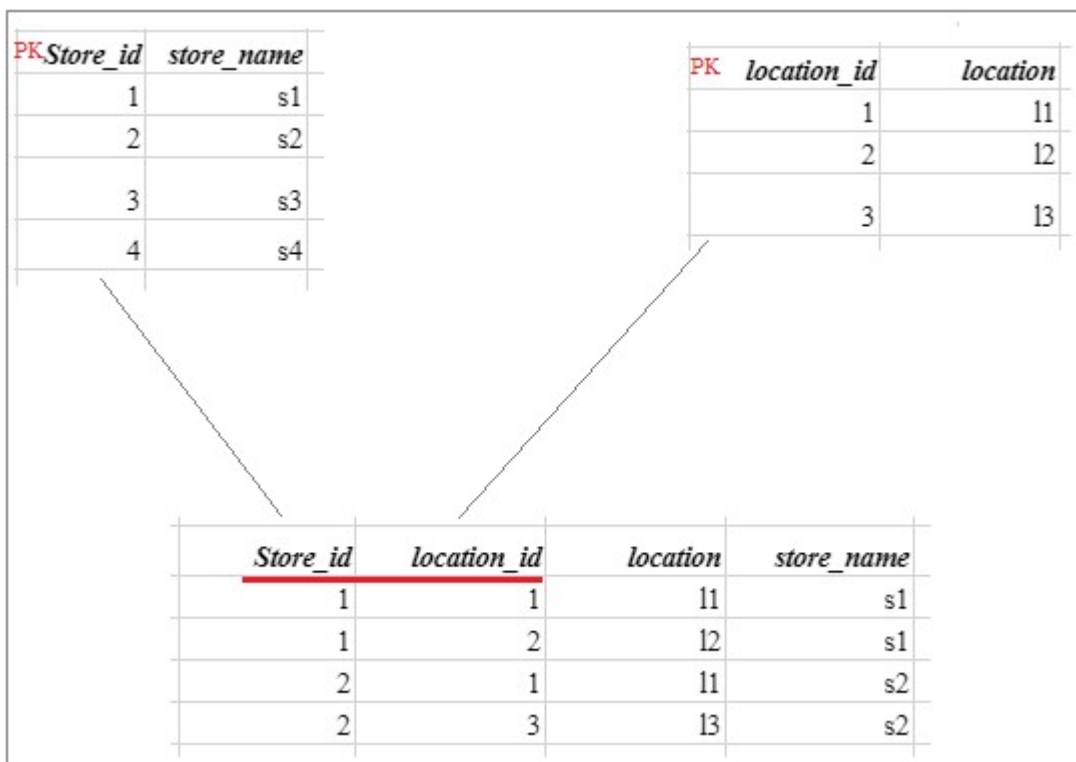
Initially

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

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.

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

4. 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.