(Big) Data Engineering In Depth

From Beginner to Professional

Mostafa Alaa Mohamed
Senior Big Data Engineer
MoustafaAlaa Moustafa Alaa @Moustafa_alaa22
mustafa.alaa.mohamed@gmail.com

[1]Big Data & Analytics Department, Epam Systems

The Definitive Guide to Big Data Engineering Tasks

Videos classification

Watching Method / Audience	Computer	Mobile/Tablet	Just listening
Developer	●		
DevOps	●		
Business	●		

Table: Video classification
The green circle ● means short video.
The blue circle ● means medium video.
The red circle ● means long video

Dimensions Types: Slowly changing Dimensions

Slowly changing Dimensions

- It the dimension which changes over time. So, for a specific date we have different value.
- It has different types as following
  - Type 0 (Fixed Dimension): We don't change the current even the source changes.
  - Type 1 (No History): No history is maintained only the latest replace the current.
  - Type 2 (History): Series of history of records are maintained.
  - Type 3 (Hybrid): Only the last Change and the Current new change is stored
  - Type 4 : We split the data into two tables, first the current record and second is the historical (most common usage).

Slowly changing Dimensions

Note

There are some other types which is a combination between the above similar than type 3 combined between 1 & 2.
You can check the chapter resources for more information about the other types.

Slowly changing Dimensions

- Type 0.

CustomerID	Name	City
123456789	Ronaldo	Madrid

CustomerID	Name	City
123456789	Ronaldo	Turin

Table: Source System Old vs New

ID	CustomerID	Name	City
1	123456789	Ronaldo	Madrid

Table: Customer Profile Dimension

Slowly changing Dimensions

- Type 1.

CustomerID	Name	City
123456789	Ronaldo	Madrid

CustomerID	Name	City
123456789	Ronaldo	Turin

Table: Source System Old vs New

ID	CustomerID	Name	City
1	123456789	Ronaldo	Turin

Table: Customer Profile Dimension

Slowly changing Dimensions

- Type 2.

CustomerID	Name	City	UpdatedDt
123456789	Ronaldo	Madrid	2018-12-12
123456789	Ronaldo	Turin	2019-06-12
123456789	Ronaldo	London	2019-08-12
123456789	Ronaldo	Porto	2019-12-12

Table: Source System Old vs New

ID	CustomerID	Name	City	effectiveDt	TerminationDt	isCurrent
1	123456789	Ronaldo	Madrid	2018-12-12	2019-06-12	false
2	123456789	Ronaldo	Madrid	2019-06-12	2019-08-12	false
3	123456789	Ronaldo	London	2019-08-12	2019-12-12	false
4	123456789	Ronaldo	Porto	2019-12-12	null	true

Table: Customer Profile Dimension We can replace null with a finite date (9999-12-31) but it needs to be consistent

Slowly changing Dimensions

- Type 3.

CustomerID	Name	City	UpdatedDt
123456789	Ronaldo	Madrid	2018-12-12
123456789	Ronaldo	Turin	2019-06-12
123456789	Ronaldo	London	2019-08-12
123456789	Ronaldo	Porto	2019-12-12

Table: Source System Old vs New

ID	CustomerID	Name	City	UpdatedDate	previousCity
1	123456789	Ronaldo	Porto	2019-12-12	London

Table: Customer Profile Dimension

Slowly changing Dimensions

- Type 4 (Split current and Historical).

ID	CustomerID	Name	City	effectiveDt	TerminationDt
1	123456789	Ronaldo	Madrid	2018-12-12	2019-06-12
2	123456789	Ronaldo	Madrid	2019-06-12	2019-08-12
3	123456789	Ronaldo	London	2019-08-12	2019-12-12
4	123456789	Ronaldo	Porto	2019-12-12	null

Table: Customer Profile Dimension Hist

ID	CustomerID	Name	City	UpdatedDate
1	123456789	Ronaldo	Porto	2019-12-12

Table: Customer Profile Dimension

Slowly changing Dimensions

- How does the Facts join SCD? We have two scenarios as following:
  - Getting the current customer information (Join with the latest).
  - Getting the historical customer information (Join with the historical table based on ***cust id & date***).

ID	CustomerID	TotalCalls	CallDate
1	123456789	30	2018-12-12
2	123456789	30	2019-12-12

Table: Customer Usage

Slowly changing Dimensions

```
--Get latest customer details from customer profile snapshot
select * from cust_usage_dly a
inner join cust_profl b
on a.CustomerID = b.CustomerID;

--Get historical customer details from customer profile hist
select * from cust_usage_dly a
inner join cust_profl_hist b
on a.CustomerID = b.CustomerID
and CallDate between effectiveDt and TerminationDt
```

Listing 1: Example to show how to use SCD