(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
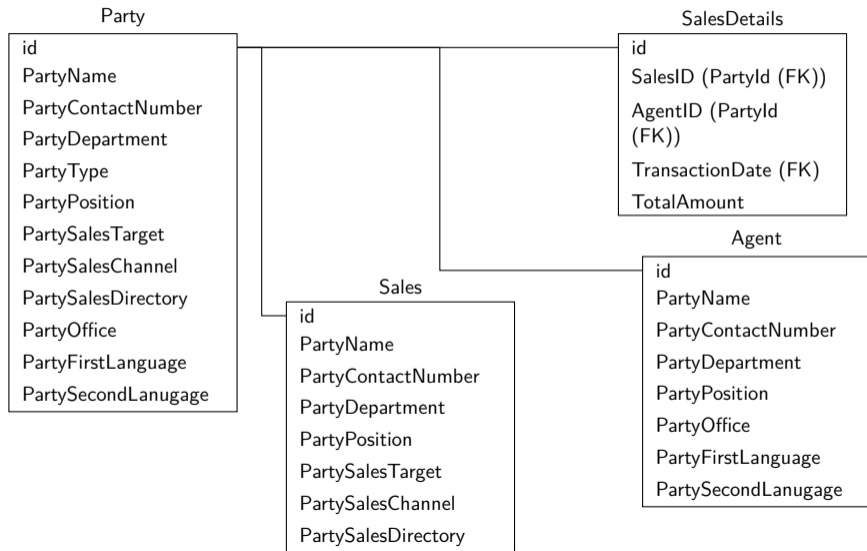The green circle ● means short video.
The blue circle ● means medium video.
The red circle ● means long video

Dimensions Types: Swappable Dimensions

Swappable Dimensions

- A dimension that has multiple alternate versions of itself that can be **swapped at query time**.
- Each version of the hot-swappable dimension (sub-types)
    - It has a different meaning
    - It has a different structure.
    - It has fewer data compared to the primary dimension (fewer rows and columns).
    - It has a different output based on the input version and its alternatives.
    - Multi versions could be used together in the same fact with different types.
    - It can act as the primary dimension and join to the same fact table.
    - It has different target users and sometimes we restrict the users to access the primary dimension and only access the swapped version to restrict the data without needs to show the whole primary attributes.

.

Swappable Dimensions

Party

id
PartyName
PartyContactNumber
PartyDepartment
PartyType
PartyPosition
PartySalesTarget
PartySalesChannel
PartySalesDirectory
PartyOffice
PartyFirstLanguage
PartySecondLanuguage

SalesDetails

id
SalesID (PartyId (FK))
AgentID (PartyId (FK))
TransactionDate (FK)
TotalAmount

Agent

id
PartyName
PartyContactNumber
PartyDepartment
PartyPosition
PartyOffice
PartyFirstLanguage
PartySecondLanuguage

Sales

id
PartyName
PartyContactNumber
PartyDepartment
PartyPosition
PartySalesTarget
PartySalesChannel
PartySalesDirectory

Swappable Dimensions

Implementation

- Direct join between Fact and Dimension with filter based on PartyType (run-time). In this case Party includes some empty columns based on the type.
- Logical views each view has its own number of columns and rows based on the type details.
    - Pros: Easy for (managing, implementation) with consistent views.
    - Cons: Performance and manage the authorization per view.
- Physical tables (Types & Sub-types).
    - Pros: Performance, better design.
    - Cons: Data redundancy, key could be duplicated (when join with fact), increase in data size, and ETL headache.

Attention: Conformed vs Role-Playing Dimension vs Swappable

- **Conformed** is the same dimension which used in different facts and has *the same meaning and value*
  ✒ ⊕ CustomerID 123 can be represented into the whole model using the same value and same meaning.

- **Role-Playing** is the same dimension which used multiple times within the same fact but *with different meanings and same value*
  ✒ ⊕ Date Dimension 20191012 can be used for different purpose order delivery date, expire date but different meanings.

- **Swappable** different version from the primary dimension each version has its own attributes and meanings based on the use case *(different meaning based on the category)*.
  ✒ ⊕ party id dimension has different version sales, agent, employee and all of them sub-types of the party but used for different purpose in different facts.