(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

Schema Types

Schema Types

- Star Schema.

Schema Types

- Star Schema.
- Snowflake Schema.

Schema Types: Star Schema

Star Schema Characteristics

- **Simplicity**: It is the simplest type of DWH schemas.

Star Schema Characteristics

- **Simplicity**: It is the simplest type of DWH schemas.
- **Query effectiveness**: Because of simplicity, It needs less join to query the data (It is optimized to query large dataset).

Star Schema Characteristics

- **Simplicity**: It is the simplest type of DWH schemas.
- **Query effectiveness**: Because of simplicity, It needs less join to query the data (It is optimized to query large dataset).
- **Data Redundancy and Large Table Size**: Due to de-normalization, it has a data redundancy, and the table size is huge.

Star Schema Characteristics

- **Simplicity**: It is the simplest type of DWH schemas.
- **Query effectiveness**: Because of simplicity, It needs less join to query the data (It is optimized to query large dataset).
- **Data Redundancy and Large Table Size**: Due to de-normalization, it has a data redundancy, and the table size is huge.
- **Most** used and **widely** supported.

Star Schema Characteristics

- Dimensions represented by one one-dimension table.

Star Schema Characteristics

- Dimensions represented by one one-dimension table.
- The dimension table are not joined to each other

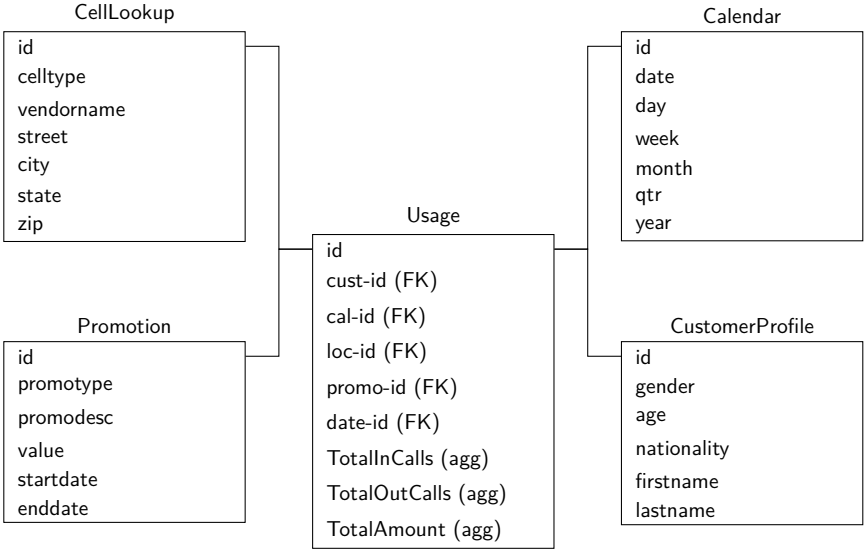
Star Schema Characteristics

- Dimensions represented by one one-dimension table.
- The dimension table are not joined to each other
- The fact table would contain key and measure.

Star Schema Characteristics

- Dimensions represented by one one-dimension table.
- The dimension table are not joined to each other
- The fact table would contain key and measure.
- Data integrity is not enforced due to the de-normalized structure.

Schema Types: Star Schema Example



Schema Types: Snowflake Schema

What is Snowflake?

Figure: Snowflake Photo taken from https://earthsky.org

What is Snowflake?

Figure: Snowflake Simple Design

What is Snowflake?

Figure: Snowflake Final Design

Snowflake Schema Characteristics

- **Extension**: Snowflake is an extension of the Star Schema.

Snowflake Schema Characteristics

- **Extension**: Snowflake is an extension of the Star Schema.
- **Normalized**: Dimension tables are normalized; this means every dimension may expand into additional tables.

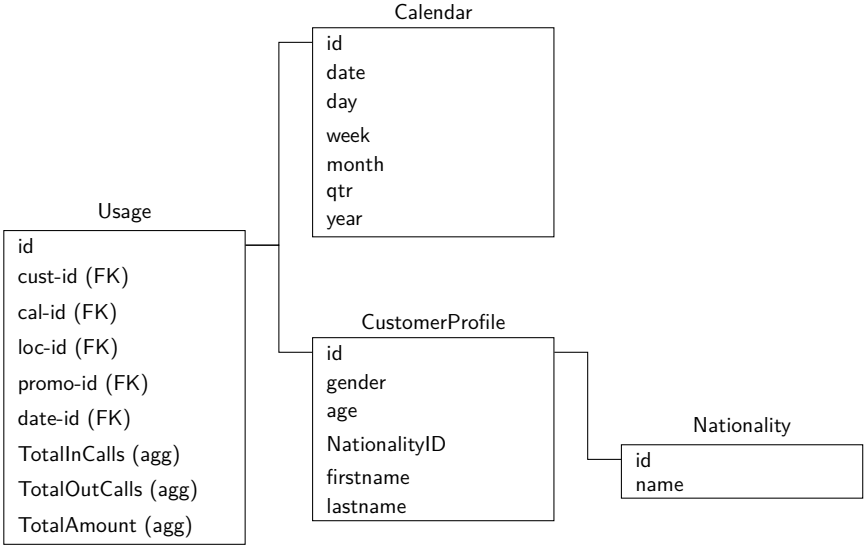
Snowflake Schema Characteristics

- **Extension**: Snowflake is an extension of the Star Schema.
- **Normalized**: Dimension tables are normalized; this means every dimension may expand into additional tables.
- **Disk Space Efficiency**: Due to its normalization methodology, it uses less desk space, which enhances the query as we scan less data size.

Snowflake Schema Characteristics

- **Extension**: Snowflake is an extension of the Star Schema.
- **Normalized**: Dimension tables are normalized; this means every dimension may expand into additional tables.
- **Disk Space Efficiency**: Due to its normalization methodology, it uses less desk space, which enhances the query as we scan less data size.
- **Complicated**: Due to the normalization query needs to join more table in some cases to get the data which reduces the performance.

Schema Types: Snowflake schema (Example)



Star Vs. Snowflake Schema

Star	Snowflake
Dimension represented by one-table	Dimension tables are expanded into multi-tables
Fact table surrounded by dimension tables	Fact table surrounded by Hierarchy of dimension tables
Less join	Requires many joins
Simple Design	Very Complex Design
De-normalized Data structure	Normalized Data Structure
High level of Data redundancy	Very low-level data redundancy
Maintenance is difficult	Maintenance is easier
Good for datamarts with simple relationships (1:1 or 1:many)	Good for core to simplify (many:many)