Data Engineering Day 22

The credit for this course goes to Coursera. Click More

Another link : Azure data Engineer

Getting started with Data Warehousing and Business Analytics

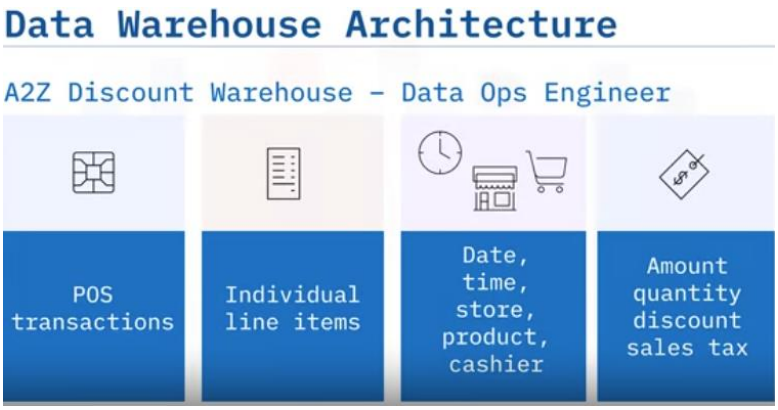
#Data Modeling using Star and Snowflake Schemas:

- Star schema is graphing whose edges are relations between facts and dimensions.
- Star schemas are used to develop a special type of data warehouse called DataMart.
- Snowflake schemas are generalized of star schema.
- Normalizations means the separations or the breaking down of the higher hierarchy tables into small tables.
- Snowflakes can also be known as a normalized star schema.

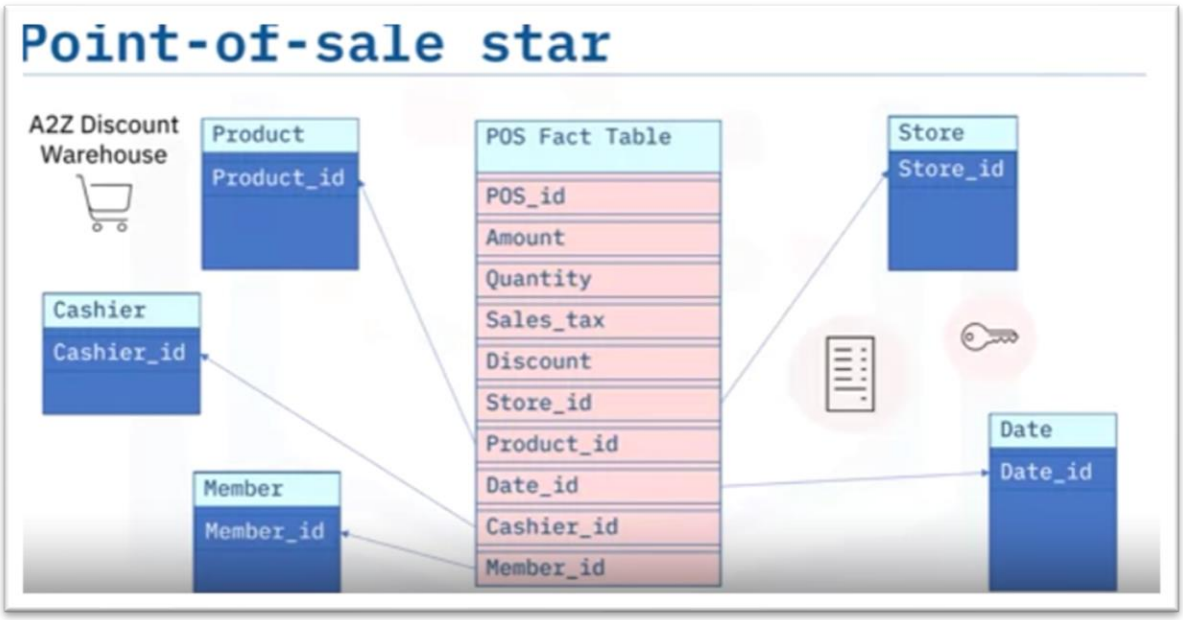
The following are the general principal that one should consider while designing a data model for a star schema.

1. Select a business process.
2. Choose a level of details [are you looking for an annually sales or a monthly sale of a company?]
3. Identify the dimensions. [it includes an attribute such as name, id, time, sale products etc.]
4. Identify the facts [sales amount, quantity sold, discounts applied, profit margins, and shipping costs, which are aggregated over various dimensions for analysis.]

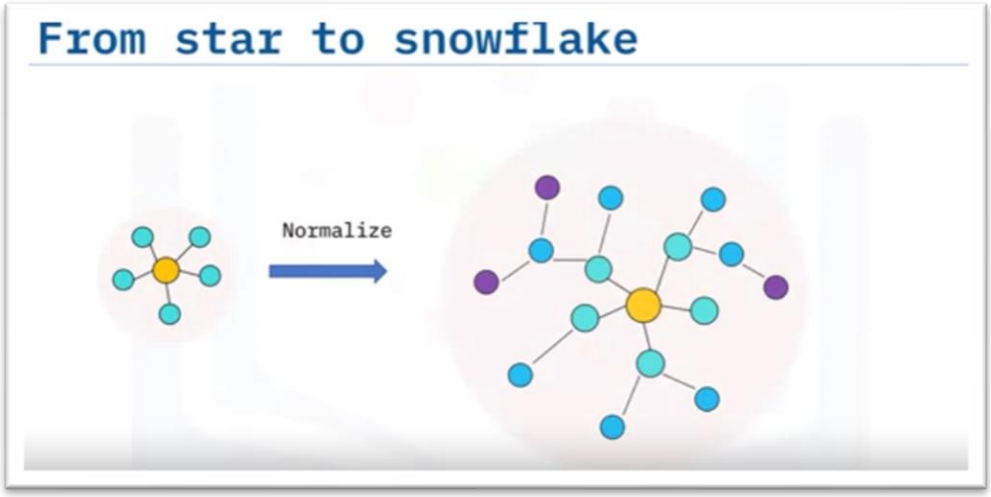
For an example if a company called A2Z asked me to design a data operation for tracking the daily customers sales through scales and pay, I should be designing my mode as follows:



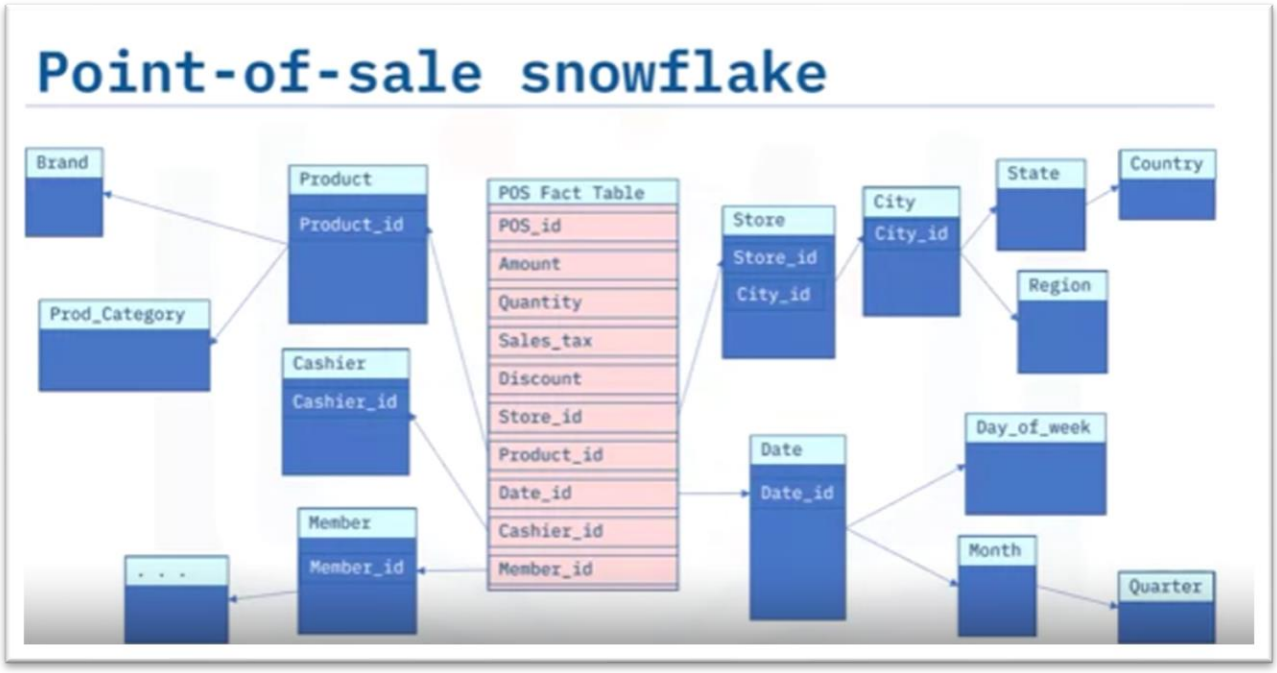
The figure below represents the possible architecture for the case mentioned above. In the center lies the fact table which is the most important for designing the star schemas. After that the facts



such as Pos_id, amounts etc. will be collected and will be implemented as a foreign key with the other tables to make a generalized relationship between them.



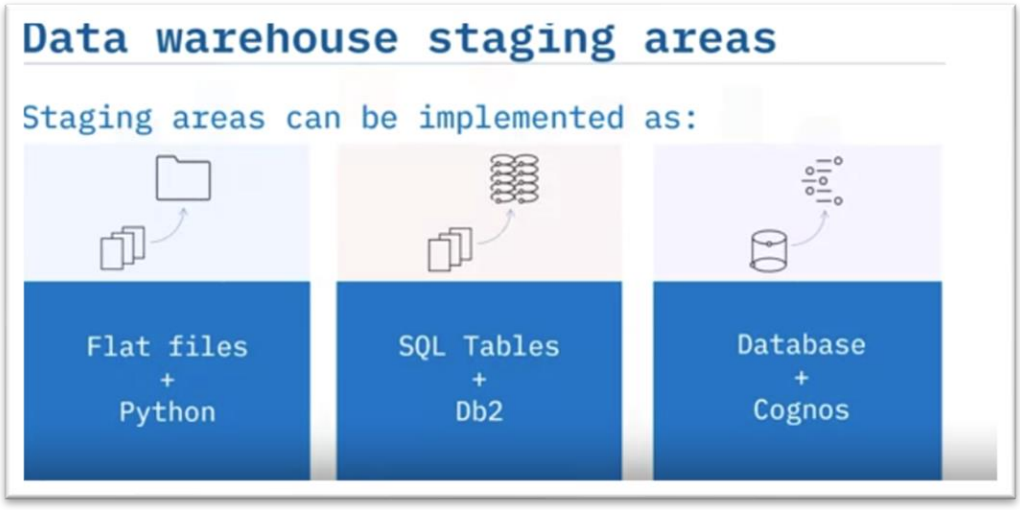
Normalizations from star to Snowflake:



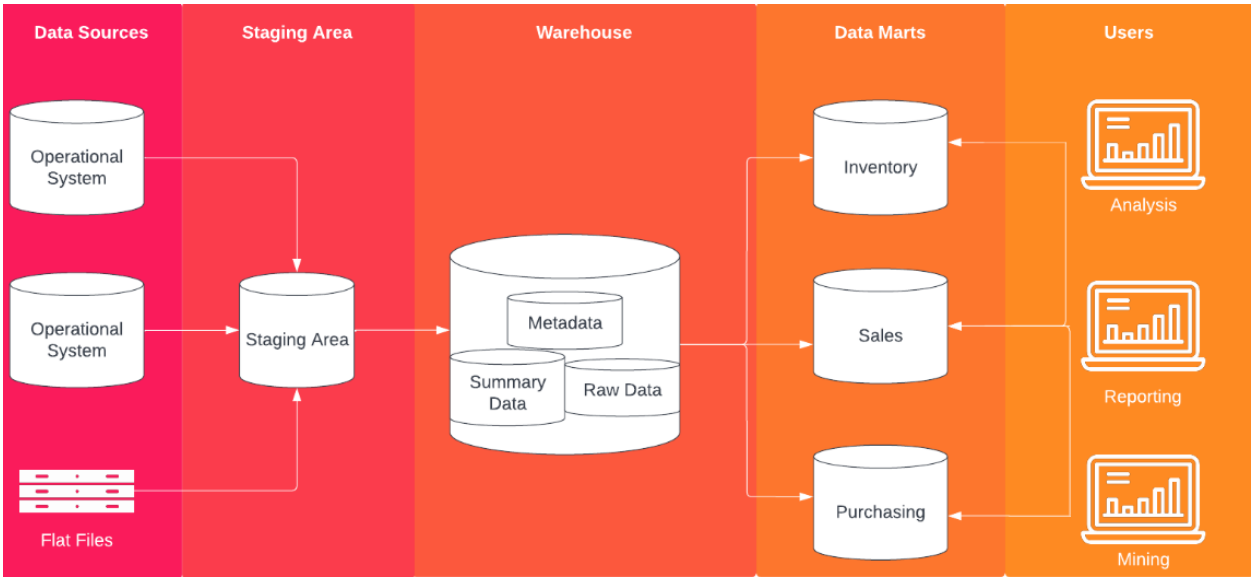
From the figure below, it is quite clear that the normalization means to make a more branches of star schema to a new structured called snowflake. Indeed, snowflakes are more detail oriented to the star schema.

 # Staging Areas for Data Warehouses:

- Data warehouse staging is an intermediate staging used for ETL processing.
- Staging acts like a connector or a bridge between a data source and the target systems where it could be a data mart or a lake as well.
- Staging is erased after execution of the code.
- One of the most beneficial of using staging is it helps to optimize and monitor the ETL jobs.
- Staging can be applied to various areas as shown below.

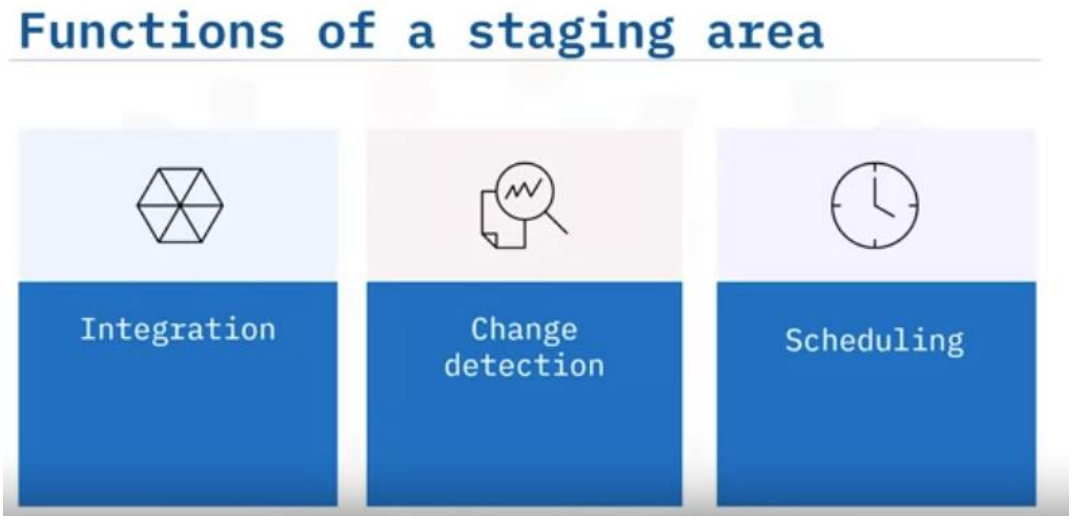


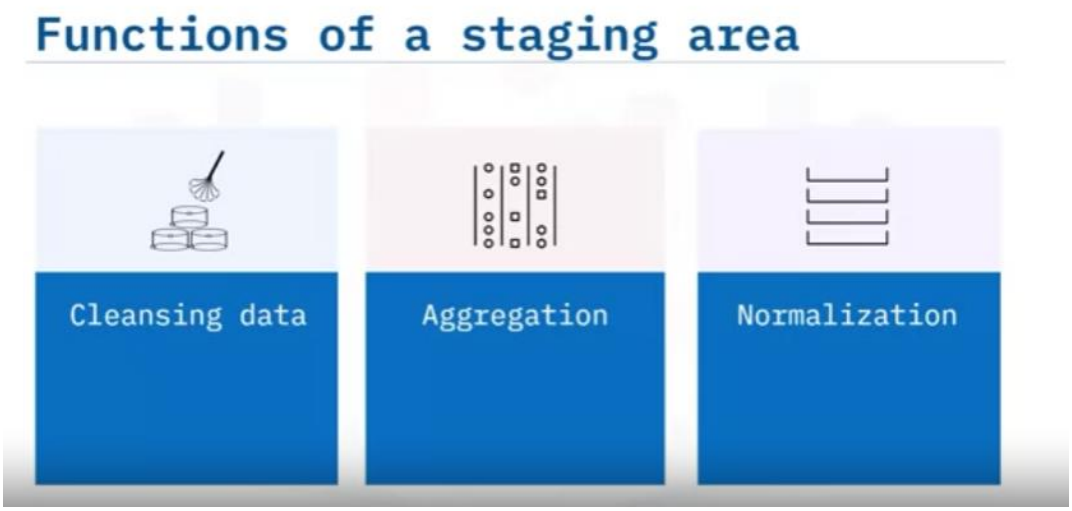
Below is an architecture of Data staging while its being transferred from the source to the target system. The figure shown below, belongs to a website called Zuar and please click me to read more.



Well, let's discuss more about the functions of a staging area. We already know that the data staging is the process of preparing and organizing the data before it is being moved to its destination ensuring cleanliness and consistency for the smooth integrations. A data staging area provides the controlled space where all these harmonizations and preparation takes place. The core reasons for using staging are as it minimizes the data corruption, simplifies the ETL works, simplifies the recovery process of the data.

The following are the main actions that are being performed while staging the data and are as shown by two figures below.





- **Verify Data Quality:**
    - Data accuracy verifications includes checking the accuracy of the data, completeness, consistency, currency (is data up to date?)
    - Data verification is about managing data quality and enhancing its reliability.
    - Data accuracy means cross checking the precision of input and the output data. Accuracy is very necessary to be concerned with because the users might put wrong data, spelling mistakes and many more possibility of human error.
    - Steps taken for managing data are as follows:
        1. Writing SQL queries for detecting and testing the data.
        2. Creating rules for correcting and managing those conditions time and again
        3. Create the scripts that runs the data validation SQL queries every night.
        4. Create automation scripts that check, correct, and process the data all the time.
        5. Review and report the issues and analyze them further.