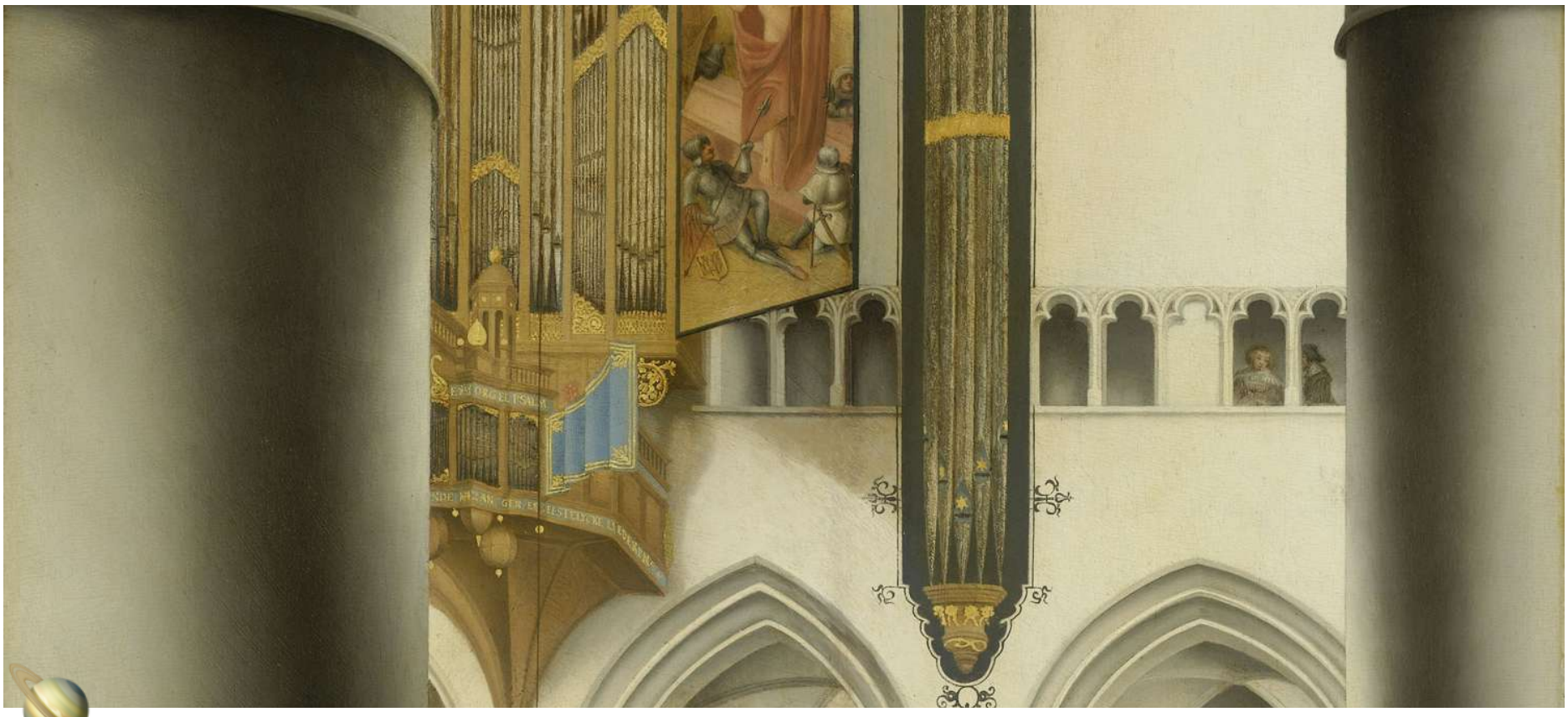


Data warehouse

Tags	Database
Date	@May 13, 2024

I. Data warehouse architectures

- Data warehouse consist of 2 side by side layer : user access and staging layer

1. **Staging layer** :

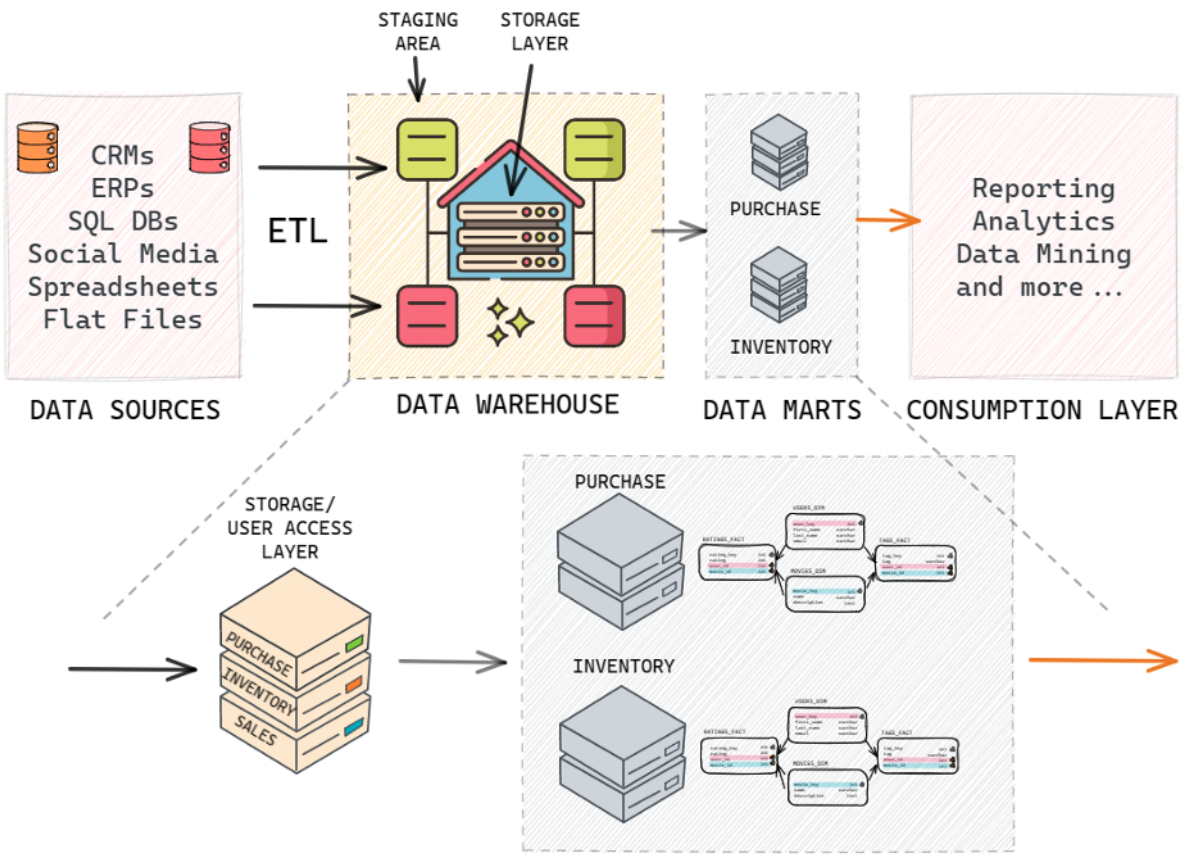
- It is where data is initially loaded before it it is transformed and fully integrated for reporting and analytical report. There are two types of staging layers : persistent and non-persistent
- Staging layer serve as a landing zone for incoming data from source applications
- **Non-persistent staging layer :** data in here is temporary. After being loaded into the user access layer, the staging layer is emptied
- A persistent staging layer retains data even it has been loaded into the user access layer

2. **User access layer**

- That is where user interacts with the data warehouse or data mart. This layer deals with design, engineering, and dimensional modelling, such as star schema, snowflake schema, fact tables, and dimension tables.

3. **Data marts** :

- There are two types of data marts : dependent and independent. The first one reply on the existence of a data warehouse to be supplied with data. The second one draw data directly from one ore more source applications and do not require data warehouse
- The difference between a data warehouse and data mart is the number of data source and business. Data warehouse have many sources (10 -50), while this independent data marts have fewer



II. Dim table

- Dimension tables add more detail to fact table
- Dimension / attributes a bout a process of :
  + Hold reference data
- The value are more likely to be duplicated (solve it by using SDC)

Table Name: Customer_Dim

Keys	ColumnName
PK	CustomerID
	Name
	AccountNum
	LoyaltyID
	Country
	Email

Legend: PK = Primary Key

III. Fact table

- Measurements, metrics or fact about an organization
- Links to **dimension tables** for more details

Table Name: Sales_Order_Fact

Keys	ColumnName
FK	CustomerID
FK	DateID
FK	ProductID
	UnitSold
	SalesAmount
	Tax

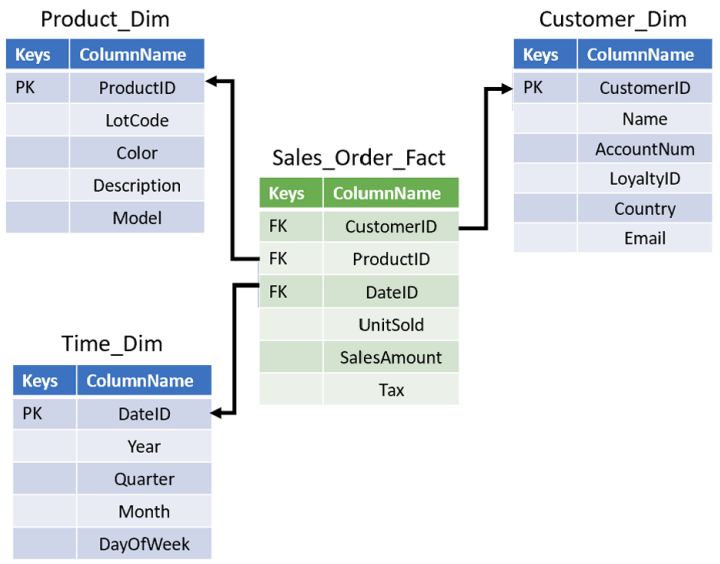
Legend: FK = Foreign Key

Different Forms of Additivity in Facts

- **Additive fact** can be added under all circumstances. Example : faculty members' salaries or students' credit hours can be added together to find a total salary or total credit hours completed.
- **Non-additive facts** differs from the above, cant be added together to produce a valid result
- **Semi-additive fact :** they can some times added together while at other times they cannot

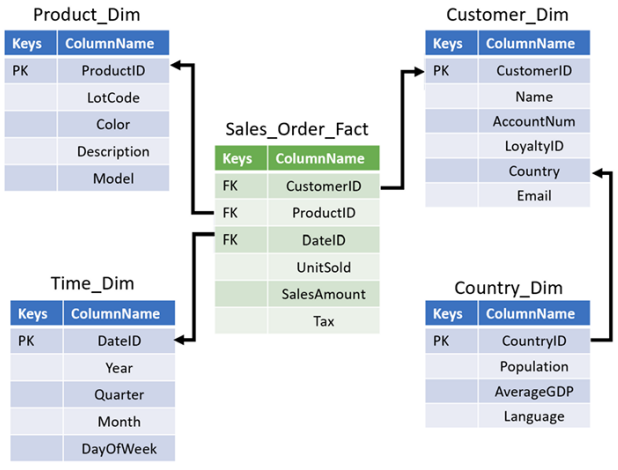
IV. Star schema

- A central fact table, with one or more dimensional tables
- Easy for business users



V. Snowflake schema

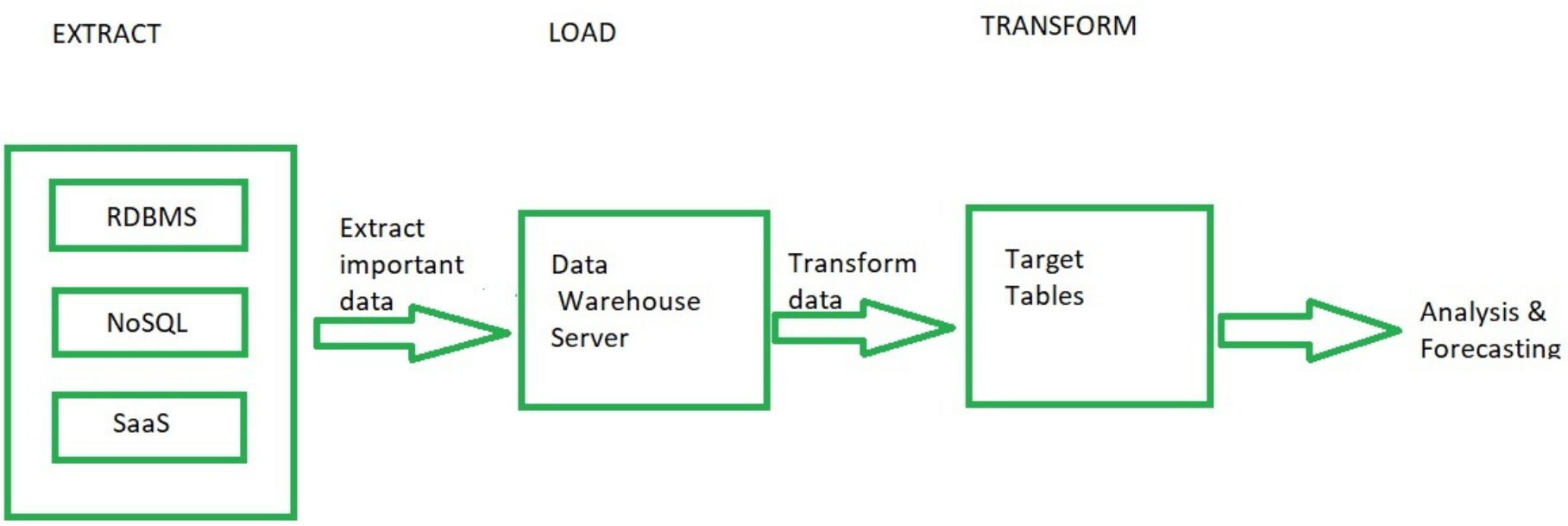
- Dimension table connected through another dimension table



VI. ELT vs ETL

1. ELT (Extract - Load- Transform)

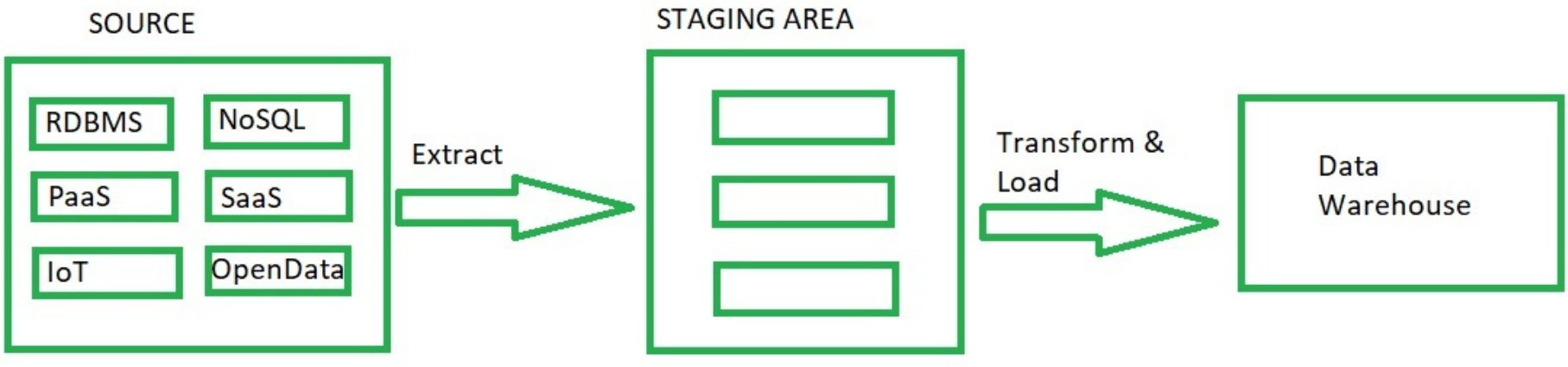
- **Extract** : the technique of identifying data from one or more sources. The sources may be database, files, ERP, CRM or others
- **Load** : the process of storing the extracted raw data in data warehouse or data lakes
- **Transform** : Data transformation is the process in which the raw data source is transformed to the



- **Drawback :** once transformed data is stored in the data warehouse, it can not be modified again, While in ELT, a copy of raw is always available and only the required data is transformed when needed

2. ETL (Extract - Transform - Load)

- **Extract** : the technique of identifying data from one or more sources. The sources may be database, files, ERP, CRM or others
- **Transform** : The extracted data is immediately transformed as required user
- **Load** : then it is loaded into the data warehouse from where user can access it



> ***Note :*** *the transformation are performed in the staging area. Once this transformed successfully, this will be stored in data warehouse*