

BUỔI 1: GIẢI NGỐ DATA

Output

- Hiểu được các khái niệm cơ bản trong ngành data
- Vai trò của SQL trong ngành Data Analytics

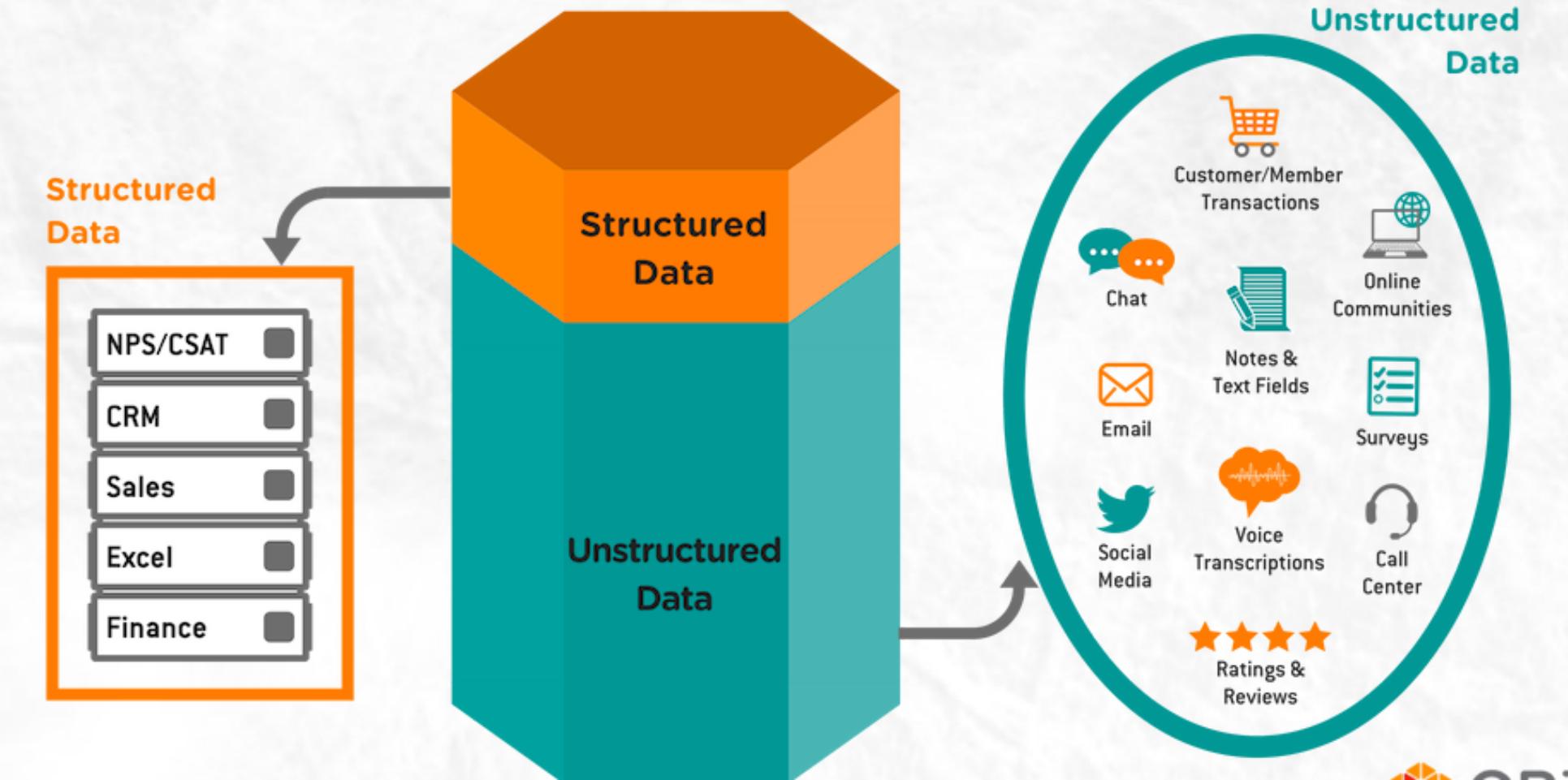
I. Data

Dữ liệu là một tập hợp các dữ kiện, chẳng hạn như **số**, **từ**, **hình ảnh**, **âm thanh** nhằm đo lường, quan sát hoặc chỉ là mô tả về sự vật



I. Data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.



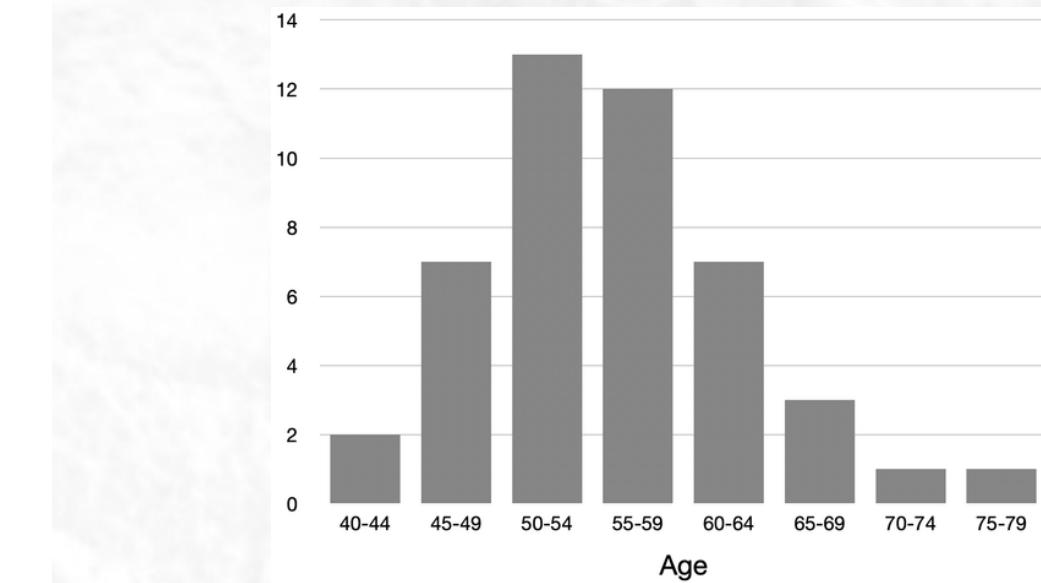
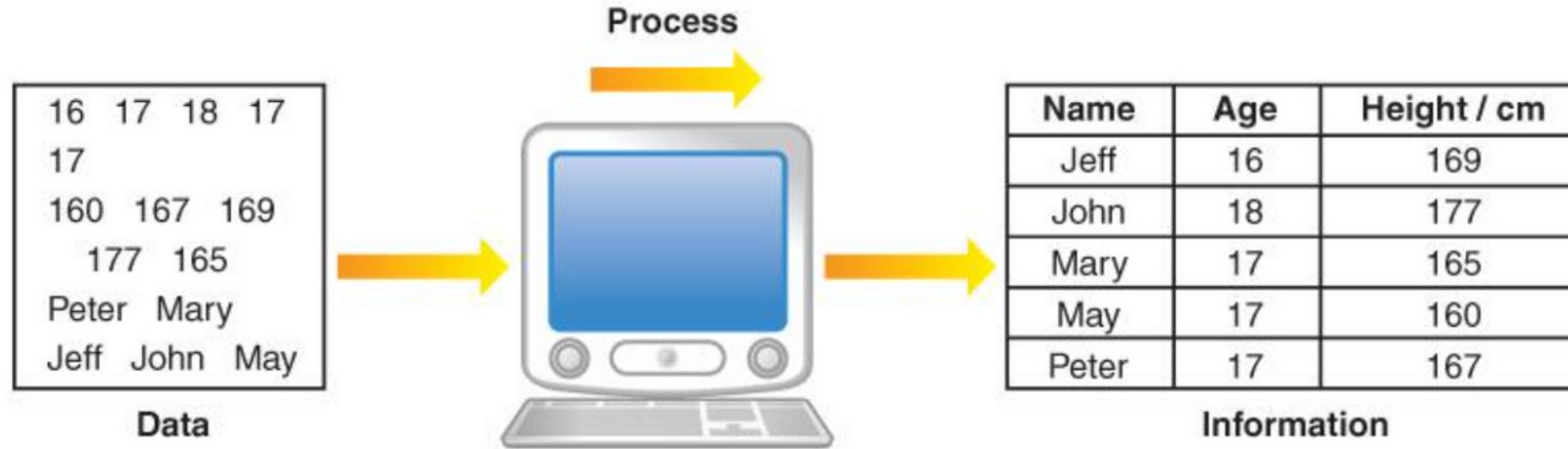
Dữ liệu có cấu trúc

- Biểu thị dưới dạng cột, dòng
- Thường là số, ngày tháng hoặc là chuỗi ký tự.
- Dữ liệu được tổ chức theo schema đã xác định trước

Dữ liệu phi cấu trúc

- Không thể biểu thị dưới dạng cột, dòng.
- Thường tồn tại dưới dạng âm thanh, hình ảnh, video,...
- Dữ liệu phải được xử lý phân tách để thấy được cấu trúc cụ thể

I. Data



Dữ liệu

Các biến định tính/ định lượng được thu thập đưa vào phân tích

Đại diện cho các thuộc tính trong thế giới thực

Dữ liệu thu thập không có giá trị nếu người nghiên cứu không đưa ra được các kết luận dựa trên dữ liệu đó

Thông tin

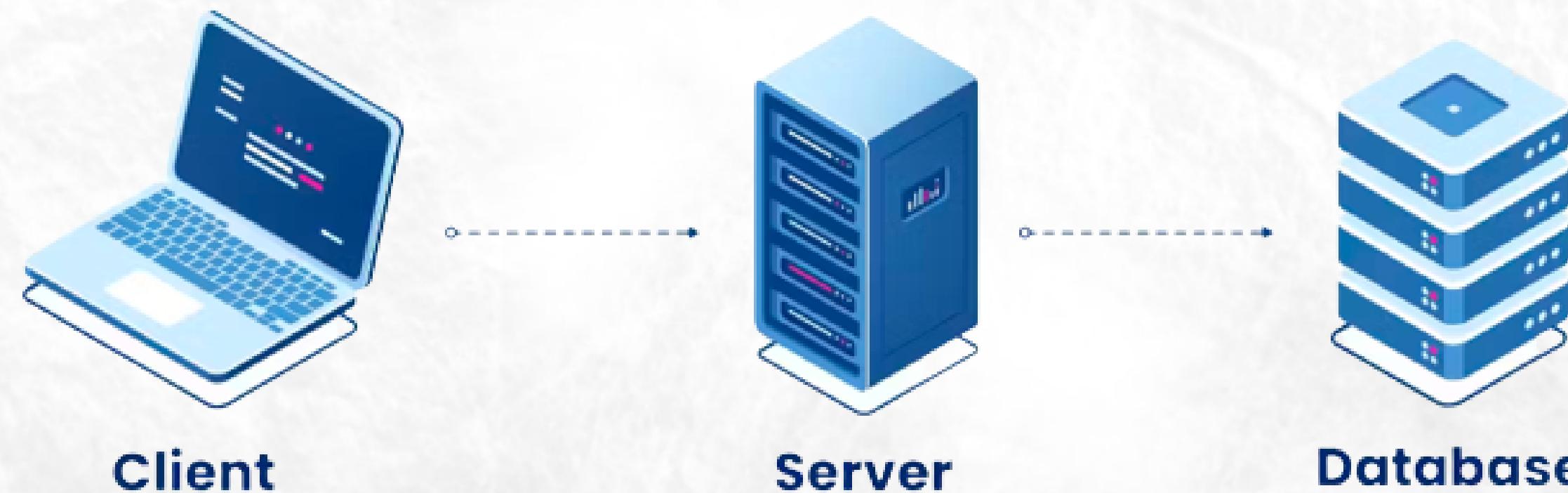
Dữ liệu được tổ chức và xử lý để đưa ra nội dung hoặc ý nghĩa nhất định

Trả lời câu hỏi trong thế giới thực

Thông tin là dữ liệu có giá trị mà từ đó người nghiên cứu tạo ra sản phẩm, kết luận có ý nghĩa

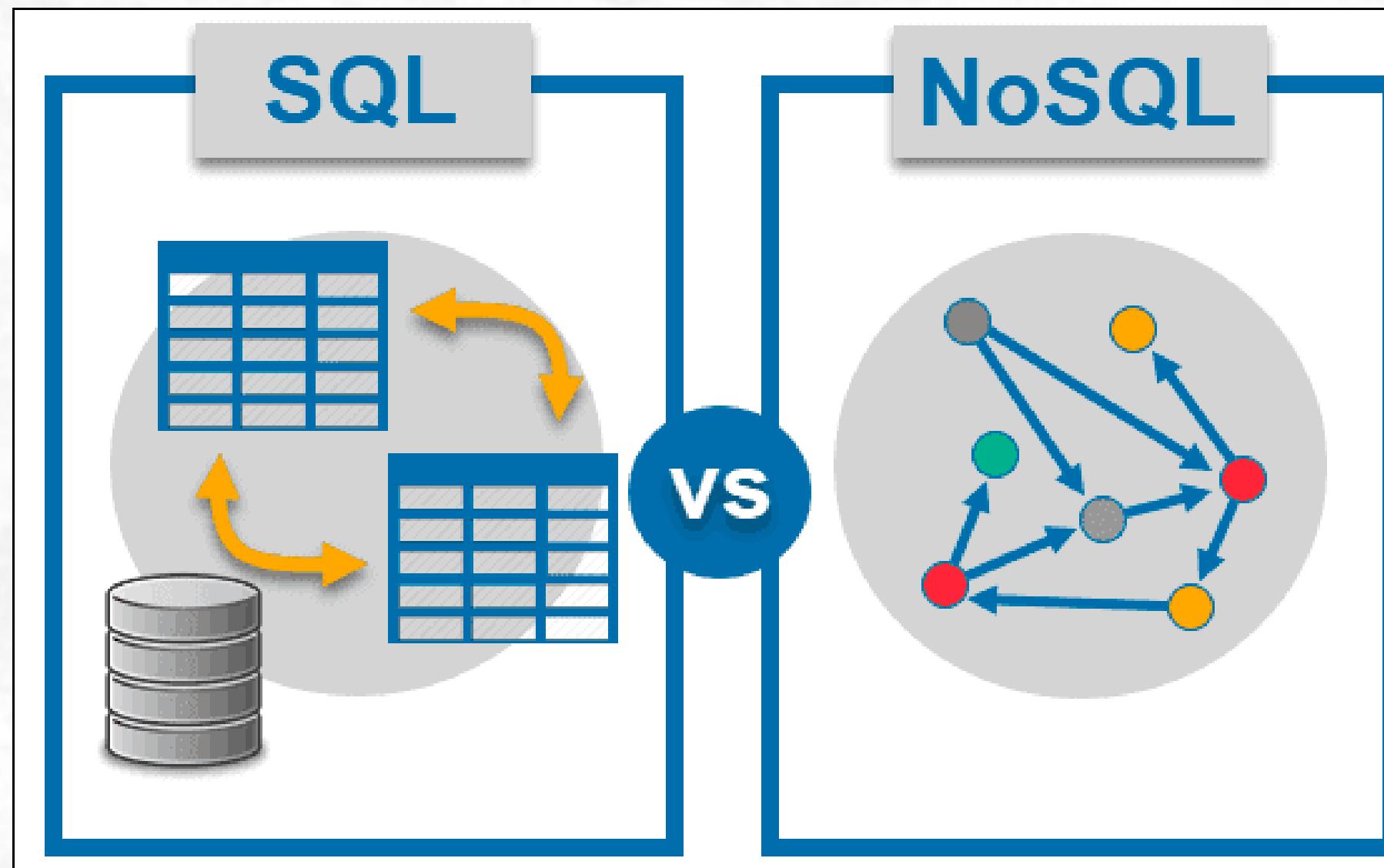
II. Database

Database (cơ sở dữ liệu) là nơi chứa dữ liệu được sản sinh từ các hệ thống, phần mềm của công ty



II. Database

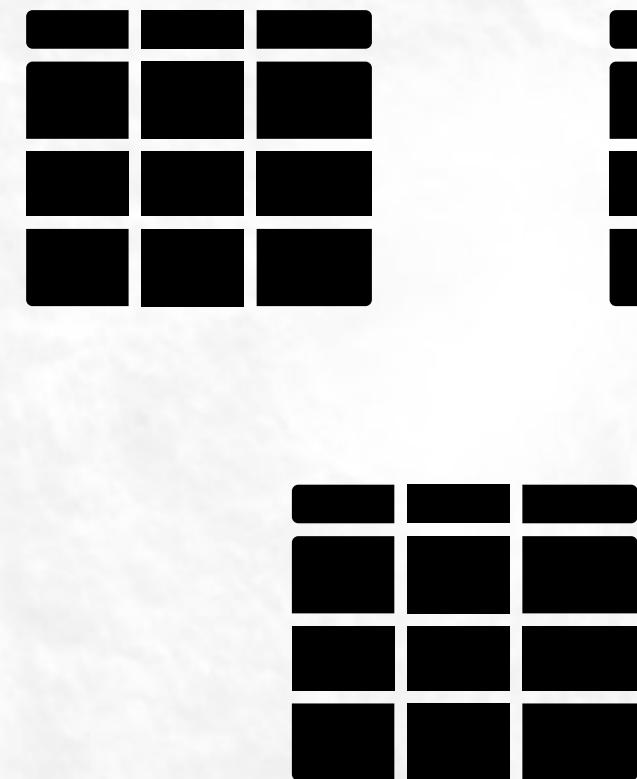
Relational Database (SQL) vs Non Relational Database (NoSQL)



- **Cơ sở dữ liệu SQL** (cơ sở dữ liệu có quan hệ): dữ liệu được sắp xếp thành các bảng và mỗi bảng có một cấu trúc cụ thể. Các bảng được kết nối với nhau thông qua các mối quan hệ.
- **Cơ sở dữ liệu NoSQL** (cơ sở dữ liệu không quan hệ): dữ liệu được lưu trữ trong một tập hợp các tài liệu. Không có cấu trúc cụ thể cho các tài liệu này và chúng không được kết nối với nhau thông qua các mối quan hệ.

II. Database

SQL database (Cơ sở dữ liệu quan hệ)



Relational Model in DBMS

Student Table (Relation)		
Primary Key →	Roll Number	Name
	001	Vaibhav
	002	Neha
	003	Harsh
	004	Shreya
		CGPA
		9.1
		9.5
		8.5
		9.3

Annotations:

- A vertical arrow labeled "Primary Key →" points to the first column of the table.
- A bracket labeled "Tuples (Rows)" points to the four rows of data.
- A bracket labeled "Columns (Attributes)" points to the three columns of data.

II. Database

Khoá chính vs Khoá ngoại

Khoá chính
(Primary key)

là giá trị định danh cho từng hàng trong bảng. Mỗi quan hệ giữa các bảng được tạo ra khi sử dụng khoá chính của 1 bảng làm khoá ngoại cho bảng khác

Khoá ngoại
(Foreign key)

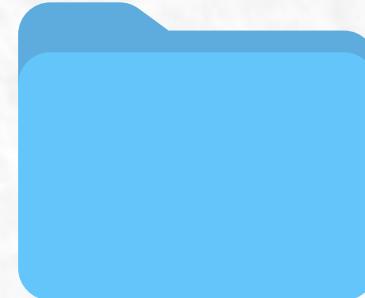
khoá ngoại biểu diễn mối quan hệ giữa các bảng với nhau

Persons Table			
PRIMARY KEY			
PersonID	LastName	FirstName	Age
1	Hansen	Ola	30
2	Svendson	Tove	23
3	Pettersen	Kari	20

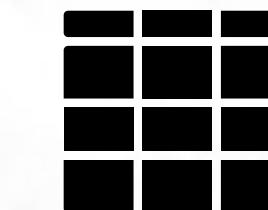
Orders Table		
OrderID	OrderNumber	PersonID
1	77895	3
2	44678	3
3	22456	2
4	24562	1

II. Database

Schema



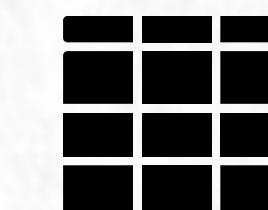
Schema 1



Schema dùng để gom nhóm các table có chung một đặc điểm nào đó để dễ dàng quản lý.



Schema 2



Bạn có thể phân quyền quản lý từng schema cho từng user khác nhau, đây chính là điểm mạnh của schema.

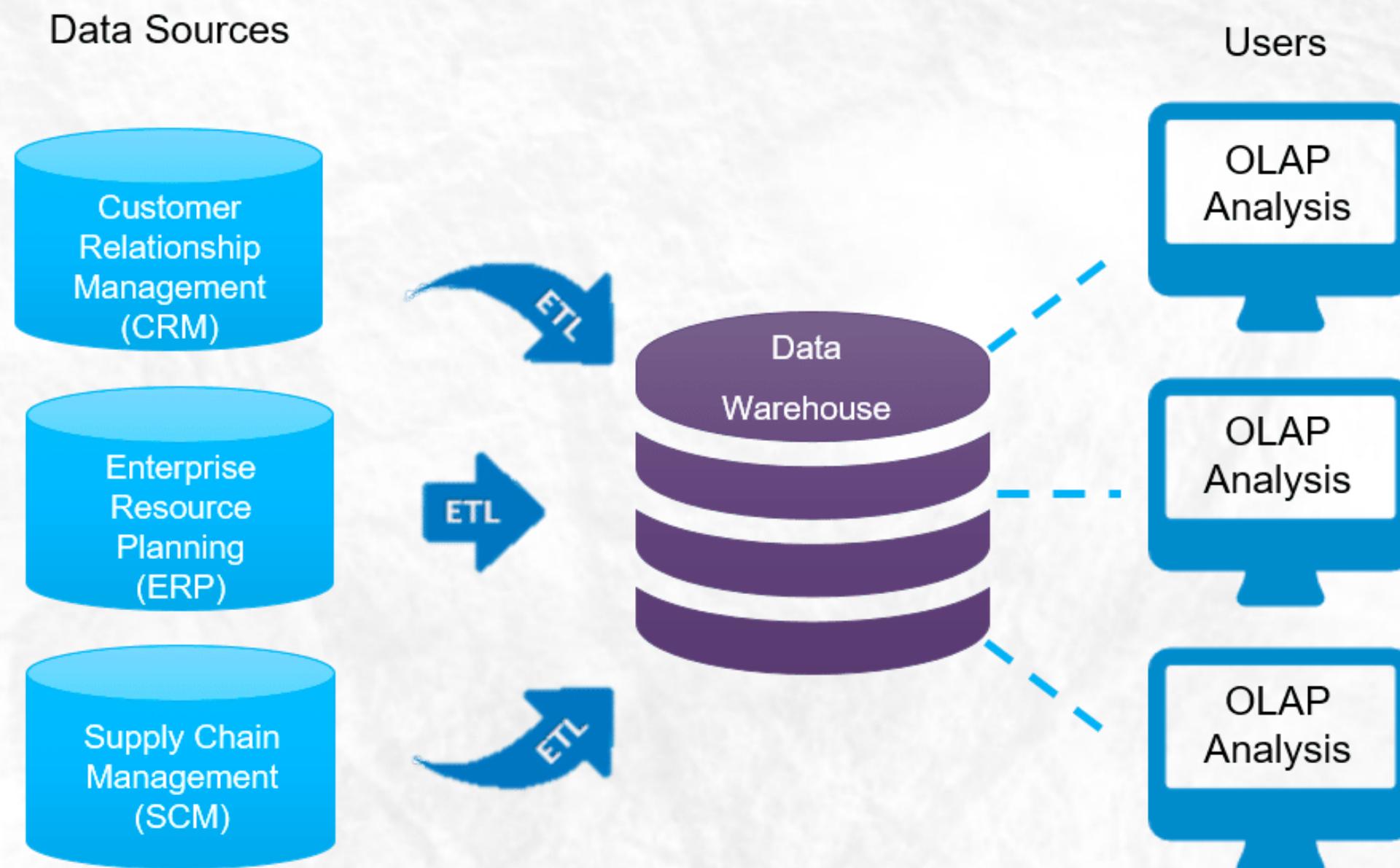
II. Database

Data type

	Data type	Ví dụ
Number	INT	11
	FLOAT	28.37
Logical	BOOLEAN	TRUE/FALSE
String	VARCHAR	"Hello"
	CHAR	"name"
	NVARCHAR	"Hình ảnh"
Date	DATE	2022-07-30
	DATETIME	7/30/2022 4:12:00 PM
	TIME	4:12:00 PM

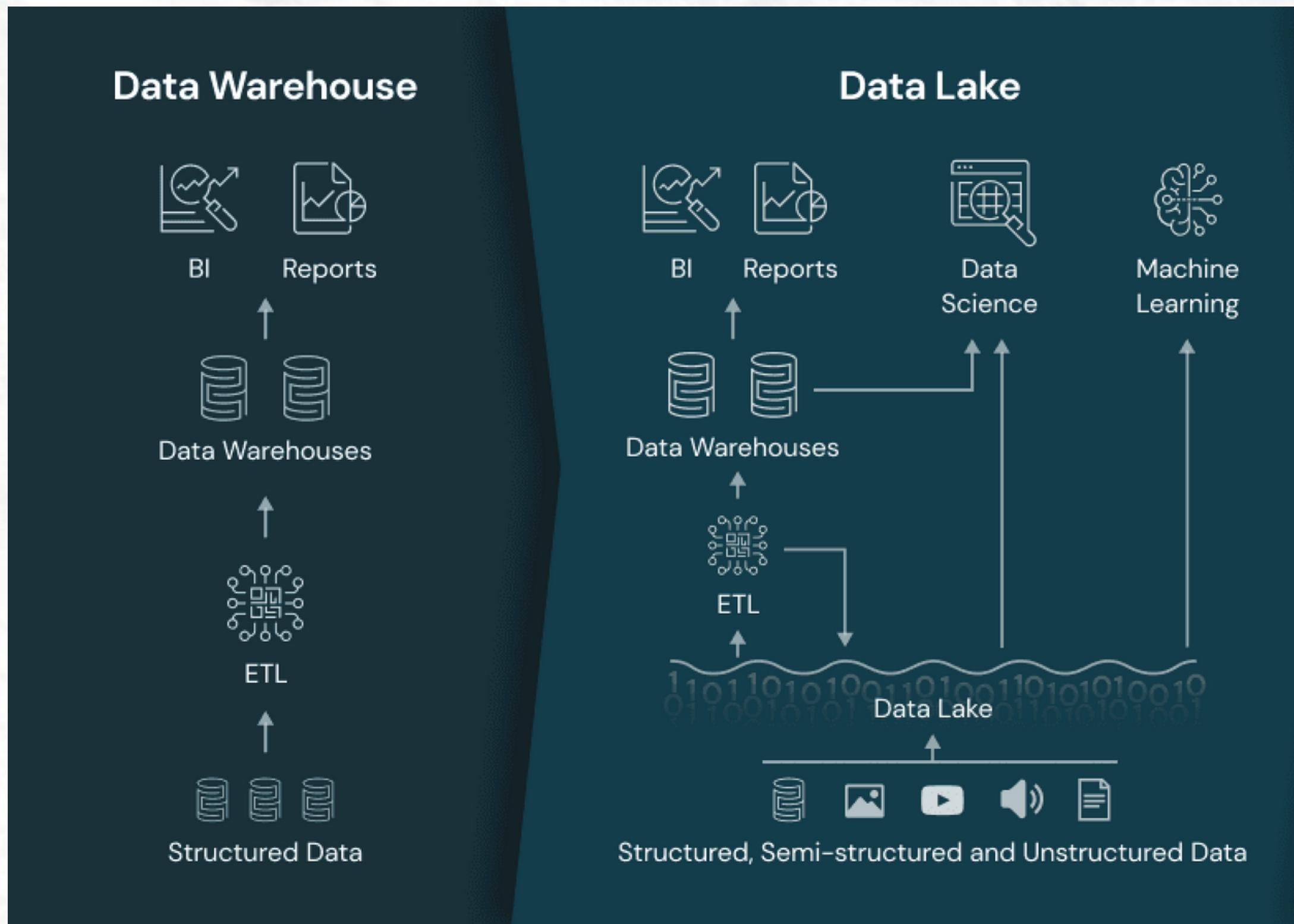
III. Data system (hệ thống dữ liệu)

DATA WAREHOUSE



Data Warehouse là kho dữ liệu tổng thể tập hợp các hệ thống dữ liệu có cấu trúc tại các phòng ban, rất phổ biến tại hầu hết doanh nghiệp, các doanh nghiệp đã có các hệ thống dữ liệu ở nhiều phòng ban, giờ tập hợp tại 1 nơi, đa phần nhiều "Business Users" có thể sử dụng dữ liệu này, đây là kho dữ liệu tổng của doanh nghiệp Data Mart nằm trong DWH ,được thiết kế riêng cho từng phòng ban

III. Data system (hệ thống dữ liệu)

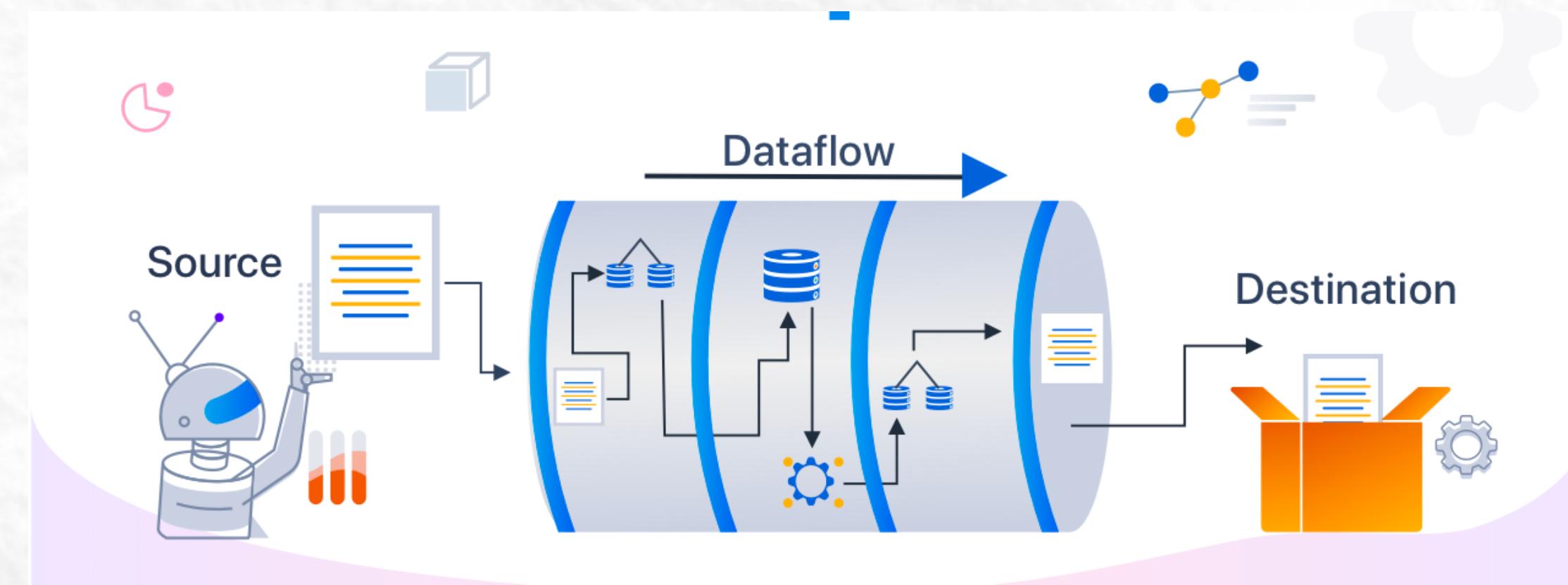


DATA LAKE

Data lake, có thể lưu trữ một lượng lớn dữ liệu có cấu trúc, bán cấu trúc và không cấu trúc. Đây là nơi lưu trữ mọi loại dữ liệu ở định dạng gốc mà không có giới hạn cố định về số lượng account hoặc file.

III. Data system (hệ thống dữ liệu)

DATA PIPELINE

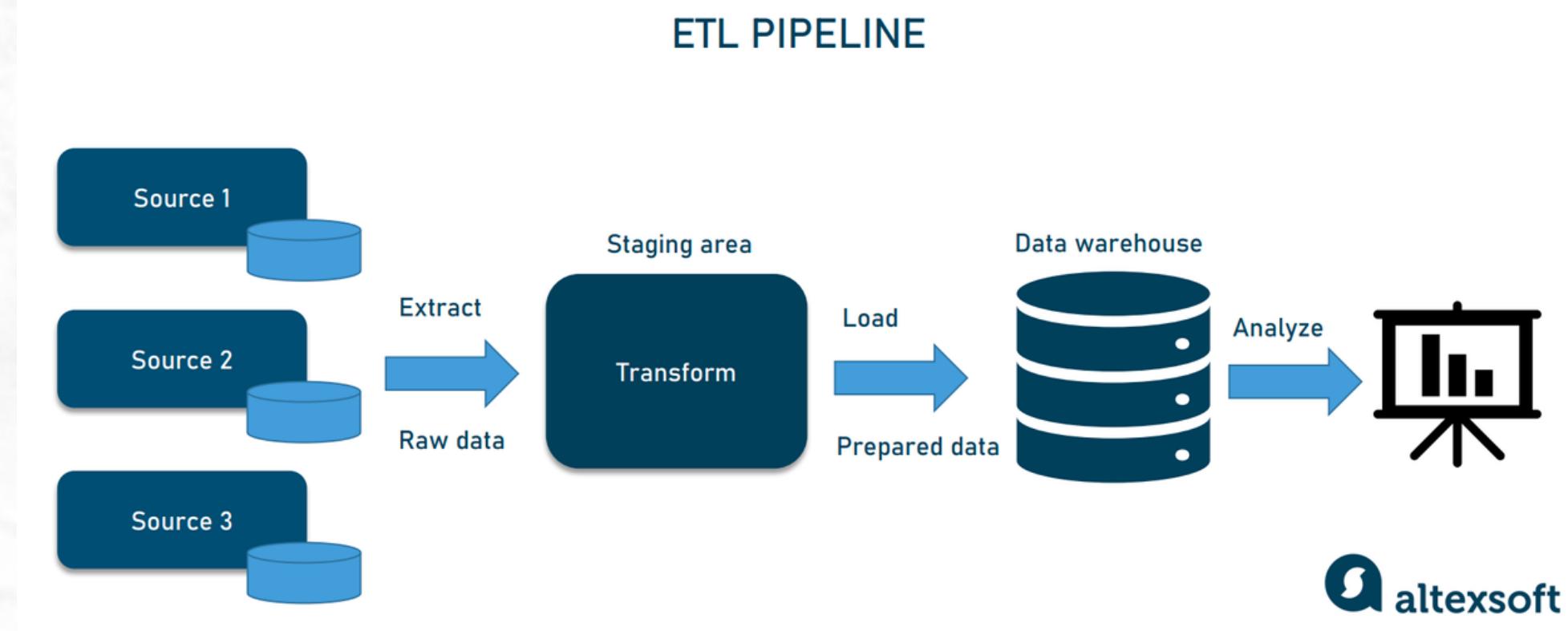


Data Pipeline (hay Đường ống dữ liệu) là một chuỗi các bước được thực hiện theo một trình tự cụ thể để xử lý dữ liệu và chuyển dữ liệu từ hệ thống này sang hệ thống khác

III. Data system (hệ thống dữ liệu)

ETL = Extract, Transform, Load

(Trích xuất, biến đổi, tải)

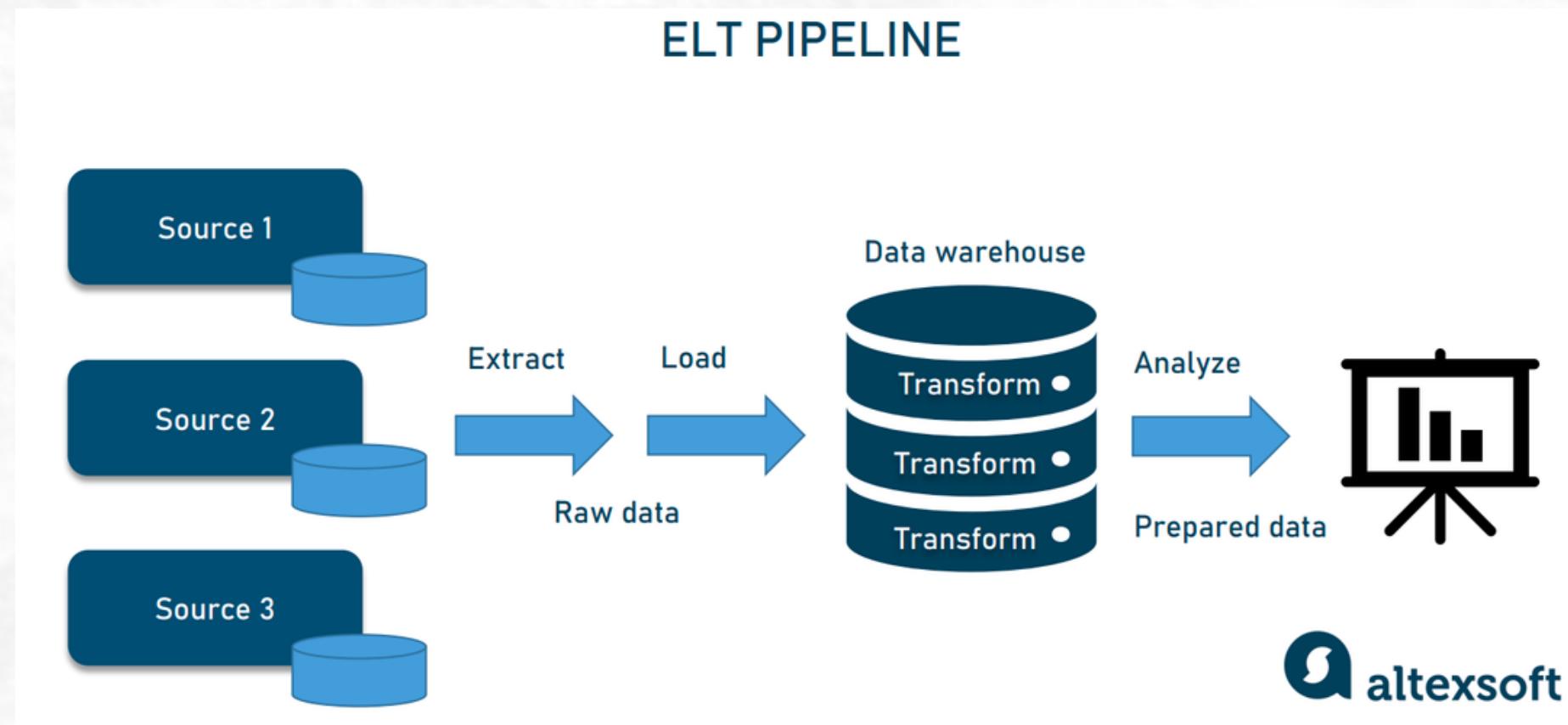


- **Extract (Trích xuất):** xác định và trích xuất các dữ liệu cần thiết, từ một hoặc nhiều nguồn khác nhau, như database, file, archives, ERP, CRM
- **Transform (Chuyển đổi):** chuyển đổi các dữ liệu trước khi được lên các database xác định (thường để trong vùng tạm trước)
- **Load (Tải lên):** tải các dữ liệu ở vùng tạm lên database xác định.

III. Data system (hệ thống dữ liệu)

ELT = Extract, Load, Transform

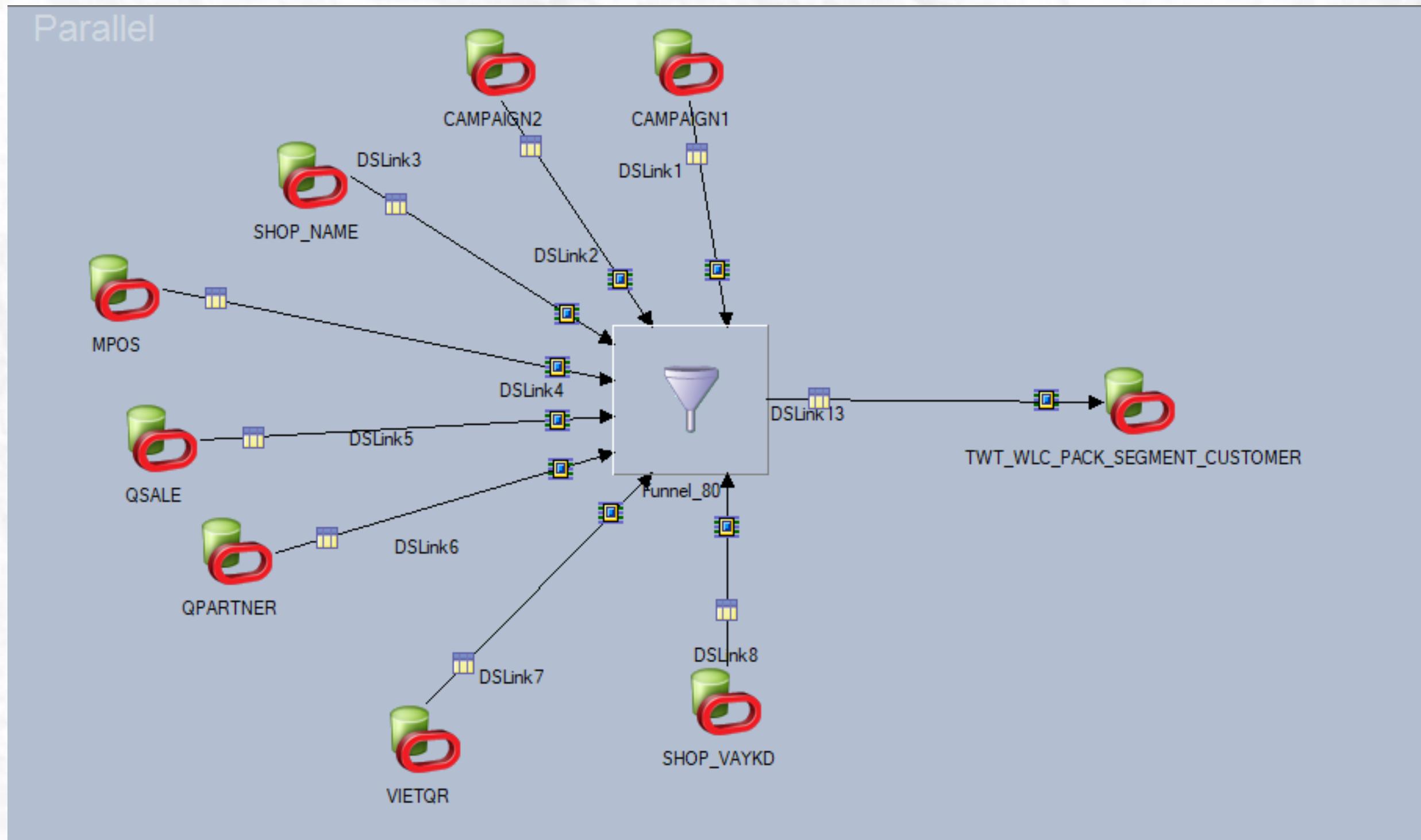
(Trích xuất, tải, biến đổi)



- **Extract (Trích xuất):** xác định và trích xuất các dữ liệu cần thiết, từ một hoặc nhiều nguồn khác nhau, như database, file, archives, ERP, CRM,
- **Load (Tải lên):** tải các dữ liệu được trích xuất sẽ được lên các database xác định.
- **Transform (Chuyển đổi):** chuyển đổi các dữ liệu để phù hợp cho việc phân tích dữ liệu.

III. Data system (hệ thống dữ liệu)

DATA PIPELINE

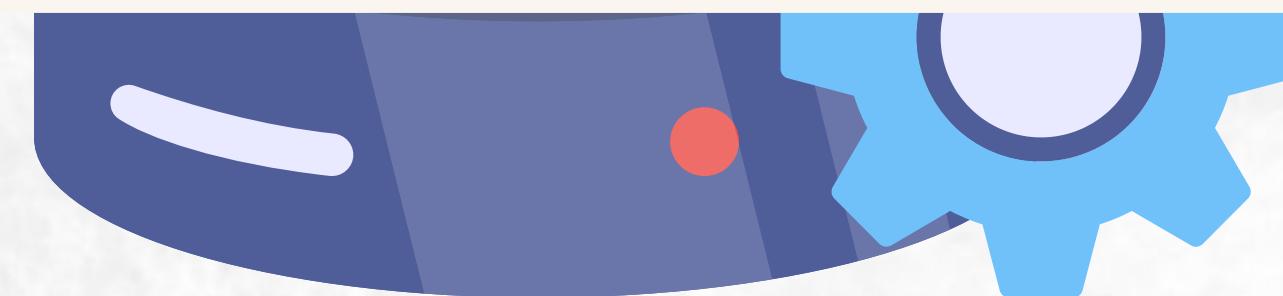


IV. SQL

SQL = Structured Query Language

Ngôn ngữ truy vấn có cấu trúc

Tương tác với dữ liệu trong database



II. Database

Hệ quản trị cơ sở dữ liệu



IV. SQL



PostgreSQL

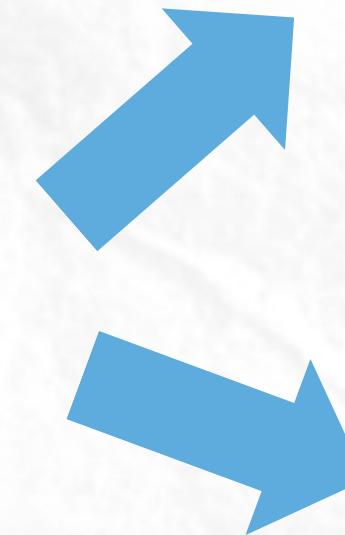
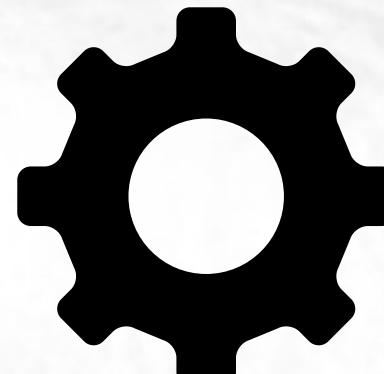
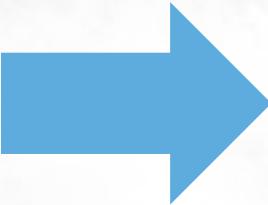
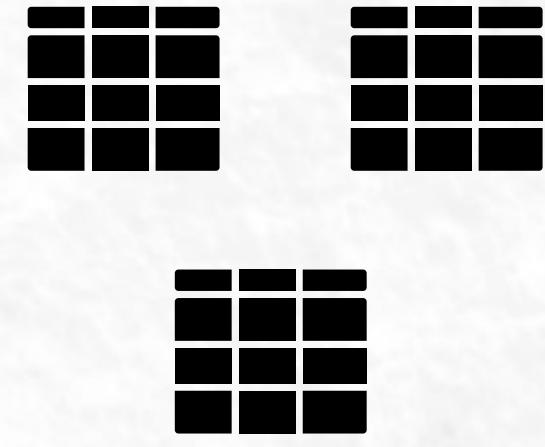
- PostgreSQL dowload và sử dụng miễn phí
- Nó có sẵn cho các hệ điều hành Windows, macOS và Linux.
- Rất phổ biến
- PostgreSQL tuân theo tất cả các tiêu chuẩn SQL (ANSI)

IV. SQL

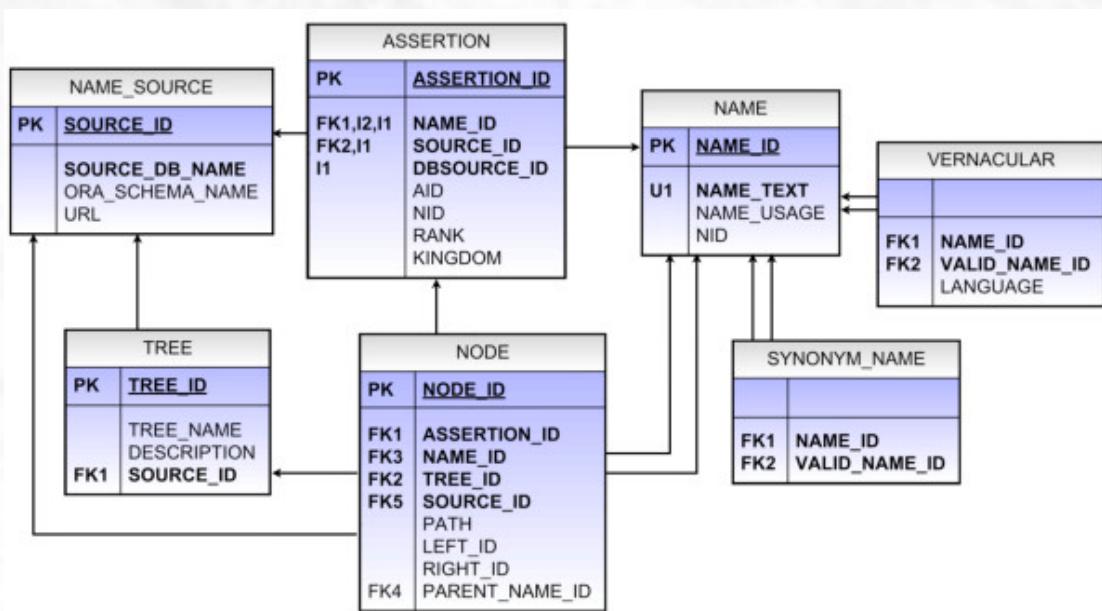
Tại sao nên học SQL

1. Dữ liệu của công ty hầu hết được lưu ở Database
2. Ngôn ngữ SQL dễ học và phổ biến
3. Là kỹ năng phải biết của:
 - Data Analyst
 - Data Scientist
 - Data Engineer
4. Thành thạo SQL có thể thúc đẩy sự nghiệp

IV. SQL



SQL

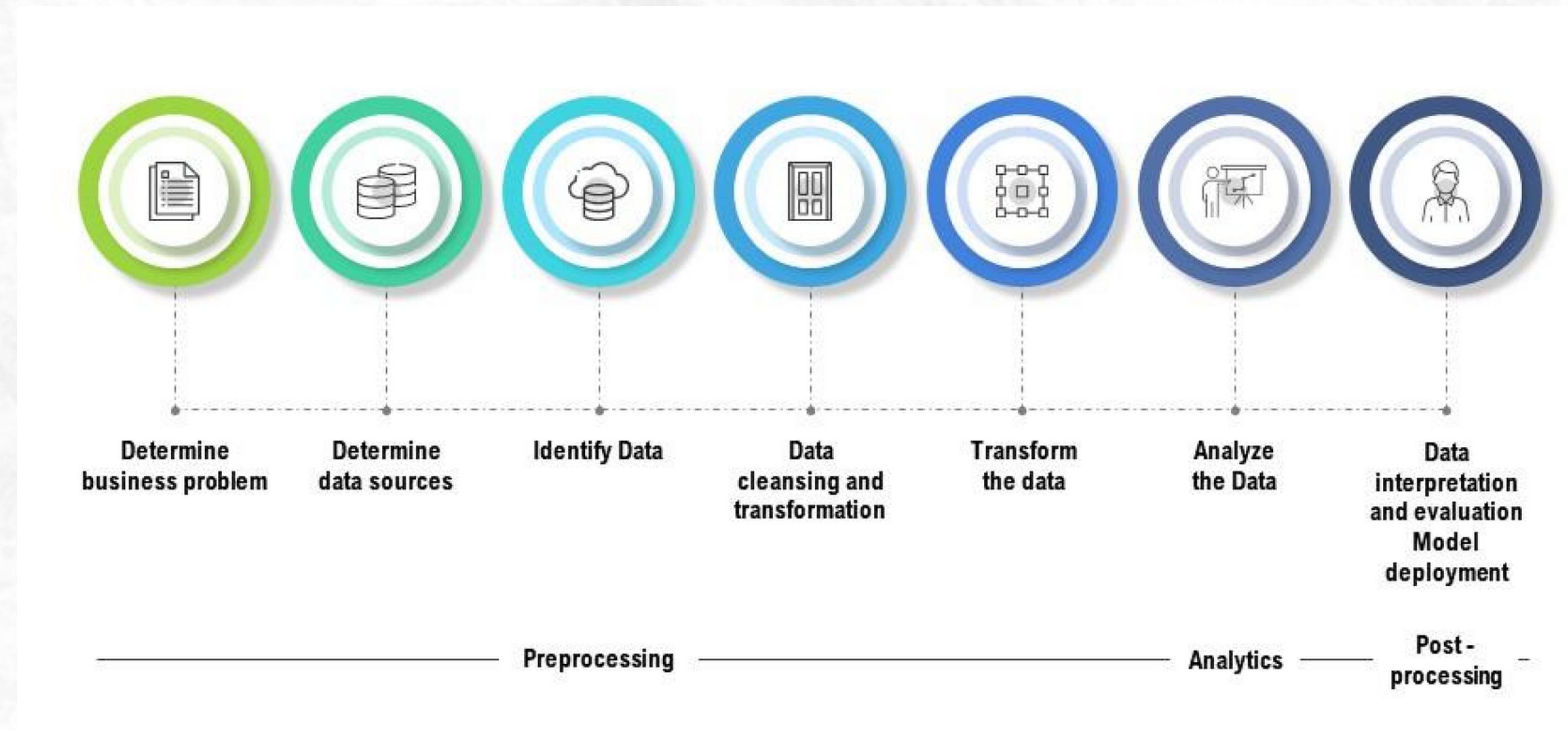


Donor ID	Type	Method	Status
S067	Donation	Credit card	Completed
S123	Shirt	Credit card	Abandoned
S345	Shirt	Paypal	Completed
S367	Donation	Cash	Completed
S121	Shirt	Paypal	Failed
S112	Donation	Credit card	Completed
S055	Donation	Credit card	Completed
S089	Donation	Paypal	Completed
S523	Shirt	Credit card	Failed
S123	Shirt	Cash	Completed
S165	Donation	Paypal	Abandoned



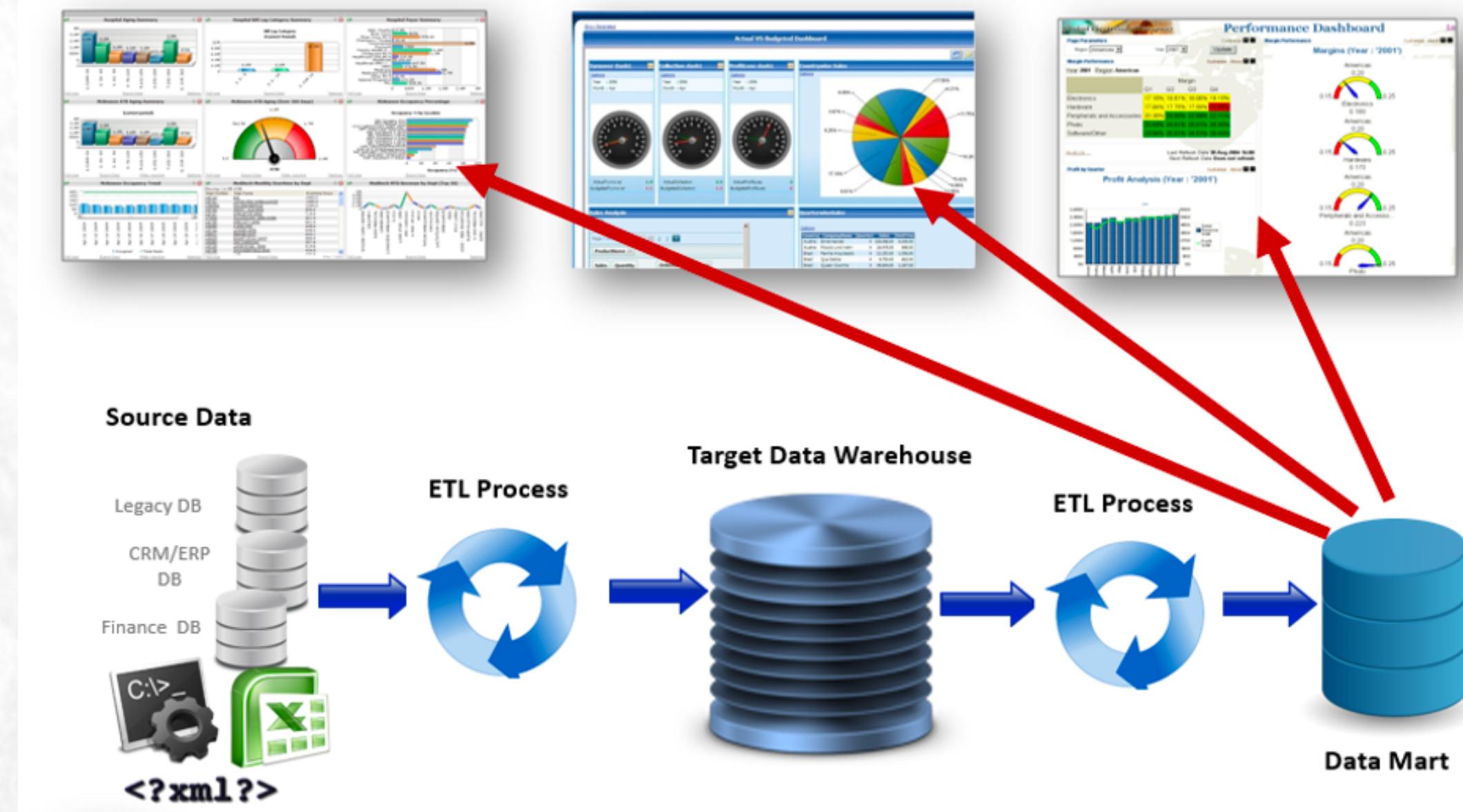
IV. SQL

SQL dùng ở đâu trong quá trình phân tích dữ liệu?



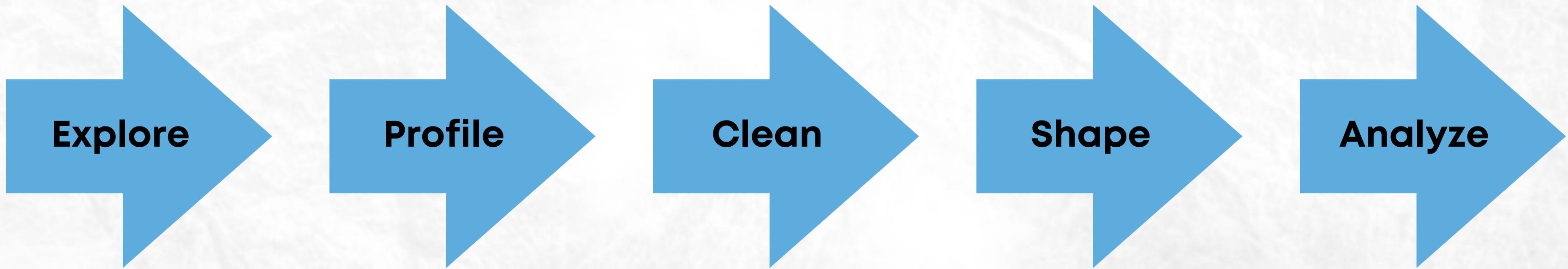
IV. SQL

SQL dùng ở đâu trong quá trình phân tích dữ liệu?



IV. SQL

SQL làm được gì trong khâu truy vấn và phân tích dữ liệu?



Làm quen với các chủ đề, bảng dữ liệu, trường thông tin

Kiểm tra tính duy nhất và phân phối của dữ liệu

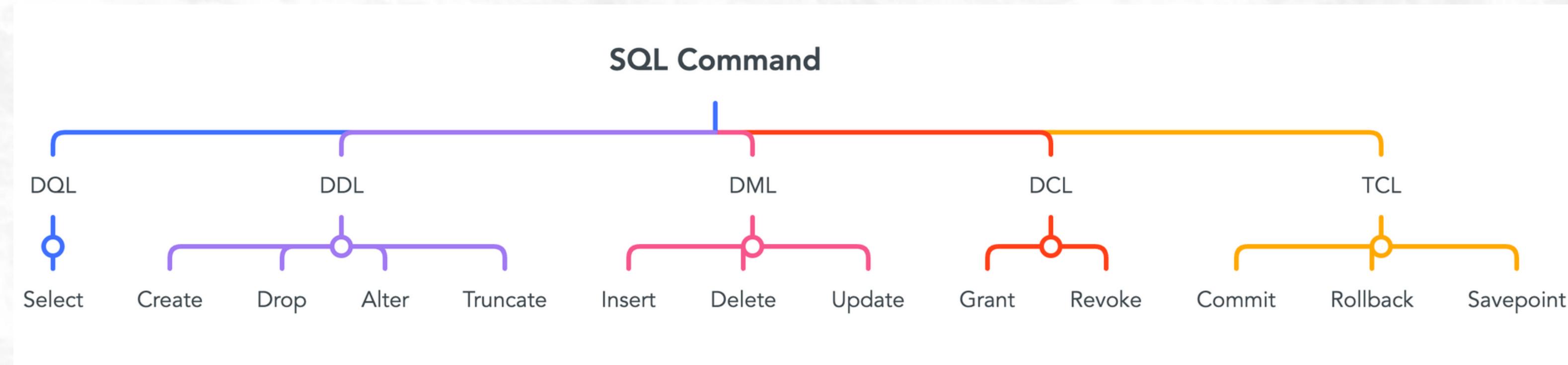
Sửa các dữ liệu ko chính xác, đầy đủ, loại trùng lặp, thêm trường cẩn thiết..

Định hình, sắp xếp dữ liệu vào các hàng, cột trong tập kết quả

Đưa ra kết luận, insight (có thể kết nối với BI Tool để trực quan kết quả)

IV. SQL

Các loại câu lệnh SQL



- DQL(Data Query Language) : lệnh truy vấn
- DDL(Data Definitine Lanaguage) : lệnh định nghĩa xây dựng và quản lý đối tượng trong DB
- DML (Data Manipulation Lanaguage) lệnh thay đổi giá trị dữ liệu trong bảng
- DCL(Data Control Lanaguage) :thao tác quản lý bảo mật dữ liệu và phân quyền đối tượng người dùng
- TCL (Transaction Control Lanaguage)