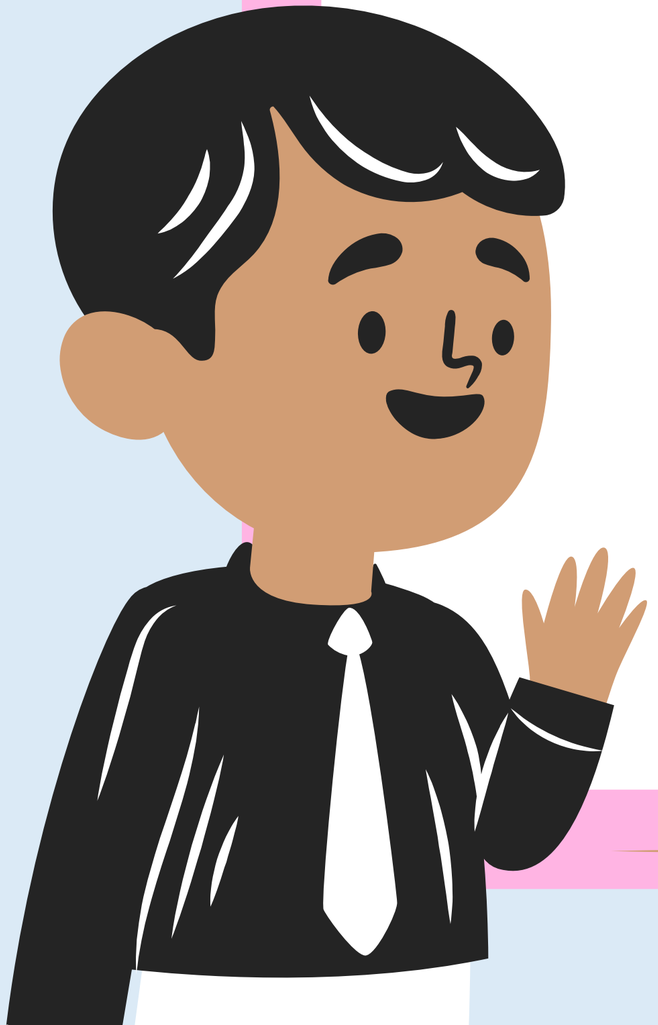
MÔN: CƠ SỞ DỮ LIỆU NÂNG CAO

CHAP 7: DATA EXTRACTION

Thuyết trình bởi Nhóm 6

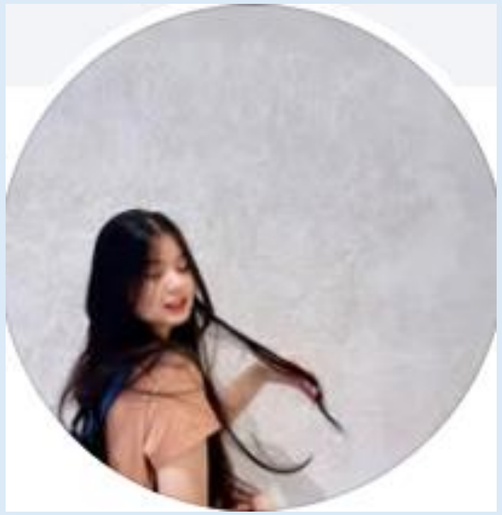
Xin chào, Thầy giáo và các bạn trong lớp KHDL1A-1!
Chúng mình là Nhóm 6!



Nhóm 6

THÀNH VIÊN NHÓM

Nguyễn Hải An
MSV: 2211090001



Đinh Lê Quỳnh Phương
MSV: 2211090031

Đinh Diệu Linh
MSV: 2211090022

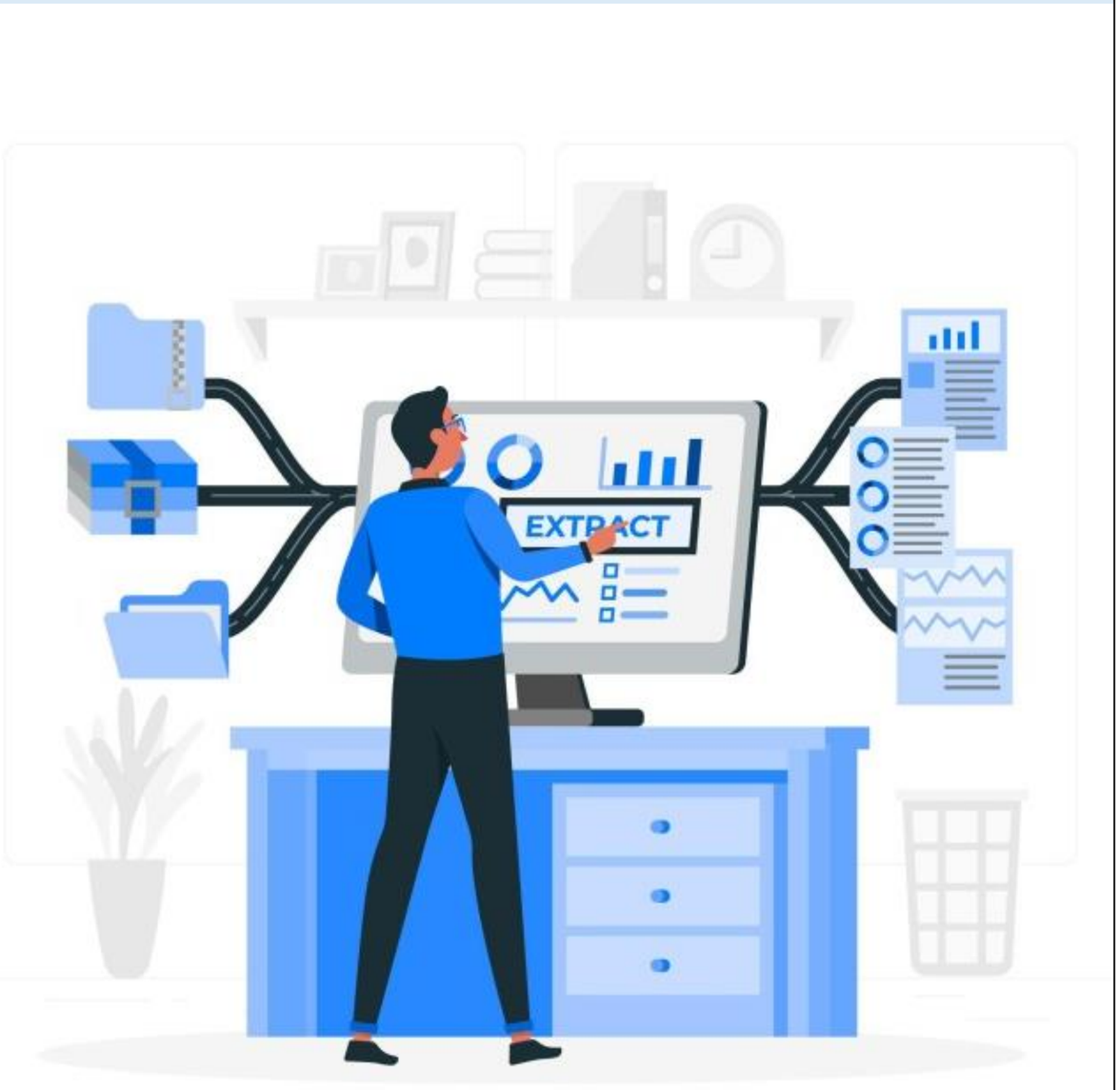
GIỚI THIỆU

Data Extraction: Đây là bước đầu tiên trong quy trình ETL (Extract, Transform, Load), có nhiệm vụ trích xuất dữ liệu từ các hệ thống nguồn để đưa vào kho dữ liệu.

Mục Tiêu Trích Xuất Dữ Liệu:

1. Thu Thập Dữ Liệu: Từ RDBMS, files, API, web logs
2. Hiệu Suất & Tốc Độ: Trích xuất hiệu quả, giảm ảnh hưởng đến hệ thống nguồn
3. Tính Toàn Vẹn Dữ Liệu: Giảm thiểu lỗi, rò rỉ dữ liệu

Vai trò và ý nghĩa: Thu thập, chuẩn hóa, làm sạch dữ liệu, đảm bảo tương thích cho phân tích và quan trọng trong tài chính, chăm sóc sức khỏe, thương mại điện tử



NGUYÊN TẮC CƠ BẢN CỦA DATA EXTRACTION

Tối ưu hóa tốc độ và kích thước dữ liệu truy suất

- Sử dụng các dấu thời gian “created” và “last_updated” => giảm tải và thời gian xử lý.
- Lập lịch truy xuất hợp lý: Chia nhỏ các khoảng thời gian trong ngày => trích xuất dữ liệu liên tục mà không ảnh hưởng lớn đến hiệu suất hệ thống chính.

Đảm bảo chất lượng và tính toàn vẹn của dữ liệu

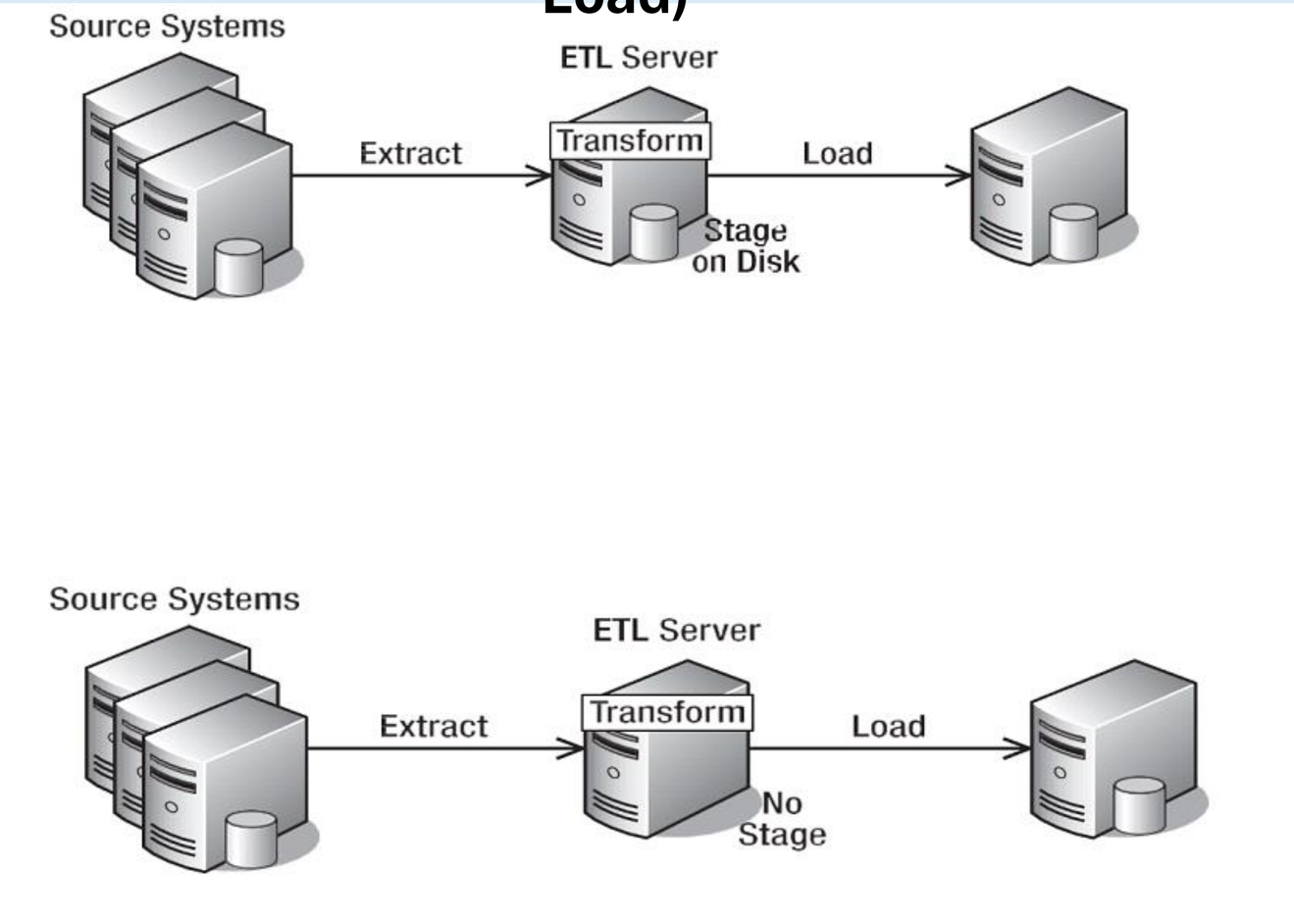
- Sử dụng các server phụ: read-only replica, Đọc dữ liệu từ log files hoặc server phụ
- Phương Pháp Đọc Incremental: Dựa vào timestamp hoặc ID tăng dần, Tránh cài đặt trigger hoặc chạy script nặng

Hạn chế ảnh hưởng đến hệ thống (OLTP)

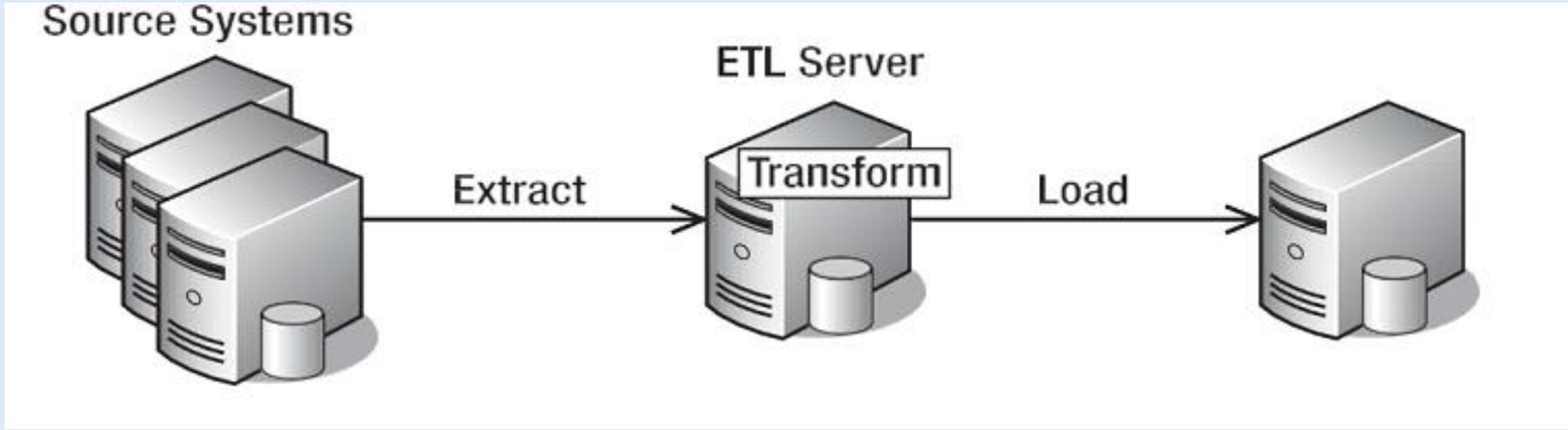
- Kiểm tra tính hợp lệ: Sử dụng checksum để đảm bảo dữ liệu không thay đổi
- Sử dụng cơ chế “watermark” => theo dõi và phục hồi bản ghi đã xử lý, đảm bảo tính toàn vẹn và khả năng khôi phục

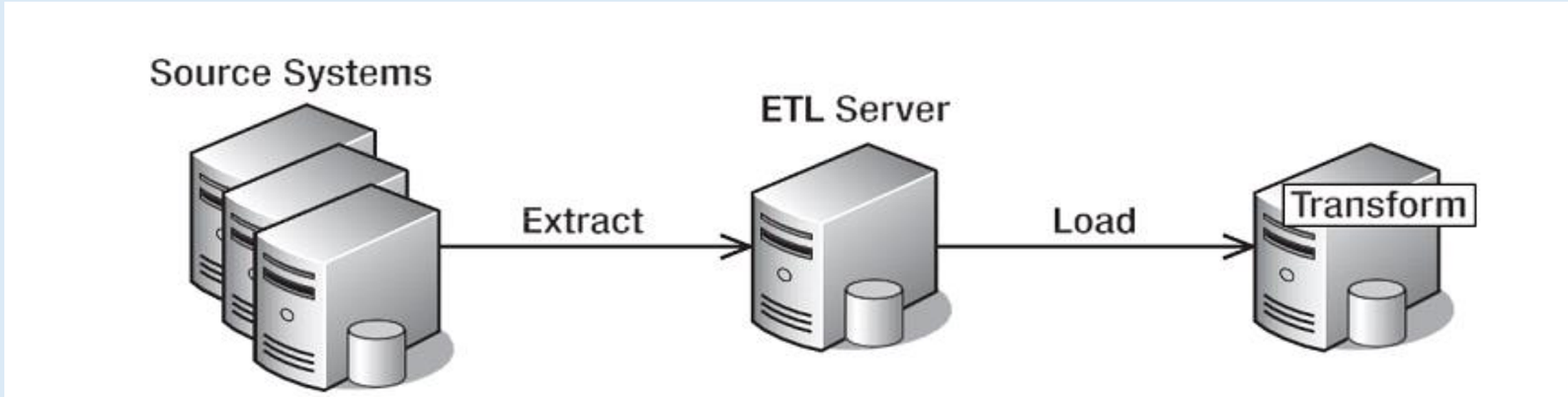
CÁC CÁCH TIẾP CẬN ETL

ETL truyền thống (Extract, Transform, Load)

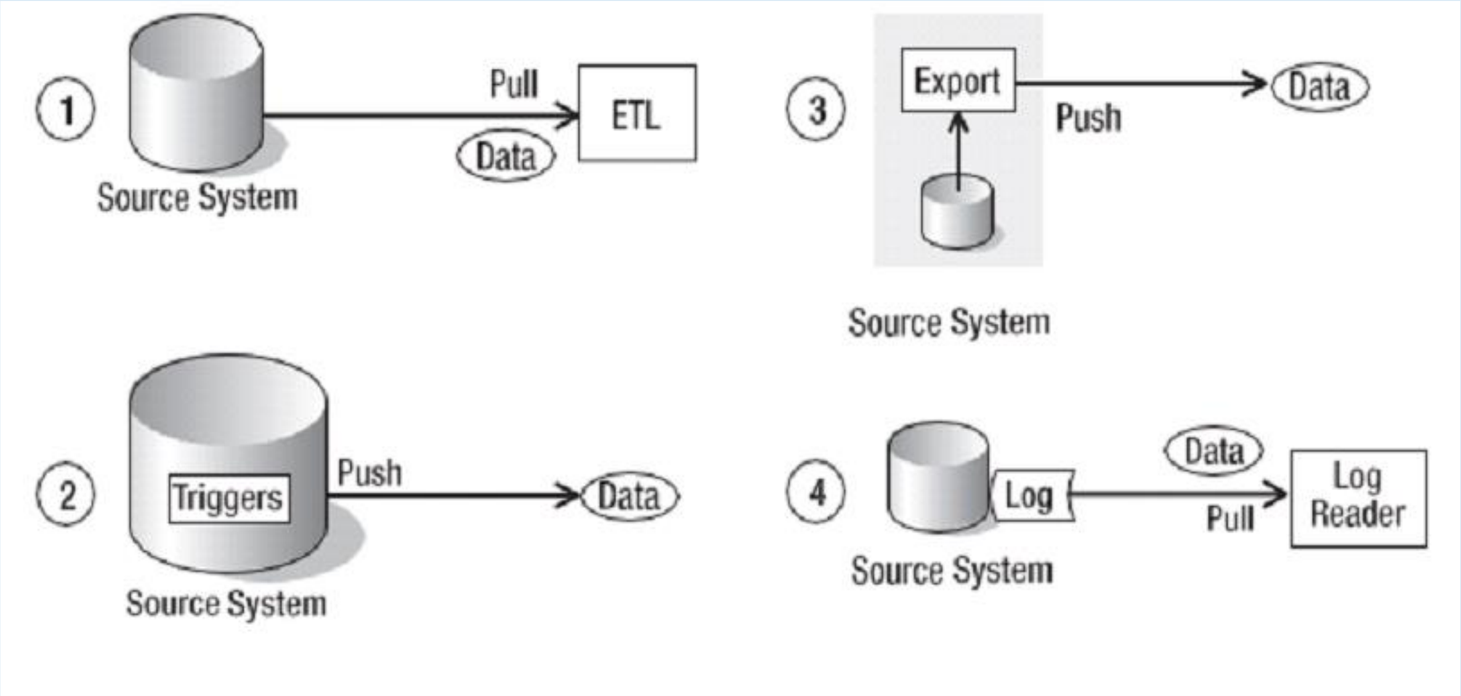


ELT (Extract, Load, Transform):





Các phương pháp ETL dựa trên việc di chuyển dữ liệu ra khỏi hệ thống nguồn



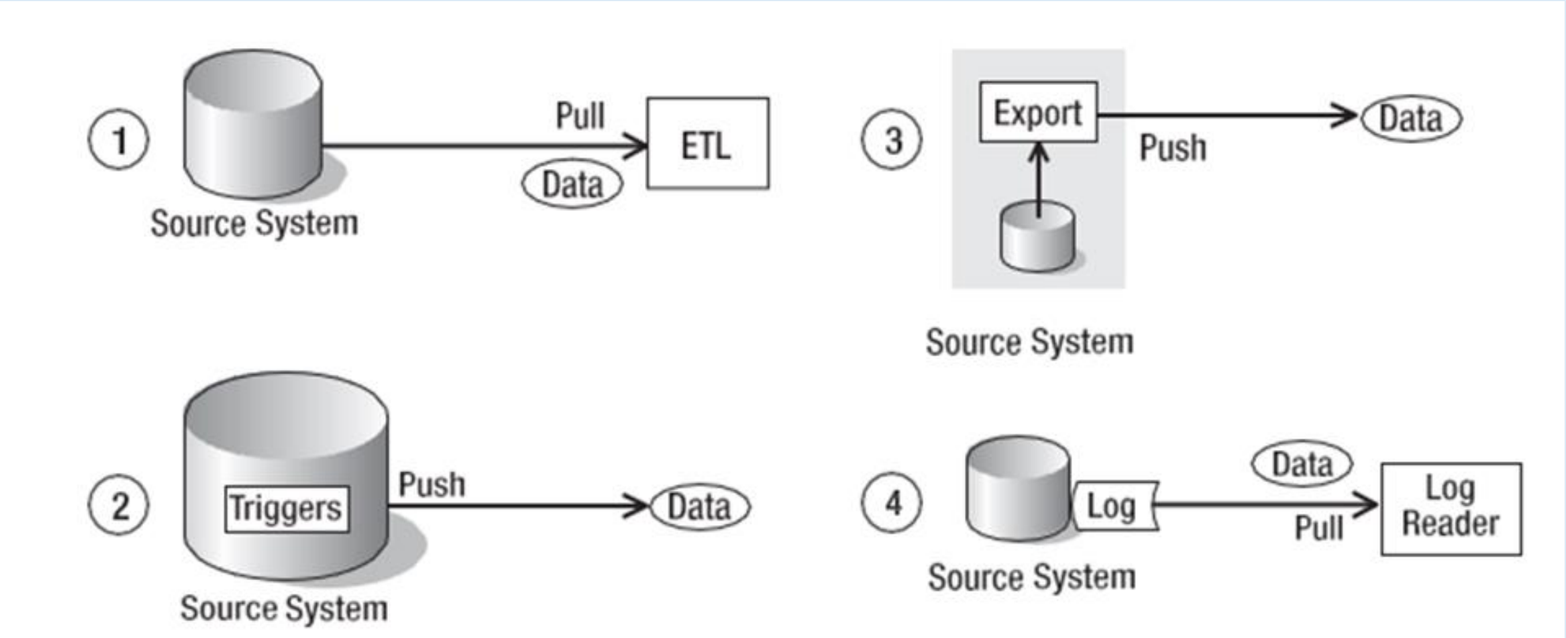
Bước 1: Truy vấn trực tiếp (Pull)

Bước 2: Đẩy dữ liệu qua Trigger (Push)

Bước 3: Xuất dữ liệu theo lịch trình

Bước 4: Đọc log file

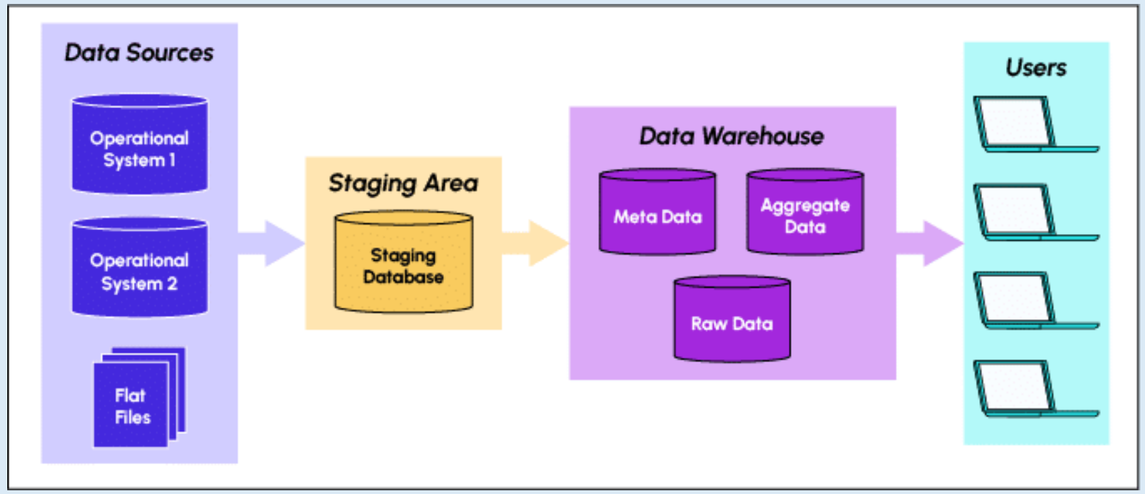
Các phương pháp ETL dựa trên nơi thực hiện quá trình ETL



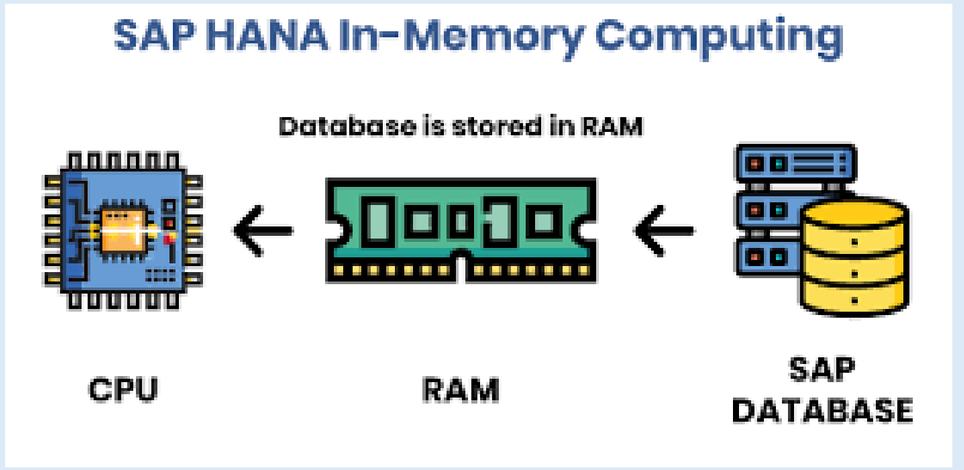
- Máy Chủ ETL Riêng Biệt: Hiệu suất cao nhất, tốn kém do mua thêm phần cứng, phần mềm
- Máy Chủ Kho Dữ Liệu: Phù hợp khi dư thừa hoặc có khung giờ không sử dụng (ví dụ ban đêm)
- Máy Chủ Hệ Thống Nguồn: Phù hợp cho kho dữ liệu thời gian thực, Dữ liệu thay đổi được truyền ngay đến kho dữ liệu

KIẾN TRÚC ETL

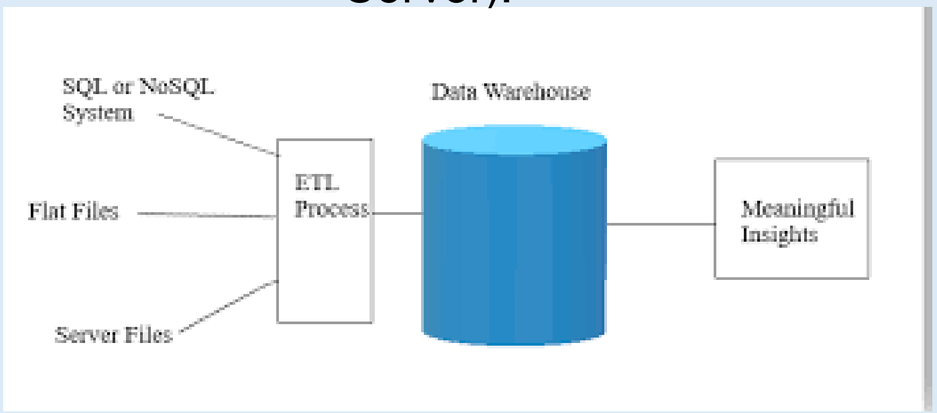
Sử dụng vùng trung gian (Staging Area):



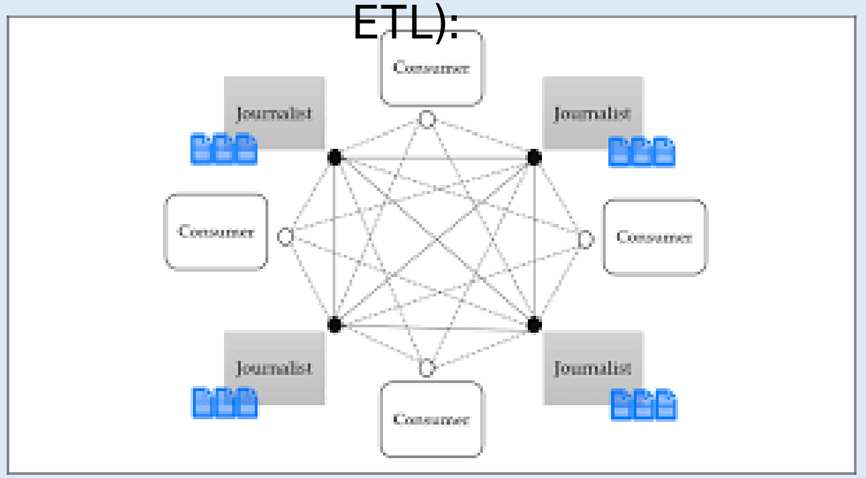
Biến đổi trong bộ nhớ (In-Memory Transformation)



Tập trung (Centralized ETL Server):



Phân tán (Distributed ETL):



CÁC PHƯƠNG PHÁP TRÍCH XUẤT DỮ LIỆU ETL (TRÍCH, XUẤT, CHUYỂN ĐỔI, TẢI):

Truy xuất toàn bộ bảng (Full Table Extraction)

Lấy toàn bộ dữ liệu từ bảng nguồn mỗi lần trích xuất, bất kể dữ liệu có thay đổi hay không.

Truy xuất gia tăng (Incremental Extraction)

Chỉ trích xuất dữ liệu đã thay đổi kể từ lần trích xuất trước (bao gồm thêm, sửa, hoặc xóa).

Phạm vi cố định (Fixed Range Extraction)

Trích xuất dữ liệu theo một phạm vi cố định (ví dụ: dữ liệu của 6 tháng gần nhất hoặc 100,000 bản ghi mới nhất).

Extract, Load, Transform (ELT):

Dữ liệu được trích xuất và tải trực tiếp vào kho dữ liệu, sau đó thực hiện biến đổi ngay trong kho dữ liệu.

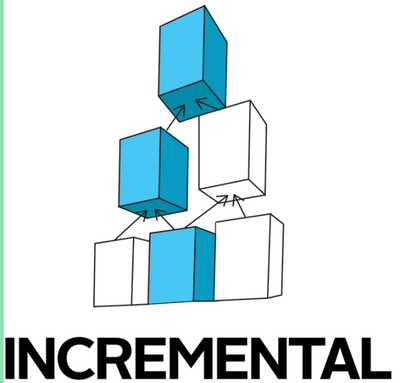
CÁC CÔNG CỤ VÀ KĨ THUẬT TRÍCH XUẤT

SQL Server Integration Services (SSIS)

Apache Kafka:

Talend Open Studio

Informatica



Incremental

CÁC VẤN ĐỀ VÀ THÁCH THỨC TRONG TRÍCH XUẤT DỮ LIỆU

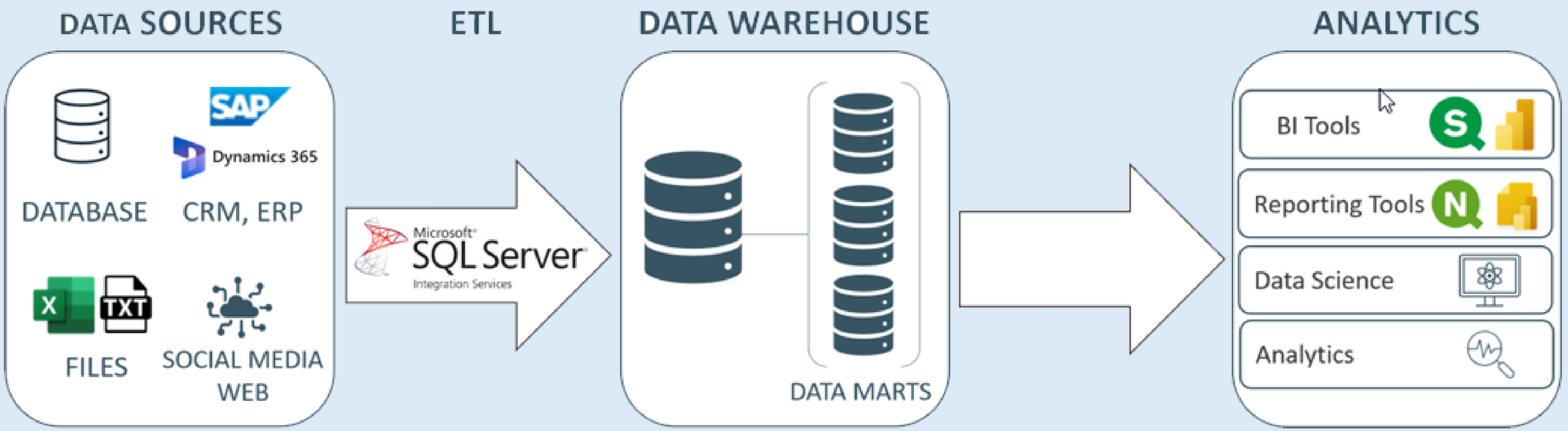
Rò Rỉ Dữ Liệu:

- Dữ liệu nhạy cảm bị truy cập hoặc tiết lộ trái phép
- Dẫn đến vi phạm bảo mật và mất mát thông tin

Ảnh Hưởng Đến Hệ Thống Nguồn:

- Tải nặng gây giảm hiệu suất
- Ảnh hưởng đến hoạt động của các ứng dụng khác

CÔNG CỤ VÀ VÍ DỤ THỰC TIỄN



ĐÁNH GIÁ ƯU VÀ NHƯỢC ĐIỂM CỦA PHƯƠNG PHÁP TRÍCH XUẤT DỮ LIỆU

ƯU ĐIỂM

- Tải Gia tăng
- Tự động hóa
- Khả năng mở rộng
- Khả năng tích hợp linh hoạt

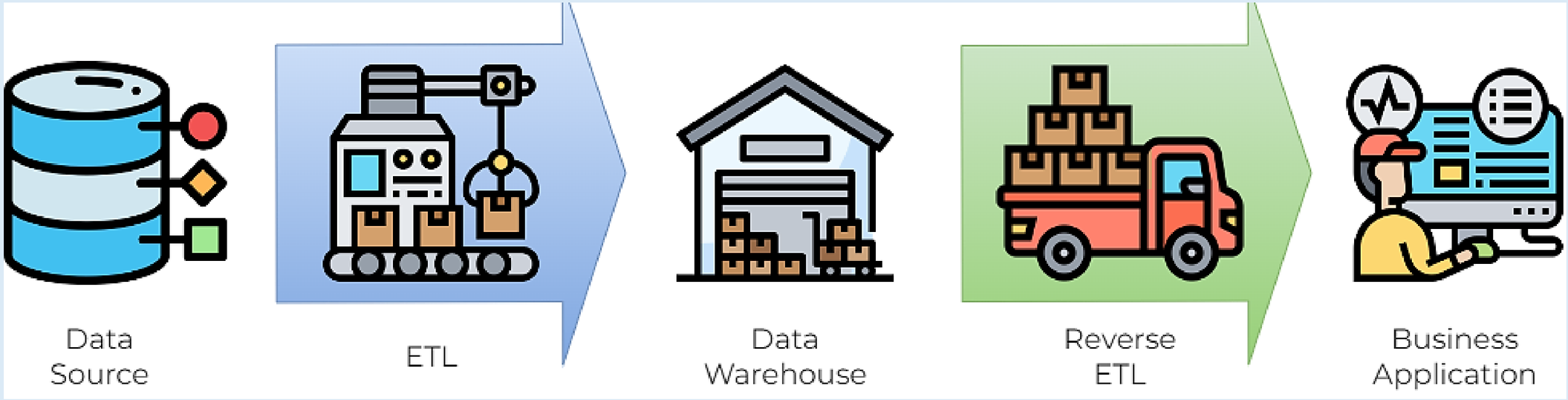
ĐÁNH GIÁ ƯU VÀ NHƯỢC ĐIỂM CỦA PHƯƠNG PHÁP TRÍCH XUẤT DỮ LIỆU

NHƯỢC ĐIỂM

- Tính phức tạp
- Phụ thuộc vào công cụ
- Hiệu suất giới hạn
- Rủi ro bỏ lỡi dữ liệu

KẾT LUẬN

Trích xuất dữ liệu là bước cốt lõi trong quy trình ETL, đòi hỏi đầu tư về công nghệ, kỹ năng, và quy trình bài bản. Những phương pháp trích xuất khác nhau mang đến đồng thời các ưu điểm và nhược điểm, đòi hỏi người thực hiện phải cân nhắc tùy theo nhu cầu và hạ tầng của hệ thống.



NGUỒN THAM KHẢO

01 Schmidt, L., Finnerty Mutlu, A. N., Elmore, R., Olorisade, B. K., Thomas, J., & Higgins, J. P. T. (2021). Data extraction methods for systematic review (semi)automation: Update of a living systematic review. doi.org/10.12688/f1000research.51117.2

02 Mykowiecka, A., Marciniak, M., & Kupść, A. (2009). Rule-based information extraction from patients' clinical data. Journal of Biomedical Informatics, 1 42(6), 1021–1032. https://doi.org/10.1016/j.jbi.2009.07.007

XIN CẢM ƠN!

Các bạn có bất kỳ câu hỏi nào cho chúng mình không?