BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÁO CÁO THỰC TẬP CƠ SỞ

BÁO CÁO TÌM HIỂU VỀ ETL VÀ DATA PIPELINE

Giảng viên hướng dẫn : Kim Ngọc Bách

Họ và tên : Ngô Vũ Minh Quý

Mã sinh viên : B22DCVT427

Lớp : E22CQCN02-B

MỤC LỤC

1. Giới thiệu chung

Trong kỷ nguyên dữ liệu hiện nay, việc thu thập, xử lý và phân tích dữ liệu là nền tảng quan trọng để doanh nghiệp đưa ra quyết định. Tuy nhiên, dữ liệu thô không có giá trị nếu không được xử lý và tổ chức hợp lý. ETL (Extract – Transform – Load) và Data Pipeline (dòng xử lý dữ liệu) là những công cụ không thể thiếu trong quy trình đó. Bài báo cáo này sẽ trình bày tổng quan từ cơ bản đến nâng cao về hai khái niệm trên, giúp người đọc có cái nhìn toàn diện và áp dụng hiệu quả vào thực tiễn.

2. Khái niệm về ETL

ETL là viết tắt của ba bước chính trong quy trình xử lý dữ liệu:

- Extract (Trích xuất): Lấy dữ liệu từ nhiều nguồn khác nhau như cơ sở dữ liệu, file CSV, API, hoặc hệ thống ERP.
- Transform (Chuyển đổi): Là bước xử lý, làm sạch, chuẩn hóa dữ liệu, tạo ra các bảng dữ liệu có cấu trúc phù hợp với nhu cầu phân tích.
- Load (Tải vào): Dữ liệu đã được xử lý sẽ được tải vào hệ thống lưu trữ đích như Data Warehouse (kho dữ liệu) hoặc Data Lake.

ETL thường được sử dụng trong các hệ thống xử lý dữ liệu theo lô (batch), nơi dữ liệu được thu thập và xử lý định kỳ.

---

3. Các bước trong quy trình ETL

3.1. Trích xuất (Extract)

- Lấy dữ liệu từ nhiều nguồn không đồng nhất.
- Các công nghệ thường dùng: JDBC/ODBC, API REST, công cụ như Apache NiFi.
- Thách thức: dữ liệu thiếu đồng nhất, dữ liệu thời gian thực.

3.2. Chuyển đổi (Transform)

- Làm sạch dữ liệu: loại bỏ null, chuẩn hóa định dạng.
- Ánh xạ dữ liệu: từ các schema khác nhau về một dạng chung.
- Áp dụng quy tắc nghiệp vụ (business rules): tính toán, phân loại, lọc.
- Công cụ: Python (pandas), Apache Spark, dbt (Data Build Tool).

3.3. Tải dữ liệu (Load)

- Ghi dữ liệu vào hệ thống đích.
- Có thể sử dụng cách tải toàn bộ hoặc tải theo dạng incremental (chỉ cập nhật phần thay đổi).
- Các hệ thống đích: PostgreSQL, Amazon Redshift, Google BigQuery, Snowflake.

---

4. Data Pipeline là gì?

Data pipeline là một hệ thống hoặc quy trình tự động hóa cho phép dữ liệu di chuyển từ nơi này sang nơi khác, có thể bao gồm các bước như:

- Thu thập
- Xử lý

- Phân tích
- Lưu trữ
- Trực quan hóa

Pipeline không nhất thiết phải có đủ ba bước ETL, mà có thể linh hoạt tùy thuộc vào mục tiêu cụ thể. Một pipeline hiện đại thường hỗ trợ dữ liệu thời gian thực (streaming).

5. So sánh ETL và Data Pipeline

Tiêu chí	ETL	Data Pipeline
Mục tiêu chính	**Trích xuất – chuyển đổi – tải dữ liệu**	**Truyền dẫn và xử lý dữ liệu từ đầu đến cuối**
Tính chất	**Batch (theo lô)**	**Có thể batch hoặc real-time**
Yếu tố xử lý	**Tập trung vào chuyển đổi**	**Có thể không cần chuyển đổi**
Tính linh hoạt	**Tương đối cố định**	**Linh hoạt, tùy theo use-case**

6. Công cụ và công nghệ phổ biến cho ETL và pipeline

6.1. Công cụ ETL truyền thống

- Informatica
- Talend
- Microsoft SSIS (SQL Server Integration Services)

6.2. Công cụ hiện đại

- Apache Airflow: orchestrator pipeline.
- Apache Spark: xử lý dữ liệu lớn.
- dbt (data build tool): thực hiện Transform trong mô hình ELT.
- Fivetran, Stitch: công cụ ETL SaaS.
- Kafka + Spark Streaming: pipeline dữ liệu thời gian thực.

---

7. ETL hiện đại và xu hướng ELT

7.1. ELT là gì?

- Thay vì chuyển đổi dữ liệu trước khi tải (ETL), ELT là mô hình mới: Extract – Load – Transform.
- Phù hợp với cloud-based data warehouse (Snowflake, BigQuery) nhờ khả năng xử lý mạnh mẽ.

7.2. Ưu điểm của ELT

- Dễ mở rộng (scalable).
- Thích hợp cho dữ liệu lớn (big data).
- Có thể sử dụng SQL để chuyển đổi trong data warehouse.

---

8. Ứng dụng thực tế trong doanh nghiệp

8.1. Lĩnh vực tài chính

- Phân tích giao dịch.
- Phát hiện gian lận.
- Tích hợp dữ liệu từ nhiều hệ thống ngân hàng.

8.2. Thương mại điện tử

- Thu thập dữ liệu khách hàng, đơn hàng, hành vi.
- Tạo báo cáo BI.
- Dự đoán nhu cầu sản phẩm.

8.3. Y tế

- Tổng hợp hồ sơ bệnh án.
- Phân tích chẩn đoán bệnh.
- Nghiên cứu lâm sàng.

---

9. Thách thức và hướng phát triển

9.1. Thách thức

- Chất lượng dữ liệu thấp.
- Hệ thống phân mảnh.

- Thay đổi yêu cầu nghiệp vụ liên tục.
- Dữ liệu thời gian thực ngày càng phổ biến.

9.2. Hướng phát triển

- Tự động hóa pipeline với AI/ML.
- ETL serverless với dịch vụ như AWS Glue.
- Real-time ETL với Kafka, Flink.
- DataOps: tích hợp DevOps vào pipeline dữ liệu.

---

10. Kết luận

ETL và data pipeline đóng vai trò quan trọng trong quá trình xây dựng hệ thống dữ liệu hiện đại. Khi dữ liệu ngày càng tăng trưởng nhanh và phức tạp, việc hiểu rõ và vận dụng đúng các mô hình ETL/ELT và data pipeline sẽ giúp tổ chức khai thác tối đa giá trị từ dữ liệu. Trong tương lai, sự kết hợp giữa pipeline tự động, dữ liệu thời gian thực và công nghệ đám mây sẽ tiếp tục định hình cách thức các doanh nghiệp xử lý và sử dụng dữ liệu.