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

I. Tổng quan về điện toán đám mây

II. Giới thiệu về Microsoft Azure

III. Giới thiệu về Amazon Web Services (AWS)

IV. So sánh Azure và AWS

V. Ứng dụng thực tiễn và kết luận

I. Tổng quan về kho dữ liệu (Data Warehouse)

Kho dữ liệu (Data Warehouse) là hệ thống lưu trữ tập trung được thiết kế để hỗ trợ việc phân tích và báo cáo dữ liệu. Nó giúp tổng hợp dữ liệu từ nhiều nguồn khác nhau, biến đổi và lưu trữ trong một cấu trúc tối ưu cho việc khai thác thông tin.

Khác với cơ sở dữ liệu giao dịch (OLTP), kho dữ liệu tập trung vào các hoạt động phân tích (OLAP) như truy vấn, thống kê, phân khúc và dự báo.

Đặc điểm chính của Data Warehouse:

- Tích hợp dữ liệu từ nhiều nguồn không đồng nhất.
- Tối ưu cho đọc và phân tích, không phục vụ ghi trực tiếp thường xuyên.
- Lưu trữ dữ liệu lịch sử để phục vụ phân tích theo thời gian.
- Không thay đổi dữ liệu thường xuyên (non-volatile).

II. Kiến trúc và thành phần của Data Warehouse

Một hệ thống Data Warehouse thường bao gồm các lớp:

1. **Nguồn dữ liệu (Data Sources):**
   Các hệ thống như CSDL giao dịch (ERP, CRM), file Excel, logs, dữ liệu IoT,...
2. **Lớp trích xuất – biến đổi – tải (ETL):**
   - **Extract:** Trích xuất dữ liệu từ nhiều nguồn.
   - **Transform:** Làm sạch, chuẩn hóa, tổng hợp dữ liệu.

    - **Load:** Đưa dữ liệu đã xử lý vào kho dữ liệu.

3. **Kho dữ liệu trung tâm (Central Data Warehouse):**
    - Lưu trữ dữ liệu đã được xử lý.
    - Thiết kế theo mô hình **ngôi sao (star schema)** hoặc **bông tuyết (snowflake schema)**.
4. **Lớp truy vấn và phân tích (BI layer):**
    - Dùng để kết nối công cụ trực quan hóa và truy vấn dữ liệu như Looker, Power BI, Tableau,...

III. Quy trình ETL trong Data Warehouse

ETL là một thành phần cốt lõi trong Data Warehouse, bao gồm 3 bước:

- **Extract (Trích xuất):**
  Thu thập dữ liệu từ các hệ thống khác nhau.
- **Transform (Biến đổi):**
  Làm sạch dữ liệu (loại bỏ trùng lặp, xử lý null), tính toán thêm các trường mới, chuẩn hóa định dạng.
- **Load (Tải dữ liệu):**
  Đưa dữ liệu đã xử lý vào kho trung tâm theo định kỳ hoặc theo thời gian thực.

Một số công cụ ETL phổ biến:

- **Apache NiFi**, **Talend**, **Informatica**, **Microsoft SSIS**, **Google Dataflow**, **AWS Glue**.

IV. Các nền tảng và công nghệ kho dữ liệu phổ biến

Hiện nay có nhiều nền tảng kho dữ liệu trên nền tảng truyền thống lẫn cloud. Một số nền tảng tiêu biểu:

Nền tảng	Mô tả
Amazon Redshift	Kho dữ liệu đám mây của AWS, tối ưu cho phân tích song song.
Google BigQuery	Kho dữ liệu serverless, có khả năng xử lý petabyte dữ liệu cực nhanh.
Snowflake	Nền tảng kho dữ liệu hiện đại, hỗ trợ lưu trữ, chia sẻ và phân tích dữ liệu đa đám mây.
Microsoft Azure Synapse	Kết hợp kho dữ liệu với khả năng phân tích nâng cao.
Oracle Exadata	Kho dữ liệu cấp doanh nghiệp, tích hợp tối ưu phần cứng và phần mềm.

Ngoài ra, các công nghệ liên quan đến Data Warehouse còn bao gồm:

- **SQL**, **OLAP Cubes**, **Data Mart**, **Data Lake**,...
- Công cụ BI như **Looker**, **Power BI**, **Tableau** giúp truy vấn và trực quan hóa dữ liệu dễ dàng hơn.

V. Ứng dụng thực tiễn và đánh giá

1. Ứng dụng thực tiễn của Data Warehouse

- **Doanh nghiệp thương mại điện tử**: Phân tích hành vi khách hàng, dự đoán xu hướng mua hàng.
- **Ngân hàng, tài chính**: Quản lý rủi ro, phát hiện gian lận, phân tích hiệu suất chi nhánh.
- **Y tế, giáo dục**: Tổng hợp dữ liệu bệnh án, quản lý học sinh – sinh viên.
- **Logistics, vận tải**: Theo dõi hành trình vận chuyển, tối ưu lộ trình.
- **Công ty công nghệ**: Phân tích sản phẩm, kiểm tra hiệu năng tính năng mới.

2. Đánh giá cá nhân

Qua quá trình tìm hiểu, em nhận thấy rằng Data Warehouse đóng vai trò quan trọng trong việc giúp tổ chức **biến dữ liệu thành thông tin hữu ích**, hỗ trợ quá trình **ra quyết định chiến lược**.

Kho dữ liệu hiện đại không chỉ lưu trữ và xử lý dữ liệu mà còn tích hợp với các nền tảng **AI/ML**, **real-time analytics**, mở ra khả năng xây dựng các **hệ thống thông minh**, thích ứng linh hoạt theo thị trường.

Việc học cách thiết kế mô hình dữ liệu, làm quen với ETL và hiểu kiến trúc Data Warehouse là nền tảng vững chắc để sinh viên theo đuổi các lĩnh vực như **Data Engineering**, **Business Intelligence**, **Data Analytics**.

3.Kết luận

Data Warehouse là một thành phần quan trọng trong hệ sinh thái dữ liệu hiện đại. Nó không chỉ giúp doanh nghiệp lưu trữ lịch sử dữ liệu hiệu quả mà còn hỗ trợ phân tích, trực quan hóa và ra quyết định. Việc hiểu và sử dụng Data Warehouse là kỹ năng không thể thiếu cho bất kỳ sinh viên CNTT nào định hướng theo lĩnh vực dữ liệu.

Trong thời đại mà **“dữ liệu là dầu mỏ mới”**, việc nắm vững kiến thức về Data Warehouse sẽ mở ra nhiều cơ hội việc làm và định hướng nghề

nghiệp đầy tiềm năng trong các lĩnh vực: **Data Engineering**, **Data Science**, **Cloud Computing** và **AI**.