Medallion Architecture

Khái niệm - Cách thực thi - Các nguyên lý - Tiêu chí đánh giá - Khuyến khích áp dụng vào đồ án

Bronze layer (raw data): data được lấy trực tiếp từ source về và chưa được chuẩn hóa
Lưu lại các mốc thời gian dữ liệu được kéo về, ID,... Mục tiêu ở layer này đó chính là khả năng CDC (Change Data Capture) và lưu lại, quay lại lịch sử thay đổi của data. Có thể làm lại quá trình preprocess mà không cần lấy lại data từ nguồn gốc.

Silver layer (cleansed and conformed): Silver data là những data Bronze được merge, làm cho dữ liệu từ các nguồn cùng form với nhau và được clean 1 cách vừa đủ để đưa ra cái nhìn tổng thể cho doanh nghiệp. Mục tiêu của layer này là nguồn để phục vụ ML và những phân tích dữ liệu. Trong lakehouse data, ELT sẽ được ưu tiên hơn ETL, họ sẽ ưu tiên biến đổi gần đủ và những rule cleaning sẽ được áp dụng khi load silver layer. Trong lớp này họ sẽ ưu tiên tốc độ nạp vào và khả năng truy xuất ra bên ngoài. Khi dữ liệu từ Silver -> Gold, dữ liệu sẽ trải qua một loạt các rule business. Thông thường dữ liệu sẽ ở dạng chuẩn 3.
VD: Python

Gold layer (curated business - level): dữ liệu được tổ chức cho những tác vụ cụ thể. Lớp này được dùng để báo cáo, sử dụng thêm những phương pháp de-normalized và cần dùng ít join hơn. Những cái rules cuối cùng cho việc chuẩn hóa data và đảm bảo chất lượng của data được áp dụng tại đây. Bình thường sẽ sử dụng các loại schema như Kimball hoặc Inmon, Snowflake
VD: Z-distribution, caching; star & snowflake schema

Áp dụng vào đồ án:

Theo Medallion Data của Databrick:
Nhóm chia ra thành 2 phần:

- Xử lí
- Lưu trữ (staging)

Mục này sẽ tập trung vào xử lí

- Bronze Data: Data lấy trực tiếp từ gg sheet, chưa được tiền xử lí hay được chuẩn hóa. Mục tiêu ở layer này là sẽ quay lại lịch sử thay đổi của data, mỗi khi có dữ liệu mới được điền vào form khảo sát, nó sẽ được trực tiếp preprocessing mà không cần phải lấy lại data từ nguồn.

- Silver Data: Mục tiêu của lớp này là khả năng được làm sạch vừa đủ để có thể quan sát tổng quan bộ dữ liệu thể hiện điều gì. Theo flow ELT (Extract - Load - Transform); khi từ silver -> gold sẽ được tiếp tục xử lí theo rule của nhóm.
  Những cái đã xử lý trong silver layer:
    - Đổi tên các cột
    - Quy đổi kiểu dữ liệu về int64
    - Thực hiện feature engineering => Tính toán thuộc tính target với 5 mức
    - Drop time_stamp, ID

        - Xử lí missing value => Xác định MNAR, MCAR => Sử dụng các cách xử lí data phù hợp

- Gold Data: Dữ liệu được dựa trên rule thang đo đánh giá của nhóm, đánh giá xếp loại của trường.

- Dựa trên thang đó đánh giá của nhóm: ở thuộc tính data, nhóm dùng thang đo gồm 5 mức ****
- Feature engineering:
    - Dựa trên thang đo đánh giá xếp loại của trường, nhóm sẽ tạo 1 cột xếp loại mới từ GPA và điểm rèn luyện, dùng đó để so sánh với cột nhóm mà học sinh tự điền để loại bỏ những dòng vô lí.
    - Thực hiện Mutual Information để giữ lại các thuộc tính có ảnh hưởng đến target **Lí do sử dụng vì bộ dữ liệu có các mối quan hệ tuyến tính và phi tuyến tính; và Mutual information là thuộc dạng filter-based, không phụ thuộc vào mô hình nào hết=>có thể dùng để lọc thuộc tính và thử trên nhiều model**
    - Thực hiện Sử dụng CountEncoding để encode các thuộc tính rời rạc không có thứ tự. **Lí do sử dụng count encoding vì sự đơn giản và không bị trường hợp bị leak dữ liệu target như target encoding**
    - Thực hiện kĩ thuật xử lí đa cộng tuyến (multicollinear) dùng VIF , Spearman, và kĩ thuật KBest(Mutual Information)

Pipeline: Airflow
Lambda Architecture, Star/Snowflake Schema
Deploy: ClickHouse