CHƯƠNG 1: MÔ TẢ ĐỀ TÀI

Tên đề tài: Xây dựng kiến trúc dữ liệu phục vụ bài toán phân tích nghiệp vụ và hoạch định chiến lược trong ngành vận tải đô thị.

1.1. Tổng quan

1.1.1. Lý do chọn đề tài:

Trong kỷ nguyên số, dữ liệu được xem như "vàng". Dữ liệu sản phẩm phụ của hoạt động kinh doanh chỉ đơn thuần là lưu trữ, mà đã trở thành sản phẩm đầu tư chiến lược, từ dữ liệu đưa ra quyết định (data-driven mindset), tăng lợi thế cạnh tranh của mọi doanh nghiệp. Tuy nhiên, dữ liệu ngày nay phát sinh dưới nhiều hình thức và từ nhiều nguồn với tốc độ chóng mặt.

- **Dữ liệu Batch (theo lô):** Là các dữ liệu được thu thập định kỳ, có kích thước lớn như báo cáo kinh doanh hàng ngày, log hệ thống, dữ liệu người dùng được kết xuất hàng tháng.
- **Dữ liệu Streaming (theo luồng):** Là các dữ liệu phát sinh liên tục theo thời gian thực như lượt click chuột trên website, giao dịch thẻ tín dụng, dữ liệu từ cảm biến IoT, tương tác trên mạng xã hội.

Các doanh nghiệp thành công là những doanh nghiệp có khả năng khai thác đồng thời cả hai loại dữ liệu này. Họ cần phân tích dữ liệu quá khứ (batch) để xây dựng chiến lược dài hạn, đồng thời phải phản ứng tức thì với các sự kiện đang diễn ra (streaming) để tối ưu vận hành, cá nhân hóa trải nghiệm người dùng, hay phát hiện gian lận.

1.2. Hiện trạng, thách thức

1.2.1. Nhu cầu, thách thức của tài xế

Một trong những vấn đề lớn nhất của tài xế taxi là "quãng đường di chuyển không có khách" (deadhead mileage). Sau khi hoàn thành một chuyến đi, tài xế phải tự quyết định nên chờ ở vị trí hiện tại hay di chuyển đến một khu vực khác để đón được khách. Quyết định của tài xế hoàn toàn dựa trên kinh nghiệm, thói quen hoặc dựa trên cảm tính. Những quyết định này thiếu cơ sở thông tin chắc chắn, dẫn đến lãng phí thời gian nhiên liệu.

1.2.2. Nhu cầu, thách thức của nhà quản lý

Khả năng quản lý hiện tại chủ yếu mang tính phản ứng thay vì chủ động. Việc thu thập và phân tích dữ liệu có độ trễ khiến Ủy ban Taxi và Limousine (TLC) không thể nắm bắt và can thiệp kịp thời trước các biến động về nhu cầu (ví dụ: các sự kiện lớn, thay đổi thời tiết đột ngột)

Từ đó, dẫn đến các hoạt động điều tiết chưa thực sự hiệu quả: phân bố dịch vụ không đều tình trạng xuất hiện nhiều "vùng trũng dịch vụ" ở khu vực ngoại vi, nơi người dân khó trong việc tiếp cận taxi, trong khi khu vực trung tâm lại quá tải; việc điều chỉnh giá cước hay ban hành các quy định mới, thiếu các dữ liệu mô phỏng để dự báo tác động có thể dẫn đến nhiều sai sót trong quyết định.

1.2.3. Nhu cầu của khách hàng

Chất lượng dịch vụ taxi truyền thống chưa hoàn toàn đáp ứng được những kì vọng về sự nhanh chóng, tiện lợi và đáng tin cậy. Vấn đề nằm ở tính có sẵn (available) và tính dự đoán (predictability) của dịch vụ. Hành khách thường không thể biết chắc chắn khi nào và ở đâu họ có thể gọi được xe.

Trải nghiệm không chắc chắn làm giảm sự hài lòng của khách hàng và làm giảm năng lực cạnh tranh của taxi truyền thống so với các nền tảng gọi xe công nghệ, vốn cung cấp cho người dùng thông tin minh bạch về thời gian chờ và vị trí của xe.

Thực trạng trên cho thấy một hệ thống vận hành mà các bên liên quan hoạt động với thông tin rời rạc và không đầy đủ đem lại hiệu quả kinh doanh không cao và việc quản lí điều phối cũng trở nên khó khăn. Do đó, việc xây dựng một nền tảng dữ liệu tập trung, có khả năng phân tích và cung cấp thông tin chi tiết theo thời gian là một giải pháp chiến lược nhằm giải quyết các thách thức, tối ưu hóa nguồn lực, nâng cao thu nhập cho người lao động và cải thiện chất lượng dịch vụ cho cộng đồng.

1.3. Mục tiêu

1.3.1. Mục tiêu kỹ thuật

1.3.1.1. Xây dựng một Đường ống dữ liệu tự động hóa:

- **Lập lịch Tự động:** Cấu hình Dagster sử dụng decorator @schedule để pipeline tự động chạy vào lúc 6:00 sáng hàng ngày.

- **Xử lý Gia tăng:** Áp dụng Partitions trong Dagster. Thay vì tải lại toàn bộ dữ liệu, pipeline sẽ chỉ xử lý dữ liệu mới của ngày hôm trước (D-1).
- **Giám sát và Cảnh báo:** Tích hợp Dagster với Slack hoặc Email: Cấu hình hệ thống để tự động thử lại (retry) tác vụ 3 lần nếu gặp lỗi tạm thời. Nếu sau 3 lần vẫn thất bại, một cảnh báo chi tiết kèm log lỗi sẽ được gửi đi.

1.3.1.2. Đảm bảo Chất lượng Dữ liệu

- **Kiểm thử Cấu trúc (Schema Tests):** Triển khai các bài kiểm thử cơ bản của dbt trên các bảng dữ liệu chính:
    - Unique và not_null: Đảm bảo khóa chính (ví dụ: trip_id) là duy nhất và không bị rỗng.
    - Relationships: Kiểm tra tính toàn vẹn tham chiếu giữa bảng Fact và các bảng Dimension (ví dụ: đảm bảo mọi LocationID đều tồn tại trong bảng dim_zones).
    - Accepted_values: Đảm bảo các cột danh mục (ví dụ: payment_type) chỉ chứa các giá trị hợp lệ (1, 2, 3, 4).
- **Kiểm thử Logic Nghiệp vụ (Custom Business Logic Tests):** Viết các bài kiểm thử SQL tùy chỉnh để xác thực các quy tắc nghiệp vụ phức tạp. Ví dụ: Kiểm tra để đảm bảo Total_Amount luôn bằng Fare_Amount + Tip_Amount + Taxes + Tolls.
- **Kiểm tra của độ mới của dữ liệu nguồn:** Cấu hình *dbt source freshness* để theo dõi thời gian cập nhật của dữ liệu thô. Nếu dữ liệu nguồn không được làm mới trong vòng 24 giờ, hệ thống sẽ cảnh báo lỗi.

1.3.1.3. Tối ưu Hiệu năng và Khả năng Mở rộng (Sử dụng BigQuery & Parquet):

- **Tối ưu hóa Định dạng File:** Lưu trữ toàn bộ dữ liệu trong Data Lake dưới định dạng **Apache Parquet**, sử dụng thuật toán nén (ví dụ: Snappy). Định dạng cột này giúp giảm I/O và tăng tốc độ đọc dữ liệu cho Spark và BigQuery.
- **Chiến lược Biến đổi Dữ liệu (dbt Materialization):** Sử dụng chiến lược *incremental* cho các bảng Fact lớn. dbt sẽ chỉ tính toán và chèn thêm các bản ghi mới, thay vì xây dựng lại toàn bộ bảng trong mỗi lần chạy.
- **Tối ưu hóa BigQuery:**

+ **Partitioning:** Phân vùng (partition) bảng Fact chính theo cột thời gian (ví dụ: pickup_datetime theo DAY). Điều này giúp các truy vấn lọc theo ngày chỉ quét một phần nhỏ dữ liệu.
+ **Clustering:** Phân cụm (cluster) bảng đã được phân vùng theo các cột thường được dùng để lọc hoặc nhóm (ví dụ: PULocationID, VendorID).

1.3.2. Mục tiêu nghiệp vụ

- **Tối ưu hóa Vận hành qua Phân tích Nhu cầu Thị trường:** Xây dựng các báo cáo tương tác để phân tích và xác định các khu vực ("điểm nóng") và khung thời gian ("giờ cao điểm") có nhu cầu di chuyển cao nhất.
- **Xây dựng Hệ thống Đo lường Hiệu suất Kinh doanh (KPIs):** Thiết kế một bảng điều khiển (dashboard) quản trị, theo dõi các chỉ số hiệu suất kinh doanh cốt lõi như tổng doanh thu, số lượng chuyến đi, và doanh thu trung bình mỗi chuyến.
- **Phân tích Chuyên sâu về Hành vi Khách hàng để Tìm kiếm Cơ hội Tăng trưởng:** Thực hiện phân tích định lượng để khám phá các mô hình và mối quan hệ ẩn trong dữ liệu, ví dụ như sự ảnh hưởng của phương thức thanh toán đến doanh thu từ tiền boa.

1.4. Kết quả dự kiến

1.4.1. Về mặt Kỹ thuật

- **Một hệ thống Data Pipeline hoàn chỉnh và tự động:** một quy trình có khả năng tự động thu thập, xử lý và nạp dữ liệu từ các nguồn thô vào kho dữ liệu. Toàn bộ mã nguồn của hệ thống và các kịch bản hạ tầng dưới dạng mã (Infrastructure as Code) sẽ là một phần của sản phẩm bàn giao, đảm bảo khả năng tái tạo và triển khai hệ thống một cách nhất quán.
- **Một kho dữ liệu được mô hình hóa và đảm bảo chất lượng:** Dữ liệu cuối cùng sẽ được lưu trữ trong một kho dữ liệu trên nền tảng đám mây. Kho dữ liệu này chứa các bảng dữ liệu đã được làm sạch, chuẩn hóa theo mô hình sao (star schema) và tối ưu hóa cho việc truy vấn. Chất lượng của dữ liệu này được đảm bảo bởi một bộ các bài kiểm thử tự động tích hợp sẵn trong đường ống dữ liệu.

1.4.2. Về mặt Phân tích và Nghiệp vụ

Báo cáo Tương tác (Interactive Dashboards): Các báo cáo này cho phép người dùng cuối tự khám phá dữ liệu một cách trực quan. Sẽ có hai loại báo cáo chính:

- Báo cáo tối ưu hóa vận hành, tập trung vào phân tích nhu cầu thị trường theo không gian và thời gian
- Báo cáo quản trị, tập trung vào theo dõi các chỉ số hiệu suất kinh doanh (KPIs) theo xu hướng dài hạn.