DiscountMate Data Engineering Pipeline Assessment.

Date: August 25, 2025

Purpose: This document provides a detailed assessment of the current state of the DiscountMate Data Engineering pipeline, explaining why no further contributions are strictly required at this stage. It is based on the project document, recent team communications, and an evaluation of the pipeline's functionality. While the pipeline is deemed "perfectly fine" for current operations, potential enhancements are noted as optional for future scalability.

Executive Summary

The DiscountMate Data Engineering pipeline is robust, functional, and meets all immediate project needs as outlined in the trimester goals. It efficiently handles data ingestion from web scrapers, storage in MongoDB, and basic transformations for downstream use by teams like Data Analysis and Machine Learning. Key strengths include automation initiation, clear documentation, and a modular design that supports current workflows without issues. As such, no immediate contributions are required from the Data Engineering team, allowing focus on other priorities or monitoring.

This assessment affirms the pipeline's stability, reducing the need for active development unless new requirements emerge.

Current Pipeline Architecture Overview

The pipeline is designed to ingest, process, and store grocery pricing data from multiple Australian supermarkets (e.g., Coles, Drakes, IGA). It integrates seamlessly with web scraping outputs and provides clean data for analysis and ML models.

Key Components:

1. **Data Ingestion**:

- Scrapers (Coles, Drakes, IGA) feed data directly into MongoDB on a weekly basis.
- Functional enhancements include timestamps, category slugs, and price per unit/special text captures.
- Automation is initiated via Amazon EC2 with $100 AWS credit, handling scheduling for basic scrapers (Drakes and IGA fully automated; Coles in progress but stable).

2. **Storage and Schema**:
    - MongoDB serves as the centralized database with timestamped collections for traceability.
    - Schema is well-structured and standardized across retailers, supporting multiple datasets without duplication or inconsistencies.
    - Variable fields are largely finalized, ensuring compatibility with SQL formats (e.g., for potential PostgreSQL integration).
3. **Processing and Transformation**:
    - Local pipeline uses Docker for containerization, Airflow for orchestration, and DBT for data build transformations.
    - Handles ingestion, cleaning, and preprocessing smoothly, with environment variables and configs for reliability.
    - Supports features like pagination handling, guest mode bypassing, and basic error detection for scraper stability.
4. **Documentation and Maintainability**:
    - Comprehensive technical docs cover setup, usage, maintenance, and handover notes.
    - GitHub repository (https://github.com/DataBytesOrganisation/DiscountMate_new/tree/main) includes clear orientation files for Data Engineering (DE), making onboarding straightforward.
    - Code reviews, pull requests (e.g., #149, #150, #163, #172), and restructuring ensure long-term reliability.

Integration with Other Teams:

- **Web Scraping**: Direct feed into MongoDB eliminates bottlenecks; updates like CAPTCHA handling (e.g., PyGui for hCaptcha) are integrated without DE intervention.
- **Data Analysis**: Provides clean, accessible data for trend analysis, dashboards, and reports (e.g., via Power BI or Plotly Dash).
- **Machine Learning**: Supplies preprocessed datasets for models like smart substitution and time-sensitive deal prediction, with aligned columns to avoid duplication.
- No reported issues in data sharing or accessibility, as per trimester updates.

Reasons Why No Contributions Are Required

1. **Meets Trimester Deliverables**:
    - Automation of scraping and database updates is already initiated and functional for core scrapers.
    - Cloud-based elements (e.g., EC2) are in use, providing monitoring and scheduling without full deployment needed immediately.
    - Database refinement (e.g., variable fields) is complete enough for current operations, with no urgent gaps identified.
2. **Stability and Reliability**:
    - Scrapers run weekly without manual intervention for most cases, handling structural changes on source websites.
    - No major failures reported; mechanisms for WAF/CAPTCHA mitigation (e.g., IP/user-agent rotation) are implemented.
    - The modular design (Docker, Airflow, DBT) ensures scalability without immediate rework.
3. **Efficiency for Current Scope**:
    - The pipeline supports real-world data from three functional scrapers, covering key supermarkets.

- Ethical boundaries are respected (e.g., halted Woolworths scraping), avoiding legal risks.
- Downstream teams (e.g., ML and Analysis) have confirmed data quality via feedback loops, with no requests for major changes.

4. **Clear Documentation and Onboarding**:
    - Existing docs and GitHub DE files provide all necessary references, reducing the need for new contributions.
    - Handover notes ensure future teams can maintain it easily.
5. **No Urgent Risks or Gaps**:
    - While Coles CAPTCHA solving is complex, it's managed manually when needed, and research is ongoing without halting operations.
    - The pipeline is production-like for a student/project environment, with no performance issues at current scale.

In summary, the pipeline is solid, self-sustaining, and aligned with trimester goals.