Data Engineer Case Study

Exercise: Debugging a Faulty Twitter API v1 ETL Pipeline

Objective
Your task is to debug and fix the provided Python ETL pipeline that interacts with the Twitter API v2. The pipeline fetches recent tweets based on a keyword, processes the data, and writes the results to a JSON file. The code is intentionally flawed to test your problem-solving and debugging skills.

Expected Outcome
After fixing the pipeline, it should:

- Fetch tweets using the Twitter API v2 with proper authentication.
- Handle rate-limiting and pagination correctly.
- Validate API responses and process only valid data.
- Write the results to a JSON file without corruption, even in multi-threaded environments.
- Log meaningful progress and error information.

Exercise: Build a Data Platform Design

Objective

Your task is to design a data platform that extracts data from the sevdesk API, ingests it into a data warehouse, and prepares it for downstream transformation and analytics using dbt. While dbt transformations and the integration with Looker are out of scope, the platform design must account for scalability, monitoring, alerting, and fault tolerance.

Instructions

Context

sevdesk provides an API for accessing accounting and invoicing data. Your data platform must:

1. Pull data periodically from the sevdesk API.
2. Load the data into a modern data warehouse (e.g., Snowflake, BigQuery, or Amazon Redshift).
3. Ensure that the ingested data is available for transformations with dbt.

Deliverables

1. **System Architecture Diagram**
2. **Technical Design Document**:
   - Describe the tools and technologies you would use
   - Detail the steps in the pipeline
   - Provide examples
3. **Possible Enhancements**