Project: Scalable Data Pipeline for U.S. University Information Aggregation

Background

One of the biggest challenges for university aspirants in the United States is finding accurate, up-to-date, and relevant information about colleges and universities. With thousands of schools nationwide, students often struggle to compare institutions based on critical factors such as tuition costs, graduation rates, admission requirements, and available programs.

To address this problem, we aim to build a **scalable data pipeline** that automates the extraction, transformation, and storage of university-related information from the **USA College Scorecard API**. The system will retrieve data for the **top 1000 schools in the U.S.** based on ranking and store the data in a structured format suitable for analytics.

Project Scope and Objectives

1. **Data Extraction:**

   - Retrieve school-related data (name, location, type, ranking, etc.).
   - Extract admission-related details (acceptance rate, SAT/ACT scores, etc.).
   - Gather cost-related information (tuition fees, financial aid, etc.).
   - Capture performance metrics (graduation rates, retention rates, etc.).

2. **Data Processing & Transformation:**

   - Clean and standardize data for consistency.
   - Handle missing or inconsistent data gracefully.
   - Normalize and structure the dataset for efficient querying and analysis.

3. **Data Storage & Management:**

   - Must be based on cloud (Use your discretion to choose the cloud platform)

4. **Scalability & Automation:**

   - Deploy an **ETL (Extract, Transform, Load) pipeline** to run on a schedule.
   - Ensure the system can handle increasing API request loads efficiently.

5. **Data Accessibility & Visualization:**

    - Develop a well-documented warehouse schema which showcases your data modelling prowess. Ensure it is comprehensive for a data analyst who intends to use the warehouse.
    - Develop **dashboards** (e.g., using Streamlit, Metabase, Tableau, Power BI, etc.) with visuals containing key metrics about schools and the admissibility criteria. Feel free to explore your creativity here.

6. **Error Handling & Monitoring:**
    - Implement logging and error handling for failed API requests or transformation issues.

Optional Score Booster :

- CI/CD Integration
- Implement Infrastructure as Code where required

Expected Deliverables

1. A **fully functional data pipeline** that can fetch, process, and store university data.
2. An analytics **dashboard**.

This project will provide valuable insights for students, analysts, and policymakers, helping them make data-driven decisions about higher education in the U.S. 🚀

GRADING METRICS

- Code best practices
- Choice of tools
- Document the thought process, the reason for architectural design and the choice of tools.
- How easy is it to run the code submitted

MEANS OF SUBMISSION

Share the GitHub repository containing the solution to the Hackathon case study. The submission link will be shared before the Hackathon ends.