Bradley Haskell

NVIDIA Data Engineering Insights

https://github.com/bradhaskell/sql-project

Project Proposal

Job Description

Data Engineering at NVIDIA focuses on building complex data platforms for advanced data analysis to enhance GPU server manufacturing. The role involves designing and maintaining ELT pipelines, ensuring data integrity, optimizing query performance and security through SQL.

This role directly aligns with my career goals in ISBA, providing an opportunity to work with large-scale datasets, optimize data processing systems, and apply SQL in real-world scenarios. Along with my strong interest in the tech industry, NVIDIA's ability to maintain technological developments, and sustain satisfied employees furthers my interest in this company and position. By the end of this project, I will have gained realistic experience in constructing ETL pipelines and querying company data in an industry and corporation I have a strong draw to.

Problem

A challenge in NVIDIA's GPU server manufacturing is maintaining efficient and reliable data pipelines. Ensuring data is safe, optimizing data cleaning, and drawing real-time insights are essential for data engineers to improve the manufacturing process. This project will simulate these real-world data challenges by constructing an ELT pipeline, leveraging SQL for data transformation, and implementing data visualization techniques to provide actionable insights.

Data Sources

To costruct a relevant dataset, I will utilize an API and web scraping. The API will help build structured industry-related data, such as server manufacturing efficiency metrics or hardware performance statistics. The web scraping component will extract insights from industry websites such as capturing trends in data pipeline efficiency or production optimizations. The API data will be accessed using Python libraries such as requests, while web scraping will be performed using BeautifulSoup. These data sources and collection techniques align with NVIDIA's real-world data engineering tasks.

Solution

To address this problem, I will develop an ELT pipeline that collects, processes, and transforms the data into a structured format. SQL queries will be used for data cleaning, aggregation, and performance optimization. The final dataset will be visualized using tools such as Excel or PopSQL to generate insights on the data collected through the API and web scraper. This solution will mimic the tasks and insights that would be taken and concluded by a data engineer at NVIDIA.