Data Engineer Intern at AIDF-CRI

Background

The Asian Institute of Digital Finance (AIDF) is a university-level institute in NUS, jointly founded by the Monetary Authority of Singapore (MAS), the National Research Foundation (NRF) and NUS. AIDF aspires to be a thought leader, a Fintech knowledge hub, and an experimental site for developing digital financial technologies as well as for nurturing current and future Fintech researchers and practitioners in Asia. The Credit Research Initiative (CRI) is a non-profit undertaking under the AIDF. Pioneering the "public good" credit risk measures, the initiative is committed to advancing big data analytics and providing directly useful credit intelligence to academic and professional communities.

AIDF-CRI is dedicated to staying updated with the latest trends and technologies, especially for AI and LLMs. Notably, efficient and seamless data collection and preparation are often overlooked but are critical components for supporting high-performing models. We are continually developing and enhancing our data pipelines by integrating new technologies to meet both project and market demands.

Responsibilities

AIDF-CRI is actively seeking candidates with a strong interest in data management and pipeline optimization. The selected candidate will play a key role in enhancing the performance of our advanced models and applications by integrating the latest technologies to build data pipelines that are not only robust but also tailored to meet evolving needs.

Particularly, the responsibilities will include:

- Data Pipeline Development & Maintenance
    - Support the development of ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform) data pipelines using Snowflake and/or Google Cloud Platform services.
    - Develop and optimize data models and schemas to support analytics, reporting, and machine learning requirements.
    - Automate workflows using tools like Cloud Composer or Cloud Functions to ensure reliable data pipelines.
- Advanced NLP Integration
    - Develop and maintain scalable data pipelines to automatically process alternative data sources such as news articles, financial filings, and social media content.
    - Apply advanced NLP techniques, such as tokenization and named entity recognition, to extract meaningful insights from unstructured data.
    - Use cutting-edge NLP models and frameworks to analyze large datasets and produce high-quality outputs.
- Team Collaboration
    - Collaborate with financial analysts and R&D team, to ensure data accessibility and usability.
- Documentation

- Maintain comprehensive documentation to promote knowledge sharing.

Minimum Requirements

- Bachelor's or Master's degree in Computer Science, Information Technology, Data Science, Computer Engineering, or a related field.
- Proficiency in at least one programming language, such as Python.
- Familiarity with Google Cloud Platform and/or Snowflake.
- Experience in database management, including working with NoSQL databases and vector databases (e.g., MongoDB, Milvus) to store, query, and manage data.
- Strong understanding of NLP techniques, including tokenization, named entity recognition, and text classification.
- Strong problem-solving skills and the ability to work independently in a fast-paced environment.
- Excellent communication and documentation skills, both written and verbal, to collaborate effectively with cross-functional teams and stakeholders.

Bonus Skills

- Strong interest in integrating LLMs into data pipelines and retrieval-based systems.
- Familiarity with web scraping tools and frameworks (e.g., Scrapy, Selenium).
- Experience working with large datasets, especially alternative data sources like news and 10-K filings.
- Familiarity with machine learning techniques in financial context.