FINM 32500 Computing for Finance in Python

Project Part 3

Problem Statements Report

Name
Andy Andikko
Harrison Zhang
Matheus Raka Pradnyatama

1. Limited Access to Raw and Processed Data in a Unified Platform

Problem: Many platforms silo raw and processed data, creating fragmented workflows. Researchers need raw data for custom processing and processed data for quicker analysis, but a lack of integration leads to inefficiencies.

Solution: Our **Data Lake** supports unified storage and retrieval of raw and processed data, ensuring seamless transitions between exploration and analysis. Using distinct directories for raw and processed datasets (raw, processed), the platform provides flexibility for custom analyses and faster workflows with preprocessed data.

2. Difficulty in Aligning Data Access with Security and Compliance Needs

Problem: Quant research often involves sensitive data, requiring careful access rights management to ensure security and compliance.

Solution: Our platform implements role-based access controls in the **Data Lake**. Access rights are verified at every operation, ensuring only authorized users can store, retrieve, or edit datasets.

secured_access = {134: "Andy", 245: "Matt", 367: "Harry"}

3. Complexity in Locating Relevant Datasets

Problem: Researchers often struggle to find the needed datasets due to poor organization and large data volumes, leading to time inefficiencies.

Solution: The **Data Catalog** organizes datasets into logical categories (e.g., Equities, Economic Data) and provides a keyword-based search functionality (search_datasets). Categorization methods like add_category and add_dataset_to_category streamline dataset discovery and improve workflow efficiency.

4. Difficulty in Preparing and Transforming Data for Analysis

Problem: Preparing data for analysis often involves extensive cleaning, aggregation, and filtering, which can be time-intensive and inconsistent.

Solution: The **Data Workbench** simplifies data preparation with the transform_data function, enabling users to apply chained transformations (e.g., VWAP, volatility calculation) efficiently. These transformations standardize workflows and reduce manual effort, making data analysis-ready.

4. Lack of Metadata and Contextual Information

Problem: Insufficient metadata makes it difficult for researchers to understand dataset attributes, sources, or collection methods, leading to errors or misinterpretation. This also hinders the speed of data exploration and the search process of relevant datasets.

Solution:

- Our **Data Lake** stores metadata for datasets, including *data_type, modification_time*, and author details.
- Our **Data Catalog** adds contextual metadata when datasets are categorized, and metadata is accessible during searches.