Project 2. Data warehouse implementation, ETL pipelines

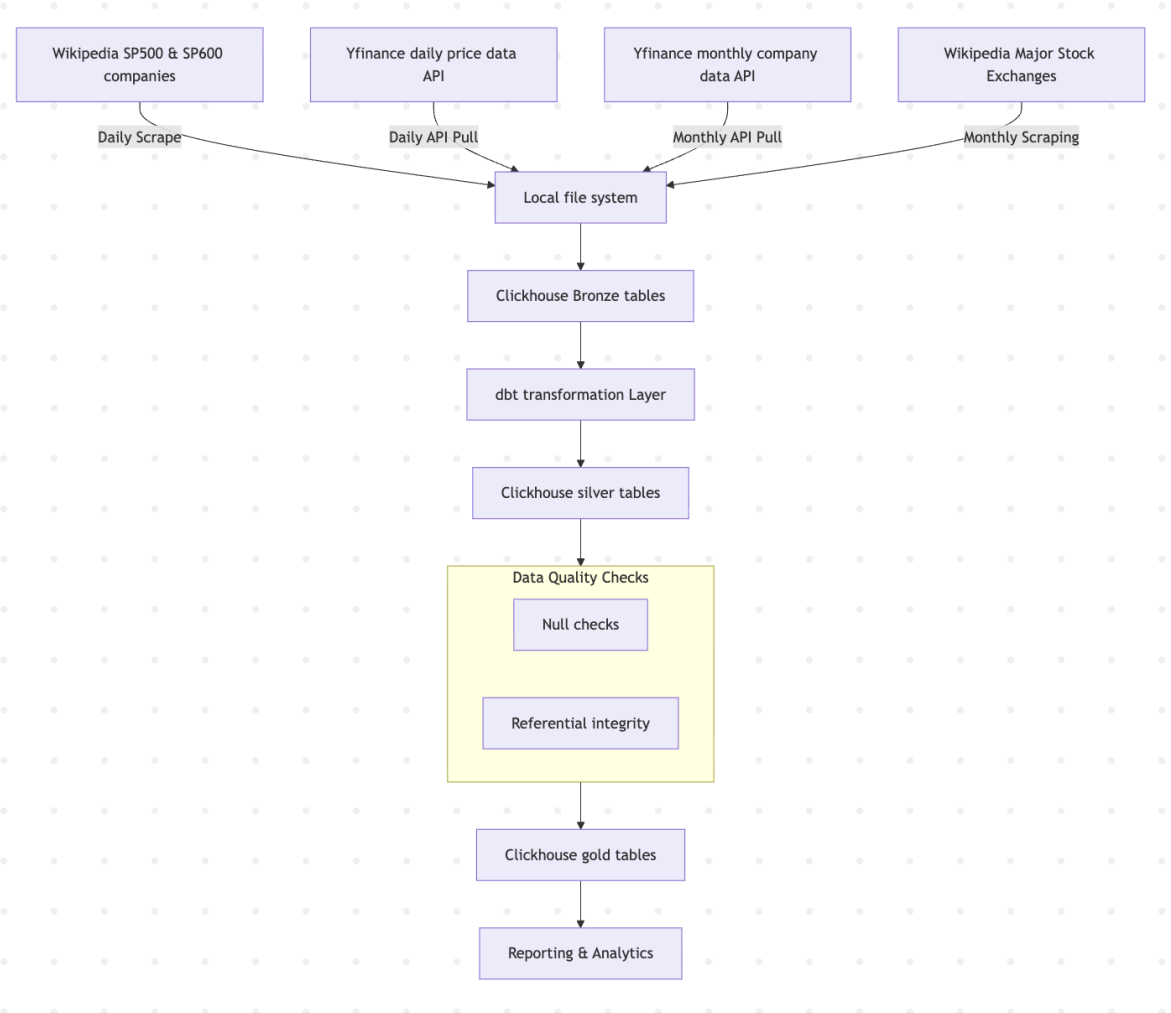
Group 19: Philip Spaier, Danni Zhang, Philipp Weissensteiner, Merlin Kägu

Changes compared to Project 1

We expanded our range from the S&P 500 to **include the S&P 600**. Since the AlphaVantage API has daily pull limits for free users, we decided instead to retrieve data using the **yfinance** library. To not be treated like a DDOS attack by yfinance, we decided to only retrieve data from the API starting with 2024 since this already results to over 500k rows.

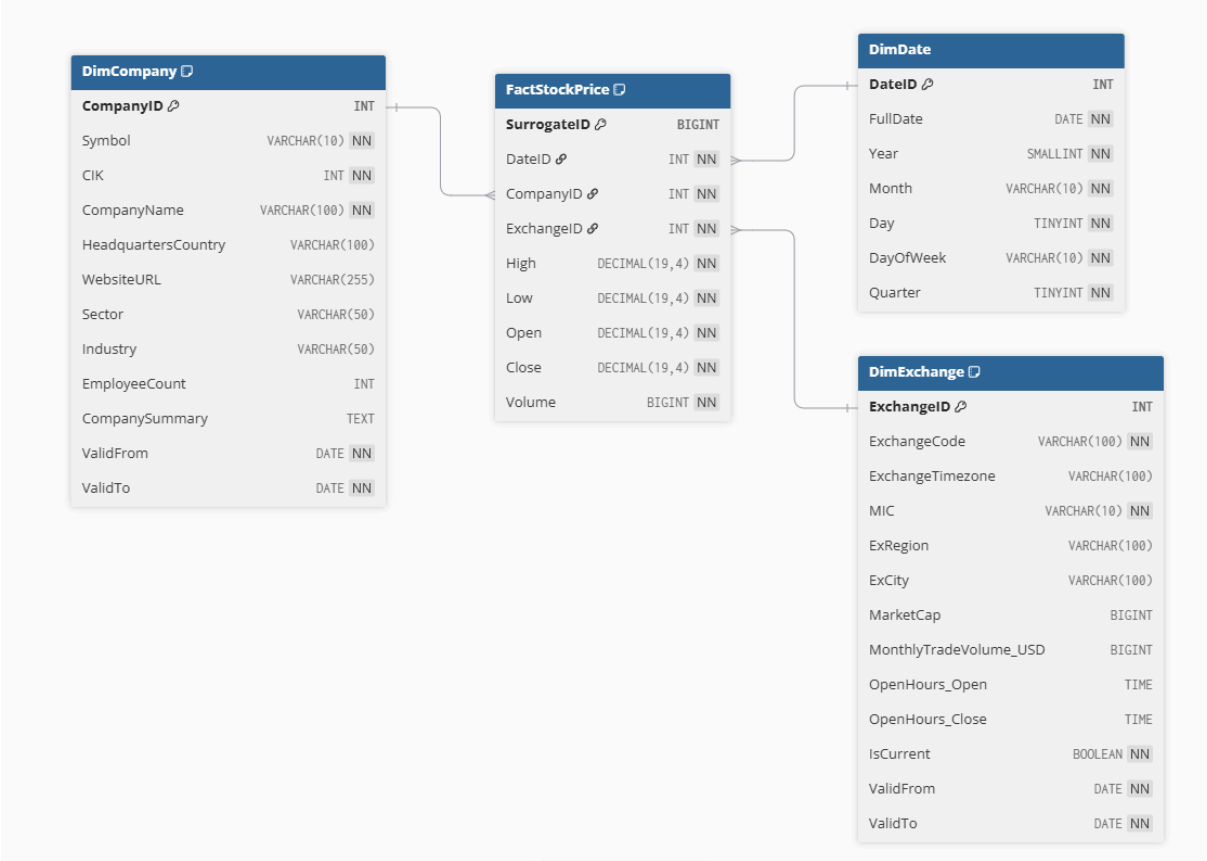
For the list of major stock exchanges, we scraped the information from **Wikipedia** into a CSV file, rather than downloading it from the official ISO website because it provides more relevant attributes.

Throughout the process, we used various tools in data architecture. For data transformation, we use **Clickhouse + dbt**. For data orchestration, we used **Apache Airflow**.



In our star schema, we removed the dim_currency table because all the stocks we're interested in are traded in USD. We modified the fact_table and three dimension_tables, and updated the data

dictionary to reflect the data extracted from the new sources. yfinance provides slightly different attributes than AlphaVantage.



AI Assistance

clickhouse ingestion: https://claude.ai/share/ef429a94-00f0-499f-beef-fad6b018044a

Environment Setup:
https://claude.ai/share/42970d37-da92-4b44-9798-d531055bde01
https://claude.ai/share/63c649d2-3116-4682-b9ae-867a74696e4d
https://aistudio.google.com/app/prompts?state=%7B%22ids%22:%5B%2218UOqiBjy9cx5LCTCX2of_4aPCM-kENCi%22%5D,%22action%22:%22open%22,%22userId%22:%22108508433267639844151%22,%22resourceKeys%22:%7B%7D%7D&usp=sharing

yfinance Assistance:
https://gemini.google.com/share/a3172fe6bb48

Orchestration and business queries help:
https://chat.deepseek.com/a/chat/s/eba9c773-94a6-431d-8072-2243c6925d84
https://chat.deepseek.com/a/chat/s/884b6949-1287-4469-93a6-d9ad47e5f242

https://chat.deepseek.com/a/chat/s/7d50661d-2d60-42dc-b224-b967188e96c0
https://chat.deepseek.com/a/chat/s/a92f63b3-925d-43a9-8144-d81ac65a113f

Transformation:
https://chatgpt.com/share/6907ca13-1a2c-8011-aefb-a7034330cfec

Github Repository

Contribution

Philip Spaier (30): Adjusted schema and selected new data sources, wrote all extraction scripts, worked on setting up the docker environment, assisted with the ingestion and orchestration
Danni Zhang (25): Ingested data into Clickhouse: created clickhouse part of compose file, designed tables' structure for loading bronze layer, then ingested by ingest_monthly/daily_data.py via Airflow. Refined business brief, data architecture, star schema, data dictionary from project 1.
Philipp Weissensteiner (30): Data transformation in clickhouse using dbt, debugging the project & fixing errors, overall architecture, data orchestration, cleanup of code repository and business queries..
Merlin Kägu (15%): Created Airflow dags, readme document for our Github and took care of the business queries. I was in charge of the data orchestration part and received a lot of help from Philip S. and Philipp W.