Анализ рынка акций

Сбор данных по курсам акций и подготовка обобщающей витрины

Цель проекта

создать ETL-процесс формирования витрин данных для анализа изменений курса акций

Решаемые задачи:

- разработать скрипты полной и инкрементальной загрузки;
- организовать хранение данных;
- сформировать витрину данных;
- настроить оркестрацию (обновление данных и витрины).

План реализации

- ознакомиться с API Alpha Vantage
- получить данные
- разработать SQL скрипт, реализующий витрину
- подготовить скрипты для полной и инкрементной загрузок
- настроить оркестрацию в Airflow

Используемые технологии

1. Загрузка – Python и Pandas т.к. позволяют выполнить преобразование данных и контроль качества.
2. Хранение – PostgreSQL как стандарт де-факто open source СУБД, csv-файлы как бэкап исходных данных.
3. Формирование витрины – SQL в Postres (materialized view) т.к. по предварительным оценкам объём данных достаточно мал, также не требуется частого обновления (витрина сводится на сутки).
4. Оркестратор – Airflow – взят “на вырост” если в дальнейшем понадобится подключать HDFS и Spark.

Архитектура

1. Docker контейнер с Airflow и Postgres
2. DAG из Airflow получает данные по Alpha Vantage API
3. Резервная копия сохраняется в ФС (csv-файл)
4. Данные загружаются в Postgres
5. Агрегация и подготовка витрины выполняется на SQL (скрипты формируются в Airflow)

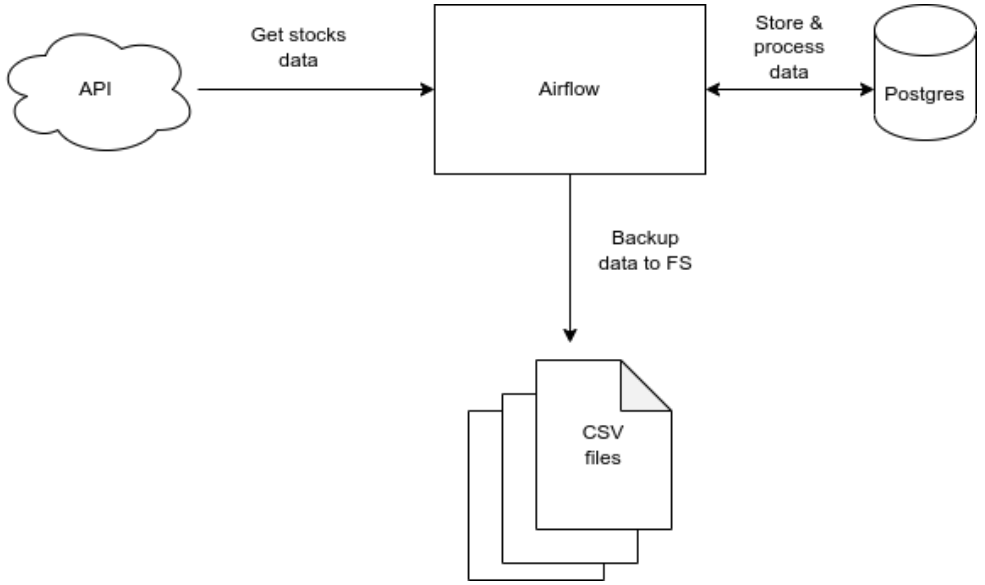
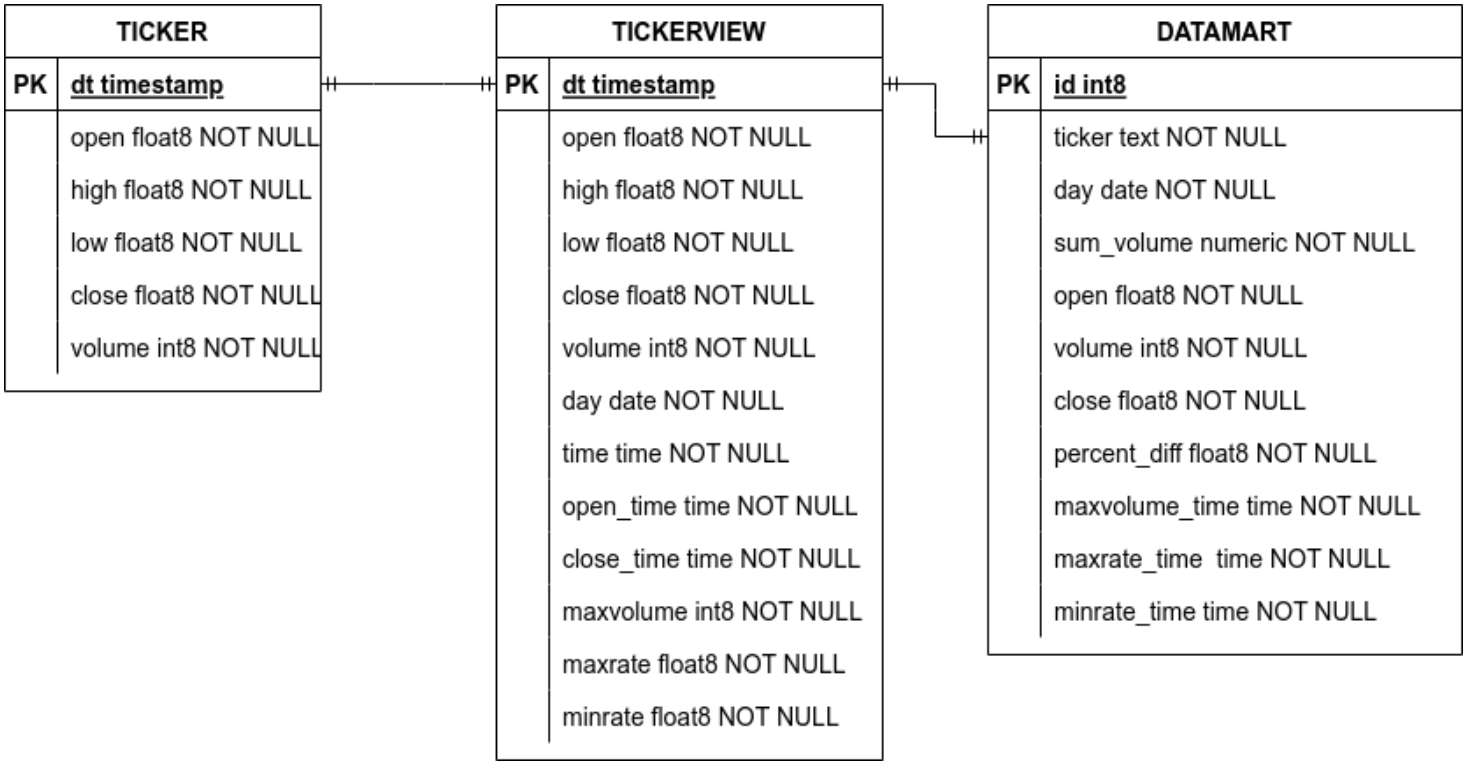


Схема данных



Результаты разработки

- Реализован ETL-процесс загрузки данных в БД по API и построения витрины с периодическим её обновлением с использованием Python, Airflow и Postgres.
- Получены практические навыки работы с Airflow, docker-compose и SQL.
- Готовый проект загружен на Github.

Выводы

Для решения задачи в том виде, в котором она была поставлена было бы достаточно использовать bash, утилиты postgres CLI, cron. Однако был выбран Airflow, как задел на будущее, чтобы перейти в дальнейшем от SQL к Spark, если перестанет хватать производительности Postgres.

Что хотелось бы добавить:

- контроль качества данных и обработка ошибок;
- бэкап в HDFS;
- веб-интерфейс для выбора тикеров (акций).