Итоговая аттестация

Проект №2

Анализ рынка акций

Курс 1Т: Data Engineer
Тарасов Сергей Евгеньевич

Задание

Общая задача: создать ETL-процесс формирования витрин данных для анализа изменений курса валют.

- Разработать скрипты загрузки данных в 2-х режимах:
  - Инициализирующий
  - Инкрементальный
- Организовать правильную структуру хранения данных:
  - Сырой слой данных
  - Промежуточный слой
  - Слой витрин
- Источник данных: API https://www.alphavantage.co

План реализации проекта

1. Анализ источника данных
2. Проектирование ER-диаграммы
3. Определение стека технологий
4. Реализация инициирующей загрузки данных
5. Реализация инкрементальной загрузки данных

Анализ источника данных

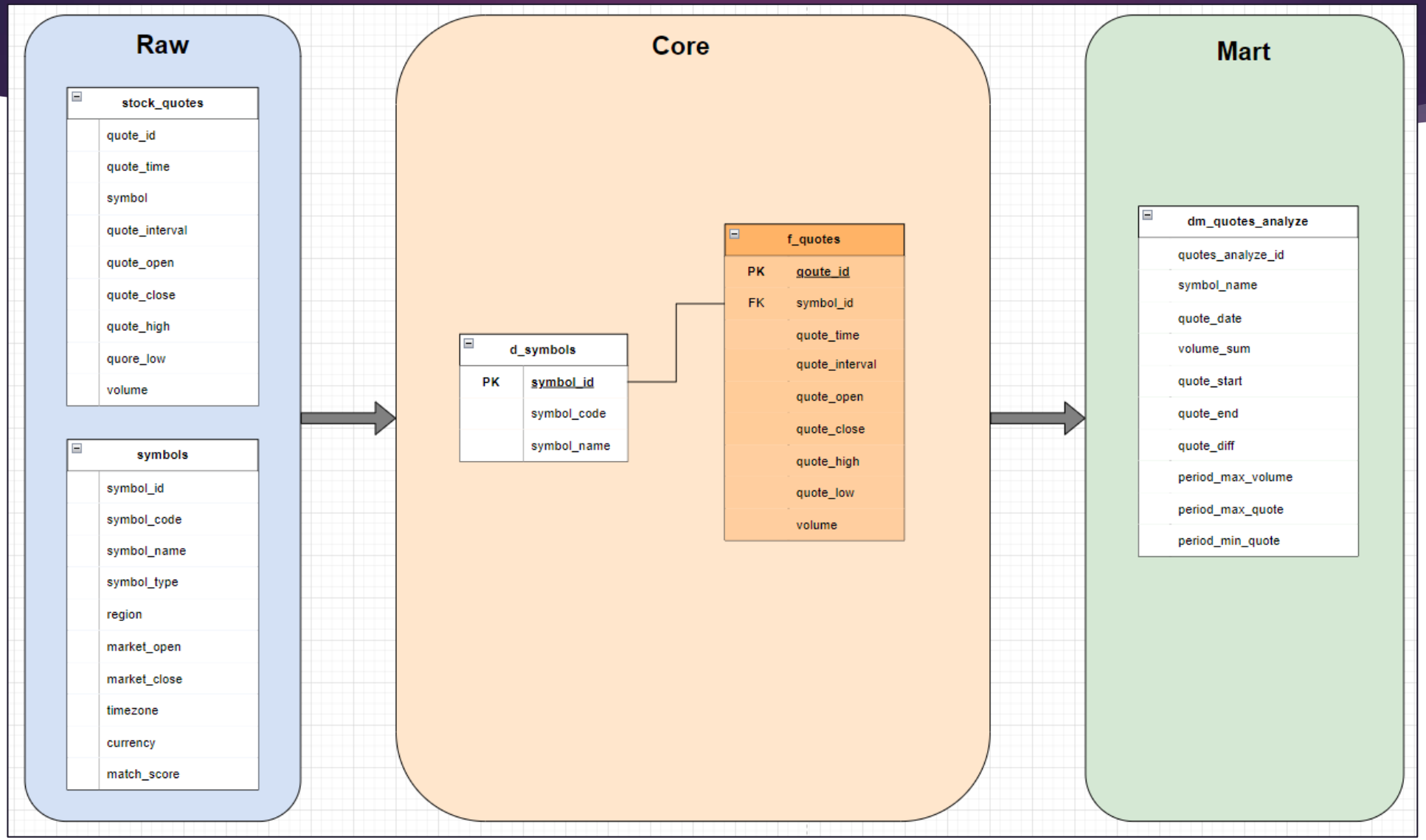
В качестве источника данных используется API сервиса https://www.alphavantage.co.

В процессе изучения API Alpha vantage, были определены параметры, которые позволяют получить необходимую для построения витрины информацию.

Параметры:

- function: TIME_SERIES_INTRADAY
- symbol: аббревиатура акции (напр. IBM)
- Interval: 5min
- month: месяц, за который необходимо получить данные (напр. 2023-12)
- apikey: индивидуальный ключ для использования API

ER-диаграмма



Стек технологий

Python: Используется для написание скриптов получения данных по API, трансформации данных а также записи данных в БД.

Postgres: Используется для хранения данных. Выбрал postgres, т.к. предполагается не значительный объем базы данных. А также из-за простоты и удобства работы для разработчика.

Docker: Используется для быстрой и удобной сборки и развертывания проекта на различных машинах.

Crontab: Очень хотелось использовать airflow для создания отдельных дагов для инициализации и инкрементальной загрузки данных, но к сожаления на моем ПК не удалось настроить работу даже версии 1.x. Пришлось воспользоваться crontab для автоматизации этих процессов.

Этап инициализации данных

- Создаются три слоя данных, и таблицы в них.
- Таблицы наполняются предзагруженными данными из csv файлов. Эти файлы были созданы, для обхода ограничения API на 25 запросов в сутки. В них собраны данные из предыдущих периодов
- Происходит актуализация данных, догружаются данные, которых не хватает в csv-файлах.
- Происходит наполнение всех слоев данными, и расчет итоговой витрины

Этап инкрементальной загрузки данных

- Инкрементальная загрузка догружает данные за последний день
- Происходит пересчет итоговой витрины
- Запускается с помощью cron, настроена на запуск в 6:00

Пример входных данных (raw слой)

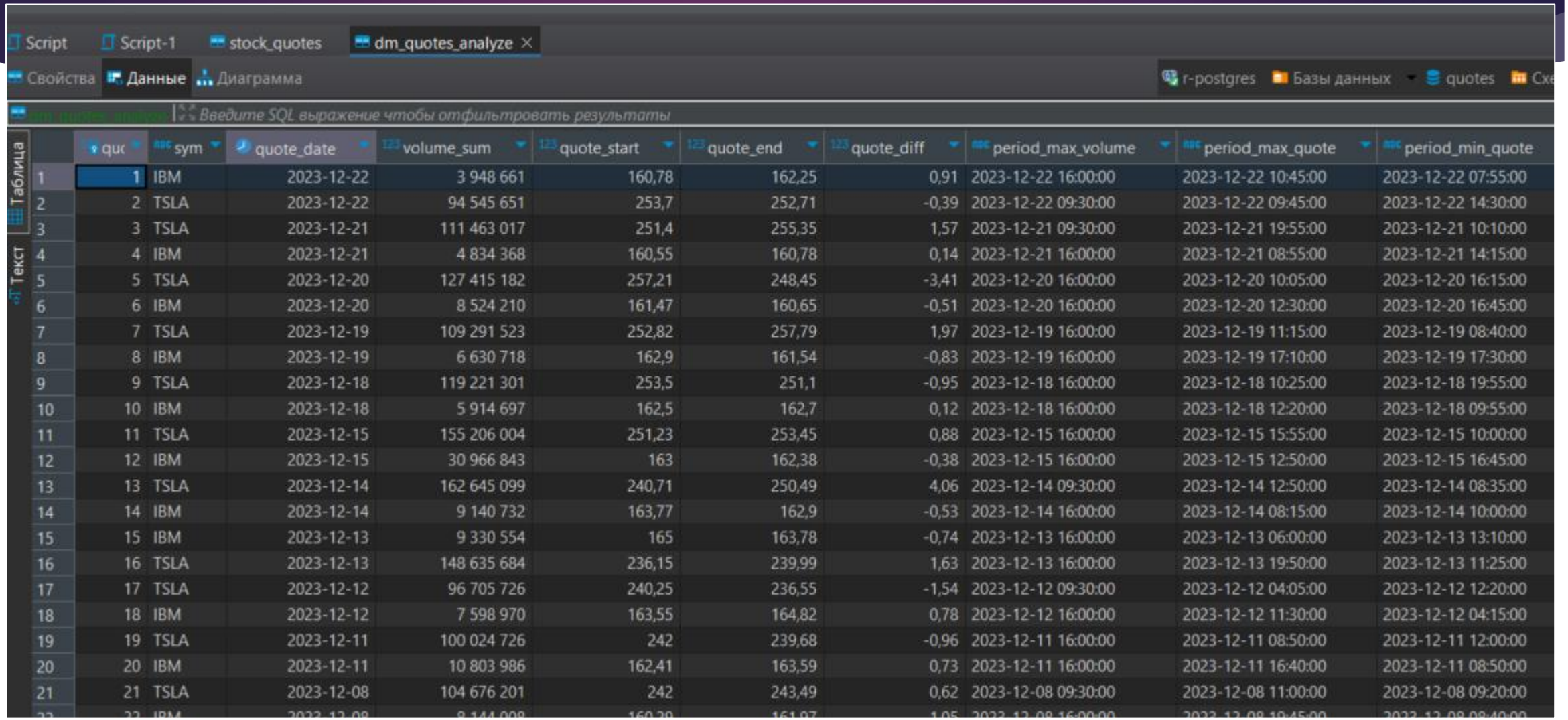
ipt | Script-1 | stock_quotes ×

ойства | Данные | Диаграмма | r-postgres | Базы дан

ck_quotes | Введите SQL выражение чтобы отфильтровать результаты

	quote_id	quote_time	symbol	quote_interval	quote_open	quote_close	quote_high	quote_low	volume
	1	2023-10-02 04:00:00.000	IBM	5min	139,174	139,88	139,16	139,876	374
	2	2023-10-02 04:05:00.000	IBM	5min	139,49	139,495	139,407	139,421	40
	3	2023-10-02 04:15:00.000	IBM	5min	139,421	139,425	139,407	139,421	176
	4	2023-10-02 04:45:00.000	IBM	5min	139,421	139,425	139,407	139,421	200
	5	2023-10-02 04:50:00.000	IBM	5min	139,431	139,435	139,417	139,431	10
	6	2023-10-02 04:55:00.000	IBM	5min	139,431	139,435	139,249	139,421	1 044
	7	2023-10-02 05:00:00.000	IBM	5min	139,322	139,327	139,308	139,323	206
	8	2023-10-02 05:05:00.000	IBM	5min	139,322	139,445	139,308	139,323	594
	9	2023-10-02 05:30:00.000	IBM	5min	139,174	139,178	138,873	138,888	255
0	10	2023-10-02 05:40:00.000	IBM	5min	139,213	139,218	139,199	139,214	16
1	11	2023-10-02 05:55:00.000	IBM	5min	139,124	139,129	139,11	139,125	16
2	12	2023-10-02 06:00:00.000	IBM	5min	138,778	138,783	138,764	138,779	3
3	13	2023-10-02 06:05:00.000	IBM	5min	139,124	139,129	139,11	139,125	3
4	14	2023-10-02 06:20:00.000	IBM	5min	139,094	139,099	139,081	139,095	7
5	15	2023-10-02 06:30:00.000	IBM	5min	138,818	138,822	138,804	138,818	1
6	16	2023-10-02 06:40:00.000	IBM	5min	138,837	138,842	138,814	138,838	20
7	17	2023-10-02 06:55:00.000	IBM	5min	138,827	138,832	138,804	138,818	2
8	18	2023-10-02 07:00:00.000	IBM	5min	138,748	138,911	138,735	138,907	3

Пример данных в итоговой витрине (mart слой)



Script | Script-1 | stock_quotes | dm_quotes_analyze

Свойства | Данные | Диаграмма | r-postgres | Базы данных | quotes

Введите SQL выражение чтобы отфильтровать результаты

	quc	sym	quote_date	volume_sum	quote_start	quote_end	quote_diff	period_max_volume	period_max_quote	period_min_quote
1	1	IBM	2023-12-22	3 948 661	160,78	162,25	0,91	2023-12-22 16:00:00	2023-12-22 10:45:00	2023-12-22 07:55:00
2	2	TSLA	2023-12-22	94 545 651	253,7	252,71	-0,39	2023-12-22 09:30:00	2023-12-22 09:45:00	2023-12-22 14:30:00
3	3	TSLA	2023-12-21	111 463 017	251,4	255,35	1,57	2023-12-21 09:30:00	2023-12-21 19:55:00	2023-12-21 10:10:00
4	4	IBM	2023-12-21	4 834 368	160,55	160,78	0,14	2023-12-21 16:00:00	2023-12-21 08:55:00	2023-12-21 14:15:00
5	5	TSLA	2023-12-20	127 415 182	257,21	248,45	-3,41	2023-12-20 16:00:00	2023-12-20 10:05:00	2023-12-20 16:15:00
6	6	IBM	2023-12-20	8 524 210	161,47	160,65	-0,51	2023-12-20 16:00:00	2023-12-20 12:30:00	2023-12-20 16:45:00
7	7	TSLA	2023-12-19	109 291 523	252,82	257,79	1,97	2023-12-19 16:00:00	2023-12-19 11:15:00	2023-12-19 08:40:00
8	8	IBM	2023-12-19	6 630 718	162,9	161,54	-0,83	2023-12-19 16:00:00	2023-12-19 17:10:00	2023-12-19 17:30:00
9	9	TSLA	2023-12-18	119 221 301	253,5	251,1	-0,95	2023-12-18 16:00:00	2023-12-18 10:25:00	2023-12-18 19:55:00
10	10	IBM	2023-12-18	5 914 697	162,5	162,7	0,12	2023-12-18 16:00:00	2023-12-18 12:20:00	2023-12-18 09:55:00
11	11	TSLA	2023-12-15	155 206 004	251,23	253,45	0,88	2023-12-15 16:00:00	2023-12-15 15:55:00	2023-12-15 10:00:00
12	12	IBM	2023-12-15	30 966 843	163	162,38	-0,38	2023-12-15 16:00:00	2023-12-15 12:50:00	2023-12-15 16:45:00
13	13	TSLA	2023-12-14	162 645 099	240,71	250,49	4,06	2023-12-14 09:30:00	2023-12-14 12:50:00	2023-12-14 08:35:00
14	14	IBM	2023-12-14	9 140 732	163,77	162,9	-0,53	2023-12-14 16:00:00	2023-12-14 08:15:00	2023-12-14 10:00:00
15	15	IBM	2023-12-13	9 330 554	165	163,78	-0,74	2023-12-13 16:00:00	2023-12-13 06:00:00	2023-12-13 13:10:00
16	16	TSLA	2023-12-13	148 635 684	236,15	239,99	1,63	2023-12-13 16:00:00	2023-12-13 19:50:00	2023-12-13 11:25:00
17	17	TSLA	2023-12-12	96 705 726	240,25	236,55	-1,54	2023-12-12 09:30:00	2023-12-12 04:05:00	2023-12-12 12:20:00
18	18	IBM	2023-12-12	7 598 970	163,55	164,82	0,78	2023-12-12 16:00:00	2023-12-12 11:30:00	2023-12-12 04:15:00
19	19	TSLA	2023-12-11	100 024 726	242	239,68	-0,96	2023-12-11 16:00:00	2023-12-11 08:50:00	2023-12-11 12:00:00
20	20	IBM	2023-12-11	10 803 986	162,41	163,59	0,73	2023-12-11 16:00:00	2023-12-11 16:40:00	2023-12-11 08:50:00
21	21	TSLA	2023-12-08	104 676 201	242	243,49	0,62	2023-12-08 09:30:00	2023-12-08 11:00:00	2023-12-08 09:20:00

Выводы

- В процессе разработки был создан ETL-процесс получения, преобразования и анализа курса акций
- Полученные результаты можно использовать для анализа курса акций и принятия решений о торговых операциях на бирже акций.
- В процессе выполнения работы закреплены навыки проектирования DWH и разработки ETL-процессов.