Путь данных

Анна Мавлютова

14.09.2024

Анна Мавлютова

TPM Data Governance

- В Т-Банке с 2017-го года
- Прошла полный путь ETL-щика
- Ушла в разработку продуктов для Data Platform

Вводная

Данные для компании

01

Обработка и получение данных

02

03

Взаимодействие систем на масштабах

04

Процесс для пользователя

Вводная

Данные для компании

01

Обработка и получение данных

02

03

Взаимодействие систем на масштабах

04

Процесс для пользователя

Вводная

Данные для компании

01

Обработка и получение данных

02

Взаимодействие систем на масштабах

03

04

Процесс для пользователя

Вводная

Данные для компании

01

Обработка и получение данных

02

Взаимодействие систем на масштабах

03

Процесс для пользователя

04

Данные

Т БАНК

Данные для компании?

Ценность данных для компании

Операционные процессы:
работа систем

Ценность данных для компании

Операционные процессы: работа систем

Мониторинг: анализ продуктов и процессов

Ценность данных для компании

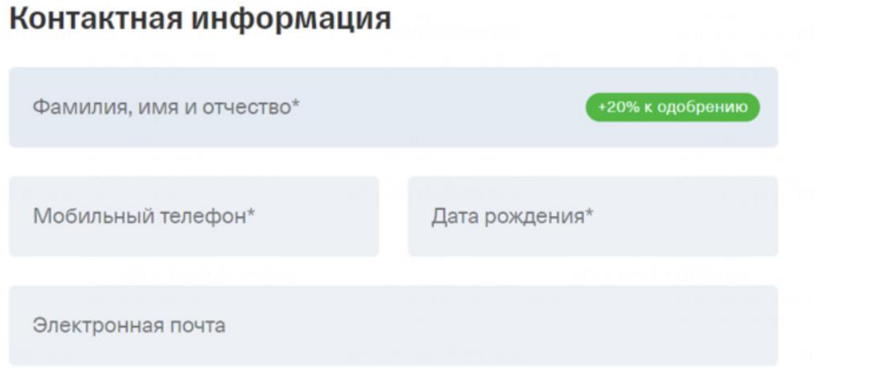
Операционные процессы: работа систем

Мониторинг: анализ продуктов и процессов

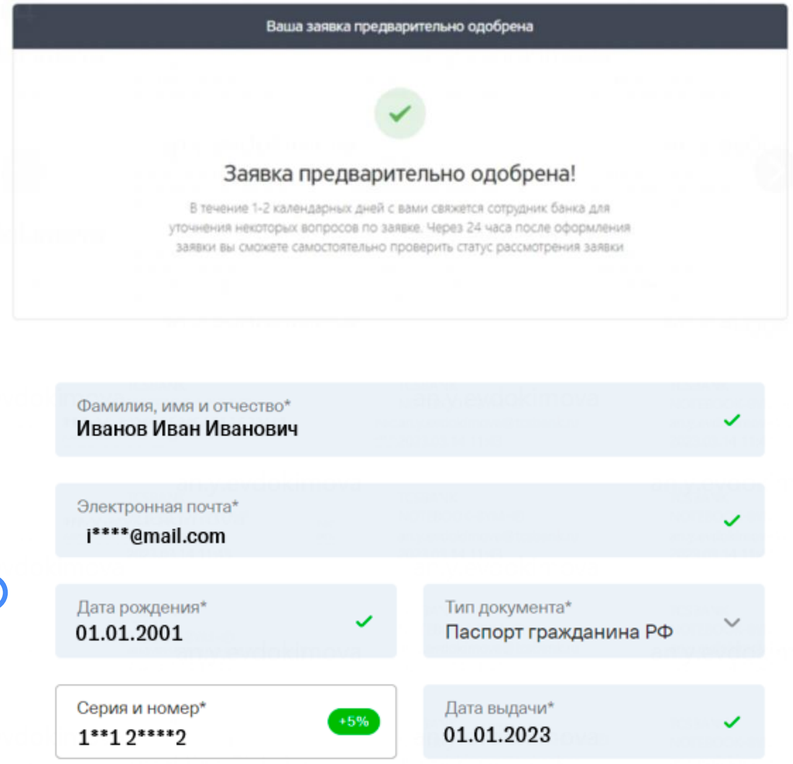
Получение ценности

Предзаполнение заявки на КК

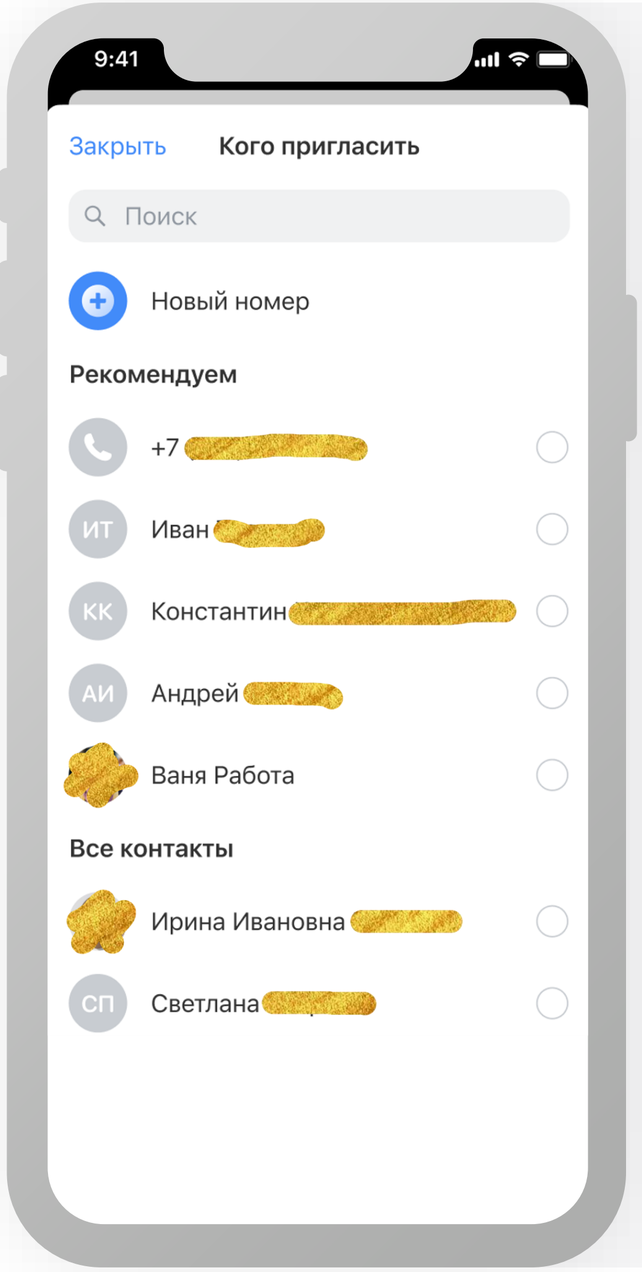
Что получаем
на вход:



Что получает клиент
на выходе:



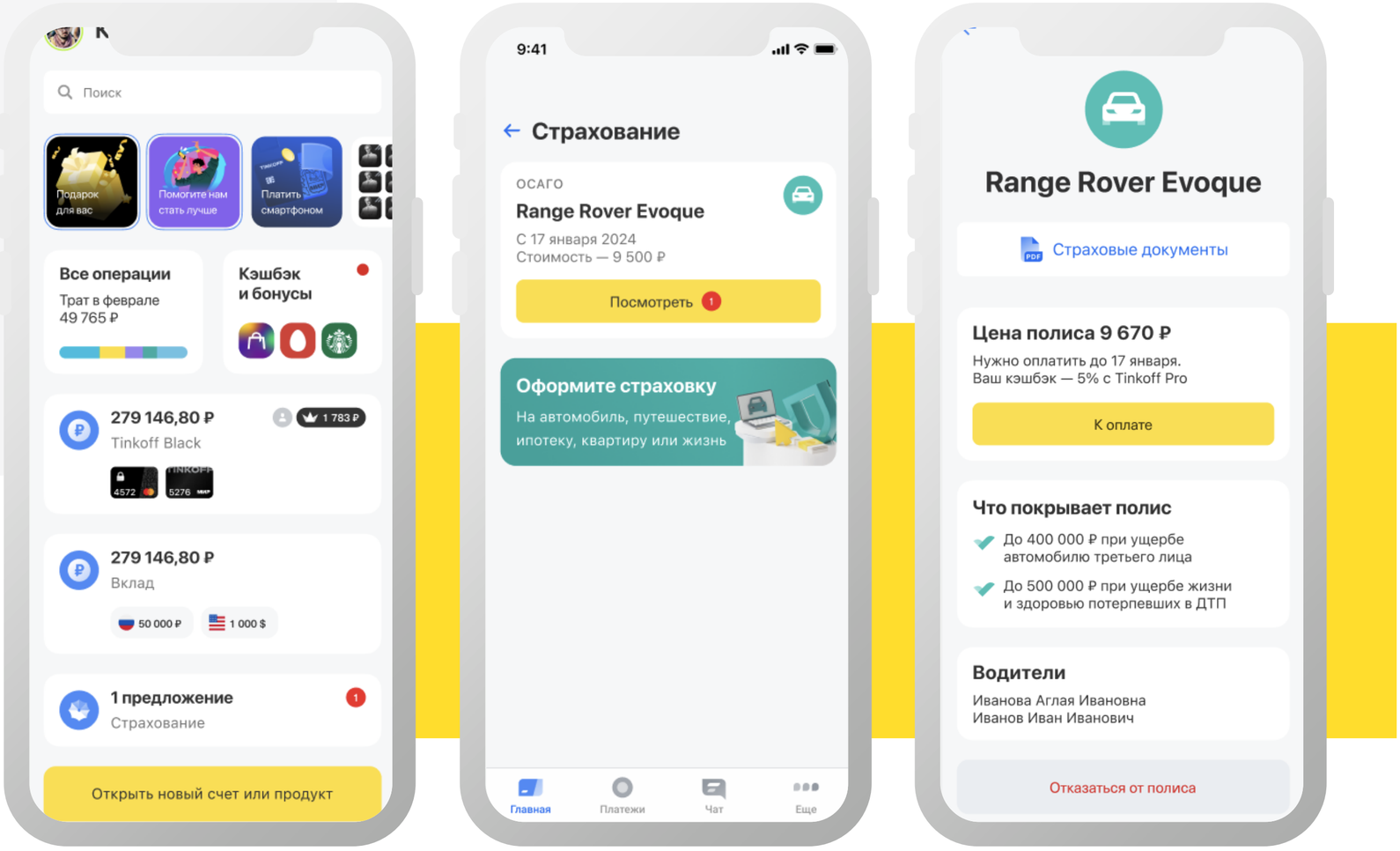
+10% конверсии
в утилизацию



Граф друзей

- Граф в рисках — **600 млн/год**
- Граф в «Приведи Друга» — увеличение конверсии в полную заявку на **5%**

OneClick – страховка в 1 клик



Конверсия 7%

25% от продаж всех полисов

200млн ежемесячно

Защитим или Вернем деньги

1. Технология Мобайла прервет звонок с мошенником
2. Банк заблокирует подозрительную операцию
3. Если мошенники обойдут защиту, мы вернем деньги

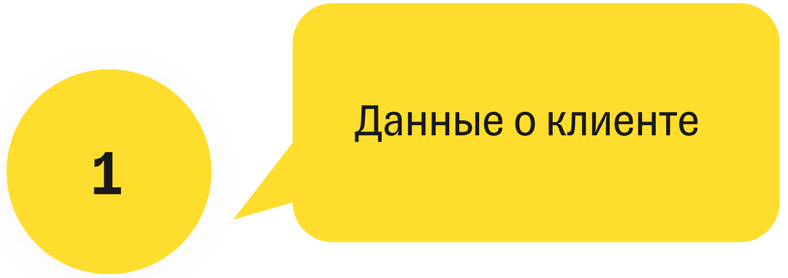
Защитили более 200млн рублей

В 10 раз ниже сумма потерь у клиентов

Блокируем 20млн/мес нежелательных звонков

Как получить аналитику?

Источники данных



Источники данных

1. Данные о клиенте
2. Заявки от клиентов

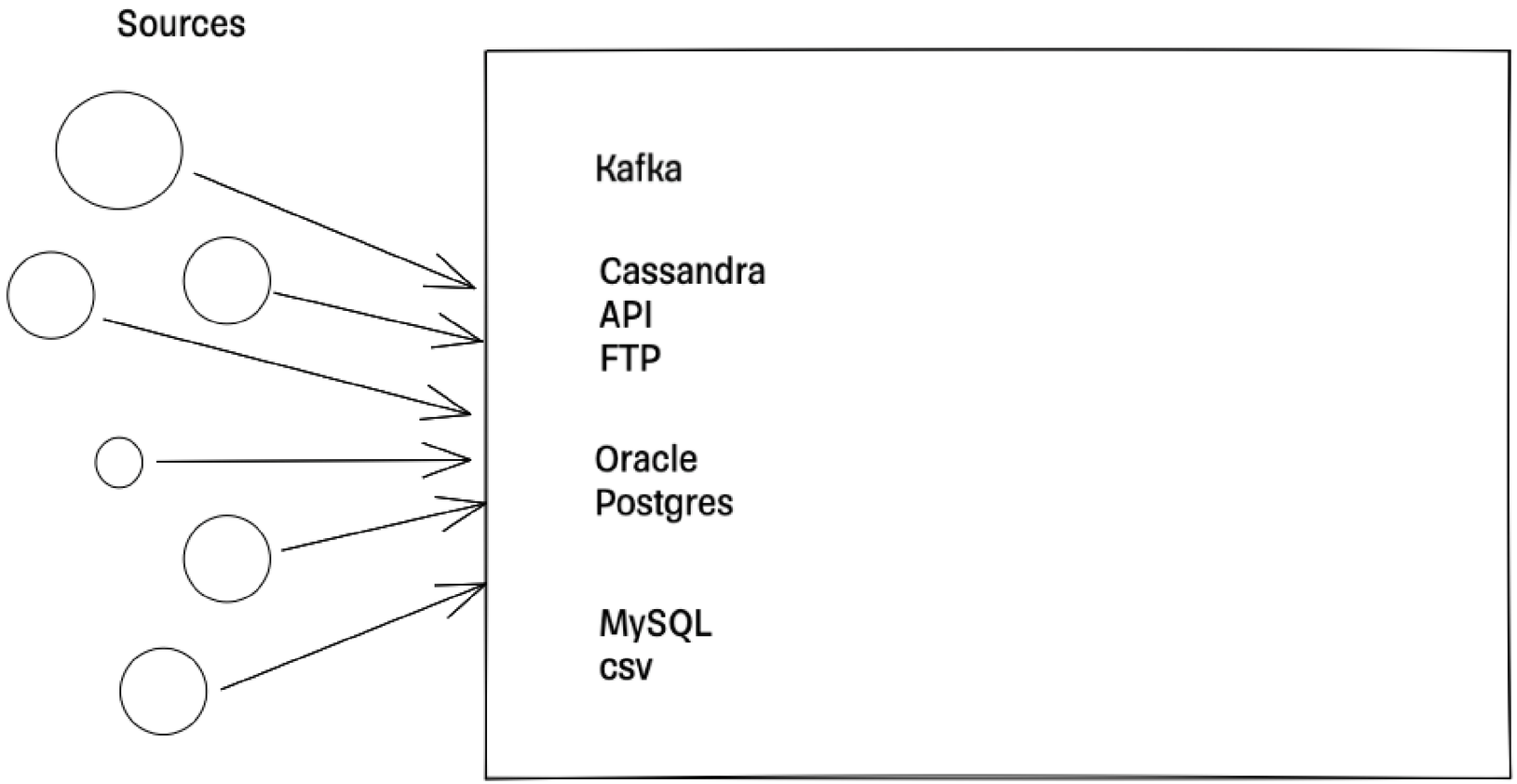
Источники данных

1. Данные о клиенте
2. Заявки от клиентов
3. Продукты

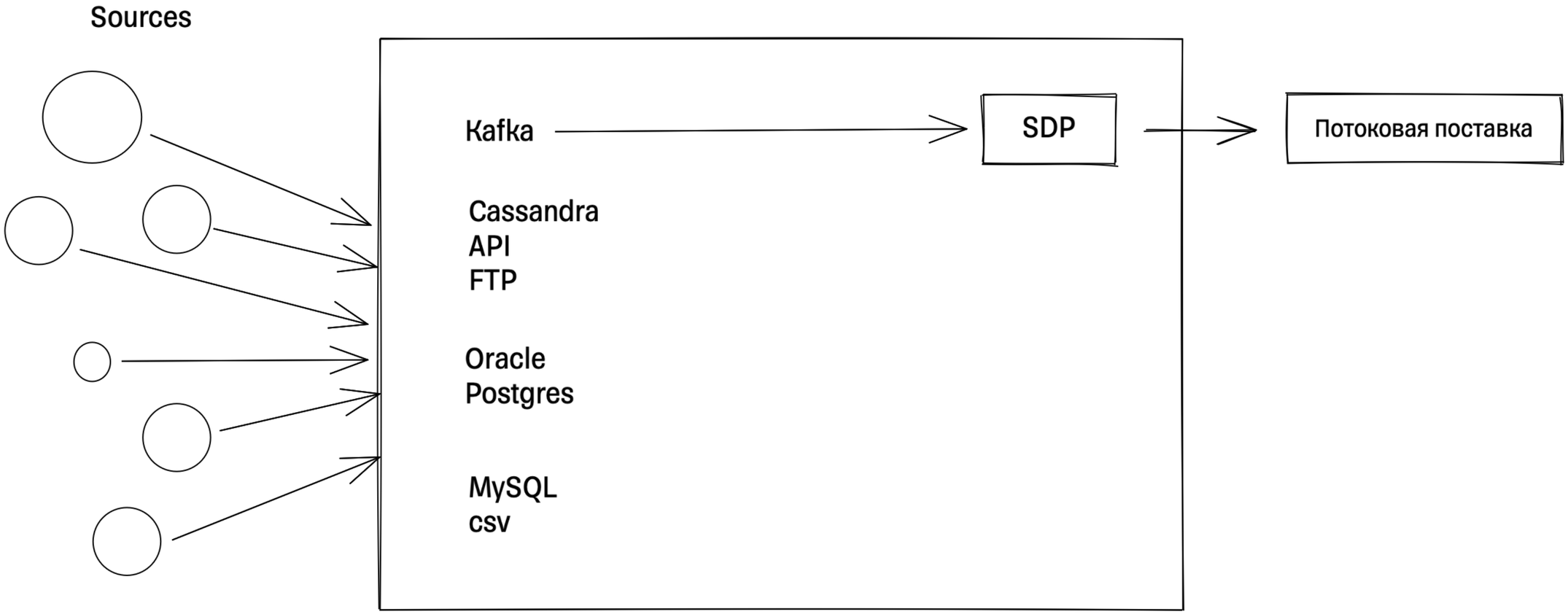
Источники данных

Источники данных

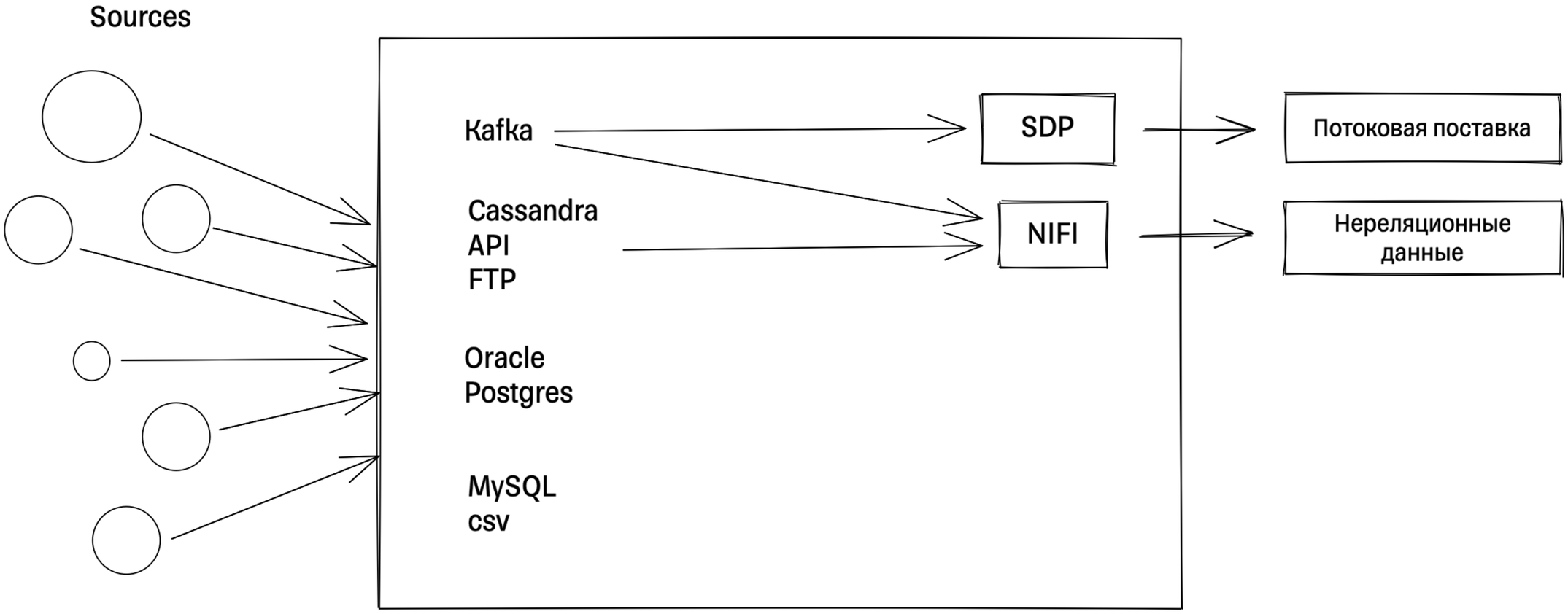
Поставка Данных в Data Platform



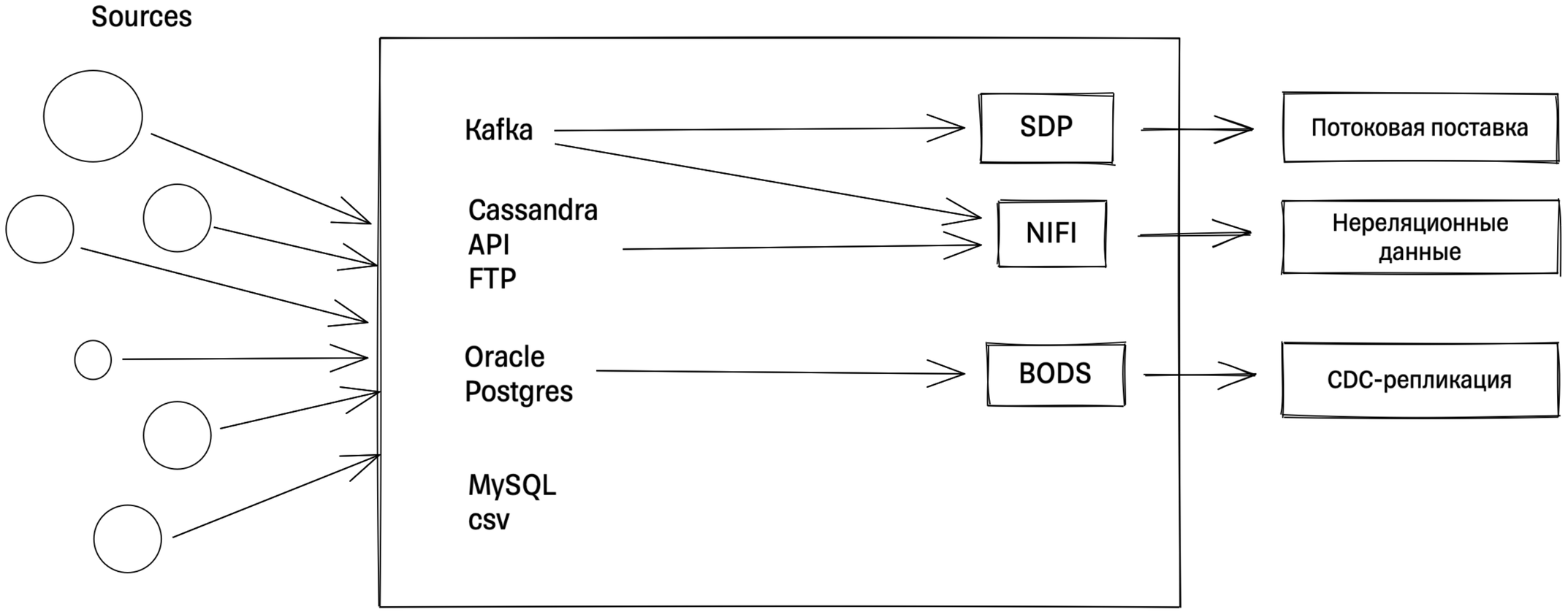
Поставка Данных в Data Platform



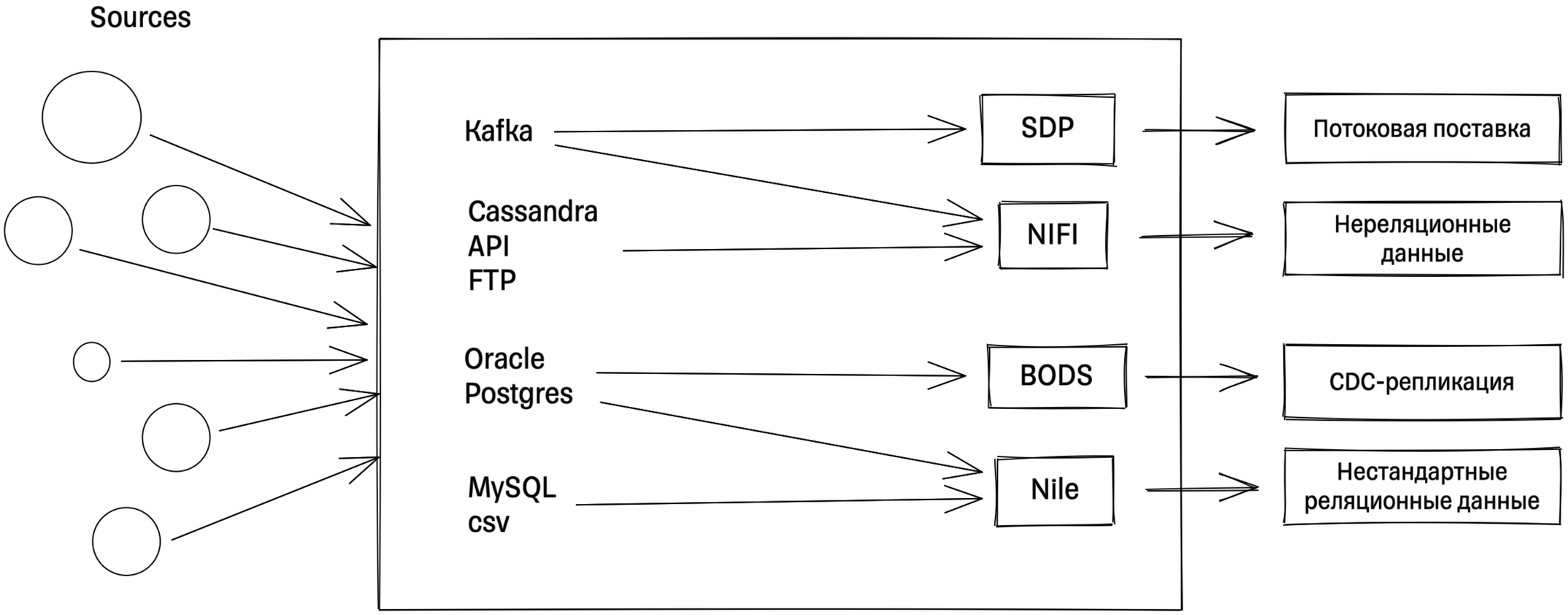
Поставка Данных в Data Platform



Поставка Данных в Data Platform



Поставка Данных в Data Platform



Поставка Данных в Data Platform

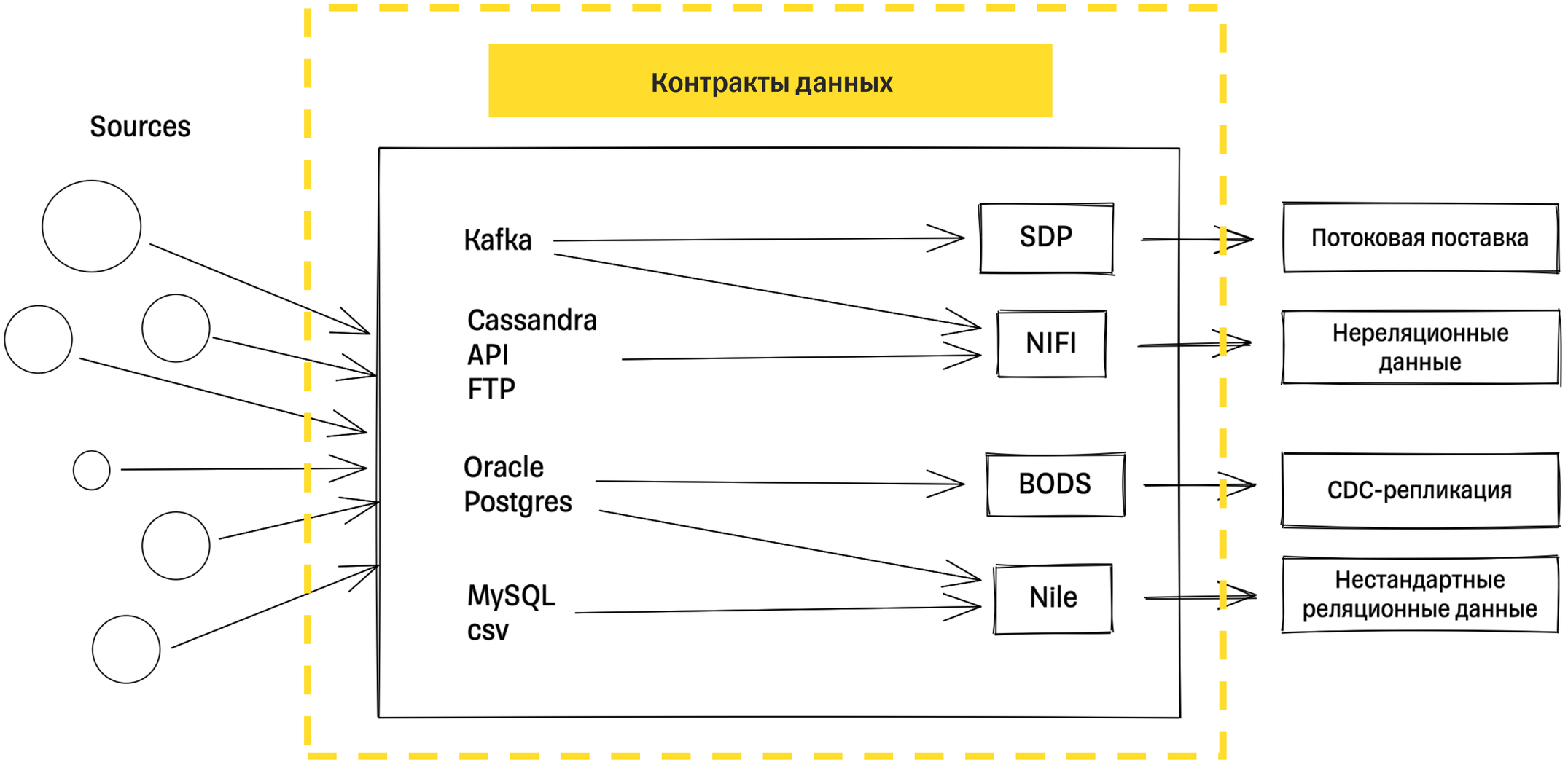


Схема загрузки данных

Схема загрузки данных

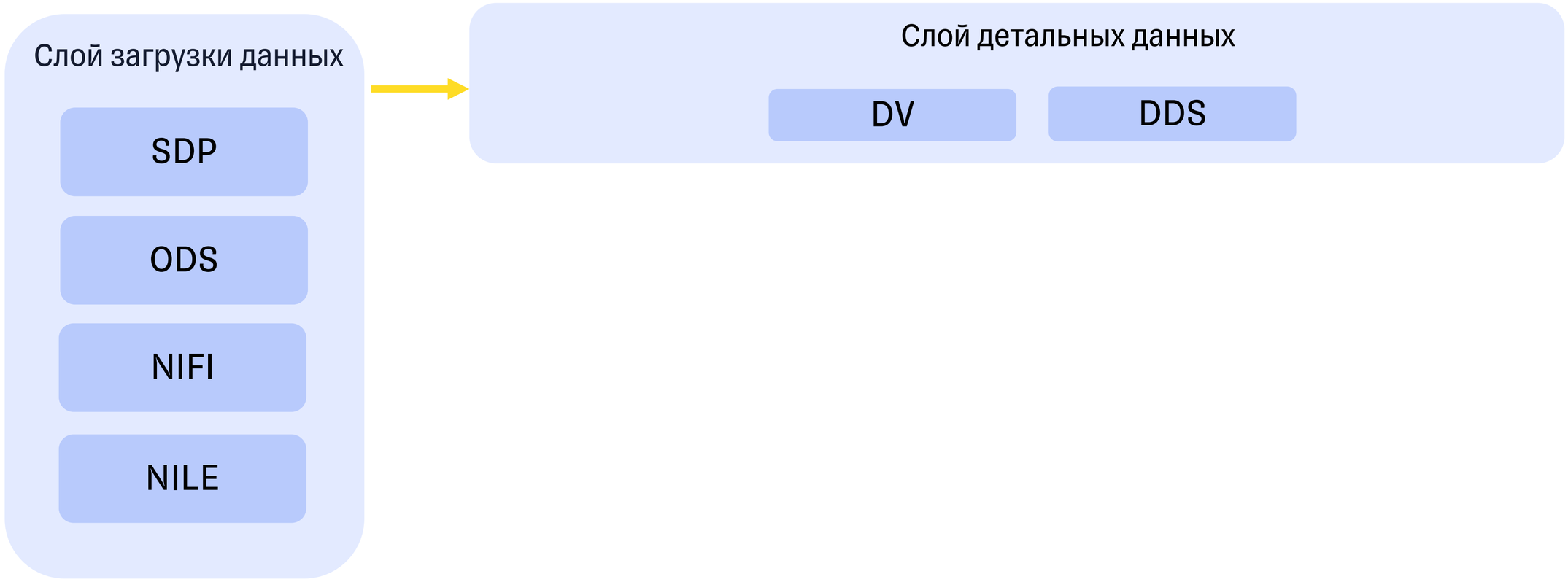


Схема загрузки данных

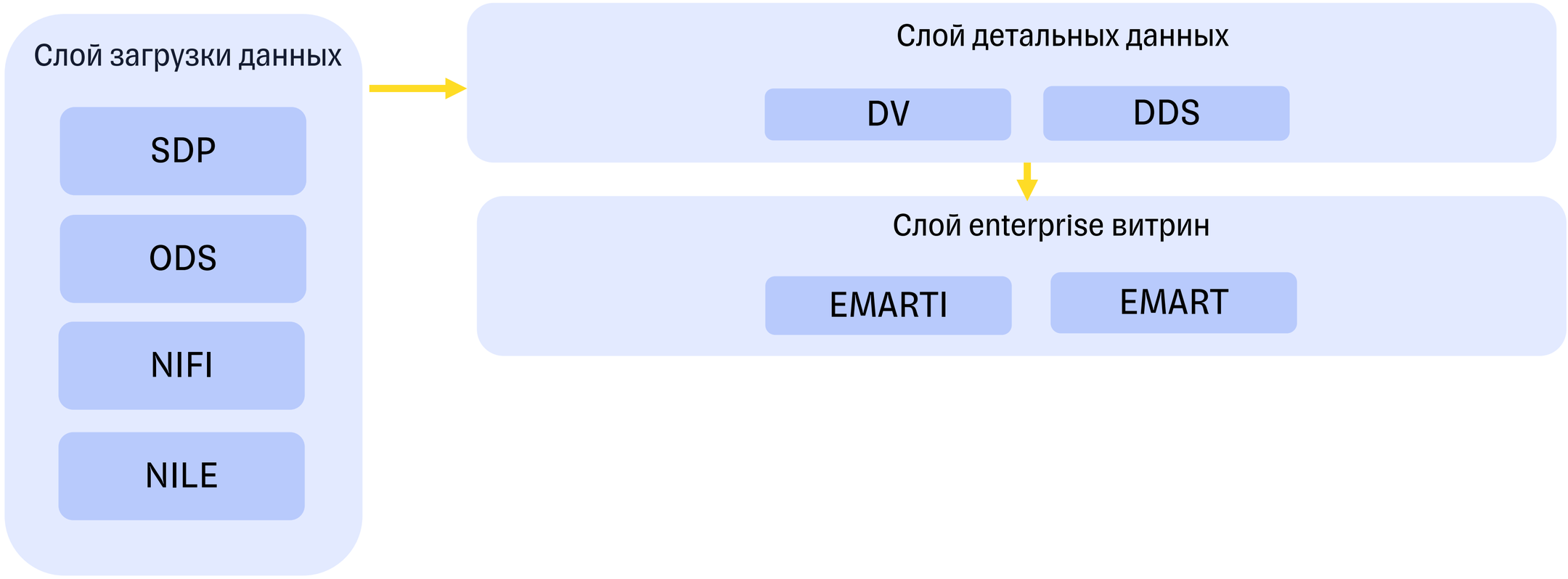
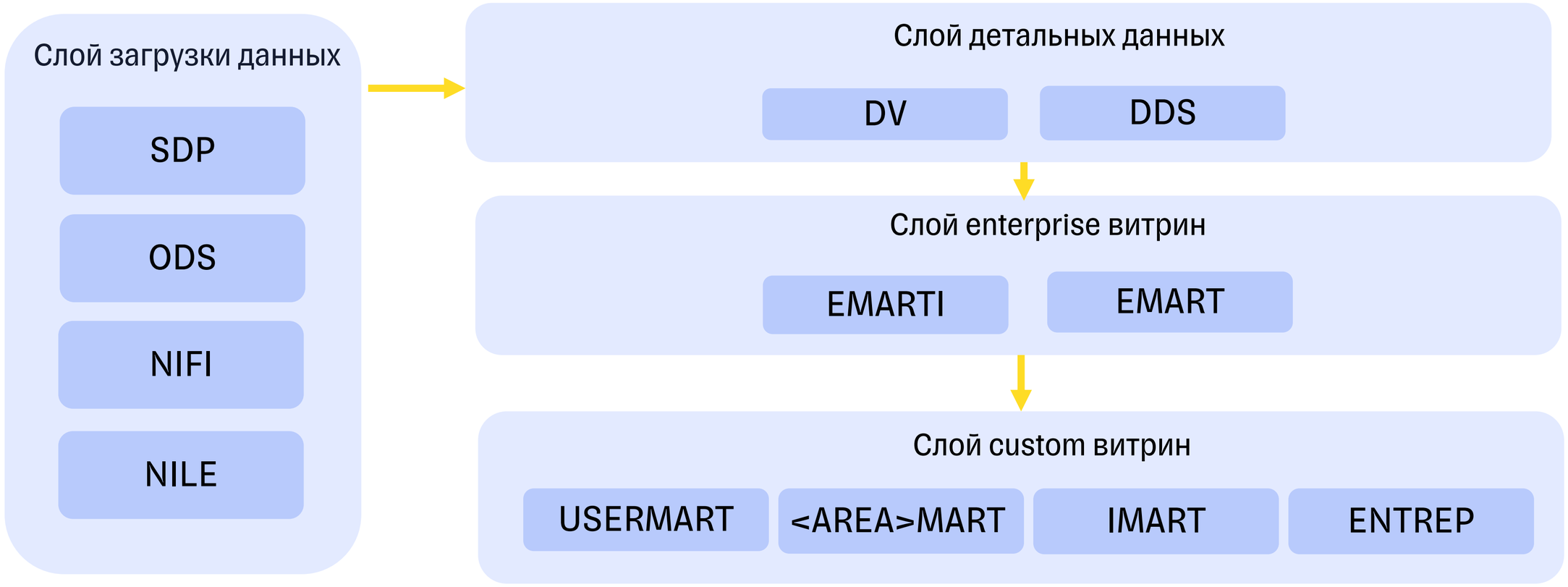


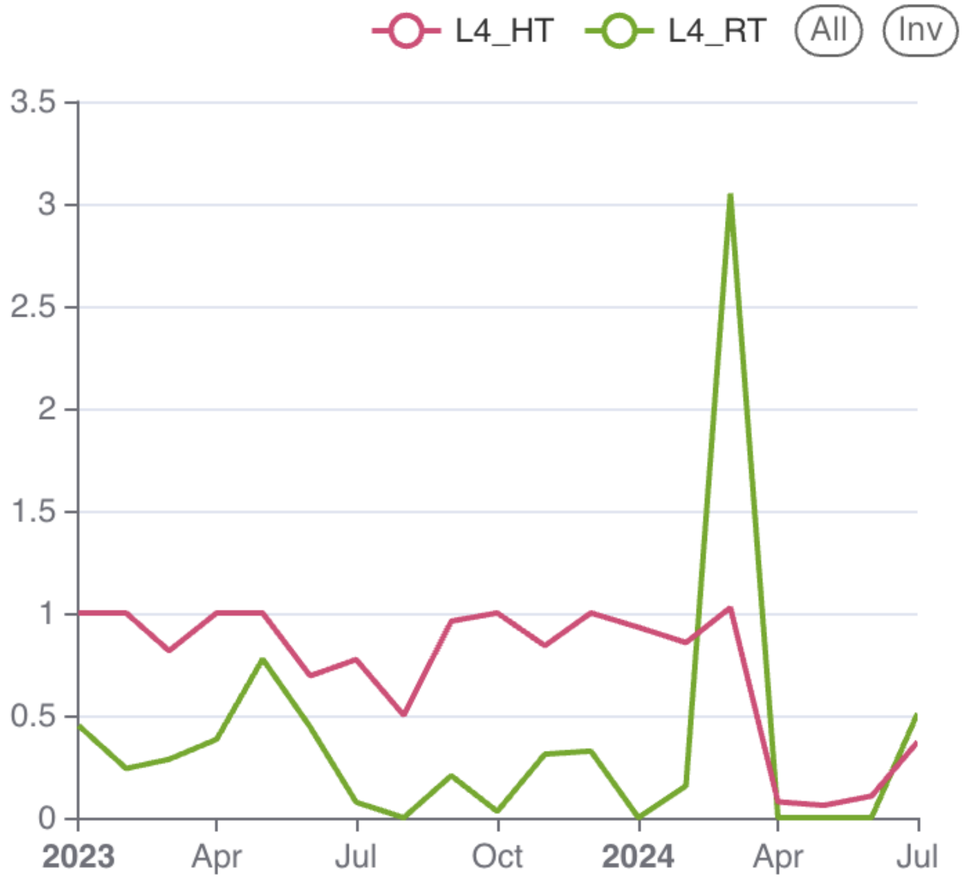
Схема загрузки данных

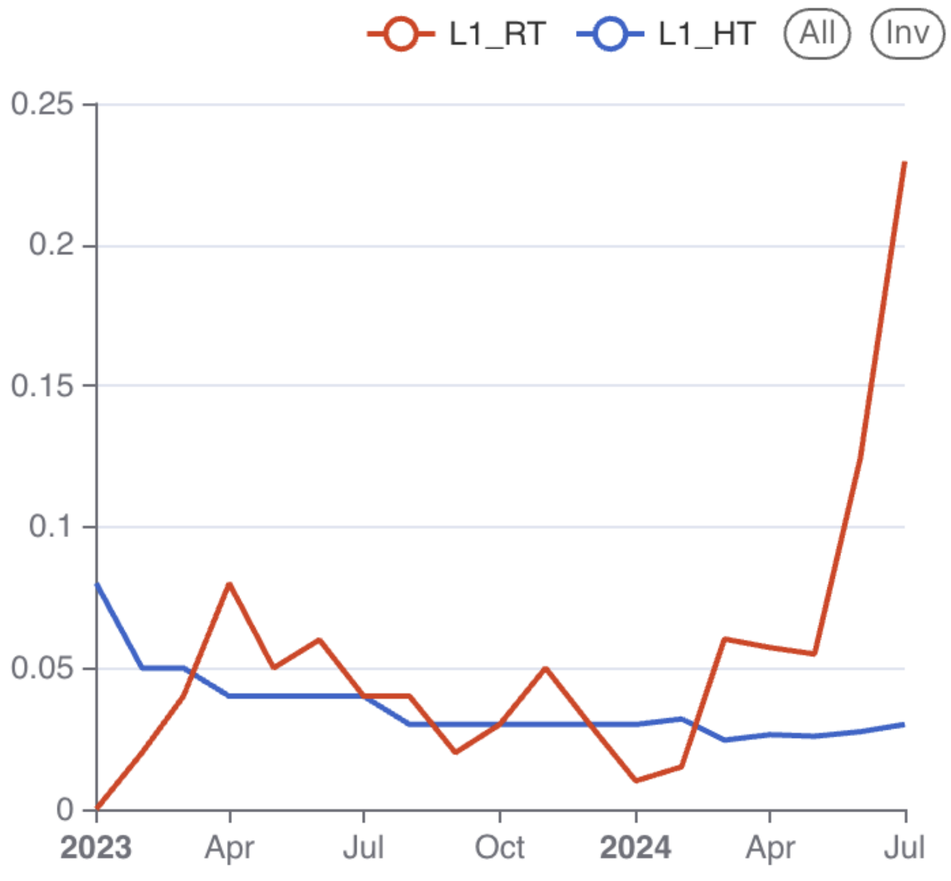


Построение аналитики

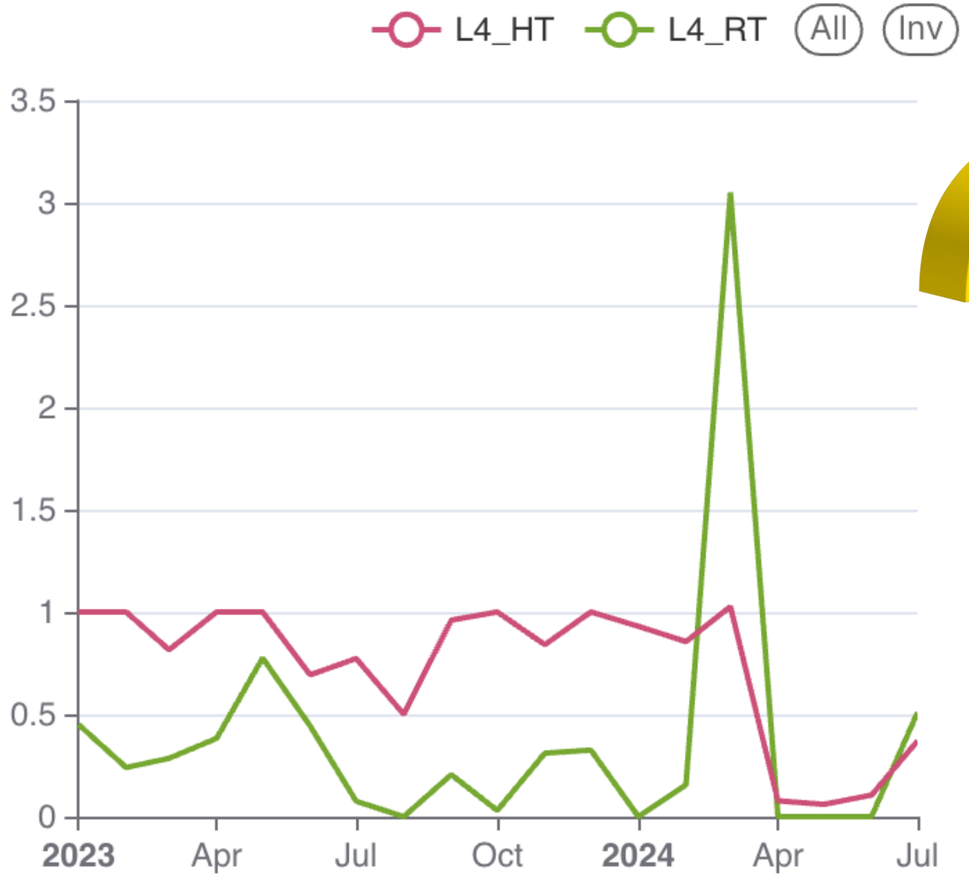
```sql
SELECT
    a.field1,
    b.field2,
    sum(field_cnt) AS field_sum,
    max(field4) AS max_field4
FROM prod_emart.client AS a
    INNER JOIN prod_dds.account AS b
        ON a.client_id = b.client_id
WHERE b.valid_to_dttm = '5999-01-01'
    AND b.deleted_flg IS DISTINCT FROM 1
GROUP BY 1,2
;



SELECT .....
FROM
LIMIT 100;
```

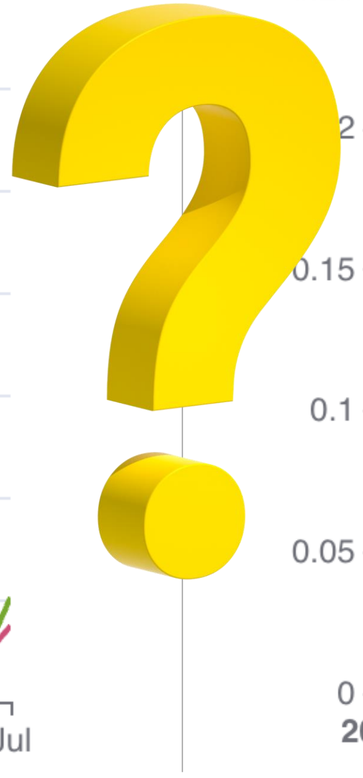
Визуализация данных

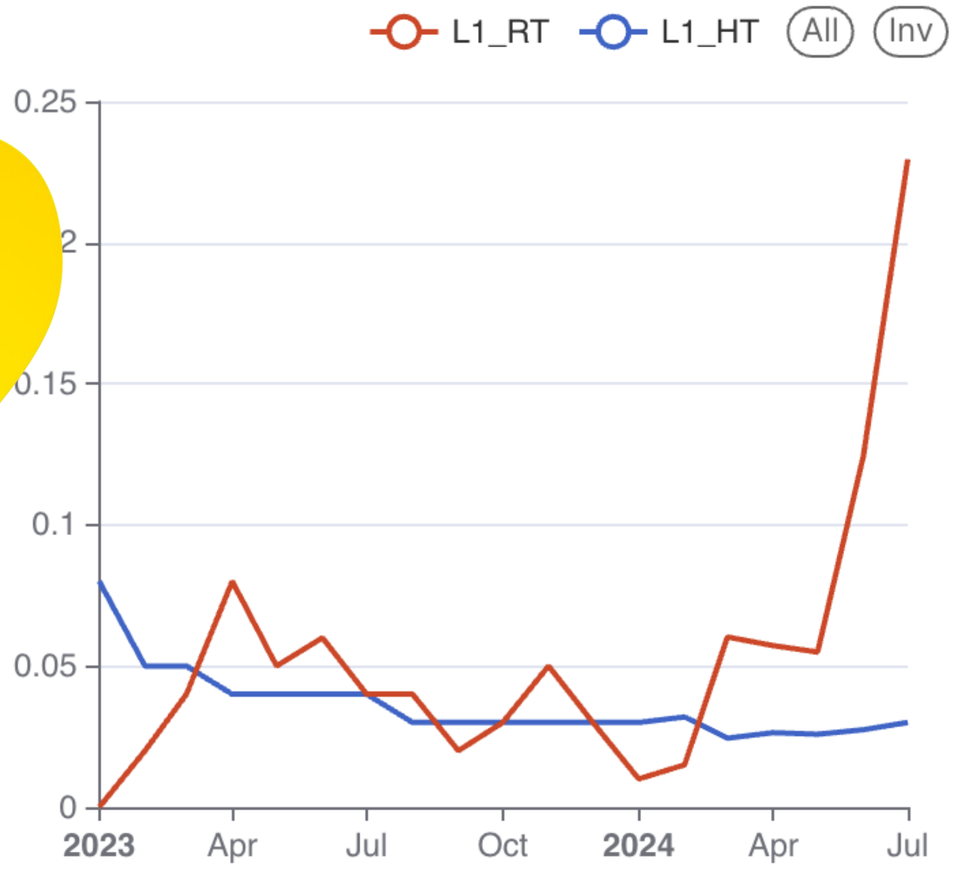




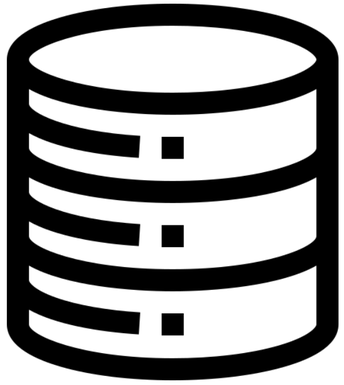
Визуализация данных





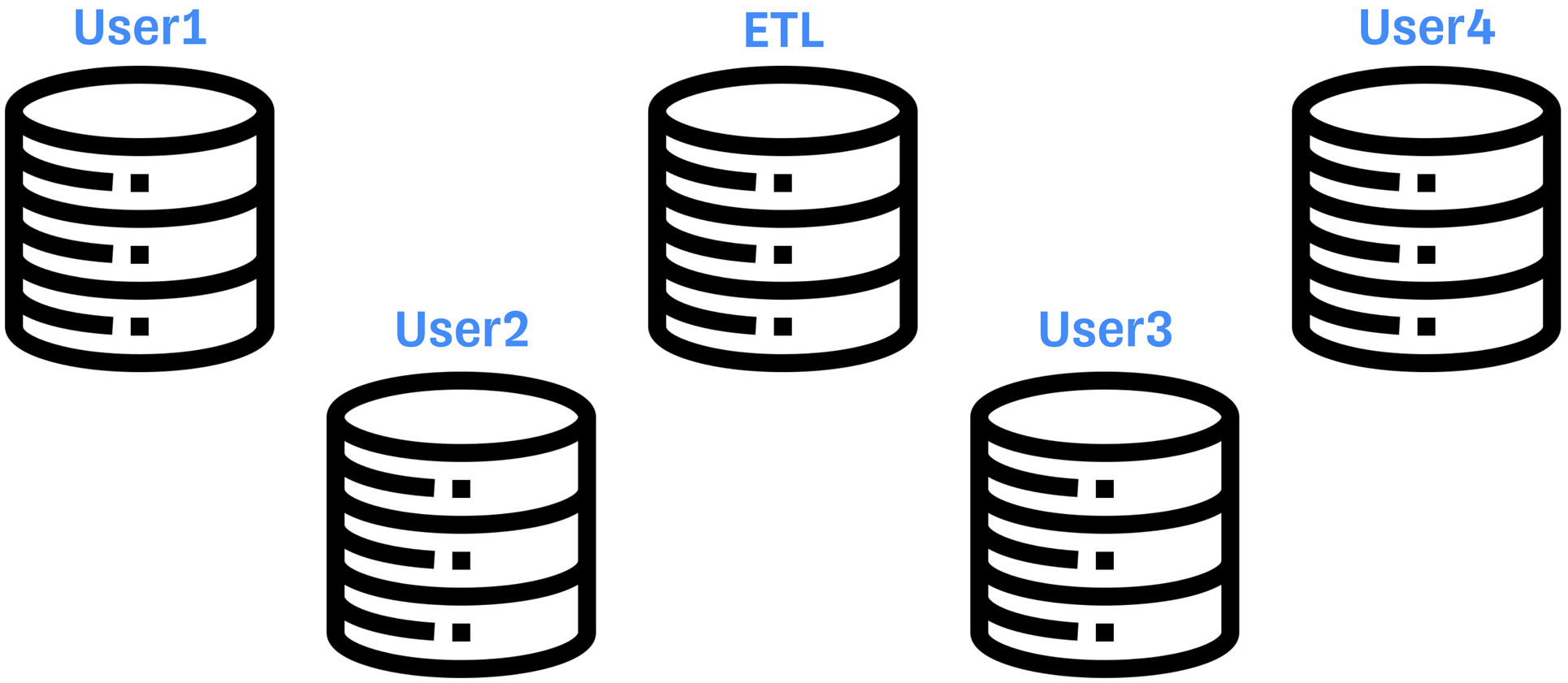


Место хранения Данных

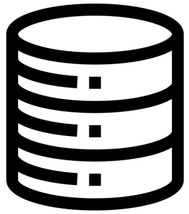


Greenplum

Место хранения Данных



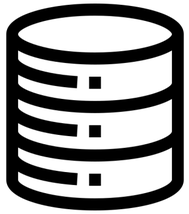
Место хранения Данных



Greenplum

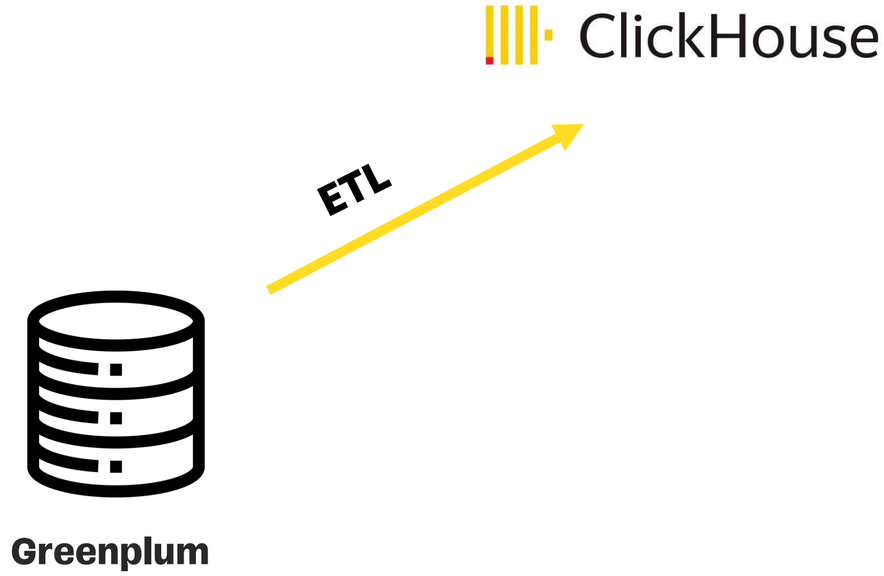
Место хранения Данных

ClickHouse

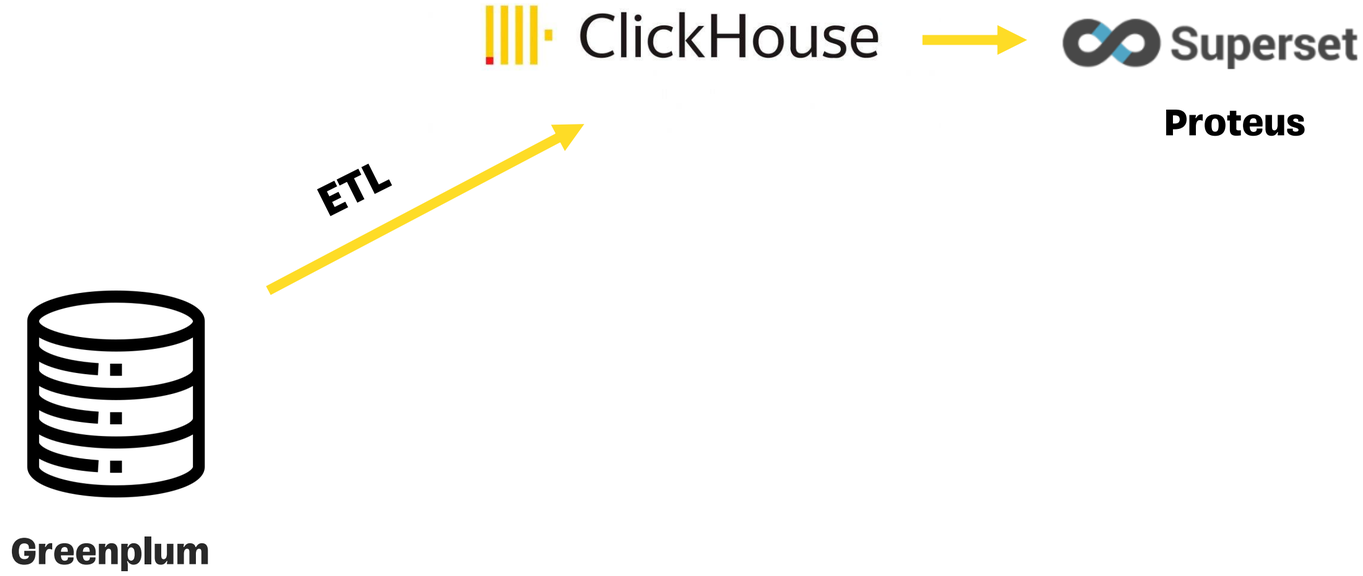


Greenplum

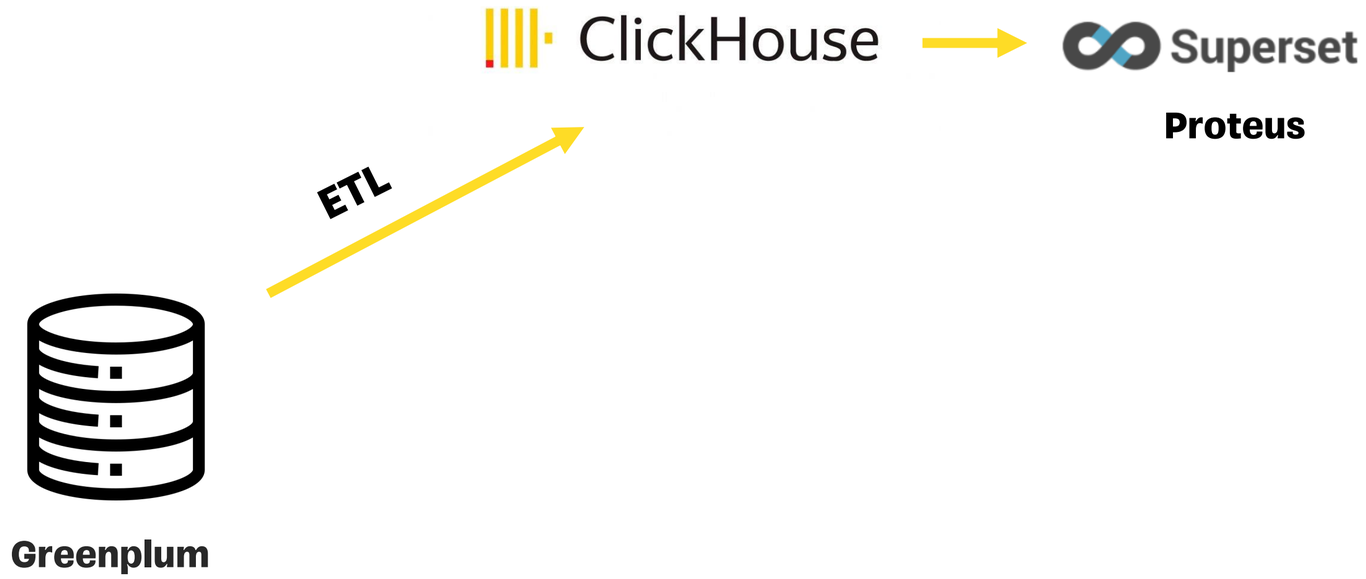
Место хранения Данных



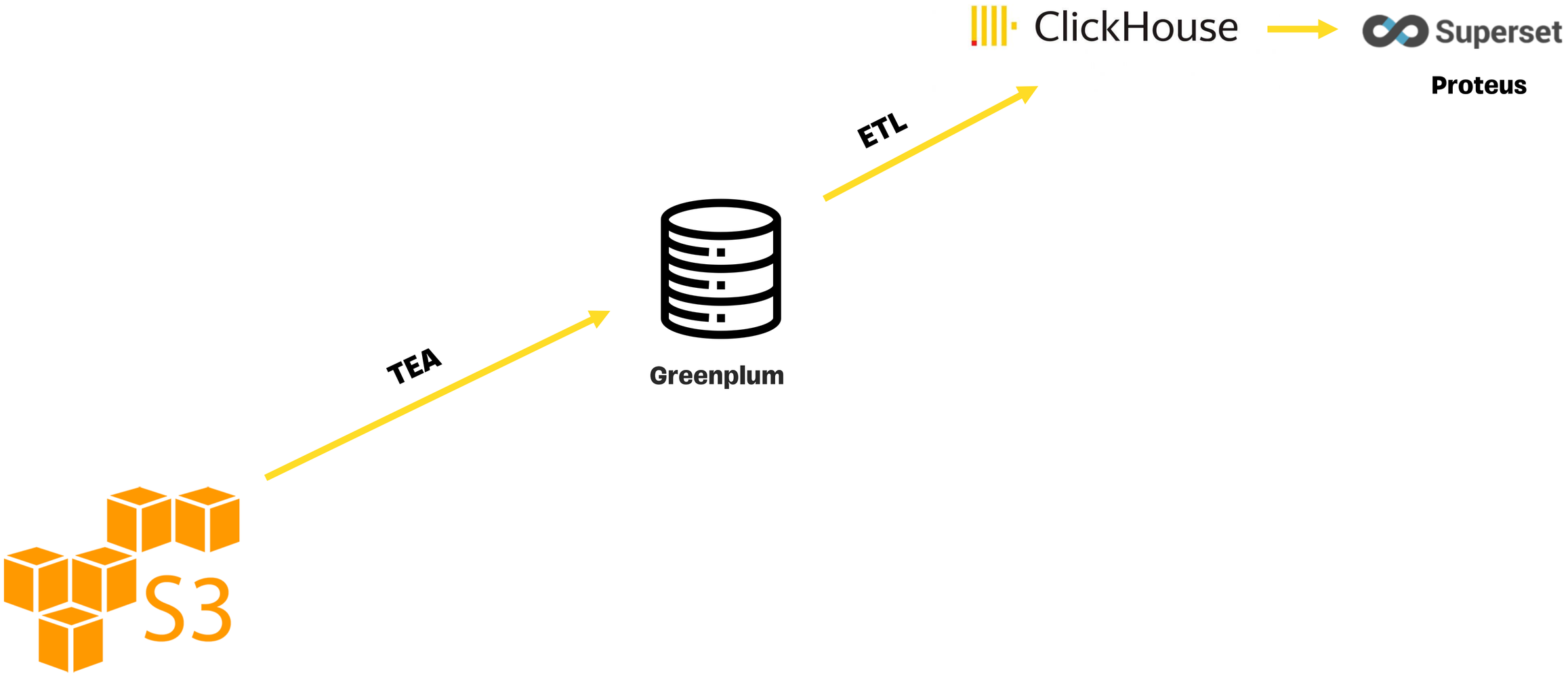
Место хранения Данных



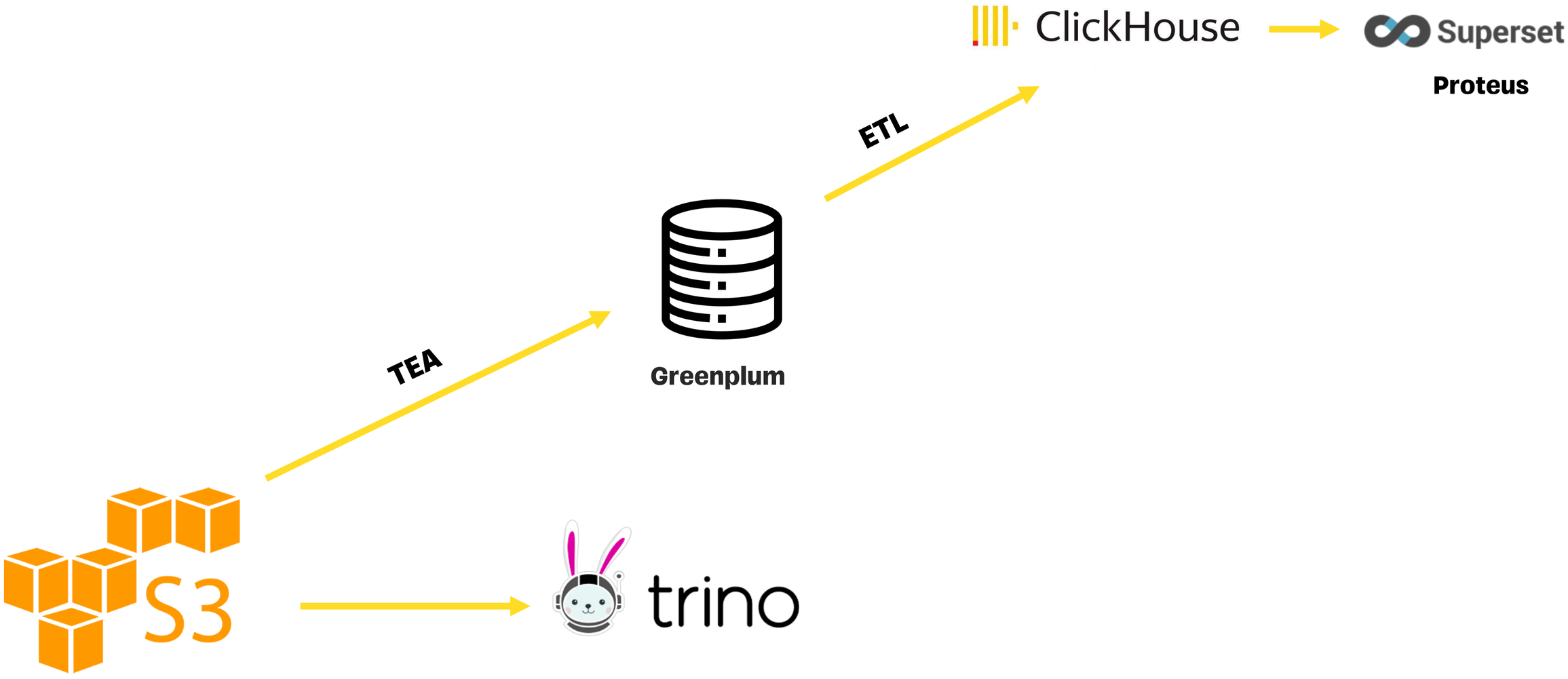
Место хранения Данных



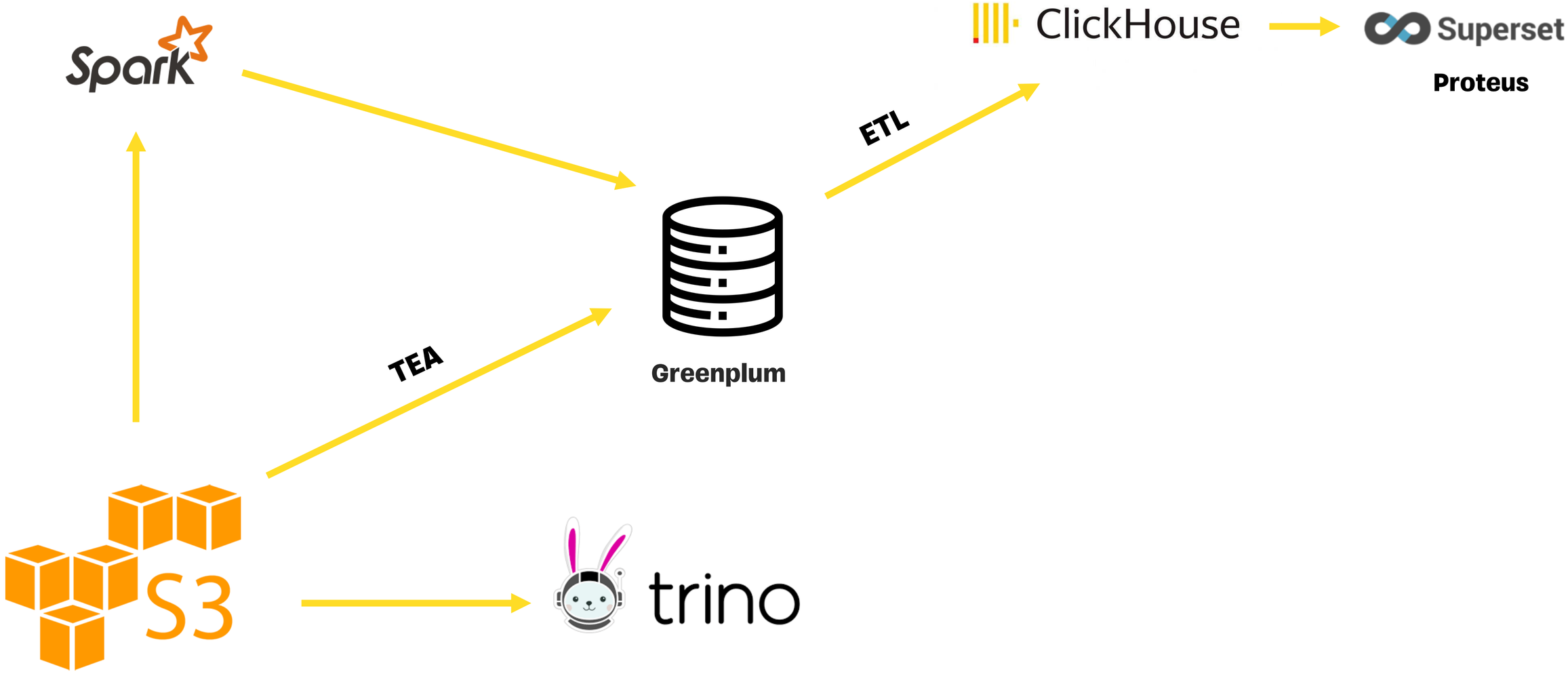
Место хранения Данных



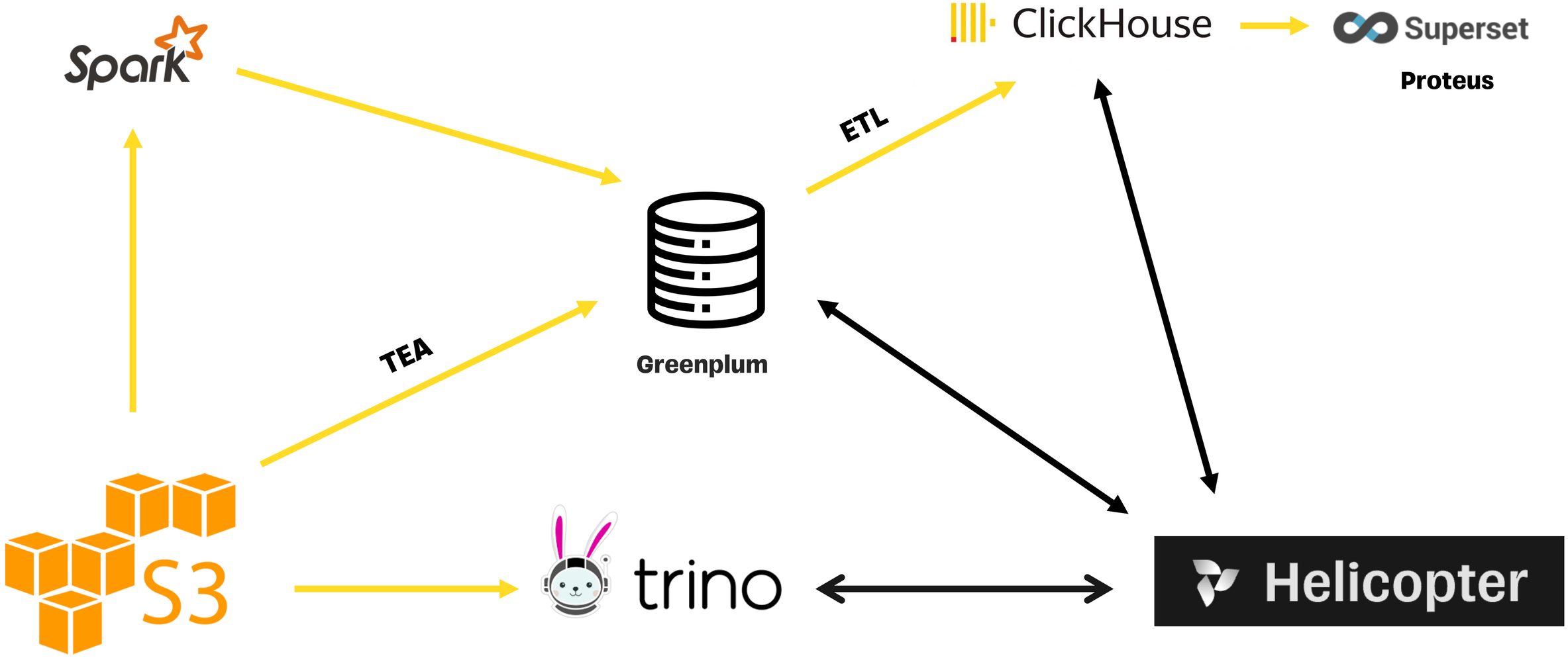
Место хранения Данных



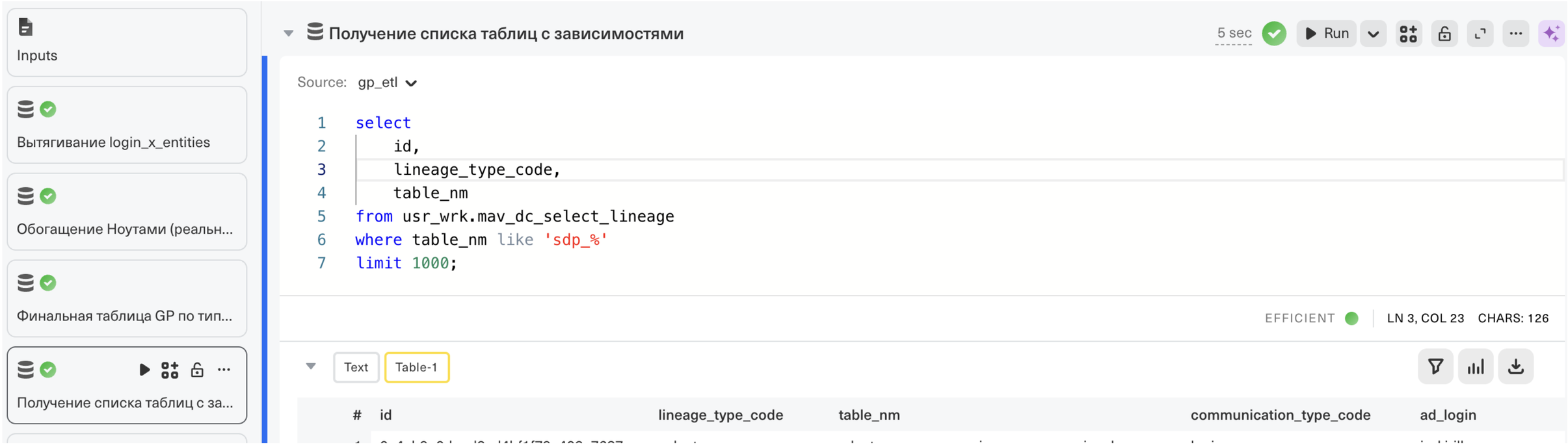
Место хранения Данных



Место хранения Данных



Helicopter


Inputs
Вытягивание login_x_entities
Обогащение Ноутами (реальн...
Финальная таблица GP по тип...
Получение списка таблиц с за...
Получение списка таблиц с зависимостями
5 sec
Run
Source: gp_etl
1 select
2 id,
3 lineage_type_code,
4 table_nm
5 from usr_wrk.mav_dc_select_lineage
6 where table_nm like 'sdp_%'
7 limit 1000;
EFFICIENT
LN 3, COL 23
CHARS: 126
Text
Table-1
id
lineage_type_code
table_nm
communication_type_code
ad_login

Flow со стороны пользователя

01

Raw-данные

Передать данные из систем-источников в Data Platform

02

Модель

Подготовить и собрать необходимые витрины в подходящих разрезах

03

Доступ и ресурсы

Определение кластера и БД, получение доступа, определение ресурсов

04

Данные на кластере

Добавить на кластер все данные, которые нужны для аналитики

05

Визуализация

Построение аналитики и отчетности

Flow со стороны пользователя

01

Raw-данные

Передать данные из систем-источников в Data Platform

02

Модель

Подготовить и собрать необходимые витрины в подходящих разрезах

03

Доступ и ресурсы

Определение кластера и БД, получение доступа, определение ресурсов

04

Данные на кластере

Добавить на кластер все данные, которые нужны для аналитики

05

Визуализация

Построение аналитики и отчетности

Profit!

Что по цифрам?

Немного фактов

Пользователи

- MAU 16k+
- 170k+ запросов к GP ежедневно

Данные

- 5+ ПБ уникальных данных
- 20+ ПБ данных всего

BI / DG

- Helicopter (Zeppelin inspired)
- Proteus (Superset based)
- Data Detective

ETL

- TEDI
- Moebius

Ingest

- CDC (GG -> Debezium)
- SDP (Kafka & Flink)
- NiFi

T-Bank Data Platform

Greenplum (с 2011)

- 15+ кластеров
- 600+ bare-metal серверов

ClickHouse (с 2019)

- 2 Bare-metal кластера
- 60+ кластеров на ВМ

S3 + Trino (с 2023)

- Разделяем storage и compute
- Будет новым ядром

Т БАНК

Оцените, пожалуйста, выступление

Спасибо!