Яндекс Еда

Delivery Club + Yandex Eats

Как мы интегрировали два хранилища

Содержание

Кому будет интересен этот доклад

1

Тем, кому предстоит проект с объединением DWH

2

Тем, кто делал что-то похожее и хочет узнать, как у других

3

Тем, кто устал от хардкора и хочет послушать кулстори

01

Мотивация.
А что случилось?

Bell сообщил о переговорах VK и «Сбера» о продаже Delivery Club «Яндексу»

Объединение Delivery Club и «Яндекс Еды»: что это значит для пользователей и ресторанов

Яндекс

Новости Вакансии Инвесторам Рекламодателям О компании Контакты

‹ Новости

Delivery Club становится частью бизнеса Яндекса

23 августа 2022

Интернет, 23 августа 2022 года. Яндекс покупает у VK сервис доставки еды и продуктов Delivery Club. Компания продолжит развивать бренд Delivery Club, приложение и сайт будут работать как раньше.

02

Что это означало для нас?

Много работы!

Исходный таймлайн

Ожидание (Сент. – Окт.)

Закрытие сделки (Окт. – Дек.)

WIN

~4 недели

Детальная проработка

Формирование команды

Подготовка скоупа и роудмэпа

Первичный анализ

~7-8 недель

Интеграция сервисов

Постепенная интеграция процессов

Переключение курьеров

Переключение вендоров

~15 Дек. – 15 Янв.

Новогодний фриз

Остановка выкатки продуктовых фич, фокус на техдолге

Продолжение подготовки к переключению

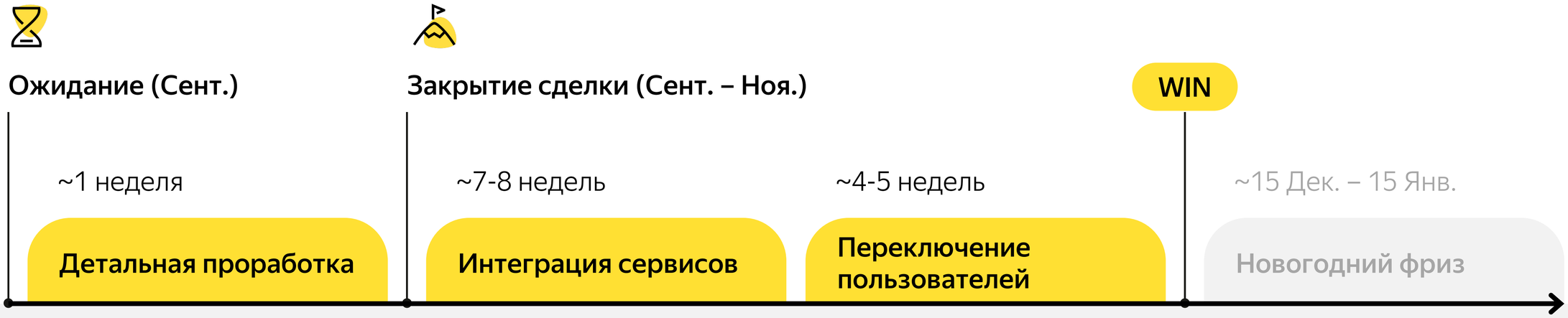
~4-5 недель

Переключение пользователей

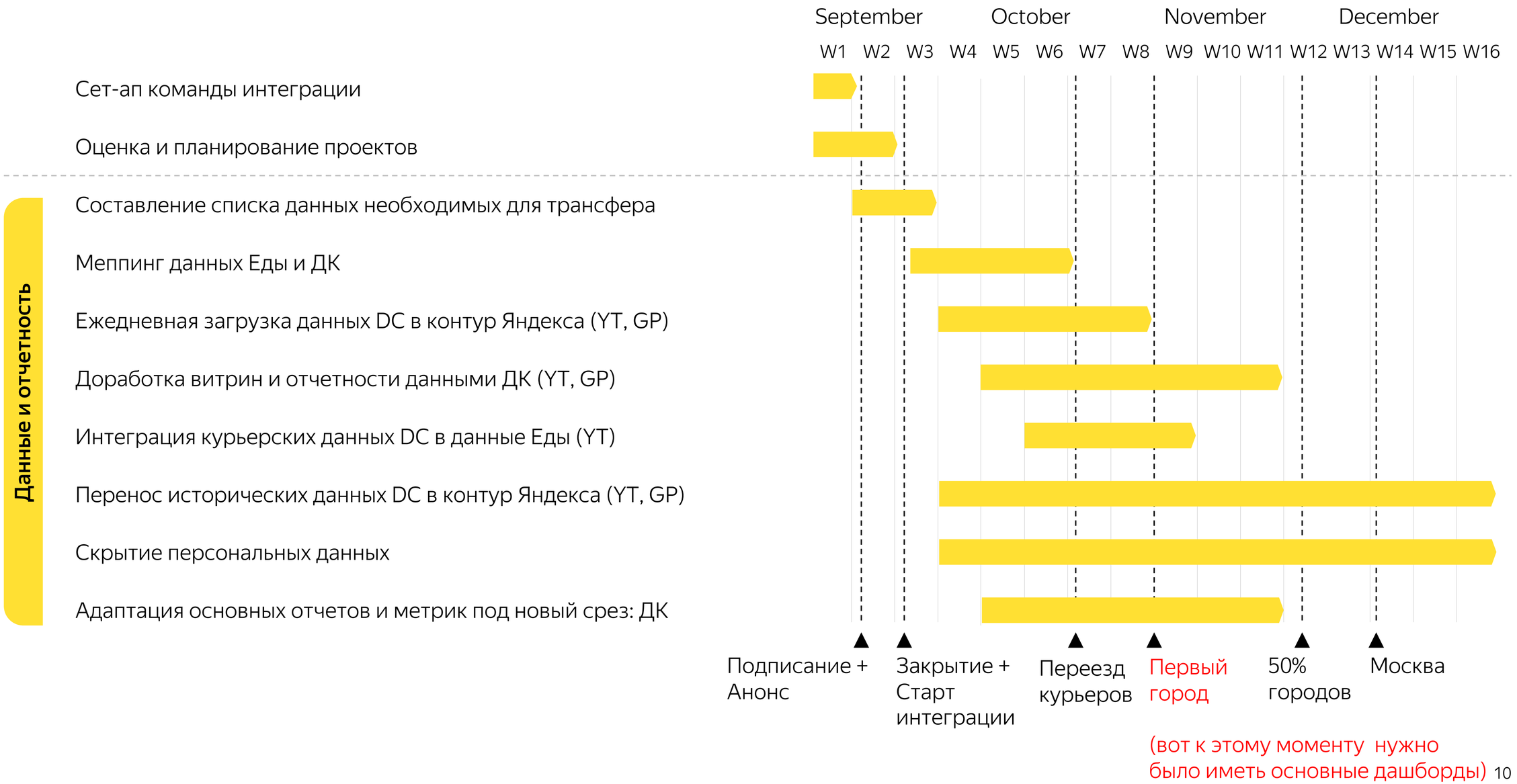
Несколько волн

- Небольшие города
- 1М+
- Москва и Санкт-Петербург

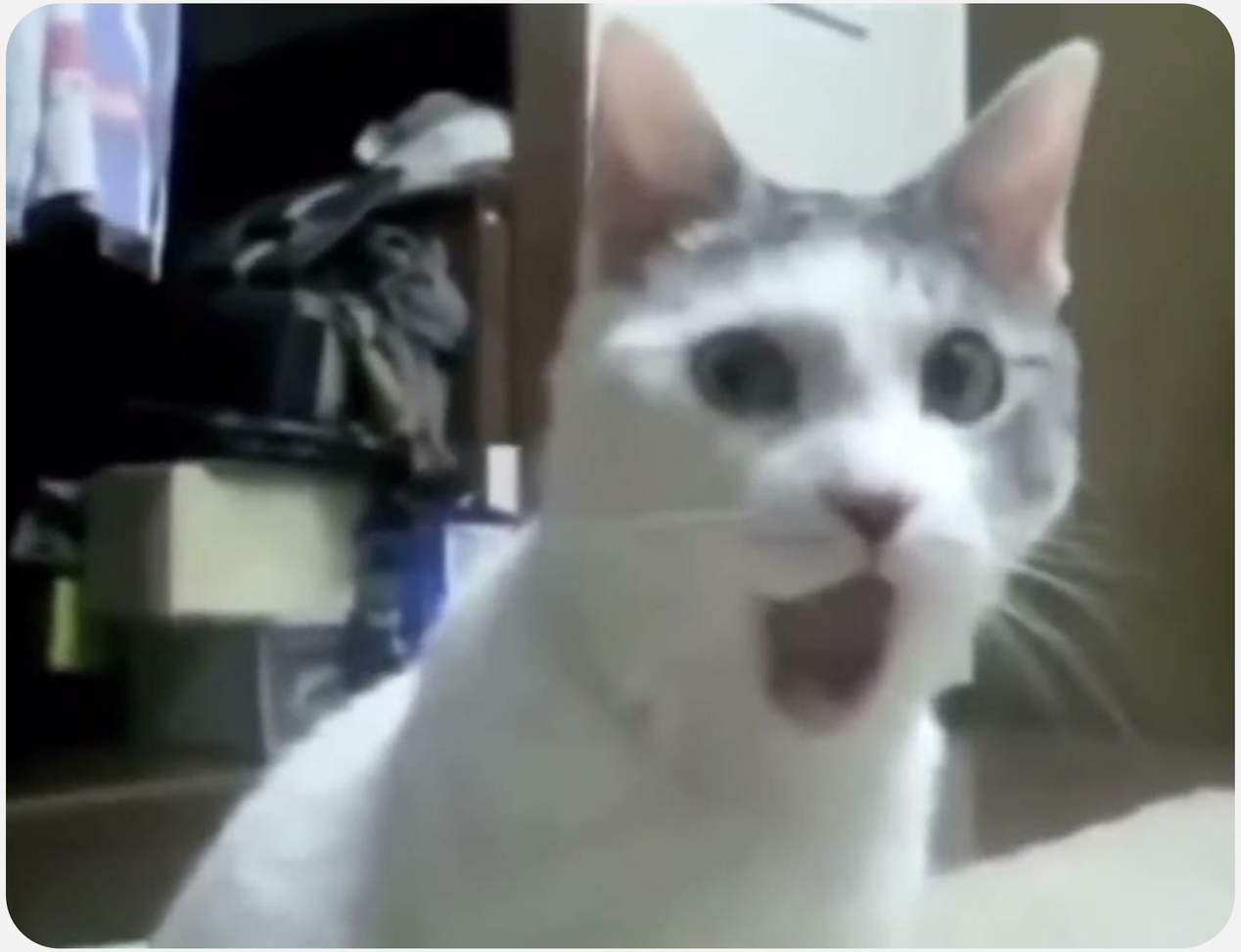
А потом что-то пошло не так



А как выглядели сроки переезда данных?

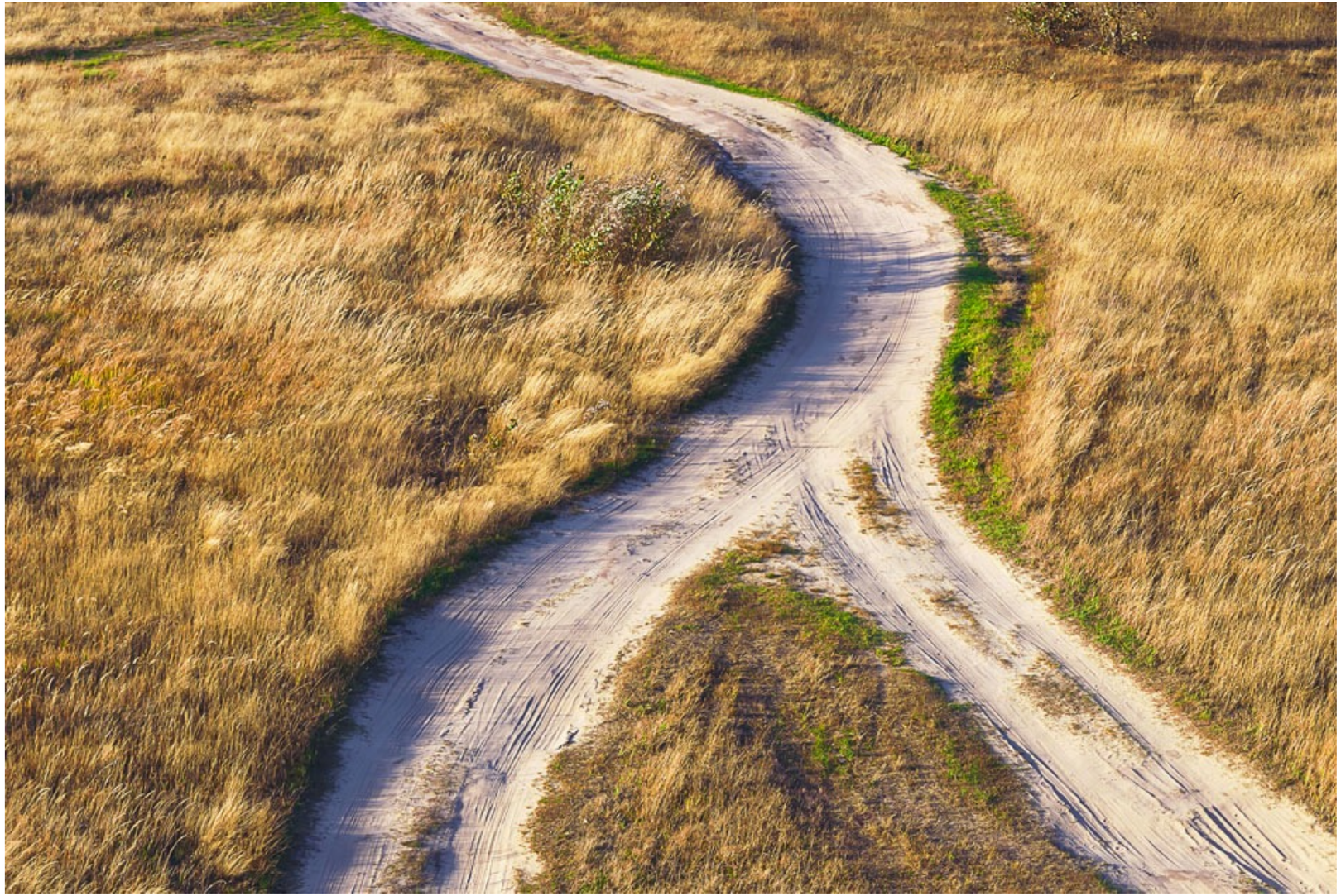


Какими глазами мы смотрели на этот Гант



04

Как можно решить эту задачу?



Как можно объединить два DWH?

Sources

Подключить источники,
пролить данные в ДДС
и витрины, повторить логику

Плюсы

- План простой и надежный
- Точно ничего не потеряется
- Долгосрочное решение

Минусы

- Реализация сложная и долгая
- Сложно и не всегда
  возможно восстановить
  логику расчета метрик

Как можно объединить два DWH?

В качестве источника взять сырой слой DWH, которое вливаем. Пролить данные в ДДС и витрины, повторить логику

Плюсы

- Ничего не теряется
- Настраивается только одно подключение

Минусы

- Все еще долго
- Сложно и не всегда возможно восстановить логику расчета метрик

Как можно объединить два DWH?

Если оба ХД имеют детальный слой, можно попробовать объединить на уровне логических моделей

Плюсы

- Данные попадут во все витрины
- Сложность не зависит от числа источников

Минусы

- Не всегда возможно восстановить логику метрик
- Подходит только если оба DWH остаются включенными
- Расходы на поддержание DDS

Как можно объединить два DWH?

Объединять уже готовые данные в витринах

Плюсы

- Быстро
- Сохраняется логика расчета метрик

Минусы

- Новые витрины
- Нужно аккуратно делать пересчеты
- Нужно придумать, как унифицировать формат данных (ID, коды)
- Подходит не всегда

Как можно объединить два DWH?

☑ Мы выбрали этот способ, пошли от витрин

05

Что нам предстояло сделать

Что нам предстояло сделать

Сетап проекта

- Команда
- Список и состав объектов, в которых требуются данные ДК
- Меппинг бизнесовых понятий и метрик

Определиться с подходом и технологиями

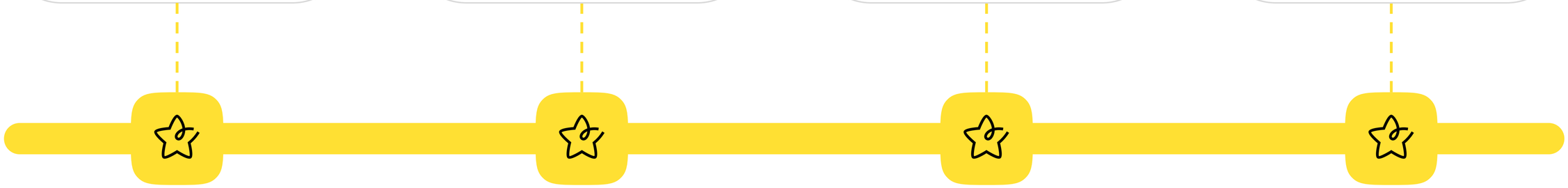
- Как забрать данные из контура VK до объединения
- Как организовать загрузку данных

Успеть

- Тут без комментариев

Не уронить качество

- Скорость поставки
- Полнота данных
- Частота падений



Как мы составляли скоуп проекта

Опрос на пользователей

Самые используемые объекты

Источники популярных дашбордов

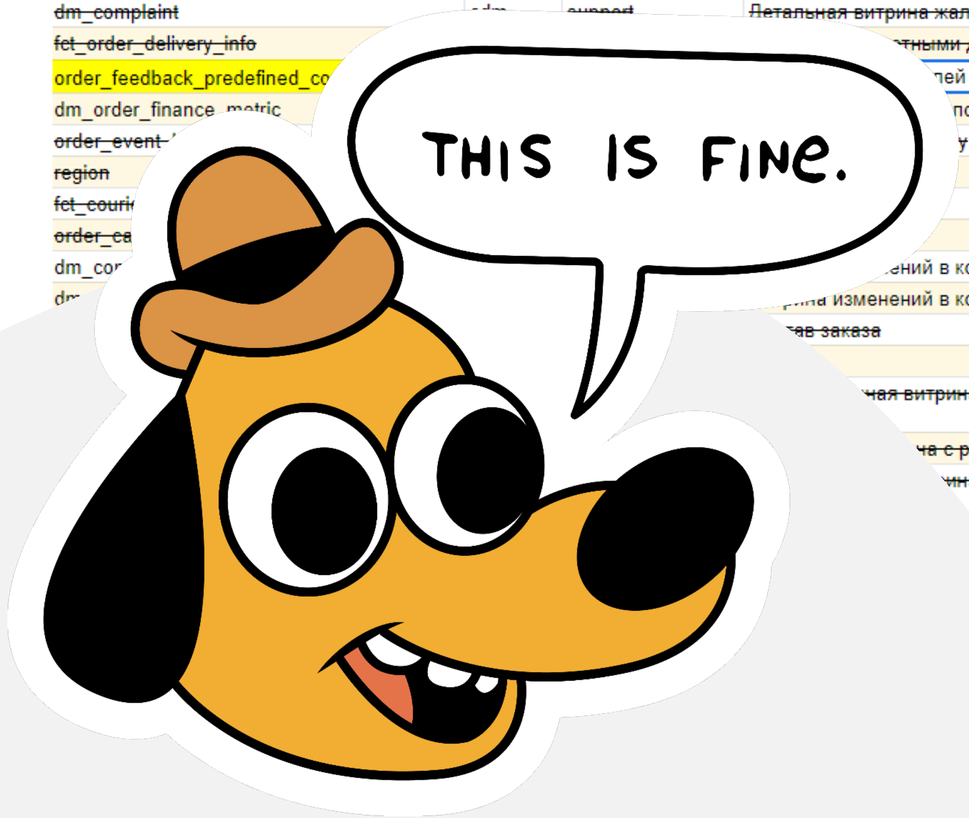
Собрали общий список ODS, витрин и отчетов

Далее

- Разделили таблицы по доменам
- Разбили по приоритетам
- Определили, что нужно делать по каждому объекту

Как мы составляли скоуп проекта

Yandex Eats										
Таблица	Слой	Домен (код)	Название бизнес-сущности	Домен данных	Tech (детальный домен)	Tech (кастомный приоритет	Приоритет	Тикет	Меппинг есть	
~~dm_order~~	~~cdm~~	~~order~~	~~Витрина заказов Еды~~	~~Заказы~~	~~orderЗаказы~~		~~1~~	~~https://st.yandex-team.~~	Да	Перв
~~dm_business_overview~~	~~rep~~	~~management_repor~~	~~Витрина под отчет business overview~~	~~Заказы~~	~~management_reportingЗак~~		~~1~~	~~https://st.yandex-team.~~	Да	Доба
~~dm_order_metric~~	~~cdm~~	~~order~~	~~Витрина с основными метриками заказа (сейчас логистическими~~	~~Заказы~~	~~orderЗаказы~~		~~1~~	~~https://st.yandex-team.~~	Да	Доба
~~dm_complaint~~	~~cdm~~	~~support~~	~~Детальная витрина жалоб~~	~~Саппорт~~	~~supportСапп~~	1	1		Нет	Мепп
~~fct_order_delivery_info~~			~~стными данными по доставке заказа~~	~~Исполнители~~	~~orderИсполн~~	~~1~~	~~1~~		Да (в dm_order)	Готов
order_feedback_predefined_co			лей на заказы (только predefined comments)	Заказы	eats_feedbac	4	4		Нет	
dm_order_finance_metric			позаказные	Финансы	financeФинан	3	3		Не нужен	
~~order_event~~			~~у~~	~~Заказы~~	~~orderЗаказы~~		1		Нет	Нужн
~~region~~				~~Общее~~	~~bigfoodОбще~~	~~1~~	~~1~~		Да	Готов
~~fct_couri~~				~~Исполнители~~	~~Исполнитель~~	1	1		Да	Ods з
~~order_ca~~				~~Заказы~~	~~bigfoodЗаказы~~		~~1~~		Да	Готов
dm_cor			ений в композиции заказа	Заказы	orderЗаказы	2	2		Не нужен	Стро
dm			рина изменений в композиции заказа	Заказы	orderЗаказы	2	2		Нет	
			~~ав заказа~~	~~Заказы~~	~~bigfoodЗаказ~~	~~2~~	~~2~~		Да	Грузи
				Заказы	eats_picker_	4	4		Нет	
			~~ная витрина метрик по партнерам~~	~~Партнеры~~	~~restappПартнеры~~		~~2~~		Не нужен	Исто учить
			~~ча с расчетом курьерского СРО~~	~~Исполнители~~	~~financeИсполнители~~		~~2~~		Нет	Анал
			~~ина метрик по партнерским акциям~~	~~Партнеры~~	~~restappПартнеры~~		~~2~~		Не нужен	Исто учить
			~~ча ресторанов~~	~~Партнеры~~	~~partner_ratingПартнеры~~		2		Есть	Исто заби дату заказ
			~~исполнителя~~	~~Исполнители~~	~~executorИсполнители~~		~~2~~		Да	Готов
			~~чки партнера~~	~~Партнеры~~	~~partnerПартнеры~~		2		Есть	Мепп пере в OD



Скоуп проекта в числах

70+

объектов
(в основном витрин)

из 8+

доменных
областей

Более 130

источников
данных

Потом мы сделали оценку...

Потом мы сделали оценку...

Статус	Оценка на доработку AN	Оценка на доработку DE	Трудозатраты
ТОТАЛ	116,5	105,5	222

Потом мы сделали оценку...

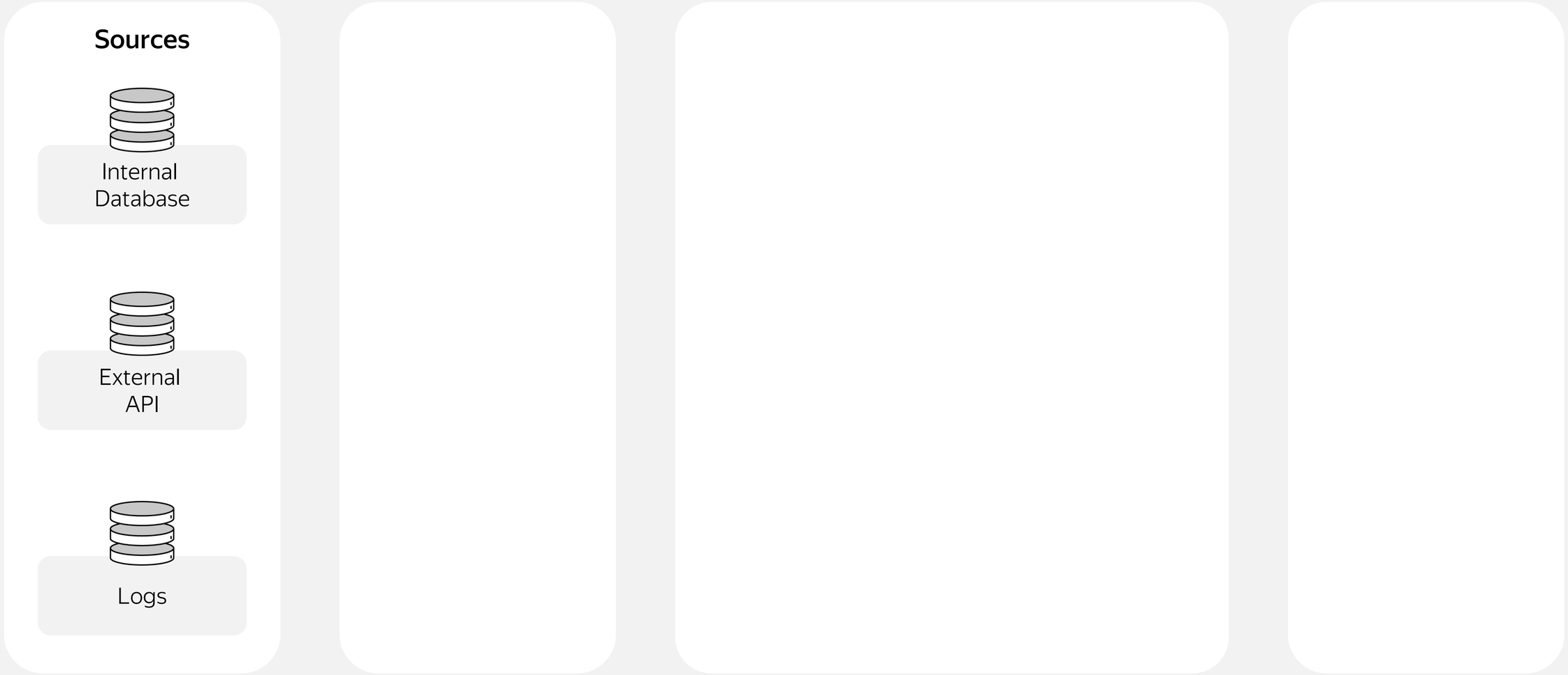
Статус	Оценка на доработку AN	Оценка на доработку DE	Трудозатраты
ТОТАЛ	116,5	105,5	222

Вот это число означало, что мы не успеем

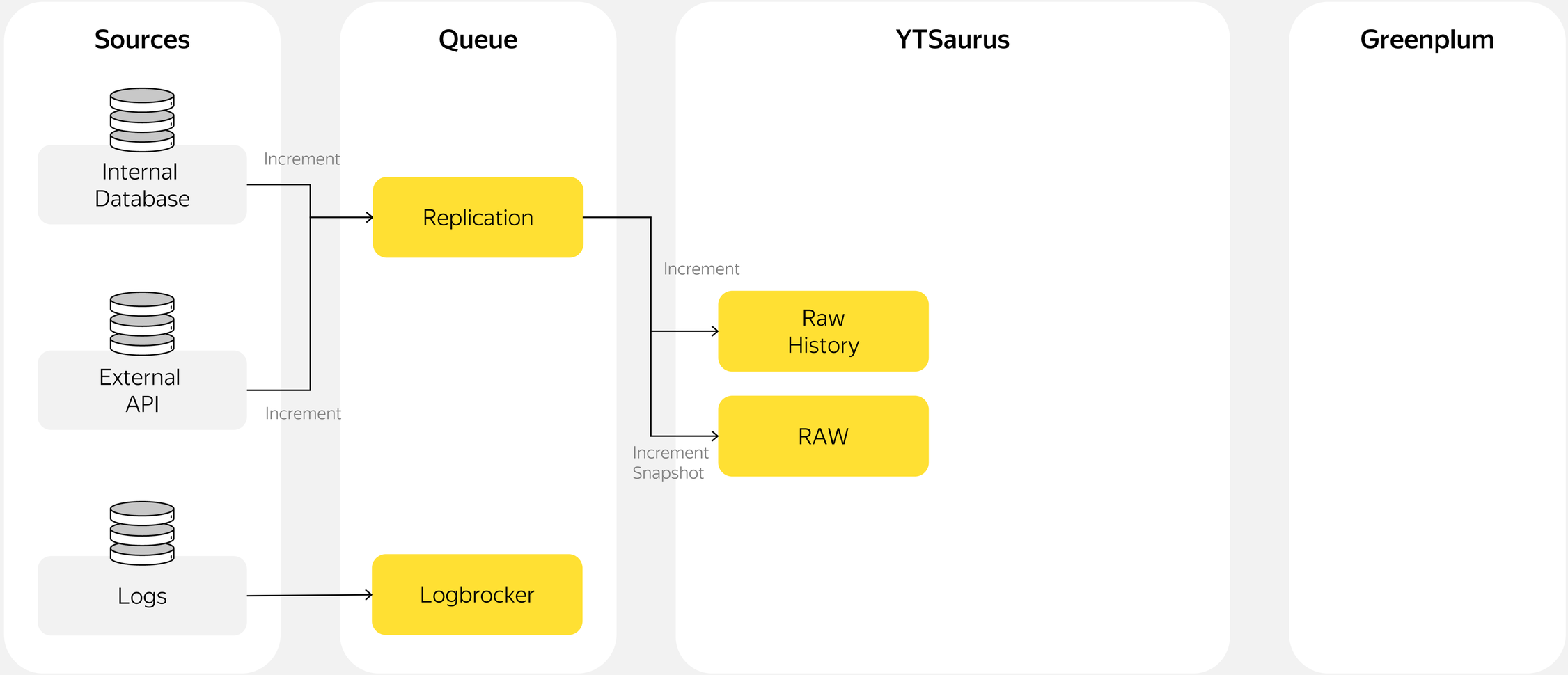
03

Что мы имели на старте

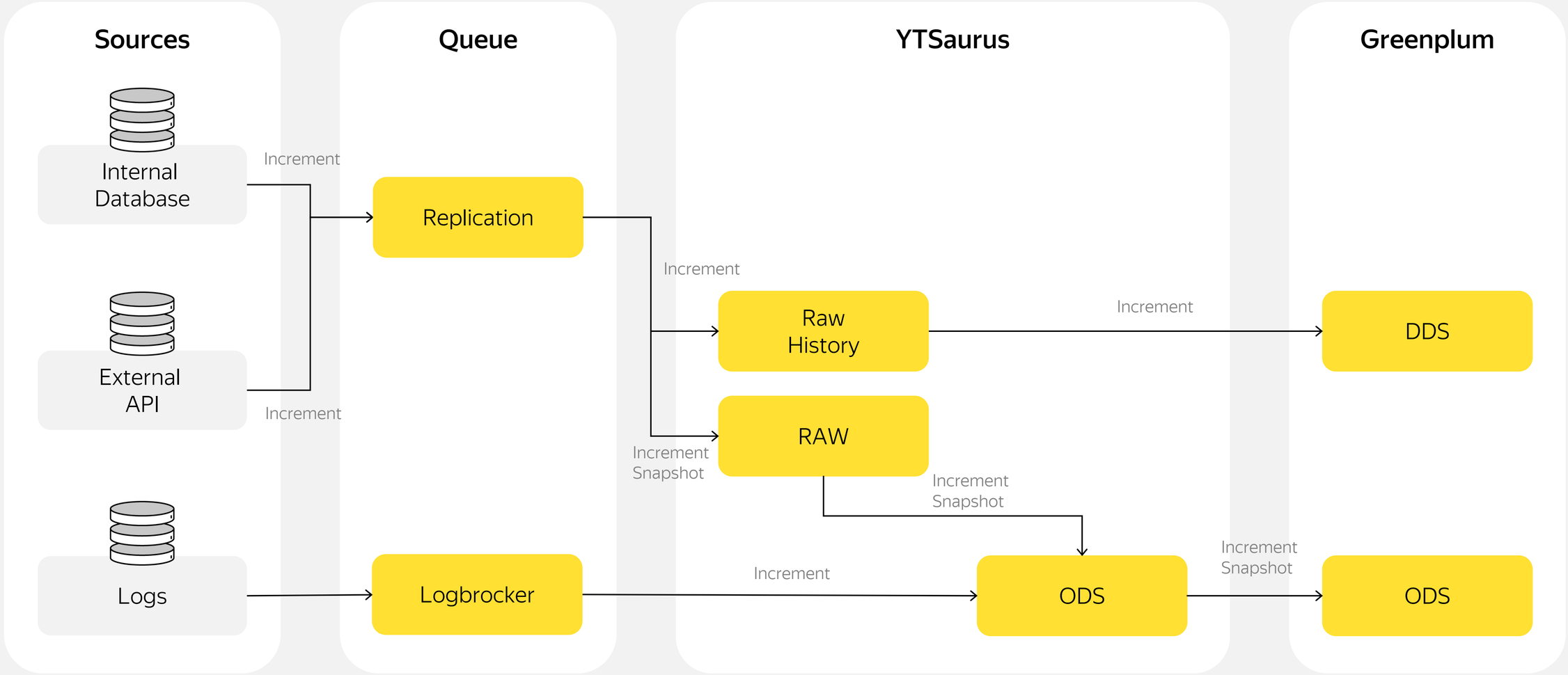
Архитектура и стек в Еде



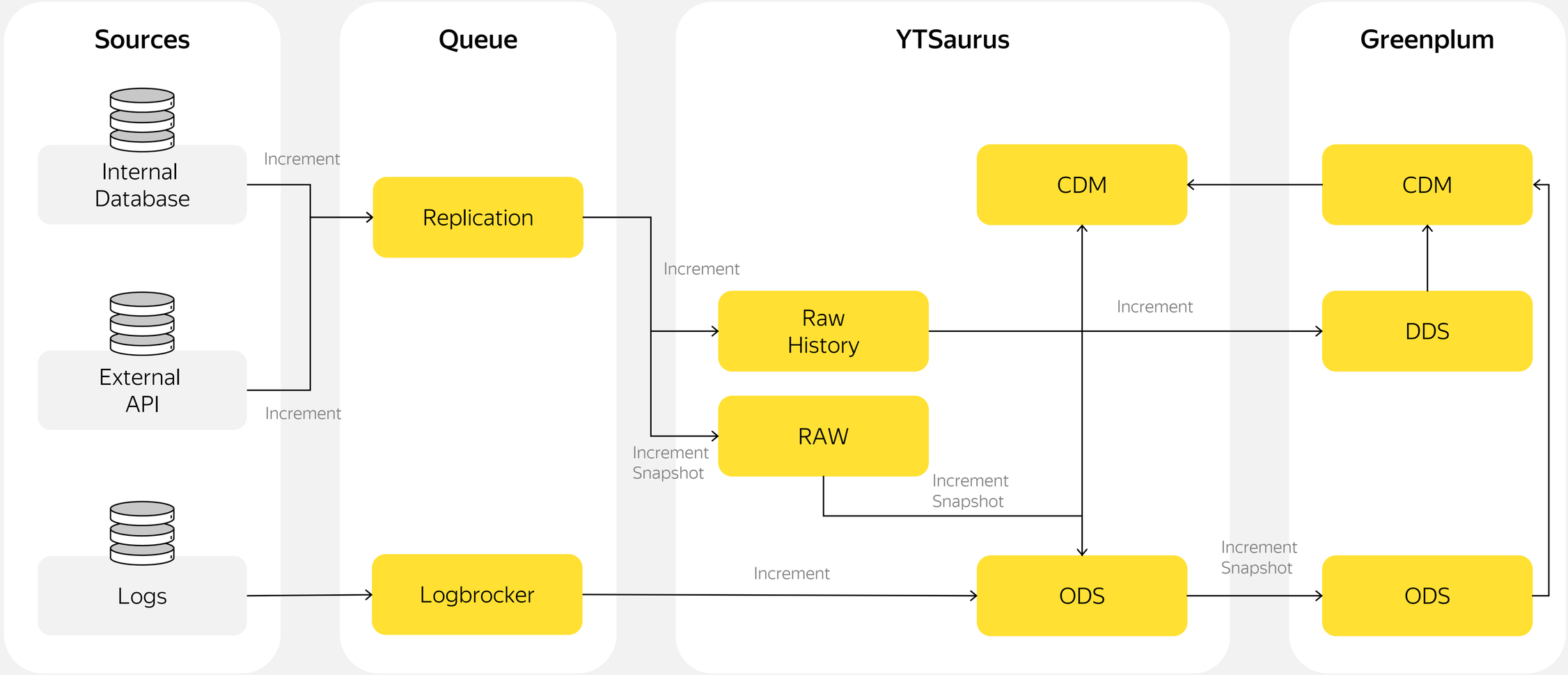
Архитектура и стек в Еде



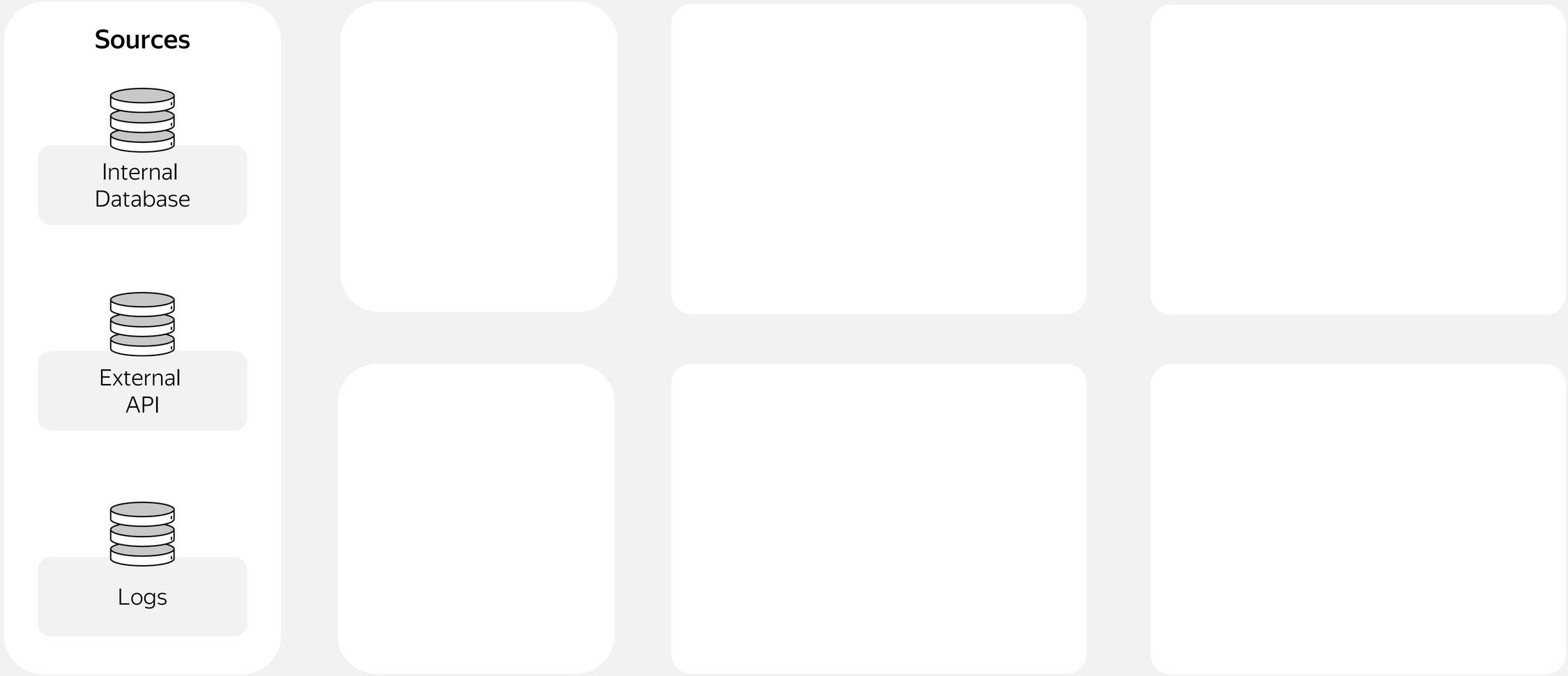
Архитектура и стек в Еде



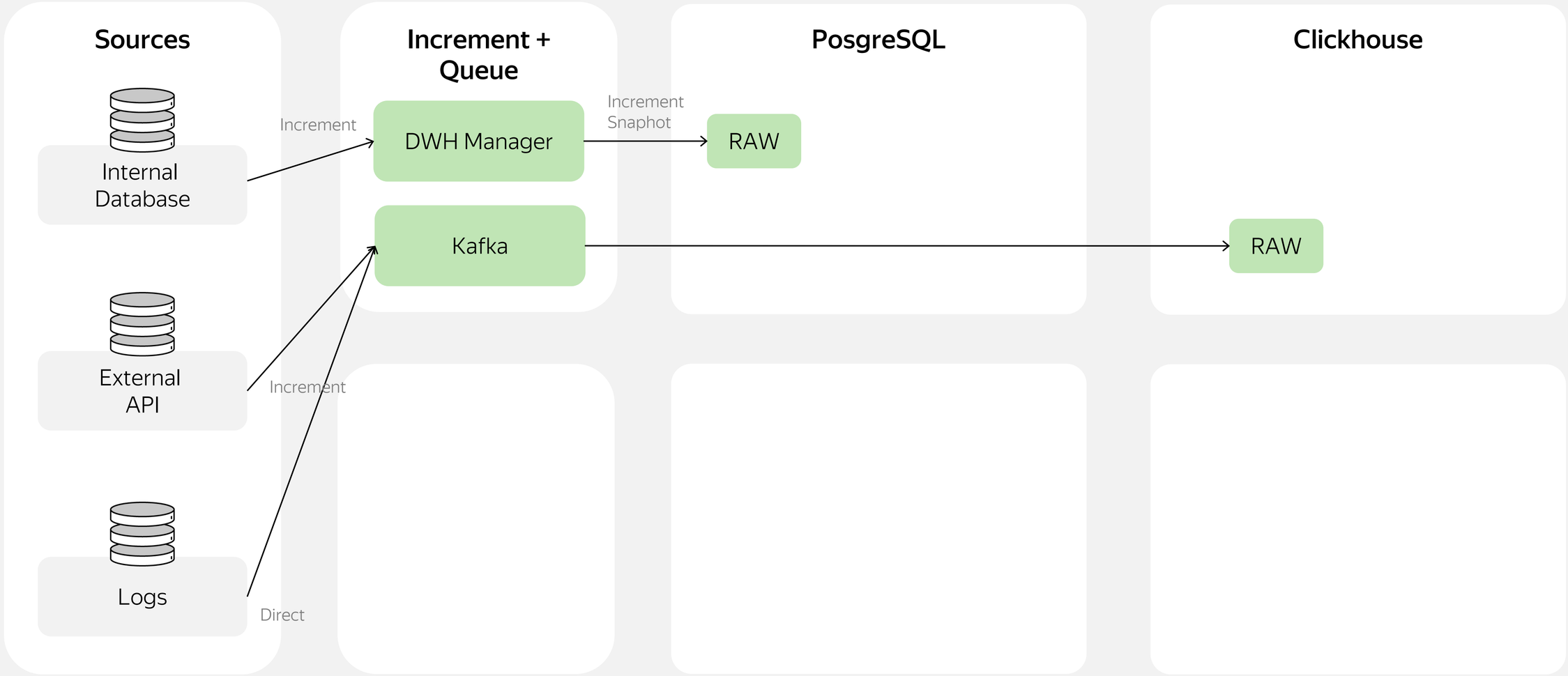
Архитектура и стек в Еде



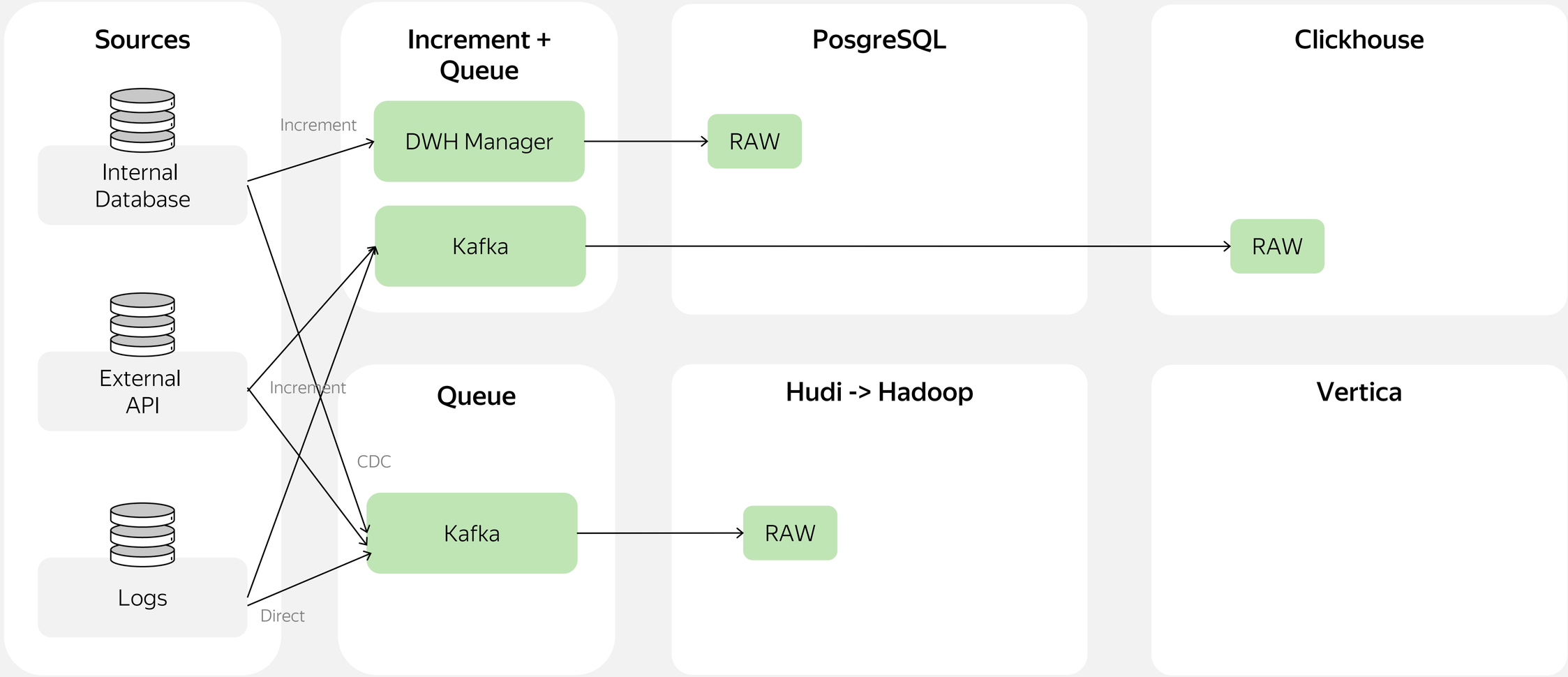
Архитектура и стек в Деливери



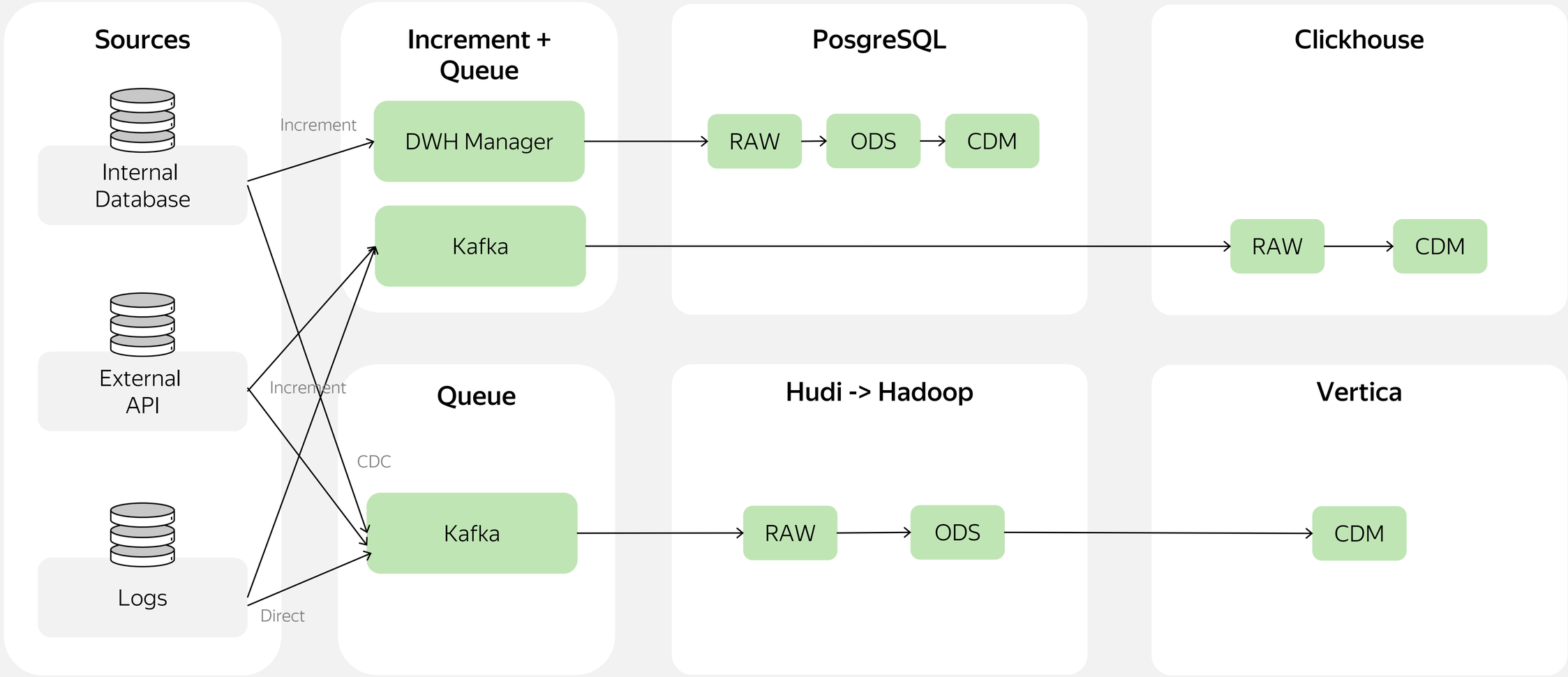
Архитектура и стек в Деливери



Архитектура и стек в Деливери



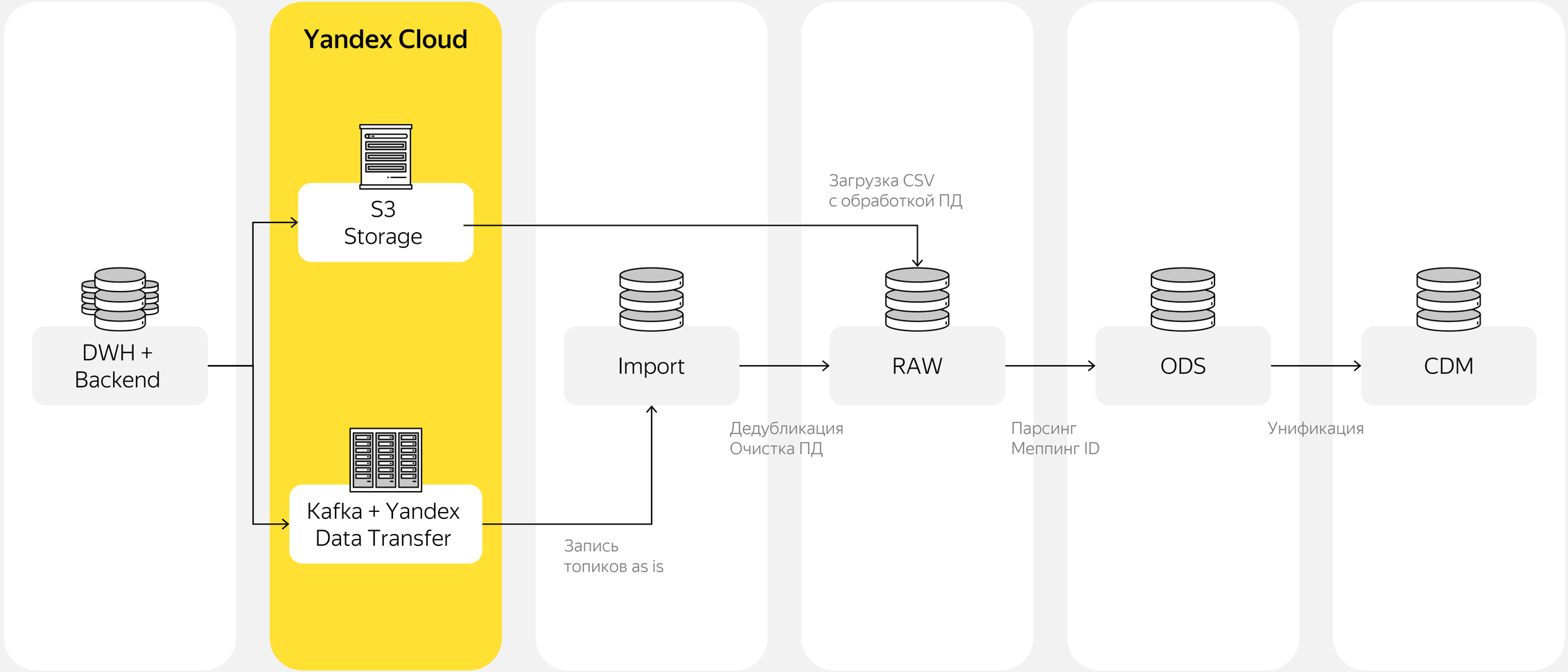
Архитектура и стек в Деливери



06

Техническая реализация

На каком решении мы остановились



Реализация загрузки из Kafka в YT

Дано

1. ~130 таблиц в Kafka
2. Очень мало времени

Решение

1. Запись в импорт через настройку DT → Все таблицы в импорте стандартизированы
2. Много загрузок → Шаблонизировали загрузку в RAW через Jinja
3. На уровне загрузки в RAW избавляемся от ПД, это тоже добавлено в шаблон

Реализация загрузки из Kafka в YT

service_query.sql.jinja2

```
{% if is_insert %}
INSERT INTO {target_path} WITH TRUNCATE
{% endif %}
    SELECT
        {%- for _, alias in primary_keys %}
        {{ alias }},
        {%- endfor %}
        MAX_BY(action, seq_no)                      AS action,
        MAX_BY(utc_insert_to_kafka_dttm, seq_no)    AS utc_insert_to_kafka_dttm,
        MAX_BY(doc, seq_no)                         AS doc
    FROM $import_table
    GROUP BY
        {%- for field, alias in primary_keys %}
        {{ field }} AS {{ alias }}{% if not loop.last %},{% else %};{% endif %}
        {%- endfor %}
```

```
# TODO: параметры для настройки таски
task_name = 'raw__delivery_club__bp__schedule__v1'
table_path = '//home/eda-dwh/import/delivery_club/eda/without_pd/bp.sched
primary_keys = (
    ('Yson::ConvertToInt64(doc.id)', 'id'),
)
query_draft = 'service_query.sql.jinja2'
target_table = RawBpScheduleV1

# TODO: неизменяемая часть кода
path = Path(__file__).parent.parent.joinpath('resources')

ENV = Environment(loader=FileSystemLoader(str(path.absolute())))

QUERY = ENV.get_template(query_draft).render(
    table_path=table_path,
    primary_keys=primary_keys,
    is_insert=True,
)

yql_query = yql_transform.YqlTaskQuery.from_string(
    QUERY,
    pool=args.use_arg('pool'),
).add_params(
    start_dttm=args.use_period_arg().start.format_datetime(),
    end_dttm=args.use_period_arg().end.format_datetime(),
).attach_file(
    YqlAttachment(**yql_sql),
)
```

Реализация загрузки с S3

Дано

Пришёл DC к Еде свататься, да не с пустыми руками. Полный S3 подарков принёс! Еда об S3 понятие имела, да только умела туда класть. А как забрать оттуда что-то – непонятно. Да ещё и подарки-то все разные! И с ПД таблицы и справочники, хорошо хоть все в CSV. 13:51

 1

Решение

1. Разрабатываем свой загрузчик, который умеет читать произвольно сжатые CSV файлы и писать в партицированные таблицы
2. Для обработки ПД написали свою трансформацию

Итого, что помогло нам прийти к успеху:

1. Засетапили команду и поддерживали общение
2. Рано сделали оценку, что помогло понять, где надо усилиться
3. Переиспользовали уже имеющиеся технологии
4. Итерационно догружали данные и по готовности дорабатывали витрины
5. Работали параллельно в стольких доменах, на сколько хватало рук

09

Грабли и выводы

На какие грабли мы наступили

1

Скорость поставки
(данные ДК доезжали только к обеду)

Какой вывод

- Разнесли загрузку данных Еды и ДК в разные пайплайны, сделали независимую запись по ключу в витрины

На какие грабли мы наступили

2

Осознали неидеальность пайплайна отчетов

Какой вывод

- Выявили критичные точки, которые можно улучшить

На какие грабли мы наступили

3

Очень медленная обработка ПД не позволяла перегружать большие объемы данных

Какой вывод

- Везде, где это возможно, разносили в разные загрузки атрибуты ПД и все остальное
- Для атрибутов ПД придумали обходной путь, использующий уже загруженные данные

На какие грабли мы наступили

3

Пересекающиеся ID

Какой вывод

- Собрали отдельные таблицы меппинга ID ДК
- Работали совместно с бекендом

На какие грабли мы наступили

5

Очень долго считали и пересчитывали историю

6

Бекенд менялся в режиме реального времени – мы не всегда успевали подстроиться

Яндекс Еда

Спасибо за внимание

Вопросы?