



Как сделать каталог данных быстро

В условиях нескольких зрелых хранилищ

Наталья Журавлёва
Руководитель группы
поддержки DWH



Обо мне

- Работаю в Ozon **больше 5 лет**
- Являюсь **лидом команды аналитиков в отделе DG**
- Застала переход от **MsSQL** к **колоночной БД**, от **DataLake** — к **Anchor DWH**
- Работала по «**обе стороны**» **данных**



Тезисы

На Data Governance
нет времени

01

Логическая модель —
фундамент каталога данных

02

Нужно думать о **внедрении
нового подхода** разработки
хранилища

03

Что можно показать
в каталоге данных, помимо
описания данных

04



Ресурсов на Data Catalog нет

Как можно быстрее

Менеджер просит
данные для принятия
решения

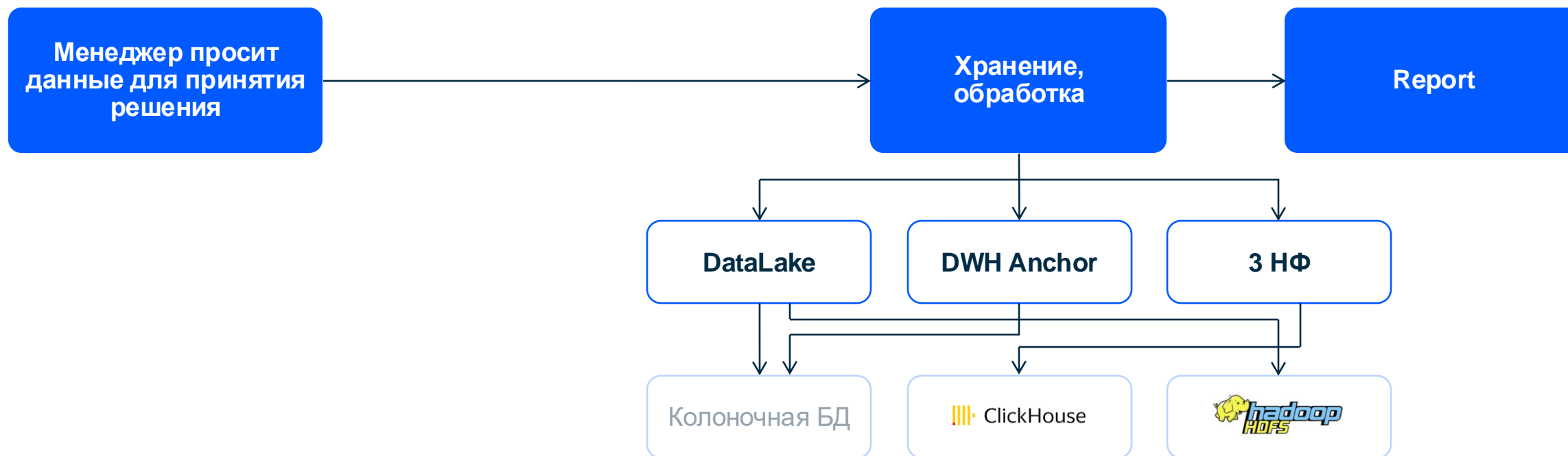
Магия



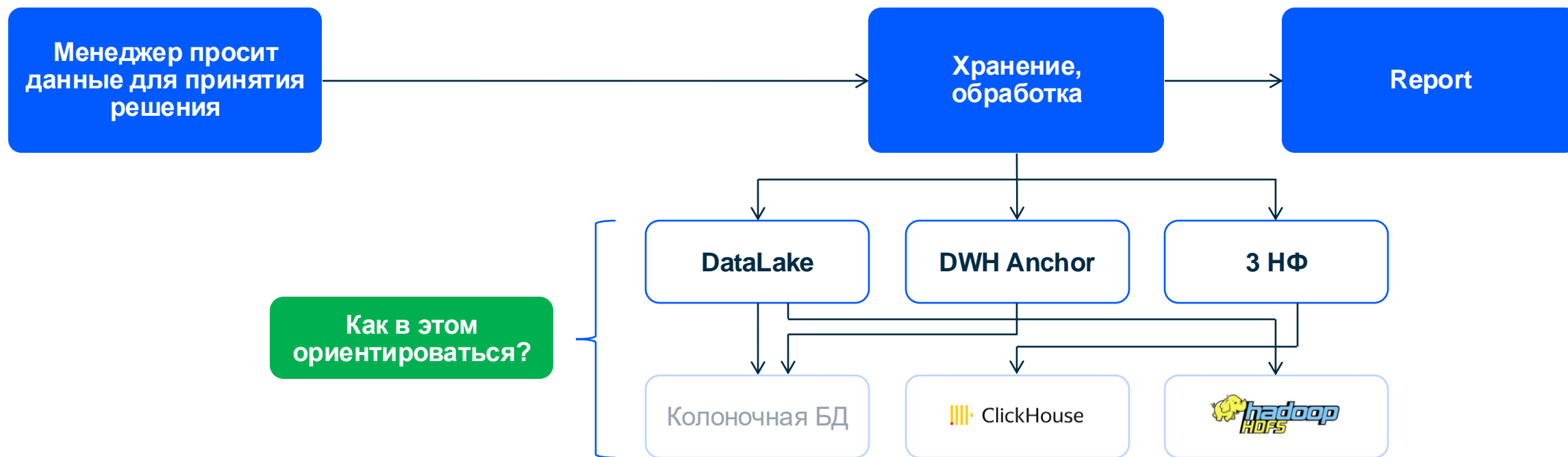
Report



Ресурсов на Data Catalog нет



Ресурсов на Data Catalog нет

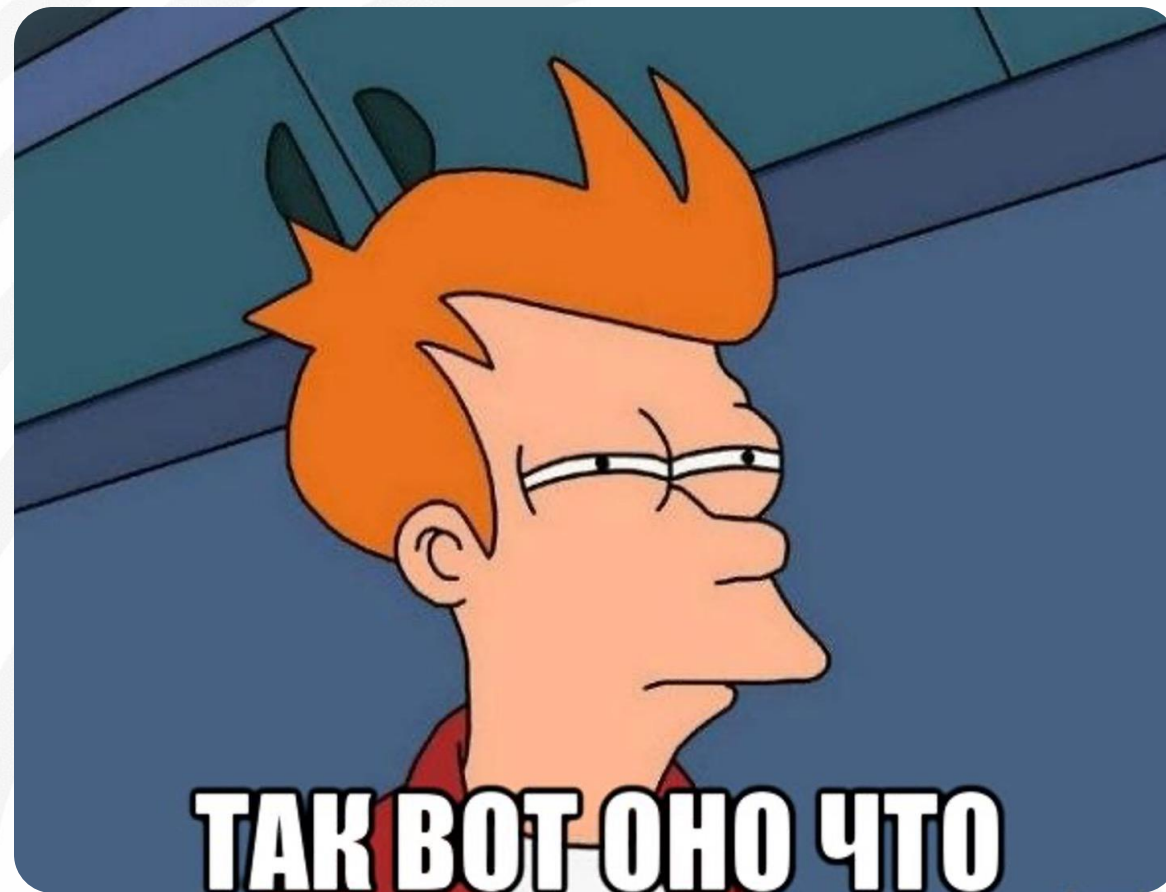


Прошло N времени

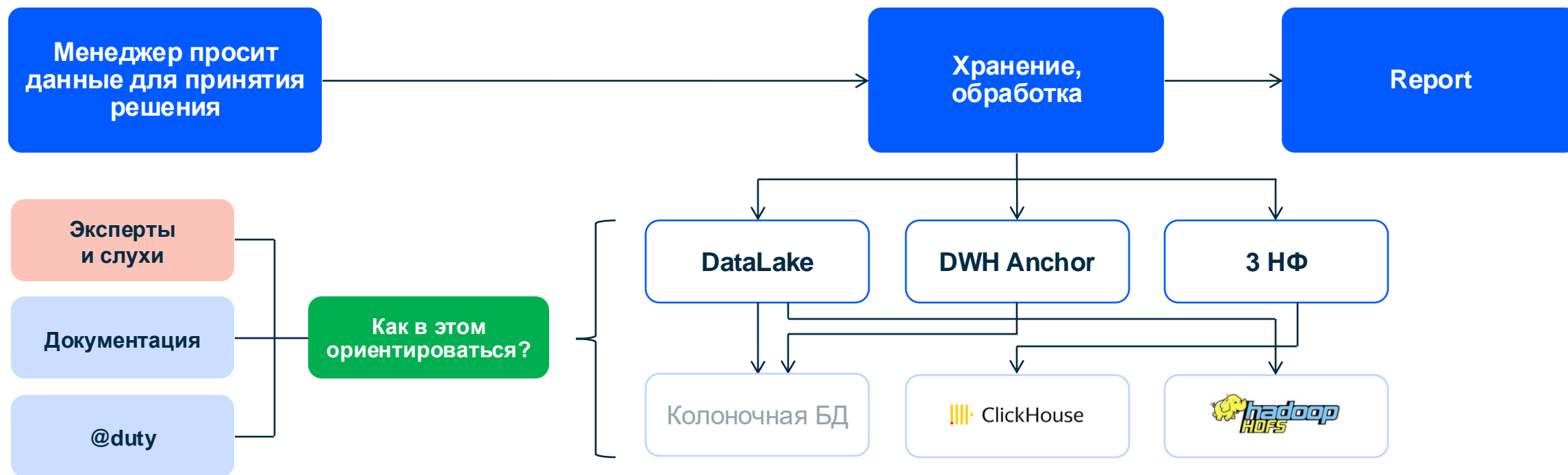
→ Данные в отчётах разные

→ Завалили вопросами

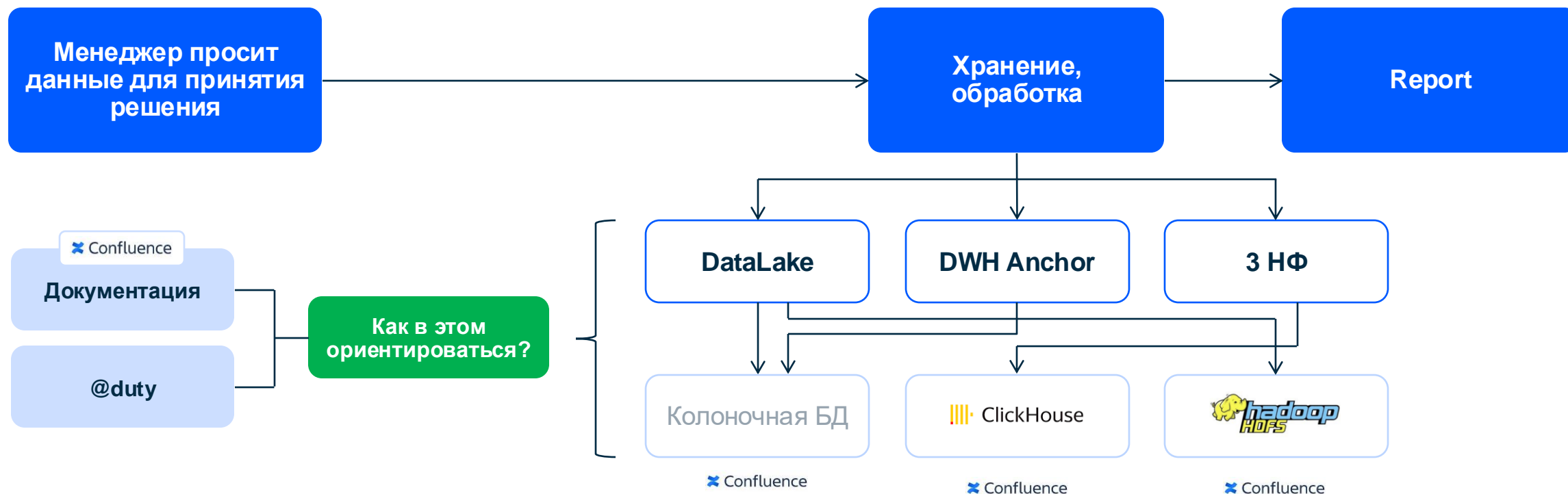
→ Сложность в разработке
НОВОГО КОНТЕНТА



Очень упрощённый flow данных



Очень упрощённый flow данных



Почему нужно идти дальше

- Сложно поддерживать

- Недостаточно информации


+ Есть база знаний на будущее



К чему пришли за полгода

Welcome back, Natalia Zhuravleva.

Analytics Govern Ingestion



View all Search Tables, Dashboards, People, & more... K

Try searching for Explore all >

[MPBX] Main Metrics [RDS-Fresh] Статистика ключевых показате... Courier on map [copy]

Pinned

Announcement
Данные Tracker уже в DataHub

Link
Ваши предложения по DataHub →

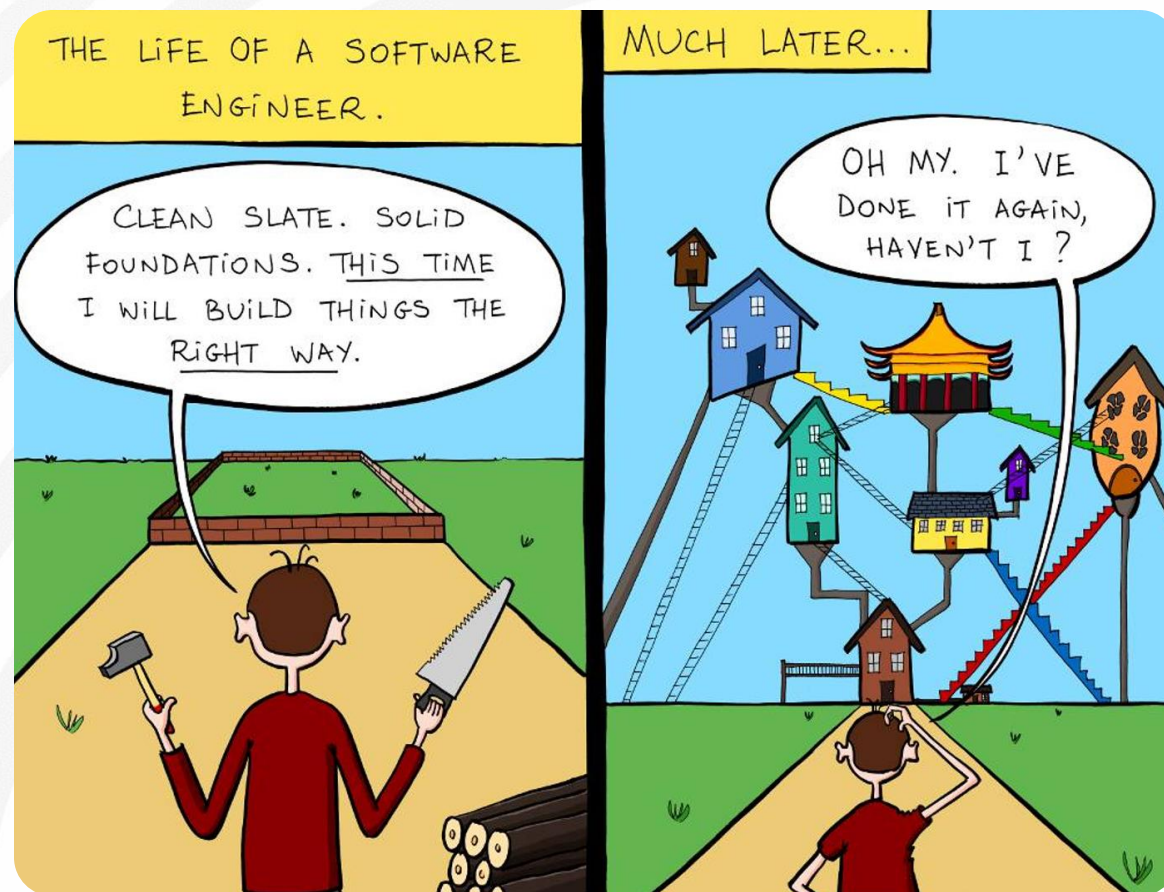
Link
DataHub FAQ →

Link
Инструкция SuperSet в DataHub →

Link
Маттермоуст Проверки DQ в →

Скажи «нет»

- Превращение каталога данных в болото
- Ручное ведение каталога
- Не интегрирован в процесс разработки хранилища



Каталог без модели = деньги на ветер



Каталог данных
без модели данных

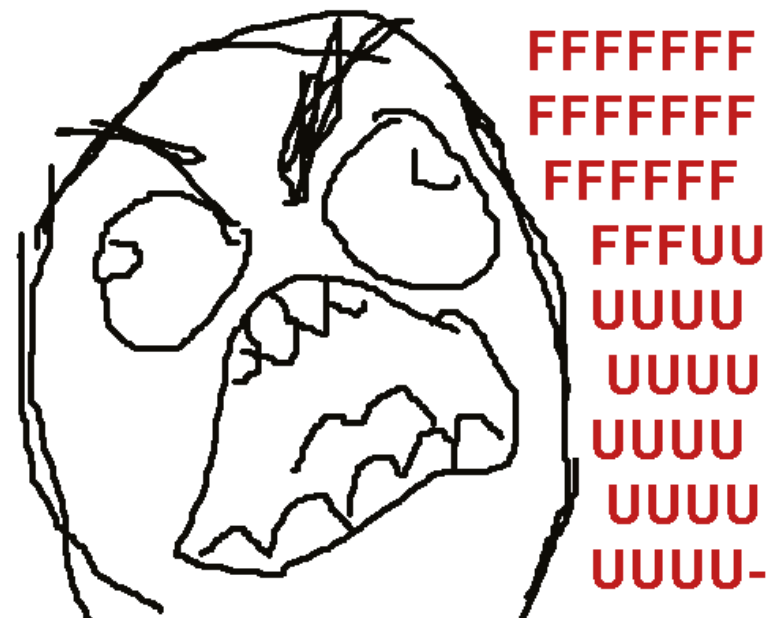


Каталог данных
с моделью данных

С точки зрения пользователя



Ввожу в поиск слово «заказ»



Как по книжке



	Концептуальная модель	Логическая модель	Физическая модель
Названия	+	+	
Связи	+	+	
Атрибутивный состав		+	
Первичный ключ		+	+
Вторичный ключ		+	+
Имя таблицы			+
Имя столбца			+
Тип данных столбца			+

Наш подход



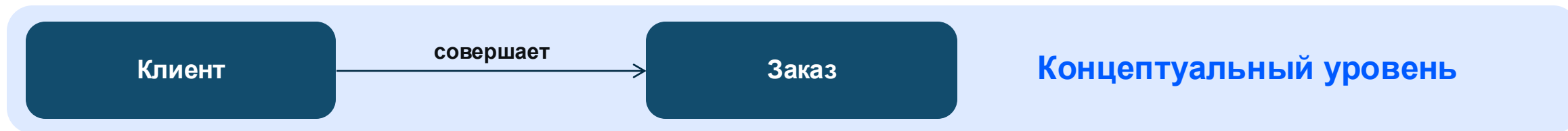
	Концептуальная модель	Логическая модель	Физическая модель
Названия	+	+	
Связи	+	+	
Атрибутивный состав		+	
Первичный ключ		+	+
Вторичный ключ		+	+
Имя таблицы			+
Имя столбца			+
Тип данных столбца			+

Наш подход

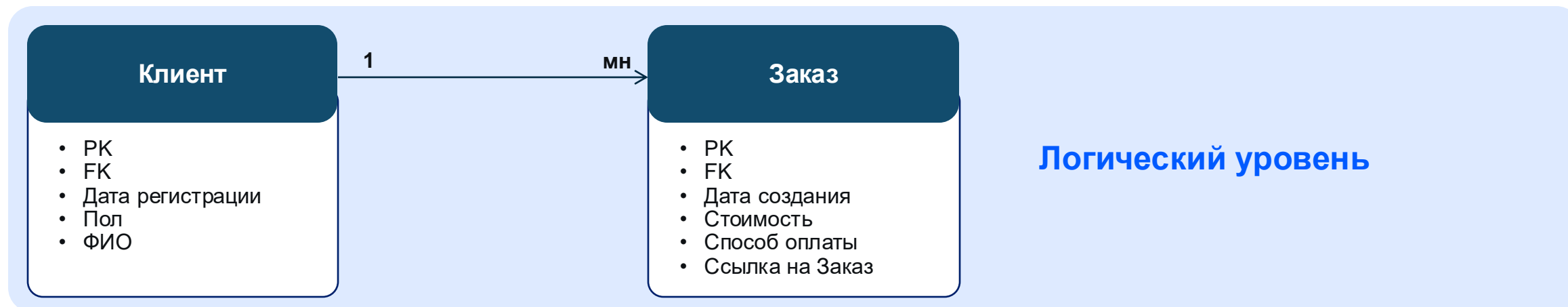
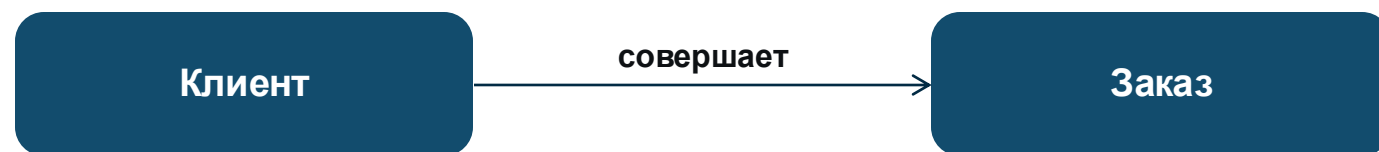


	Концептуальная модель	Логическая модель	Физическая модель
Названия	+	+	
Связи	+	+	
Атрибутивный состав		+	
Первичный ключ		+	+
Вторичный ключ		+	+
Имя таблицы			+
Имя столбца			+
Тип данных столбца			+

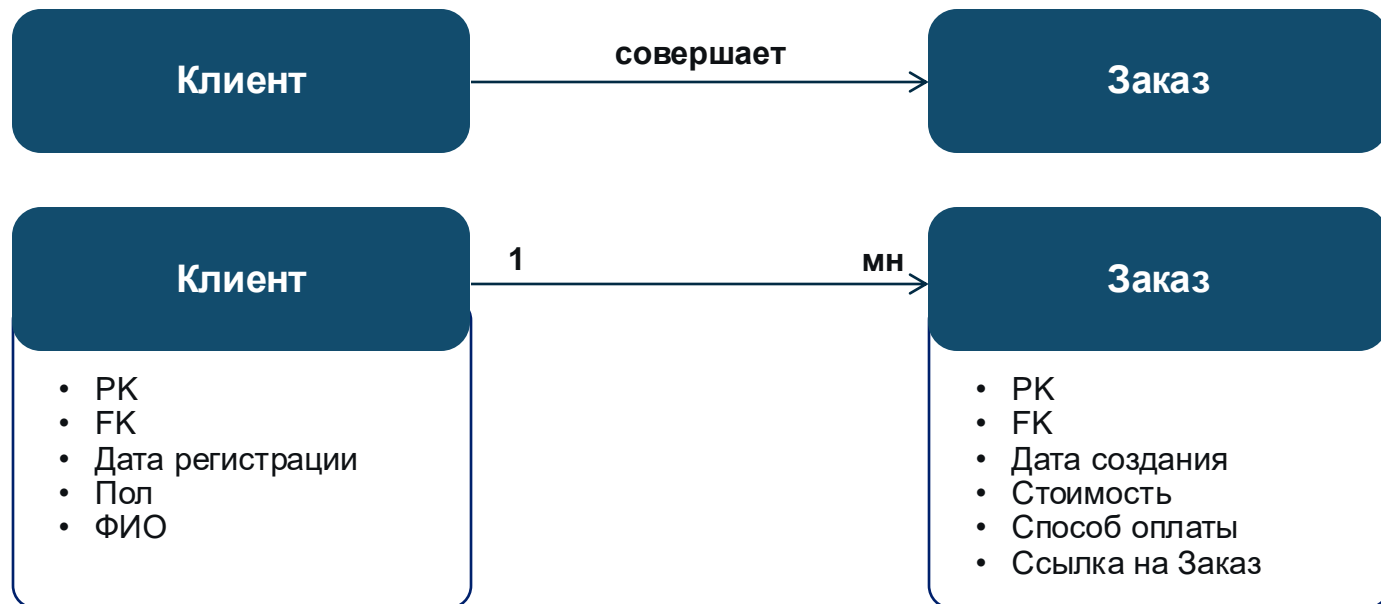
Разница модулей на примере



Разница модулей на примере



Разница модулей на примере



dwh_data.Hub_Client

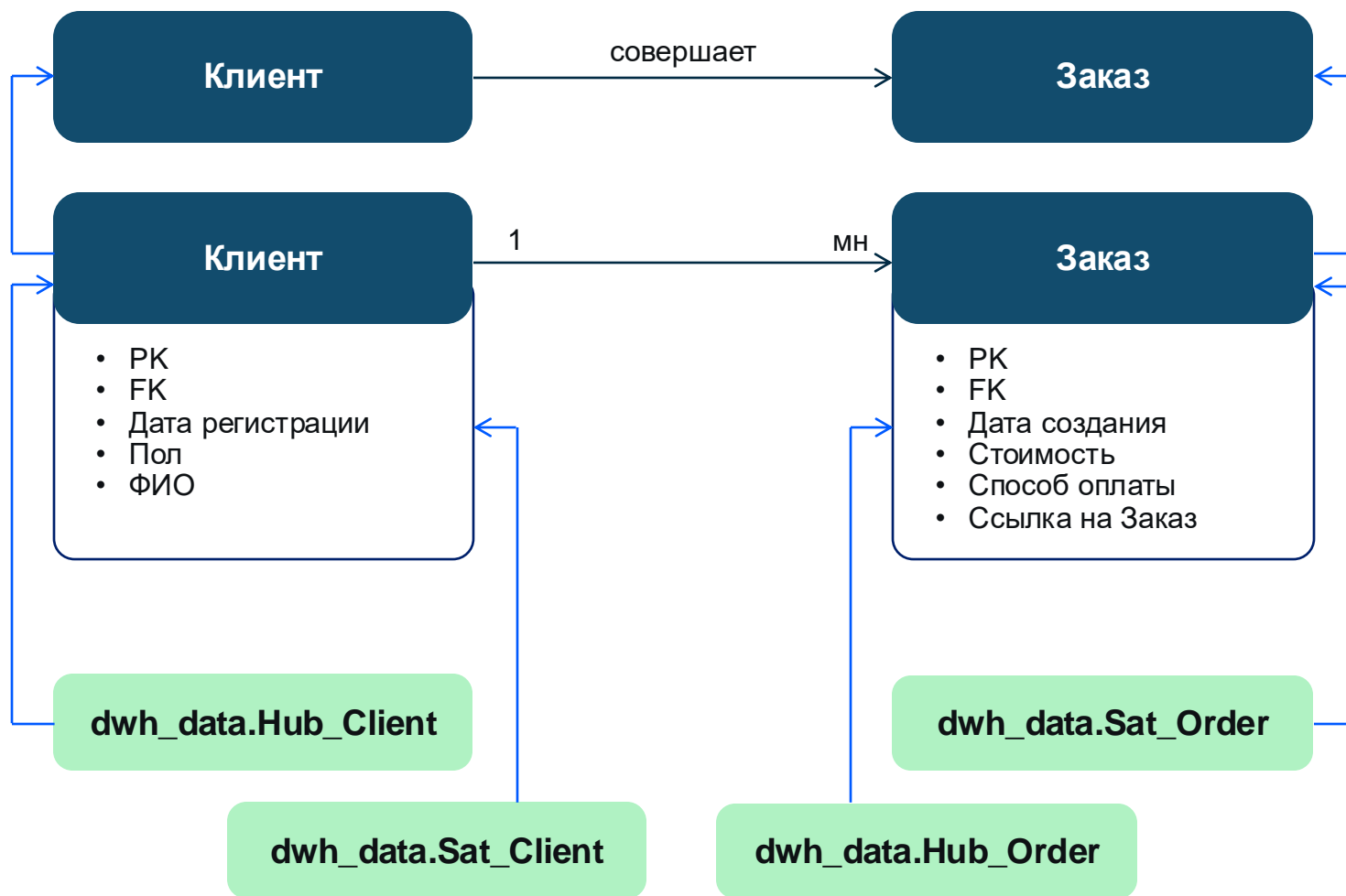
dwh_data.Sat_Order

dwh_data.Sat_Client

dwh_data.Hub_Order

Физический уровень

Связь между уровнями



Неправильный,
но жизненный
подход

В чём сложность? ⚡

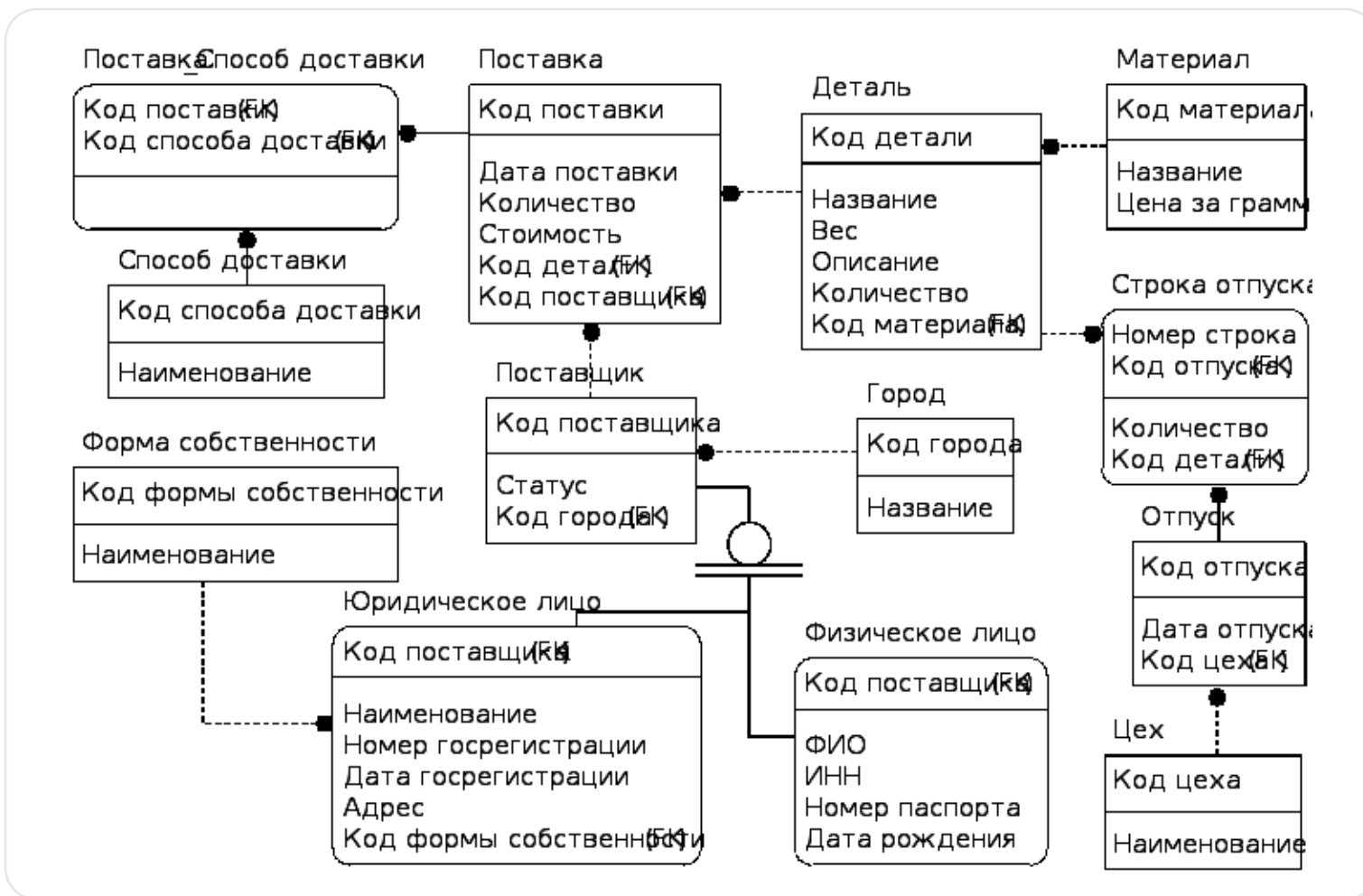
→ Много времени

→ Экспертные кадры

→ Творческий процесс

→ Как её хранить?

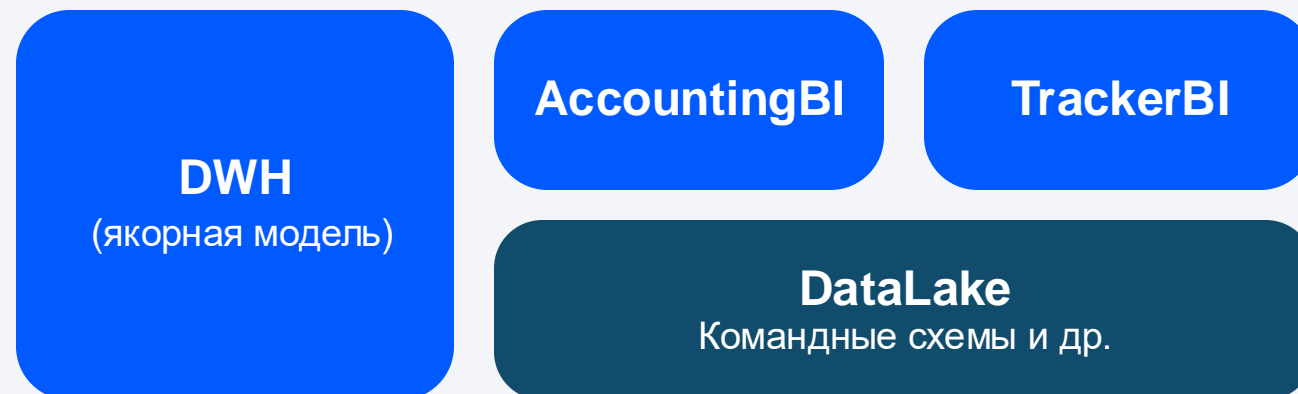
→ Как её изменять?



Как мы сделали модель

На входе:

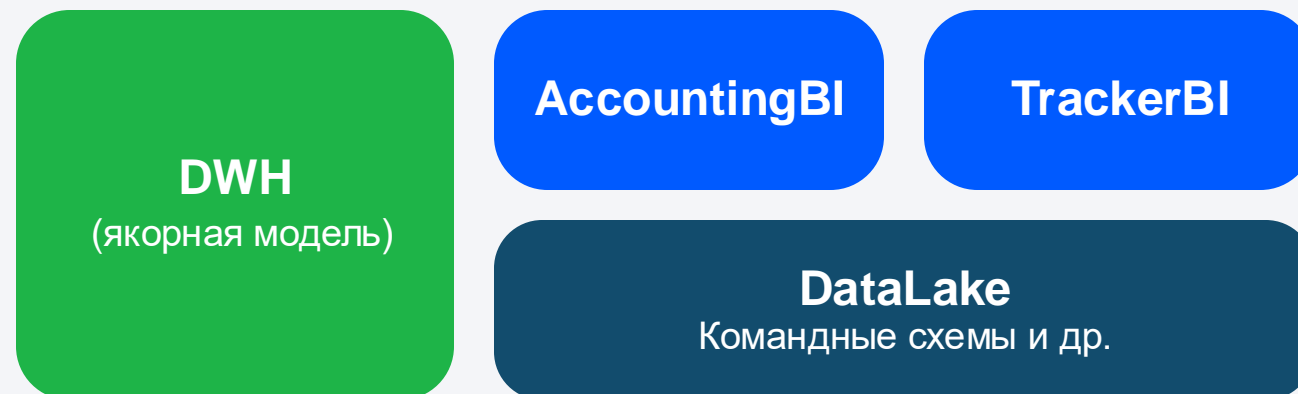
- Три зрелых хранилища
- Петабайты данных
- Несколько баз данных
- **750** якорей, **2500** атрибутов, **1000+** связей и др.



Начали с DWH

Распарсили названия таблиц в детальном слое

- Якоря = сущности
- Атрибуты = атрибуты сущностей
- Связи = ссылки на другие сущности



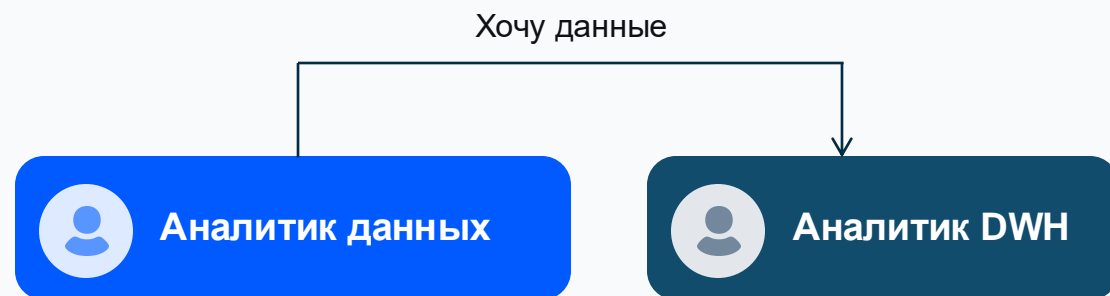
Получили модель

```
AbVariant:  
  Description: вариант A/B теста  
  Concept: AB  
  Attributes:  
    Name:  
      Description: человеко-читаемое название варианта A/B теста  
    IsControl:  
      Description: признак того контрольной группы  
  AbVariantId:  
    Description: ключ AbVariant  
    PK: true
```

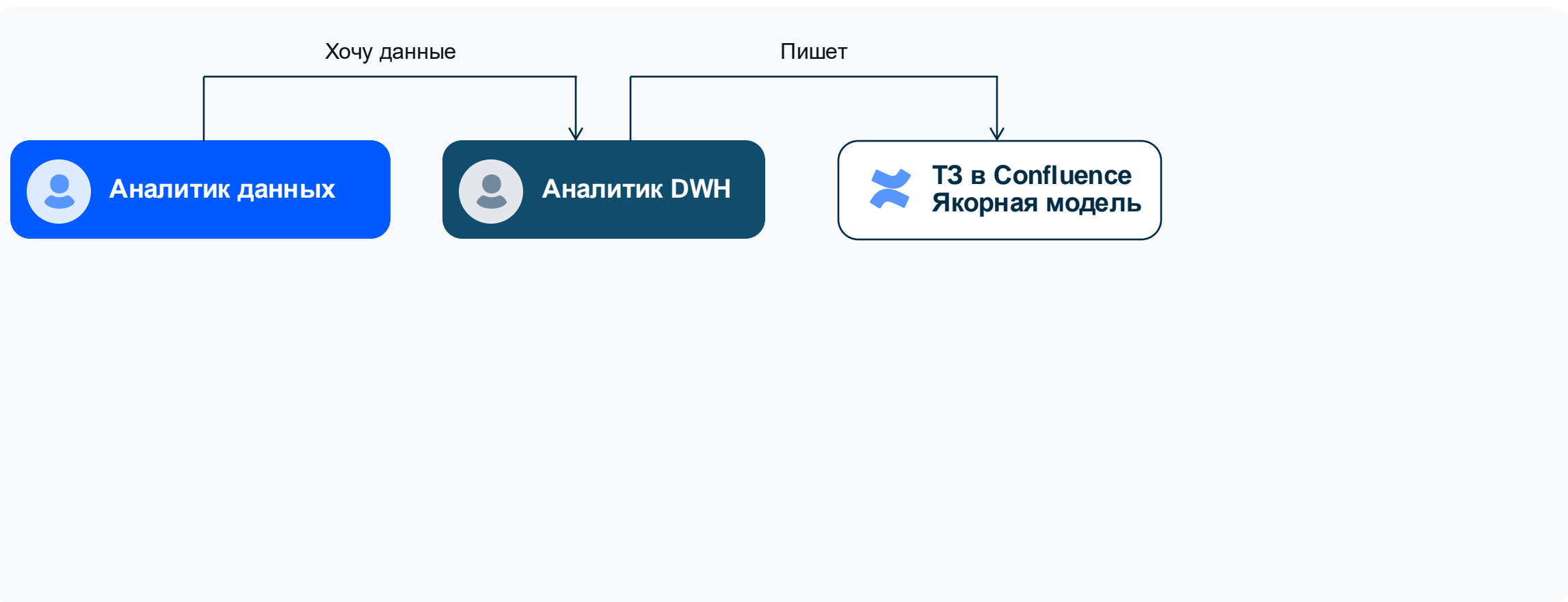
Место хранения — [yaml-файл](#)

Изменение yaml [через MR](#)

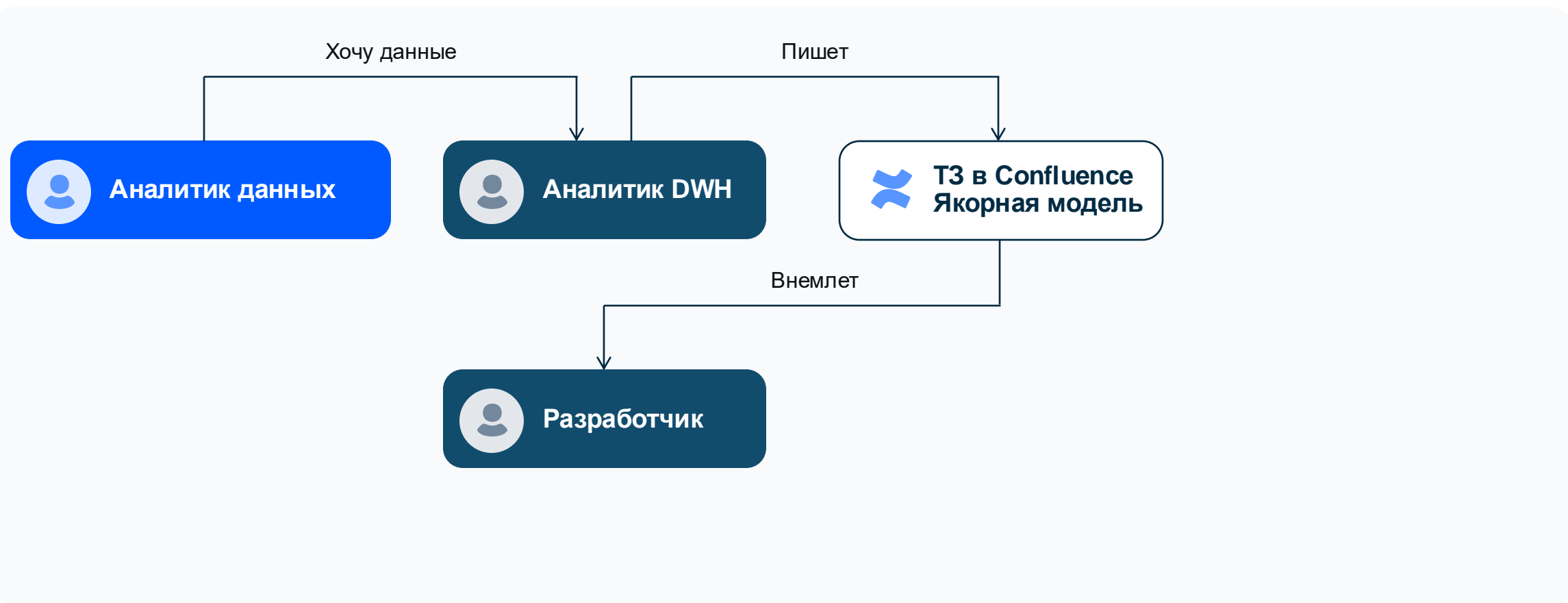
Как было «ДО»



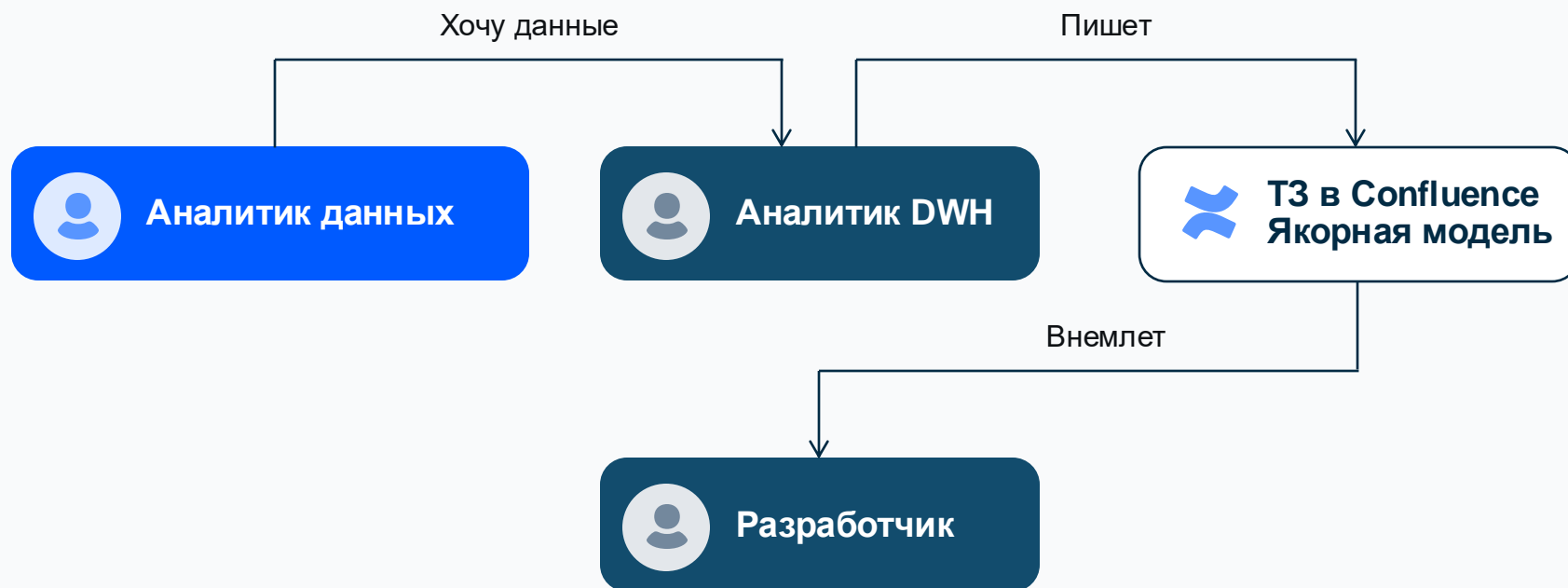
Как было «ДО»



Как было «до»



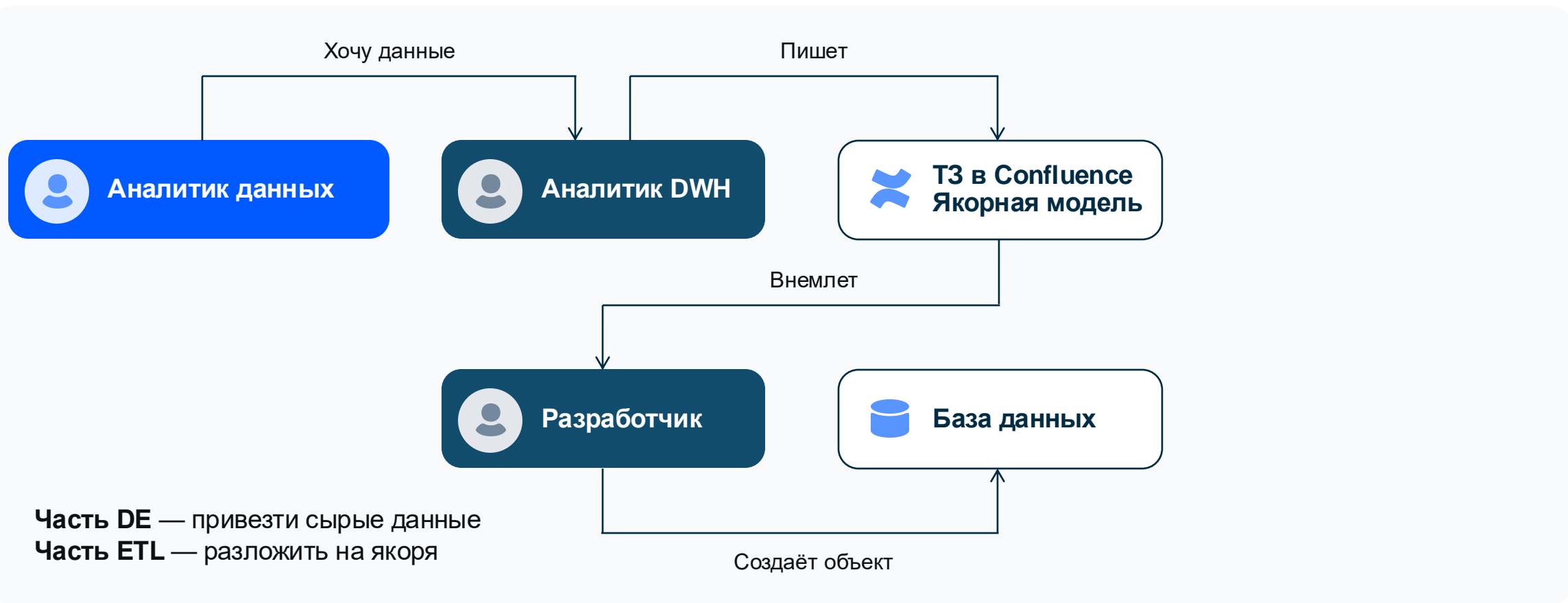
Как было «до»



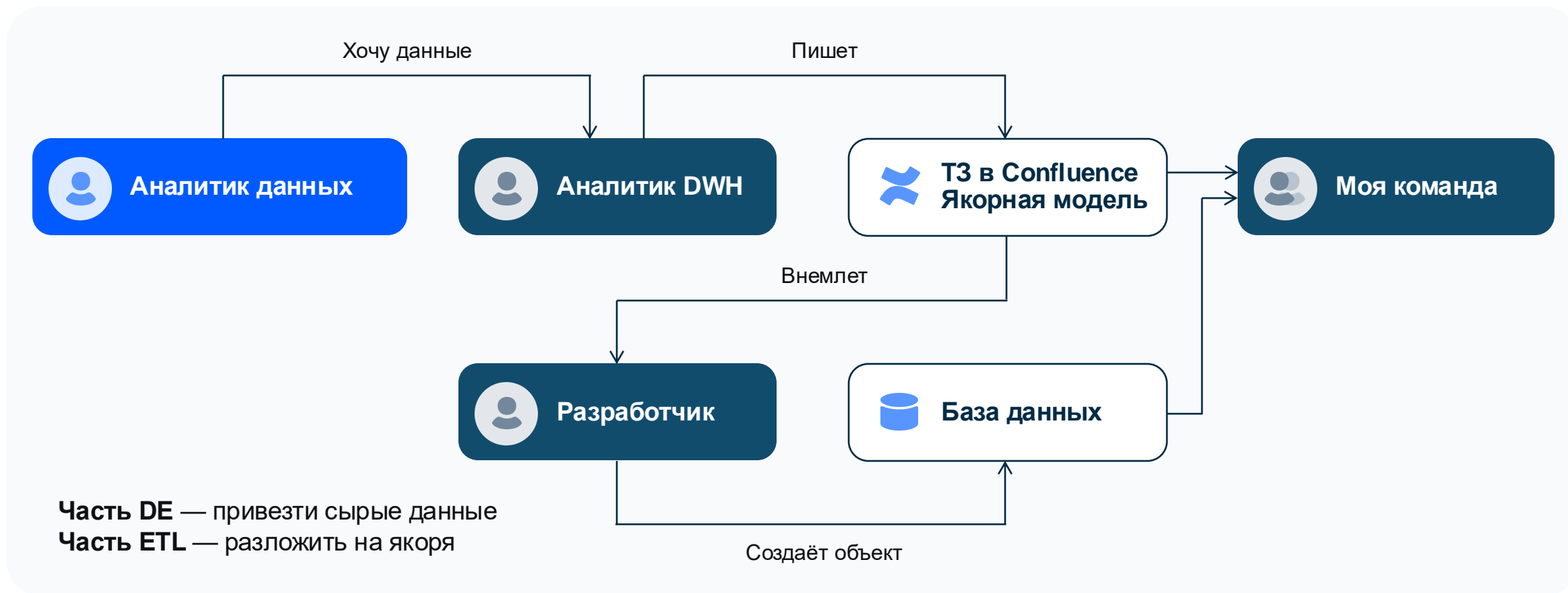
Часть DE — привезти сырые данные

Часть ETL — разложить на якоря

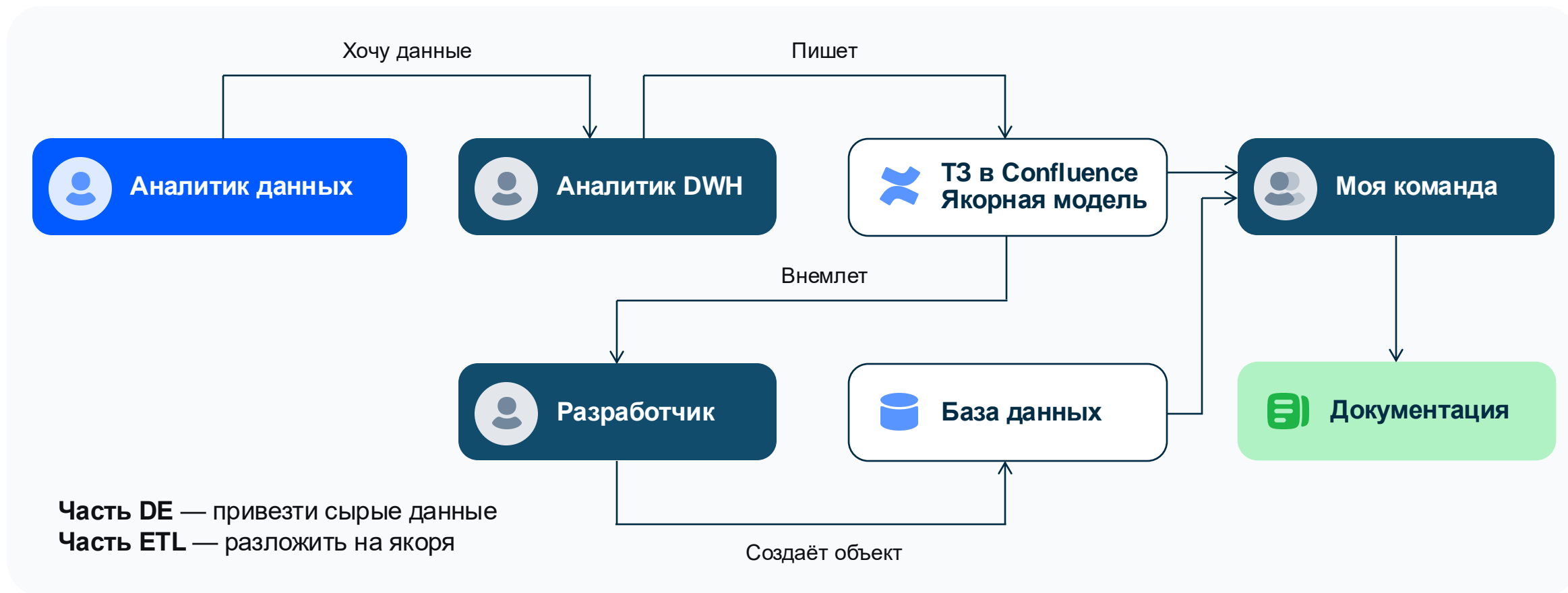
Как было «ДО»



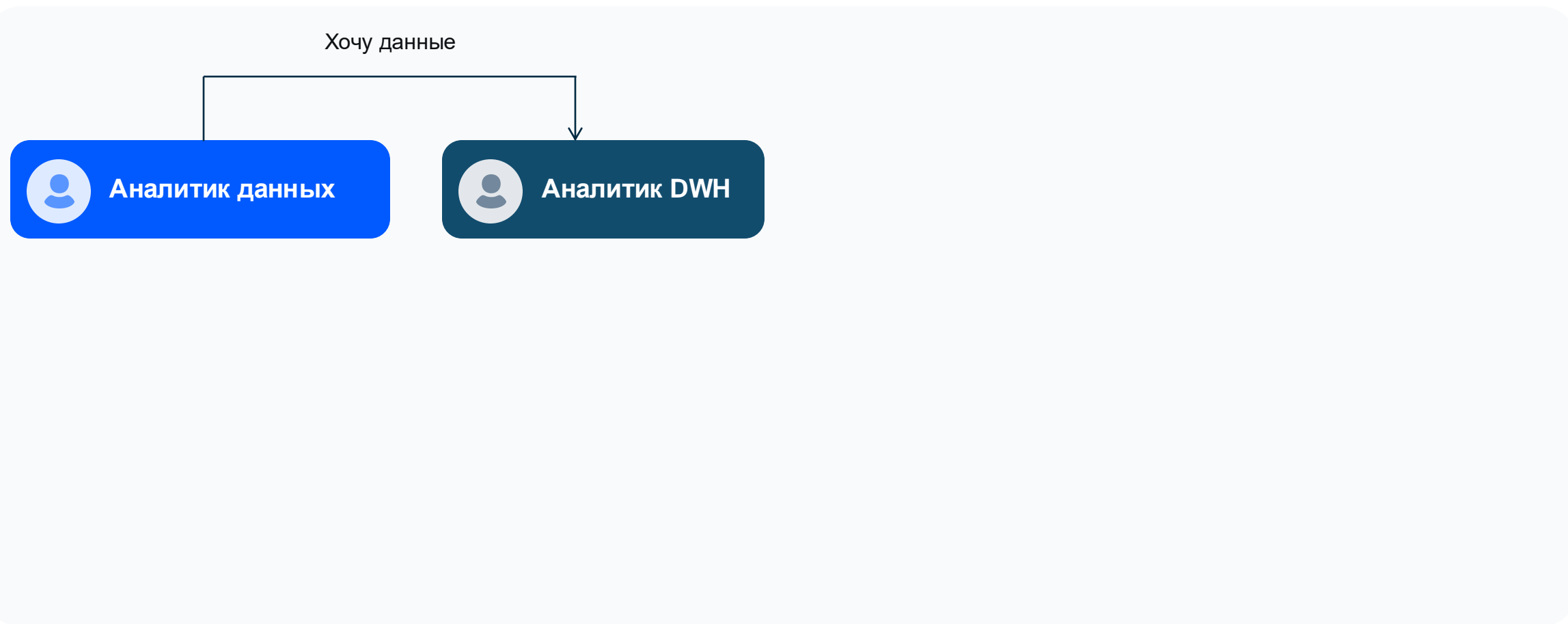
Как было «до»



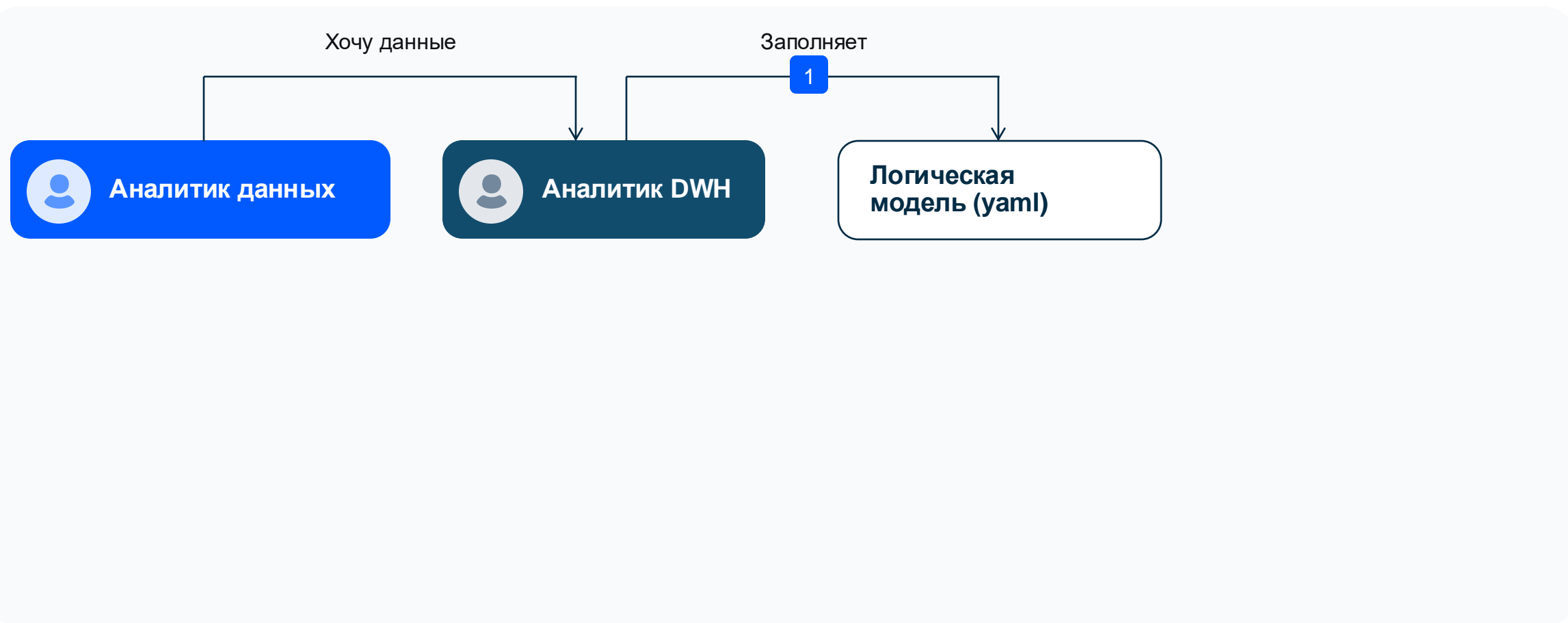
Как было «ДО»



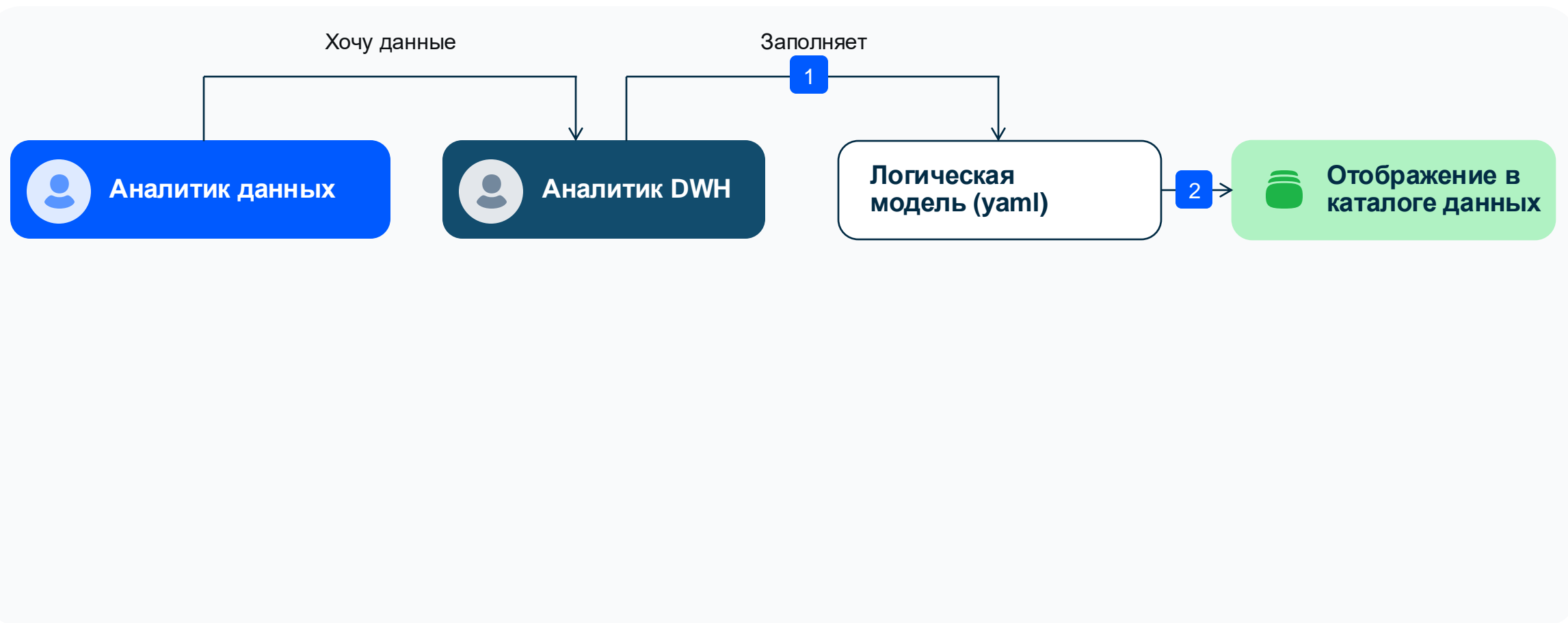
Flow работы с лог. моделью



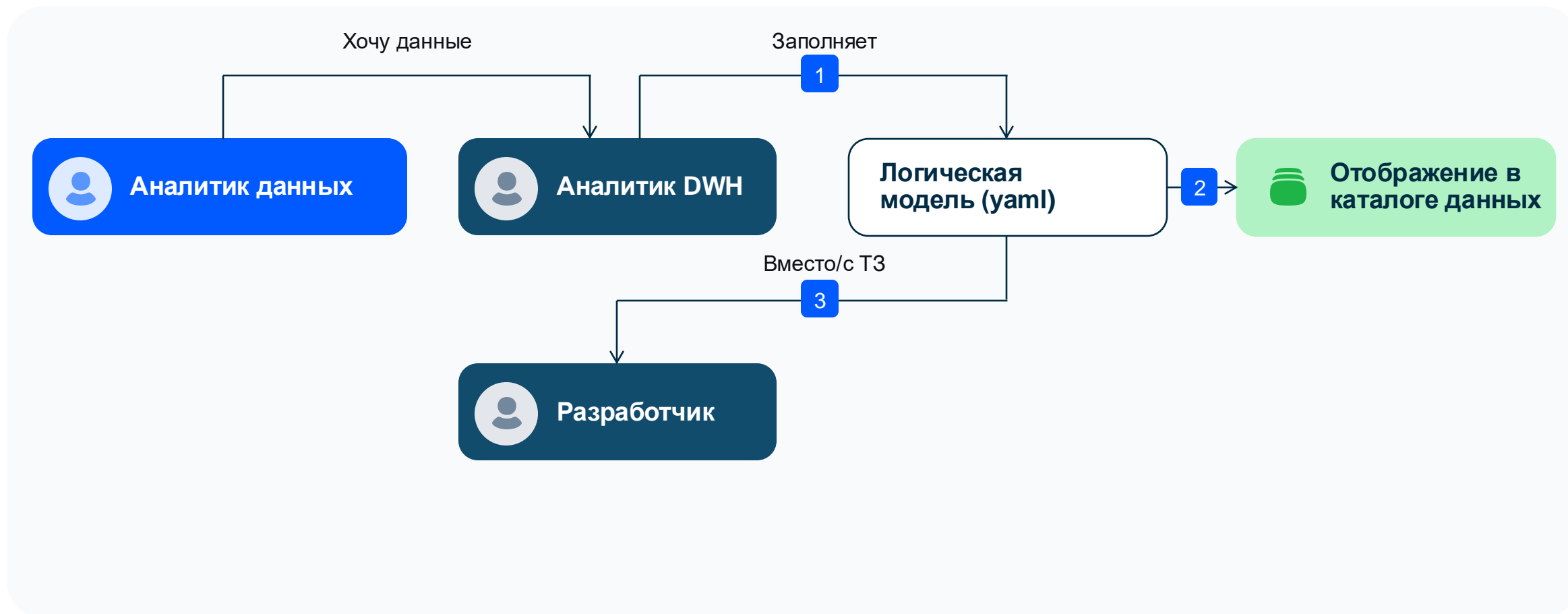
Flow работы с лог. моделью



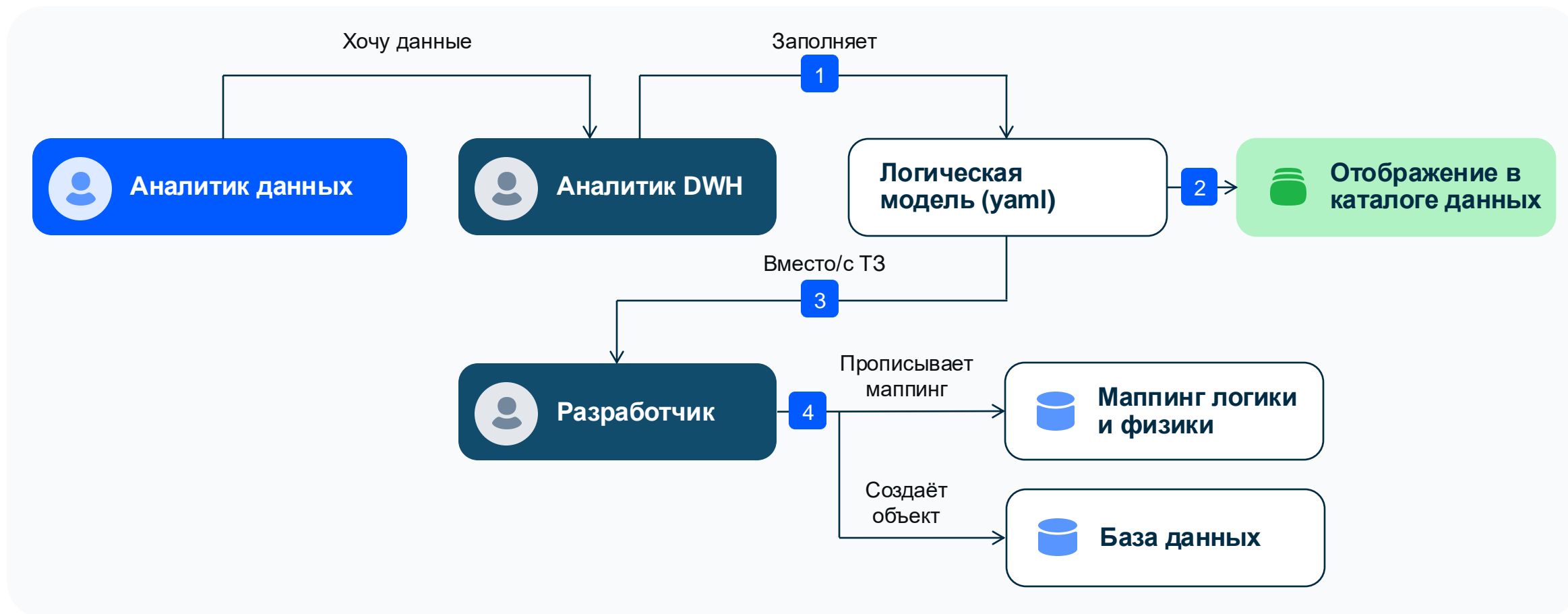
Flow работы с лог. моделью



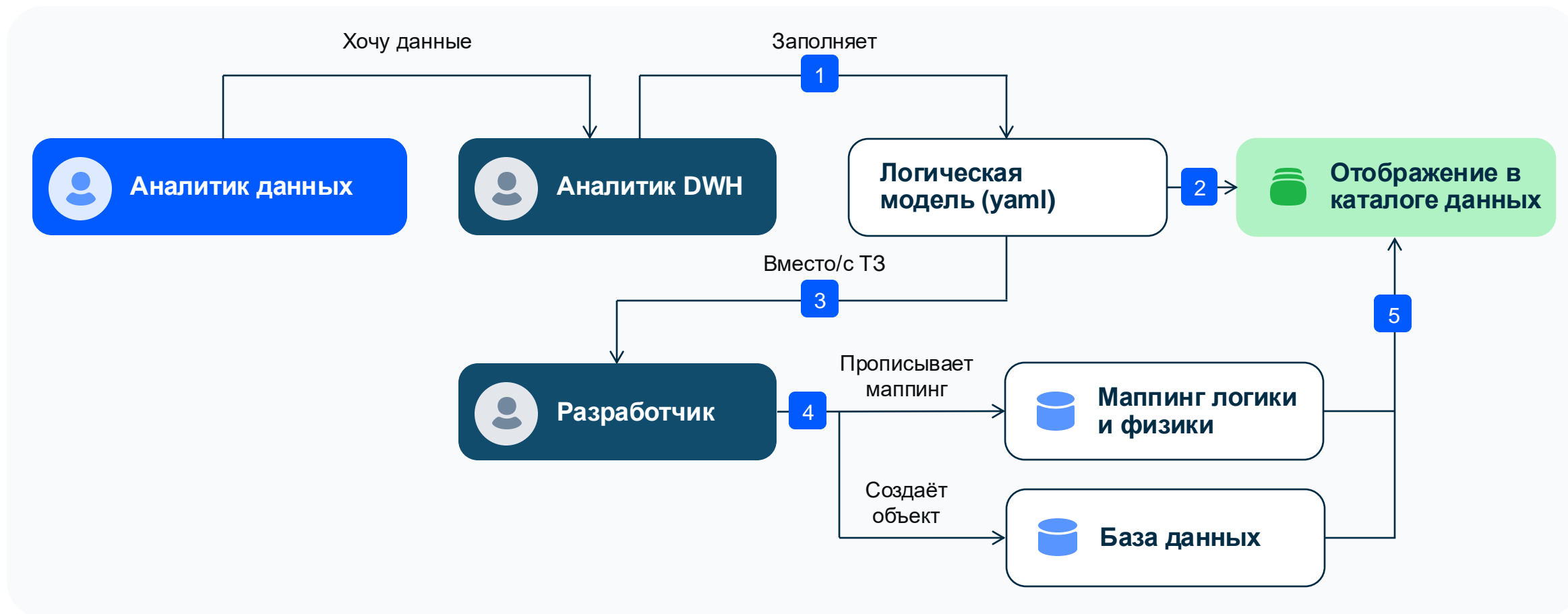
Flow работы с лог. моделью



Flow работы с лог. моделью

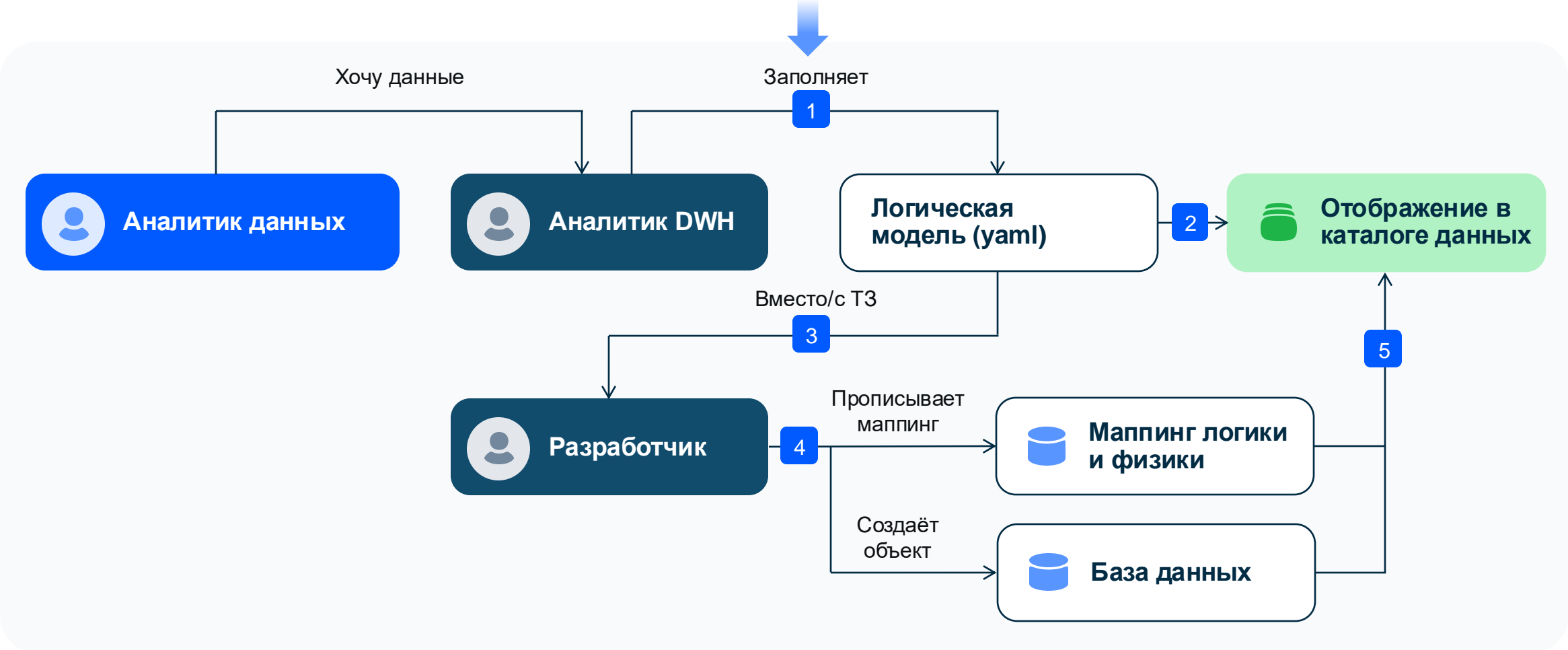


Flow работы с лог моделью

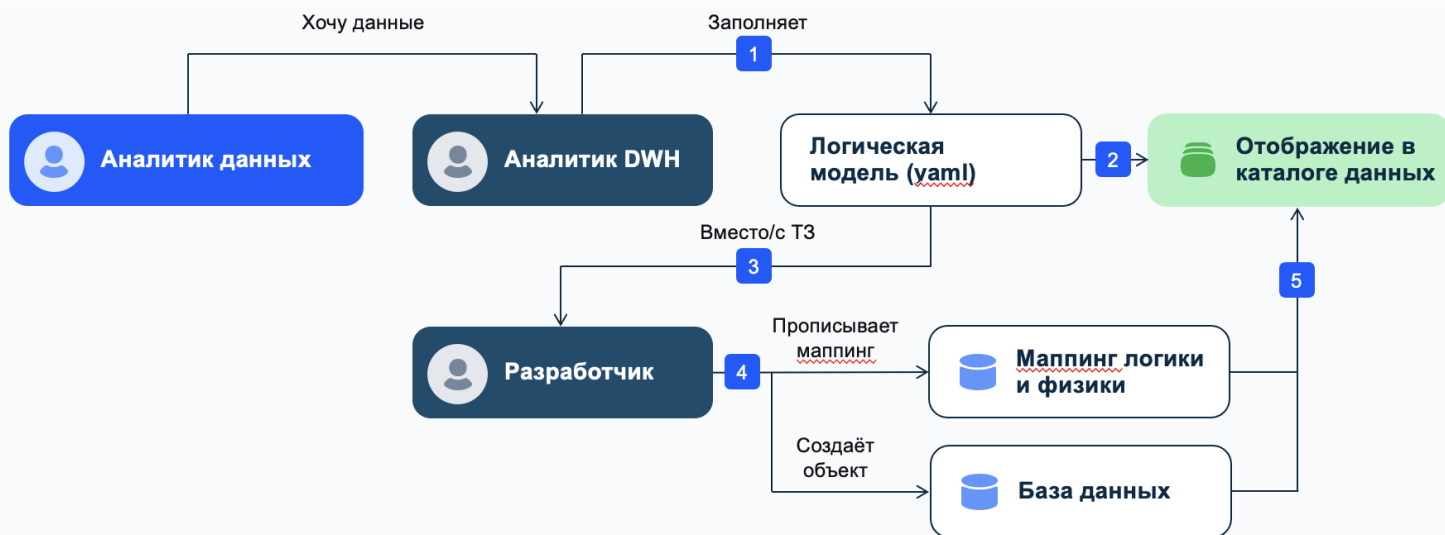


Flow работы с лог. моделью

Инструменты будущего



Преимущества процесса



- Каждый оставляет «след» в каталоге данных
- Отсутствие ручного заведения объектов в каталоге данных
- Каждый отвечает за свою часть

Уроборос построения модели

- + Быстрее
- + Каталог обновляется самостоятельно без ручных корректировок
- + Прошли весь путь полностью, с остальными хранилищами будет проще


- Процесс перестроить сложно без доп. инструментов моделирования
- Сработало только для якорной модели
- Другие хранилища придется «дружить вручную»



План работы с другими хранилищами



План работы с другими хранилищами



Лог. модель
для всех
хранилищ

План работы с другими хранилищами

Лог. модель
для всех
хранилищ



Линк между
хранилищами



План работы с другими хранилищами

Лог. модель
для всех
хранилищ



Линк между
хранилищами



Объединить
хранилища

К чему стремиться

Glossary Term

Адрес [↗](#) Add to

[Documentation](#) [Related Entities](#) [Related Terms](#) [Properties](#)

[↗ Edit](#) [+ Add Link](#)

Определение

Адрес доставки. Строка адреса или точка на карте (географические координаты).
Куда нужно доставить отправление.

Синонимы

Address, Адрес доставки.

Категории термина

Тег	Расшифровка тега
Тег:Общее	Описание



Glossary Term

Адрес [↗](#) Add to assets

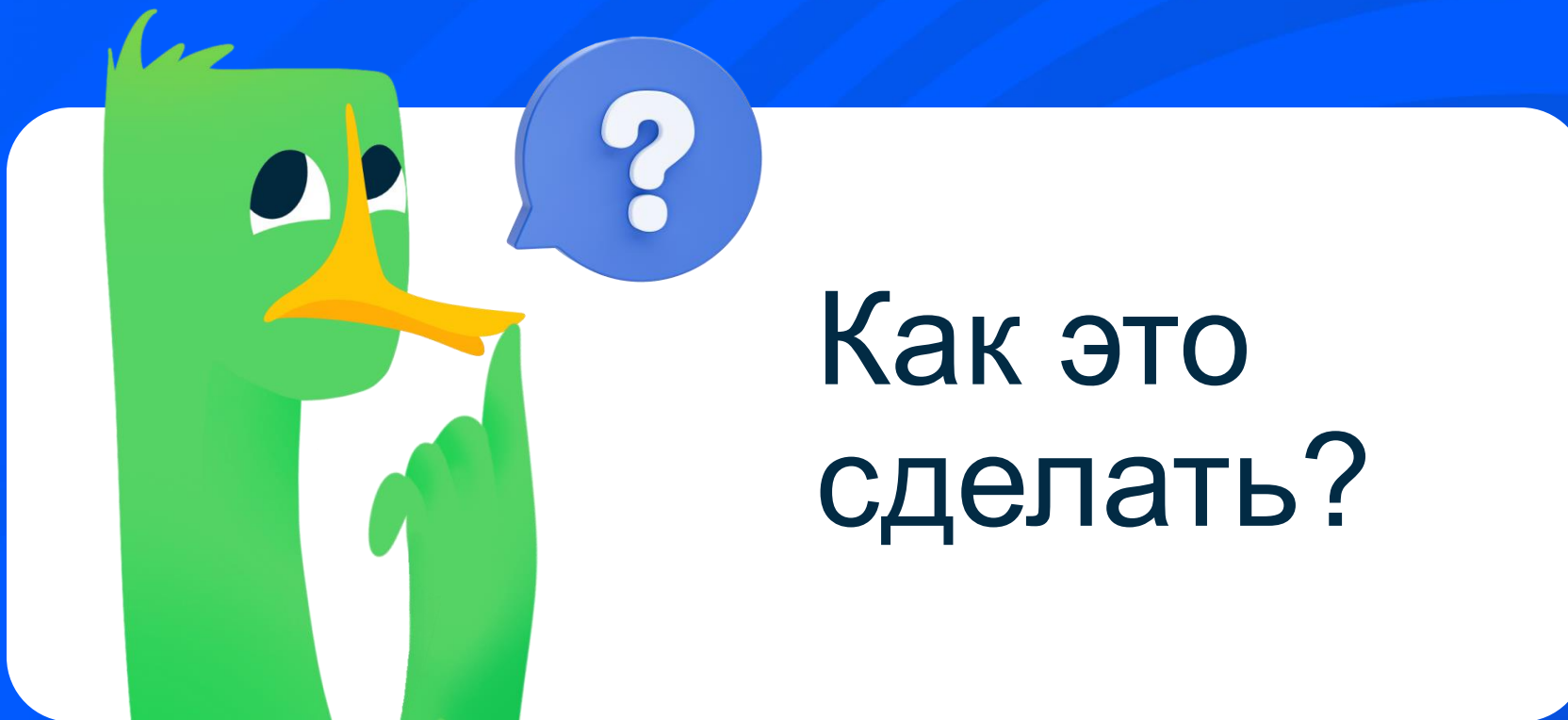
[Documentation](#) [Related Entities](#) [Related Terms](#) [Properties](#)

Filters Filter entities..

Table **clickhouse** > dictionary
dictionary.area_full_raw
"Адрес [Описание]" "Доставка [Описание]" Адрес

View **Vertica** Vertica
dwh.Dim_AddressRegion
Адрес

Table **Vertica** Vertica > dwh_data
dwh_data.Anc_AddressRegion
AddressRegion - геобъект от уровня "Страна" до уровня "Город" [Сущность] Адрес



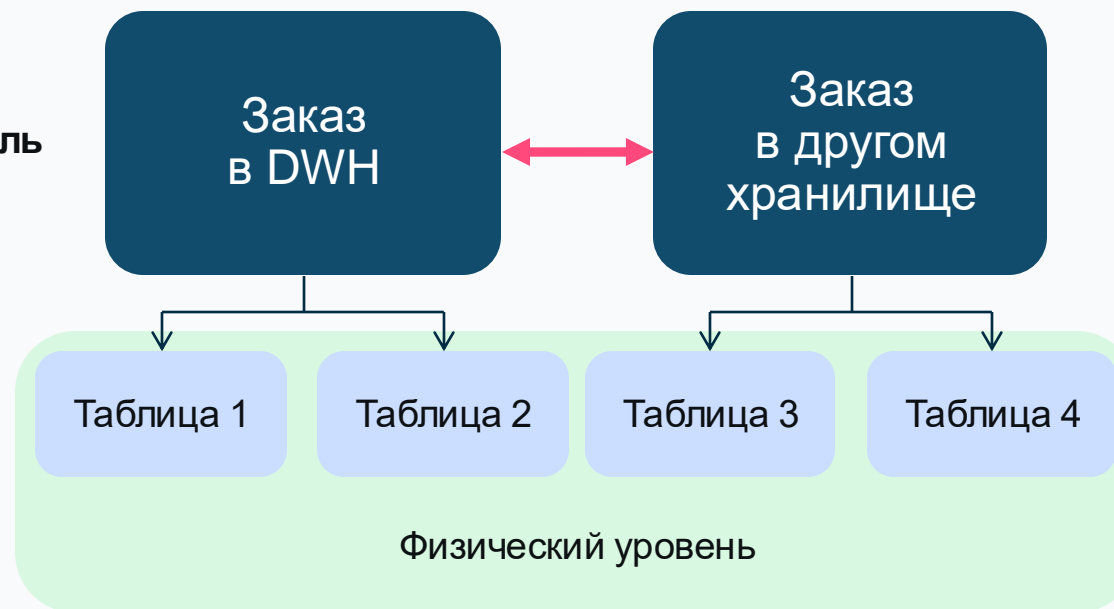
Как это
сделать?

Идеально

→ Сделать лог. модель др. хранилищ

→ Найти связи между объектами лог. моделей разных хранилищ

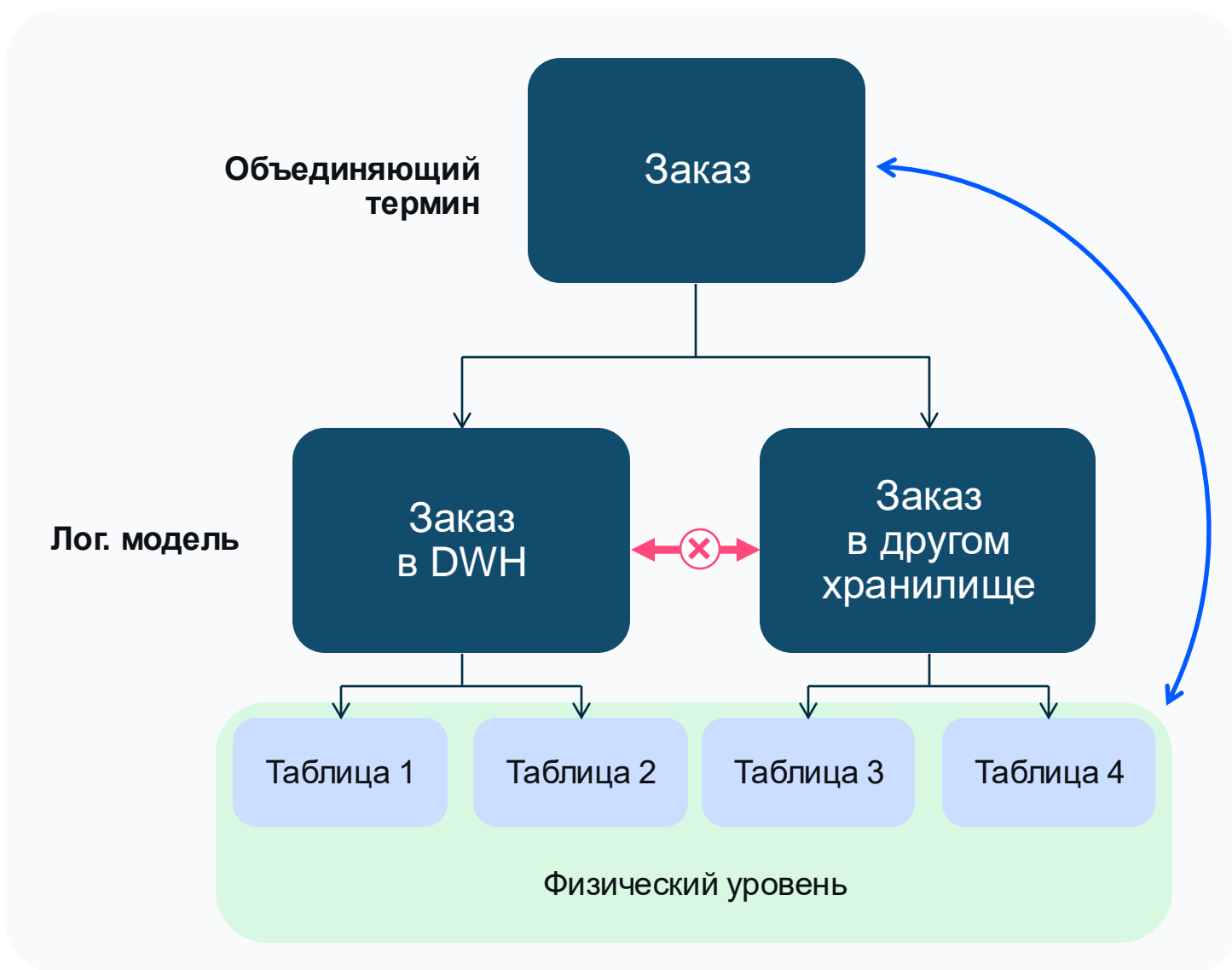
Лог. модель



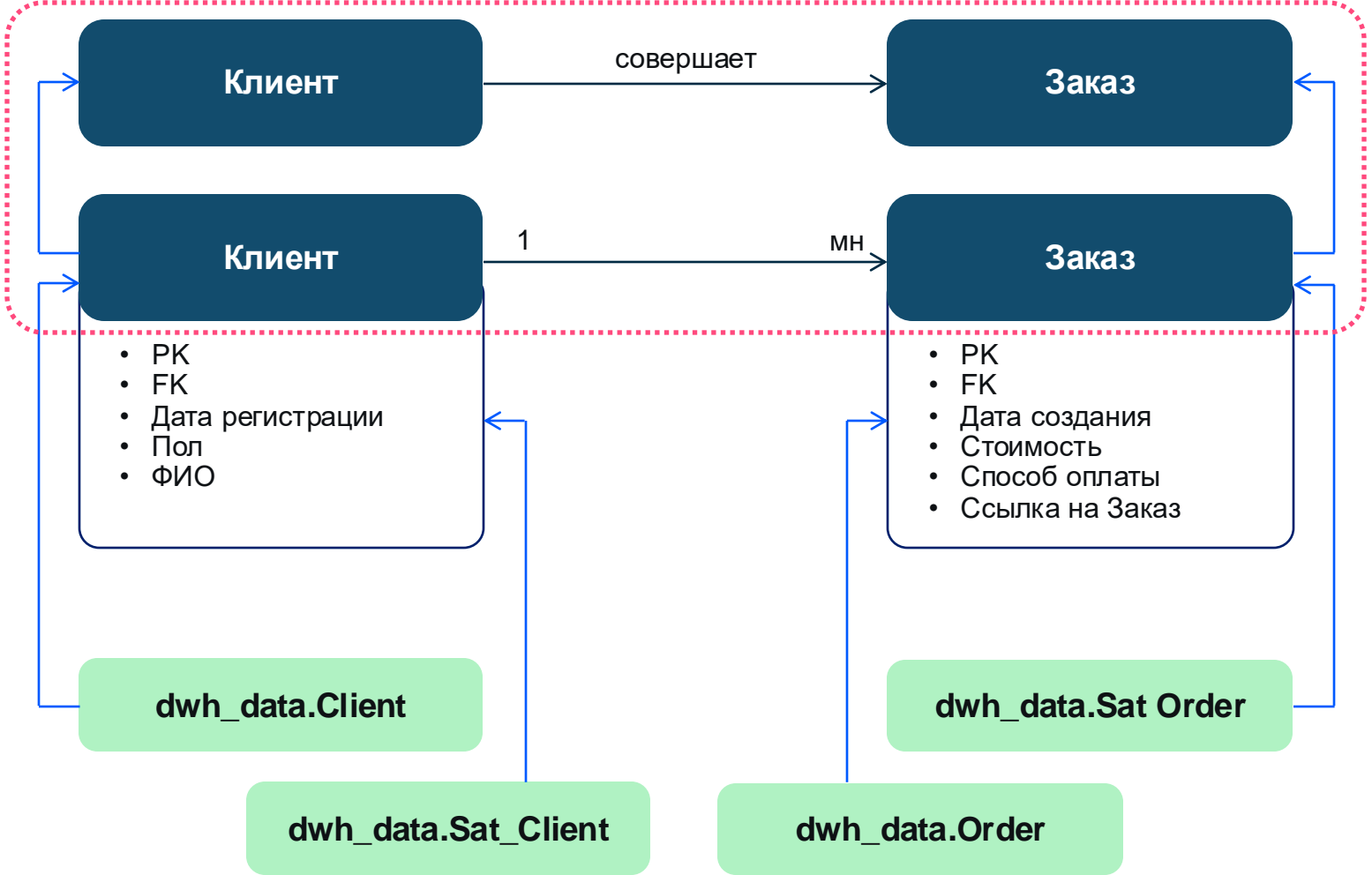
Идеально

➔ Пользователю больше кайфа

➔ Много ручной работы, увы и ах



От логики к концепции



**Неправильный,
но жизненный
подход**

Пользователь находит термин в глоссарии

Glossary Term Add to

Адрес [↗](#)

[Documentation](#) [Related Entities](#) [Related Terms](#) [Properties](#)

[↗ Edit](#) [+ Add Link](#)

Определение

Адрес доставки. Строка адреса или точка на карте (географические координаты)
Куда нужно доставить отправление.

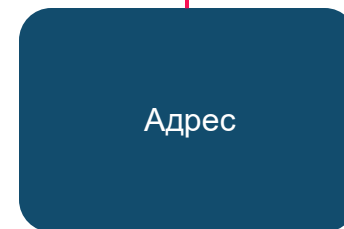
Синонимы

Address, Адрес доставки.

Категории термина

Тег	Расшифровка тега
Тег:Общее	Описание

Объединяющий термин

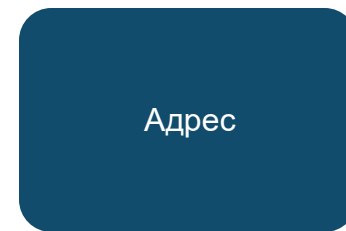


В каких лог. моделях присутствует

The screenshot shows a web interface for a glossary term. At the top, it says 'Glossary Term' and 'Адрес'. There are buttons for 'Add to assets' and 'Share'. Below this, there are tabs for 'Documentation', 'Related Entities', 'Related Terms', and 'Properties'. The 'Related Terms' tab is selected and highlighted with a pink rounded rectangle. Underneath, there is a section titled 'Contains' with a '+ Add Terms' button. This section lists two related terms:

- "Адрес [Описание]" with a link to 'View Related Entities'.
- "AddressRegionV2 - геообъект от уровня "Страна" до уровня "Город" [Сущность]" with a link to 'View Related Entities'.

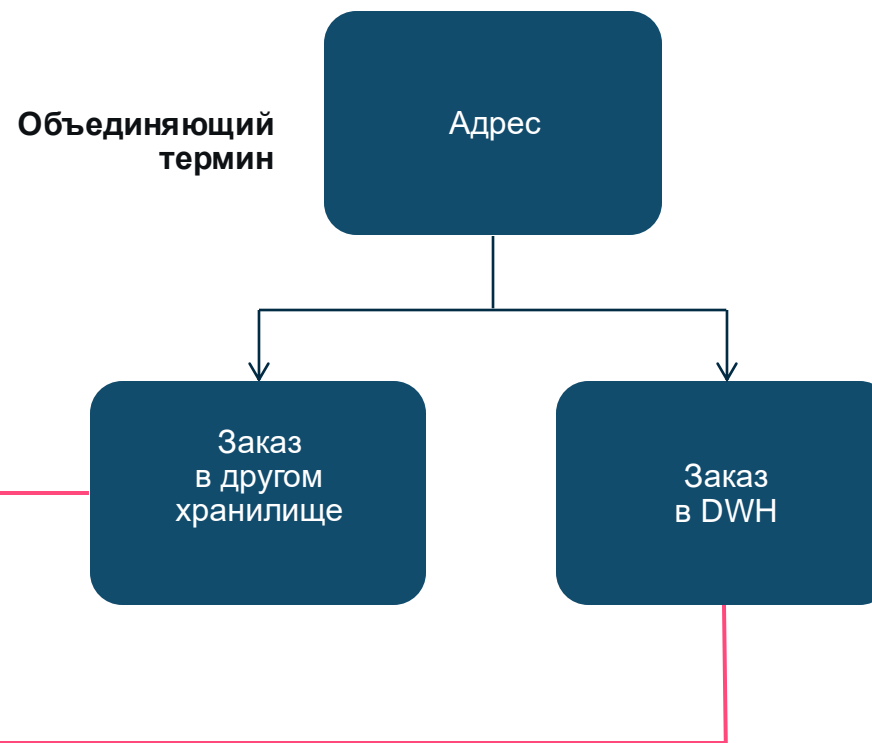
Объединяющий термин



Пользователь находит термин в глоссарии

The screenshot shows a web interface for a glossary term. At the top, it says 'Glossary Term' and 'Адрес'. Below this are tabs for 'Documentation', 'Related Entities', 'Related Terms', and 'Properties'. The 'Related Terms' tab is active. On the left, there are sections for 'Contains', 'Inherits', 'Contained by', and 'Inherited by'. The main content area shows two related terms, each highlighted with a pink box and a red label:

- Tracker BI**: A term with a description starting with 'Описание · События · addaddressaddressinaddressbookchangesafead...' and a 'Tracker' icon.
- DWH Anchor**: A term with a description 'AddressRegionV2 - геообъект от уровня "Страна" до уровня "Город" [Сущность]' and a 'View Related Entities' link.



Сразу видит таблицы с нужными данными

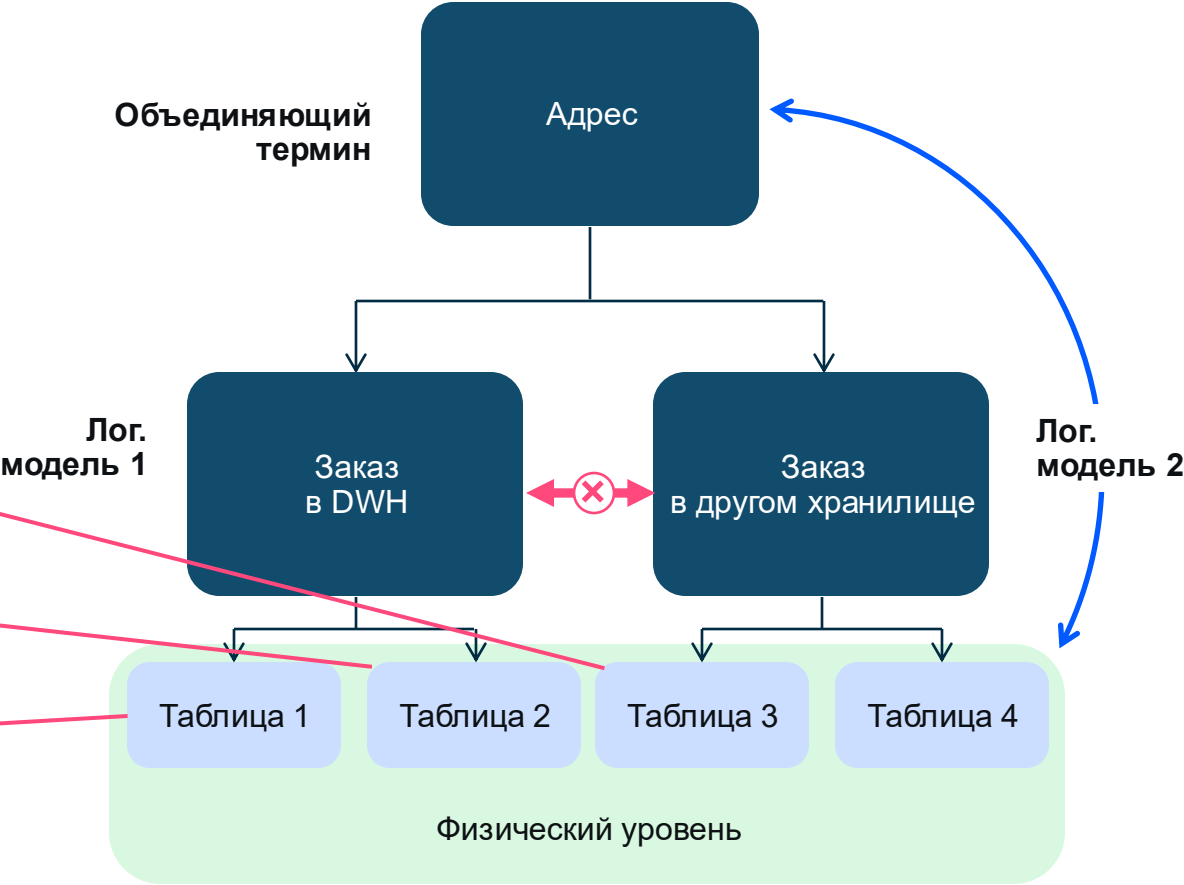
Glossary Term Add to assets

Адрес [↗](#)

Documentation **Related Entities** Related Terms Properties

Filters Filter entities..

- Table **clickhouse** dictionary
dictionary.area_full_raw
"Адрес [Описание]" "Доставка [Описание]" Адрес
- View **VERTICA Vertica**
dwh.Dim_AddressRegion
Адрес
- Table **VERTICA Vertica** dwh_data
dwh_data.Anc_AddressRegion
AddressRegion - геообъект от уровня "Страна" до уровня "Город" [Сущность] Адрес



Какие **фичи** были добавлены в каталог

Фичи, которые были сделаны для удобства **пользователя**

01

Фичи, которые были сделаны для удобства **стюарда**

02



Сборная документация

Contract

Введение

Bank - Справочник Банков РФ

Bank - Справочник Банков РФ [Сущность]

Bank_Address - Адрес отделения банка РФ

Bank_Name - Наименование отделения банка РФ

Contract_Bank

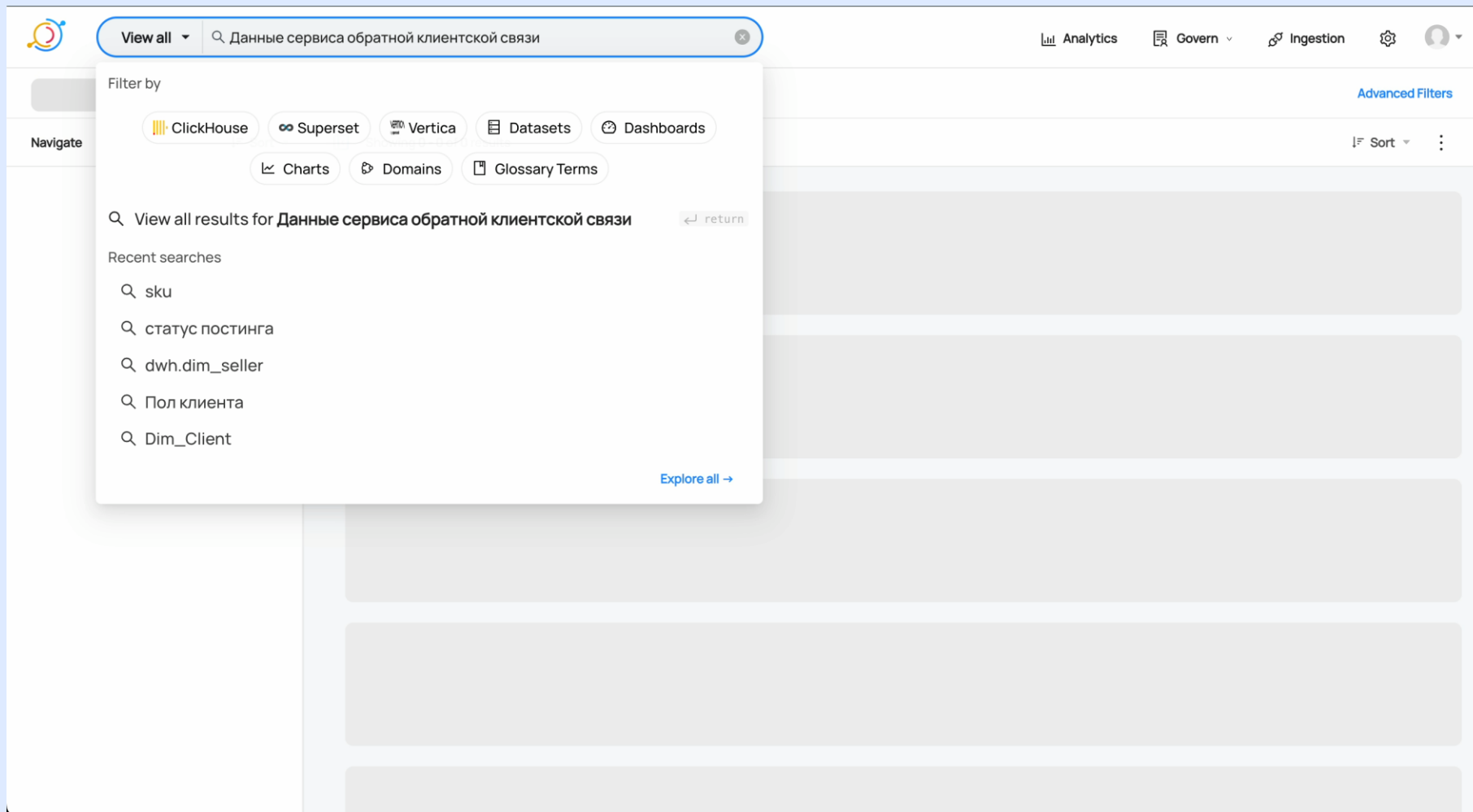
PayProps_Bank

Contract - Договор

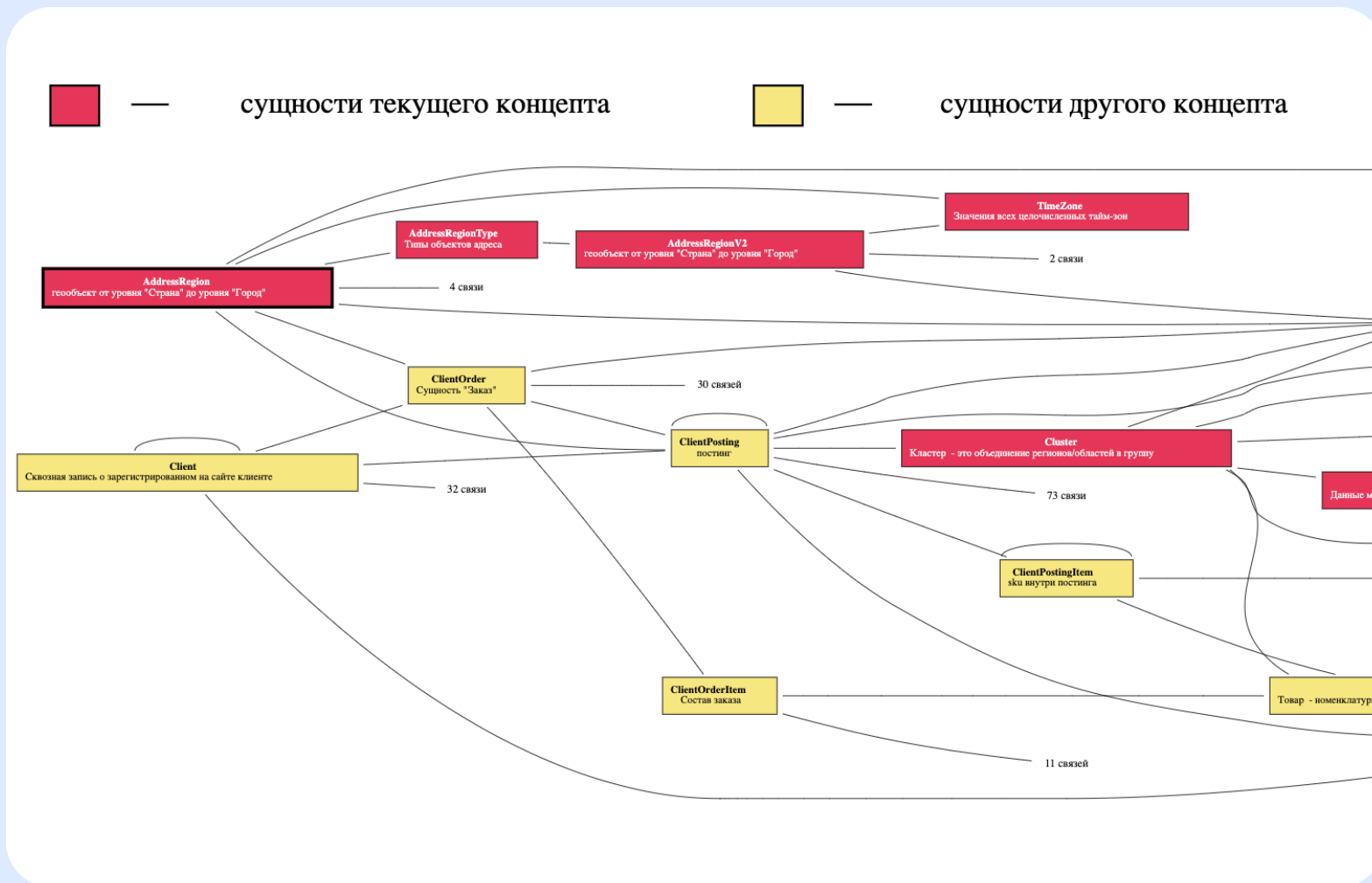
Attribute

Сущность	Атрибут	Описание
AbExperiment	AbExperiment_CreatedTime	Время создания эксперимента.
AbExperiment	AbExperiment_ExperimentEndTime	Время окончания влияния на пользователей.
AbExperiment	AbExperiment_ExperimentStartTime	Время старта эксперимента (когда он начинает влиять на пользователей).
AbExperiment	AbExperiment_Name	Человеко-читаемое название варианта A/B эксперимента.
AbVariant	AbVariant_IsControl	Признак того, что данный вариант является контрольной группой.
AbVariant	AbVariant_Name	Человеко-читаемое название варианта A/B теста.

Сборная документация



Графы модели



➔ Для тех, кто любит смотреть, а не читать

➔ Дополнительная навигация

Примеры запросов

```
SELECT ClientOrderId,  
       aa.SourceKey,  
       a.ClientPostingId,  
       a.ParentClientPostingId  
FROM   dwh_data.Anc_ClientPosting aa  
INNER JOIN dwh_data.Tie_ClientPosting_ClientPosting_Parent a  
ON a.ClientPostingId = aa.ClientPostingId  
INNER JOIN dwh_data.Tie_ClientPosting_ClientOrder b  
ON a.ClientPostingId = b.ClientPostingId  
WHERE SourceKey ilike '66068576-0038-%'
```

- Неактуальные запросы не добавляются
- Один запрос может быть примером на N объектов

Примеры запросов

```
SELECT ClientOrderId,  
       aa.SourceKey,  
       a.ClientPostingId,  
       a.ParentClientPostingId  
FROM   dwh_data.Anc_ClientPosting aa  
INNER JOIN dwh_data.Tie_ClientPosting_ClientPosting_Parent a  
ON a.ClientPostingId = aa.ClientPostingId  
INNER JOIN dwh_data.Tie_ClientPosting_ClientOrder b  
ON a.ClientPostingId = b.ClientPostingId  
WHERE SourceKey ilike '66068576-0038-%'
```

**3 таблицы
в запросе =
±20 страниц в
каталоге, где он
находится**

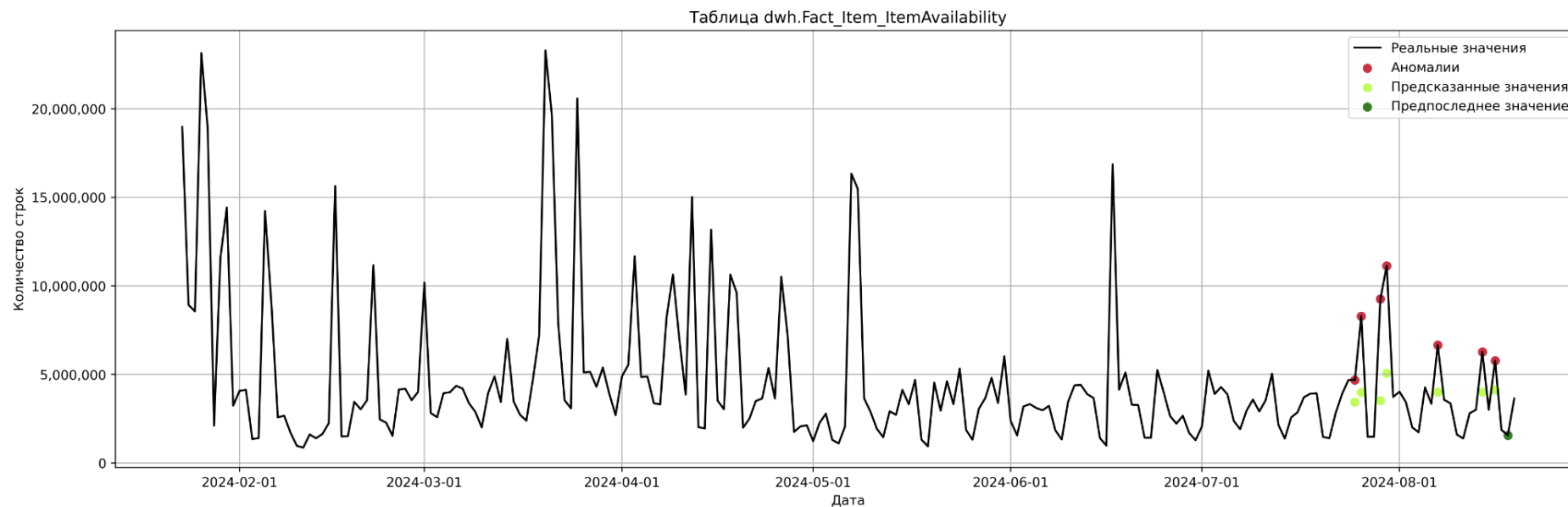
Информация о Data Quality

Schema View Definition **Documentation** Lineage Properties Queries Stats Validation Incidents

Основная проверка на аномальность

- **Метод проверки:** тройное экспоненциальное сглаживание.
- **Время проверки:** 20.08 в 09:11 МСК.
- **Расписание:** 9:00 МСК.

В таблице по полю StartTime за предыдущий день (19.08) всё в порядке. Ожидалось 3 807 274 строк, получено 3 644 664 строк.



Затрачиваемые ресурсы и результаты



Полгода на описание
1,5 хранилищ



6 аналитиков,
1 разработчик, 1 девопс



Вопросы в чатах
ушли на 70%

- **Нашли ошибки в логике хранилища**
- **Время на документацию 4 часа в неделю — меньше часа в неделю**
- **Ответ дежурного — в течение нескольких минут**

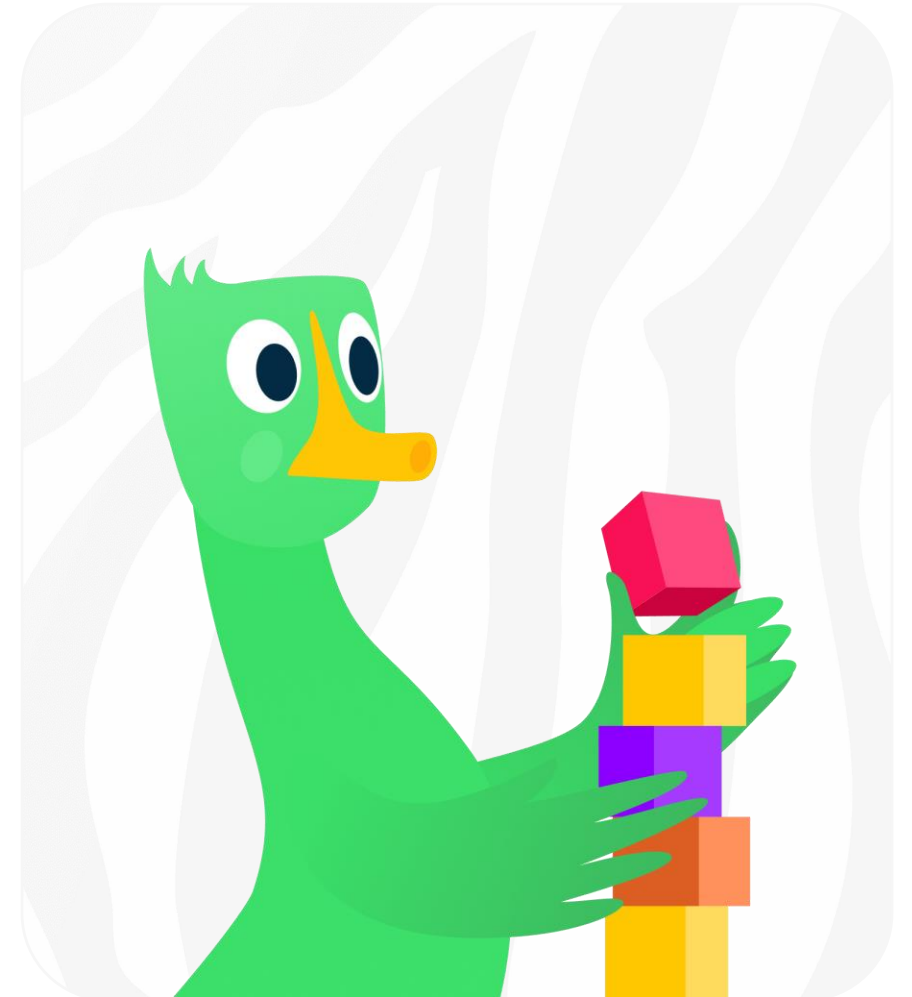
Дельные мысли



- Начать с **логической модели**
- Собирайте документацию **уже сейчас**
- Не пытаться объединять разные хранилища на старте, если это не было заложено изначально архитекторами хранилищ
- **Снизу вверх** — подход, которого не нужно стыдиться 😊
- MVP обошелся малой кровью, а вот вспомогательные инструменты — дорогое удовольствие

Темы для следующего раза

- Сколько времени ушло на сбор и консолидацию единой модели для всех хранилищ
- Процесс управления единой моделью и мониторинг ее чистоты
- Какие трудности были с добавлением «метрик»





Спасибо
за внимание!

Наталья Журавлёва
Должность
tg: @omon_ra



Ссылка QR-кода

