

ТЕХТ2SQL В РЕАЛЬНЫХ ЗАДАЧАХ: УЛУЧШАЕМ РАБОТУ В СЛОЖНЫХ ДОМЕНАХ

ГАЗПРОМ НЕФТЬ – ЦИФРОВЫЕ РЕШЕНИЯ

ВЫЛЕГЖАНИН ЕВГЕНИЙ
16.05.2025



ОБО МНЕ



ЕВГЕНИЙ ВЫЛЕГЖАНИН

ГЛАВНЫЙ РАЗРАБОТЧИК И TEAM LEAD
КОМАНДЫ МАШИННОГО ОБУЧЕНИЯ (LLM)
В КОМПАНИИ «ГАЗПРОМ НЕФТЬ –
ЦИФРОВЫЕ РЕШЕНИЯ»

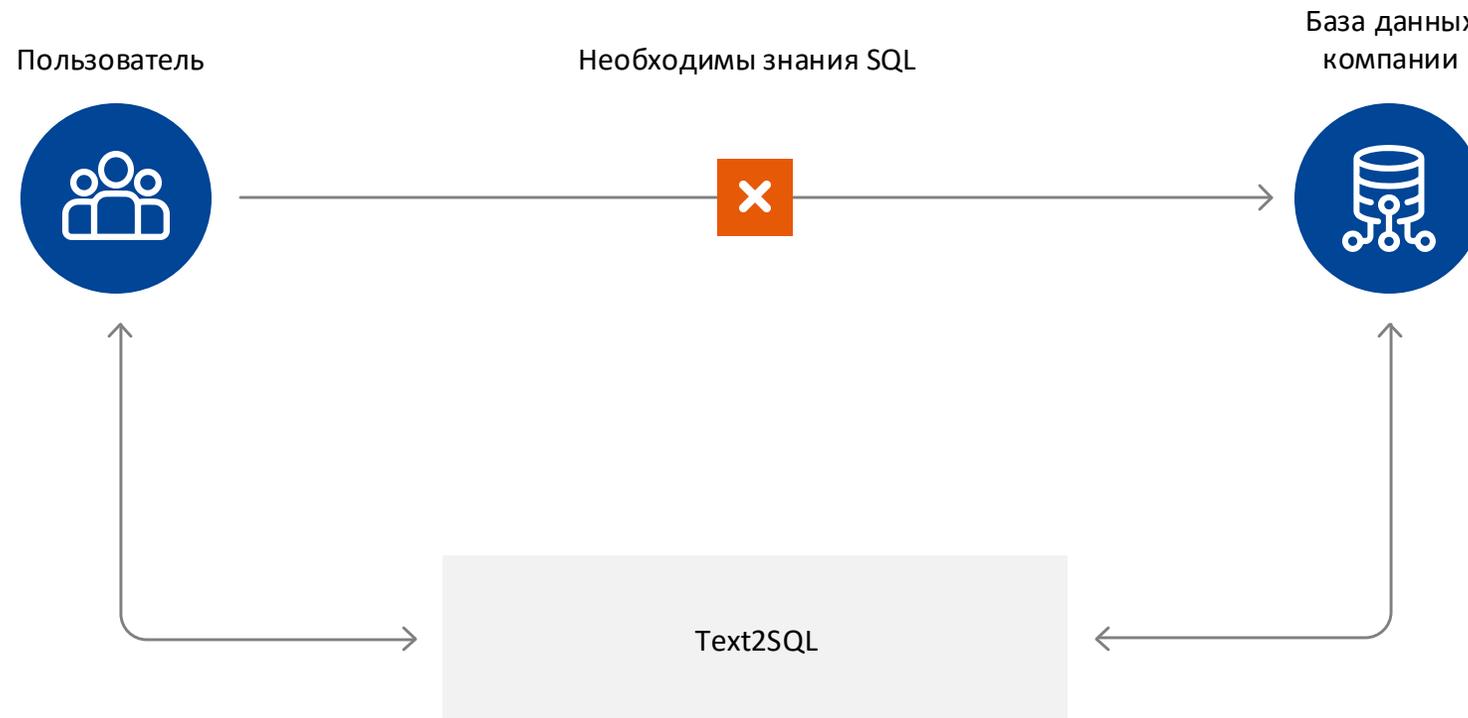


ПЛАН ДОКЛАДА

НАВИГАЦИЯ ПО СЛАЙДАМ

01	Постановка задачи	4
02	Классический подход и предыстория	5
03	Схема типового решения с LLM	6
04	Описание данных	7
05	Проблемы типового решения	12
06	Улучшаем решение	13
07	Генерируем ответ на естественном языке	33
08	Итоговая схема	36
09	Масштабирование системы	37
10	Какую LLM использовать?	38
11	Выводы	39

ПОСТАНОВКА ЗАДАЧИ



КЛАССИЧЕСКИЙ ПОДХОД (ПРЕДЫСТОРИЯ)



Вопрос
пользователя

01



Выделение
именованных
сущностей

02



Конструктор SQL
на основе правил
и алгоритмов

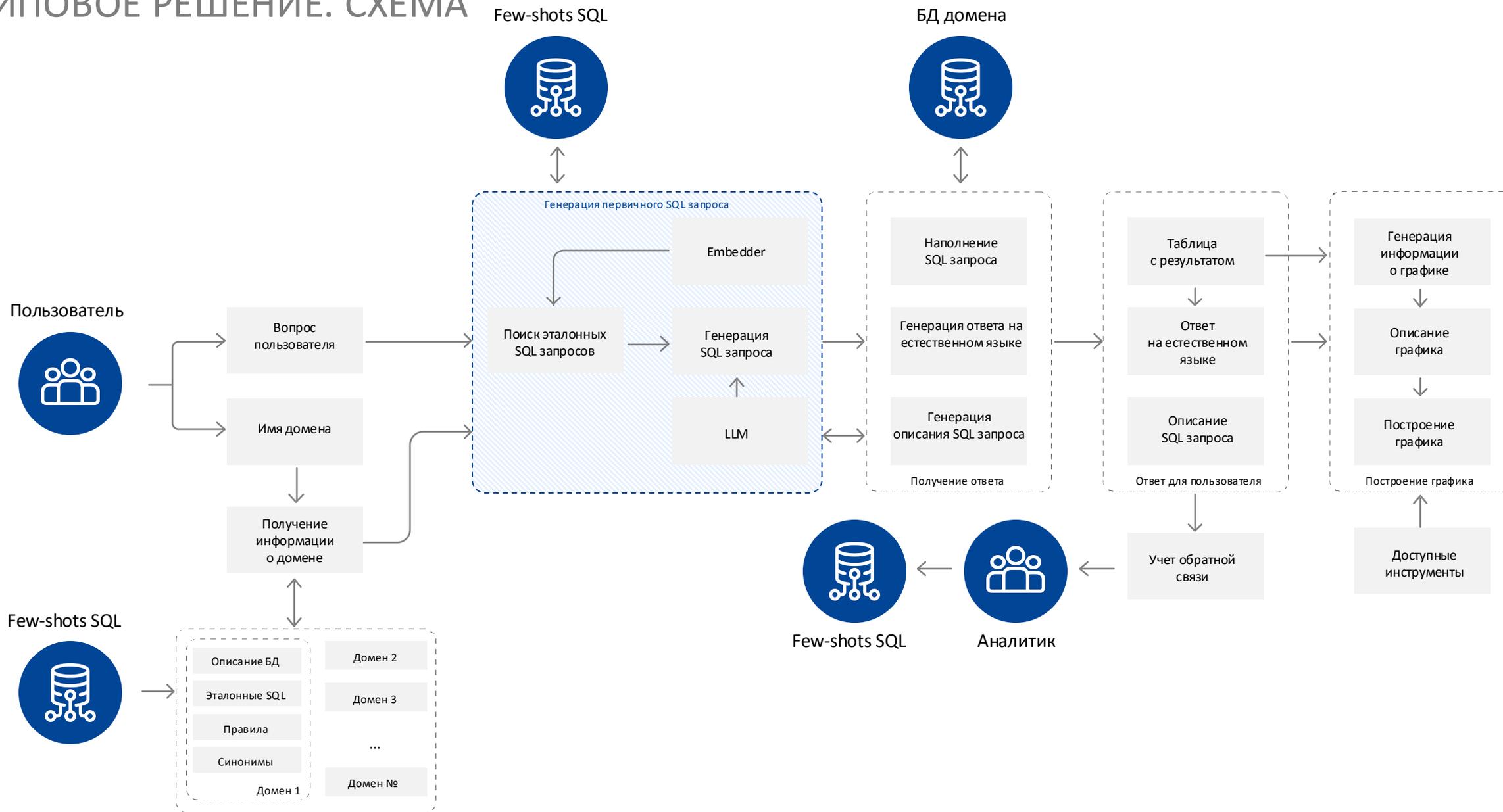
03



SQL-запрос

04

ТИПОВОЕ РЕШЕНИЕ. СХЕМА



НАБОР ДАННЫХ. ОПИСАНИЯ БД

- roles

Хранит информацию о ролях и должностях сотрудников в проектах.

```
CREATE TABLE roles (  
  role_id SERIAL PRIMARY KEY, -- Автоинкрементный ID.  
  role_name VARCHAR(255), -- Название роли / должности сотрудника в проекте. Примеры значений: [...].  
  type VARCHAR(10) CHECK (type IN (member, resource)), -- Тип роли сотрудника.  
);
```

- employee

Хранит информацию о ФИО сотрудников

```
CREATE TABLE employee (  
  employee_code VARCHAR(20) PRIMARY KEY, -- Код работника. Ключ для связи с таблицей project_employee_role.  
  organization_code VARCHAR(10), -- Уникальный идентификатор компании сотрудника.  
  f_io VARCHAR(50), -- Фамилия и инициалы сотрудника.  
  fact_company VARCHAR(100), -- Наименование организации сотрудника.  
  fio VARCHAR(100), -- ФИО. Полные фамилия, имя и отчество сотрудника.  
);
```

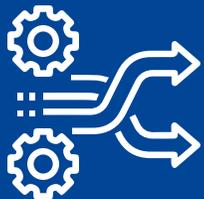
КЛАССИЧЕСКИЙ ПОДХОД (ПРЕДЫСТОРИЯ)

ДАТЫ ПРОЕКТА ДОМЕН 1?

```
SELECT
  p.project_num,
  p.project_fact_start_date,
  p.project_fact_complete_date,
  p.project_long_name
FROM
  project p
WHERE
  p.project_long_name IN ('<Полное название Домен 1>');
```

```
SELECT DISTINCT
  p.project_num,
  p.project_fact_start_date,
  p.project_fact_complete_date,
  p.project_plan_start_date,
  p.project_plan_complete_date,
  p.project_long_name
FROM
  project p
WHERE
  p.project_long_name IN ('<Полное название Домен 1>');
```

НАБОР ДАННЫХ. ПРАВИЛА



ФУНКЦИЮ EXTRACT

нужно использовать
при необходимости отбора
или группировки по дате

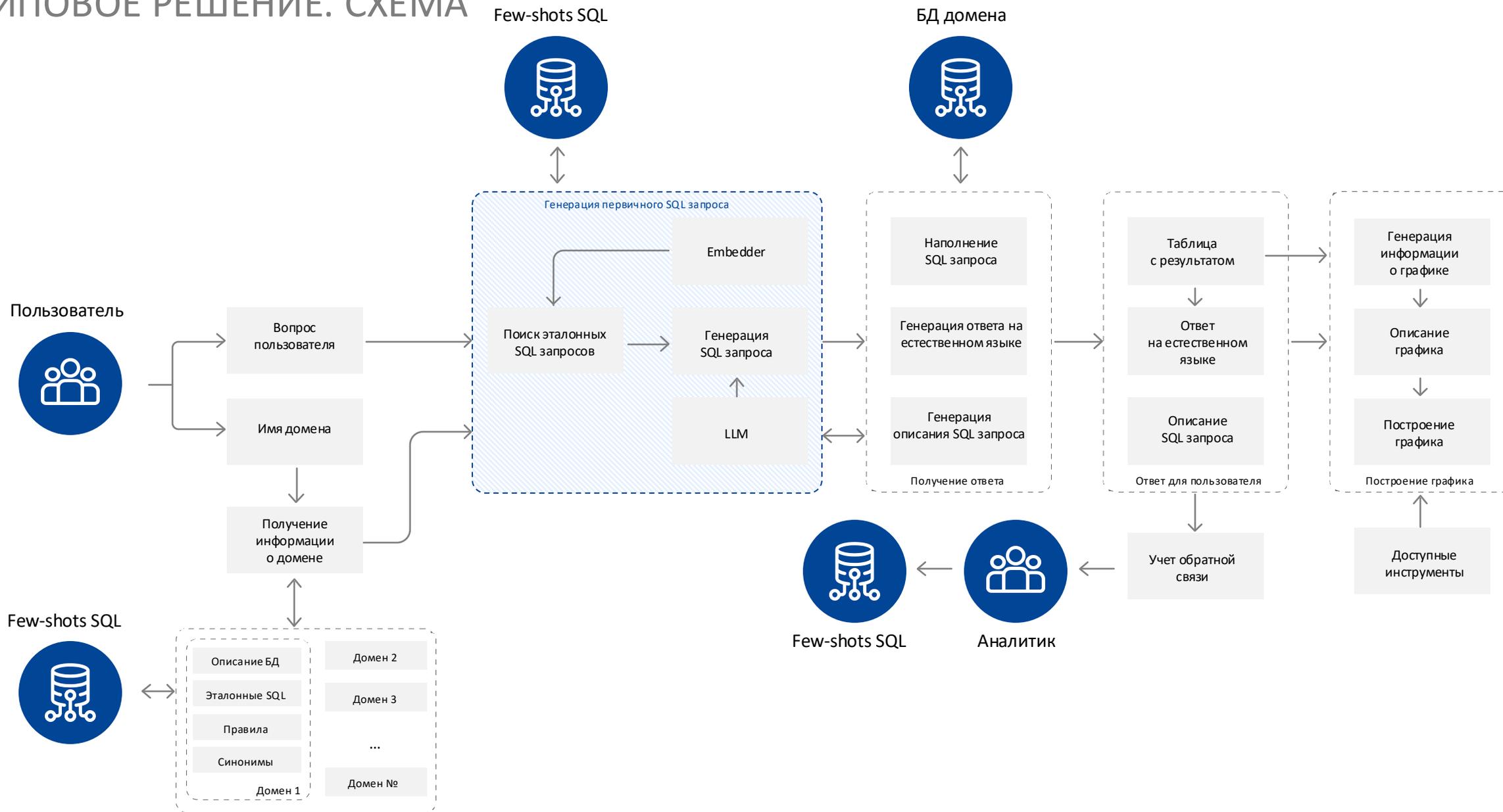
НАПРИМЕР

«Какие проекты завершены до 2020 года» ->
““where extract (year from p.project_fact_complete_date) < 2020””

НАПРИМЕР

«Какие проекты начались в 2023 года» ->
““where extract (year from p.project_fact_start_date) = 2023””

ТИПОВОЕ РЕШЕНИЕ. СХЕМА



ПРОМПТ ДЛЯ ГЕНЕРАЦИИ SQL

SYS_TEXT

Вы являетесь экспертом по SQL. По заданному вопросу создайте синтаксически правильный запрос SQL и верните его в ответе.

Никогда не запрашивайте все столбцы из таблицы. Вы должны запрашивать только те столбцы, которые необходимы для ответа на вопрос и те, которые участвуют в фильтрации / отборе строк.

Оберните каждое имя столбца в двойные кавычки (*) для обозначения их как ограниченных идентификаторов.

Обратите внимание, что используются только имена столбцов, которые вы видите в таблицах ниже.

Будьте внимательны, чтобы не запрашивать столбцы, которых не существует.

Также обратите внимание, какой столбец находится в какой таблице.

Правила формирования запроса:

{rules}

Используйте только следующие таблицы:

““sql

{db_context}

““

Используйте примеры запросов от пользователя и соответствующие им SQL-запросы, если посчитаете их релевантными:

{sql_examples}

ПРИМЕР ИСПОЛЬЗОВАНИЯ ТИПОВОГО РЕШЕНИЯ С LLM



Даты старта проекта КП?



```
SELECT project_long_name, project_fact_start_date  
FROM project  
WHERE project_long_name IN ('КП');
```

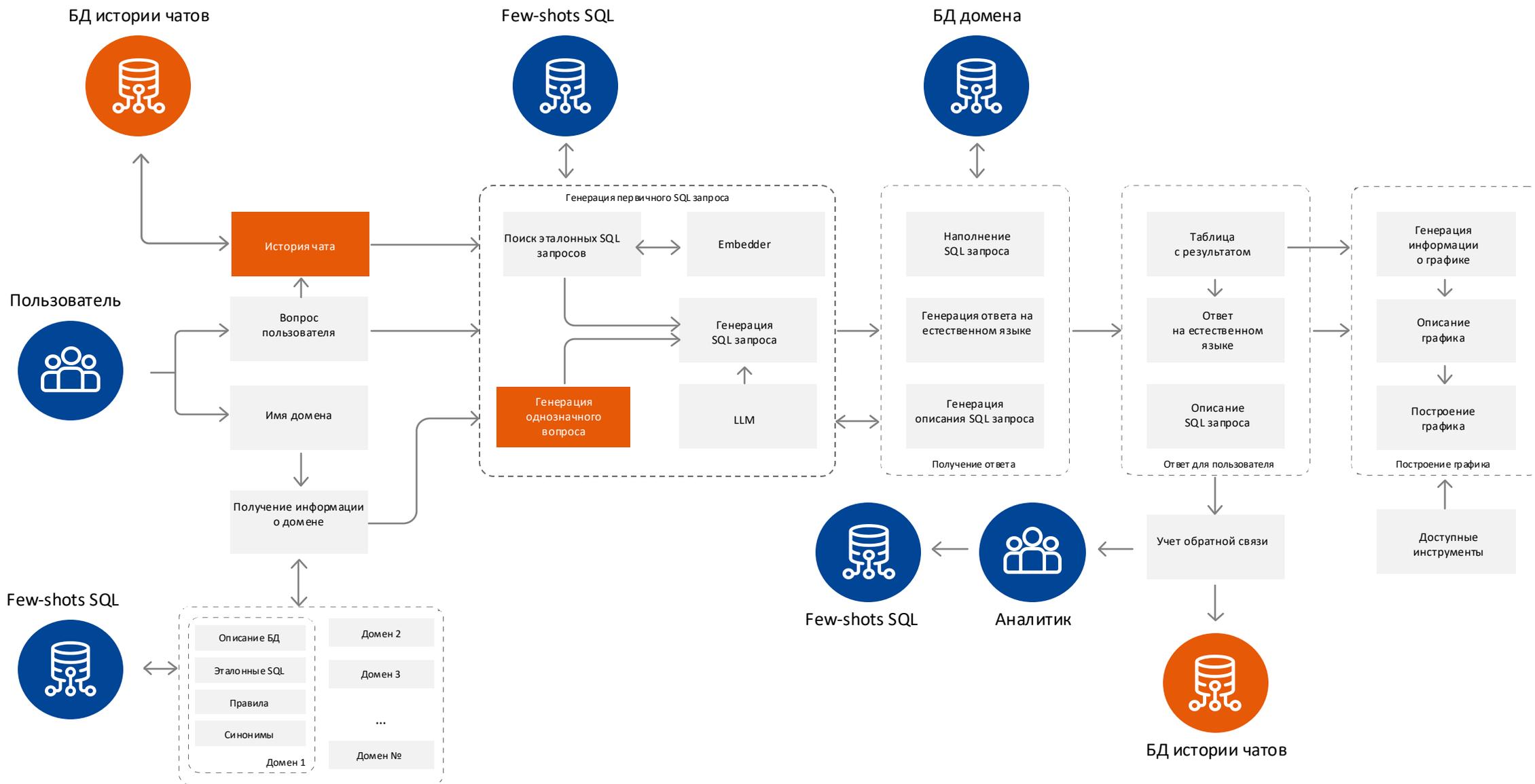


БД домена



Error

СХЕМА РЕШЕНИЯ



ДОБАВЛЯЕМ ВОЗМОЖНОСТЬ ВЕСТИ ДИАЛОГ



Даты старта проекта КП?



```
SELECT  
    project_long_name, project_fact_start_date  
FROM project  
WHERE project_long_name IN ('КП')
```



БД домена



Error



А окончания?



```
SELECT  
    project_long_name, project_fact_complete_date  
FROM project  
WHERE project_long_name IN ('КП')
```

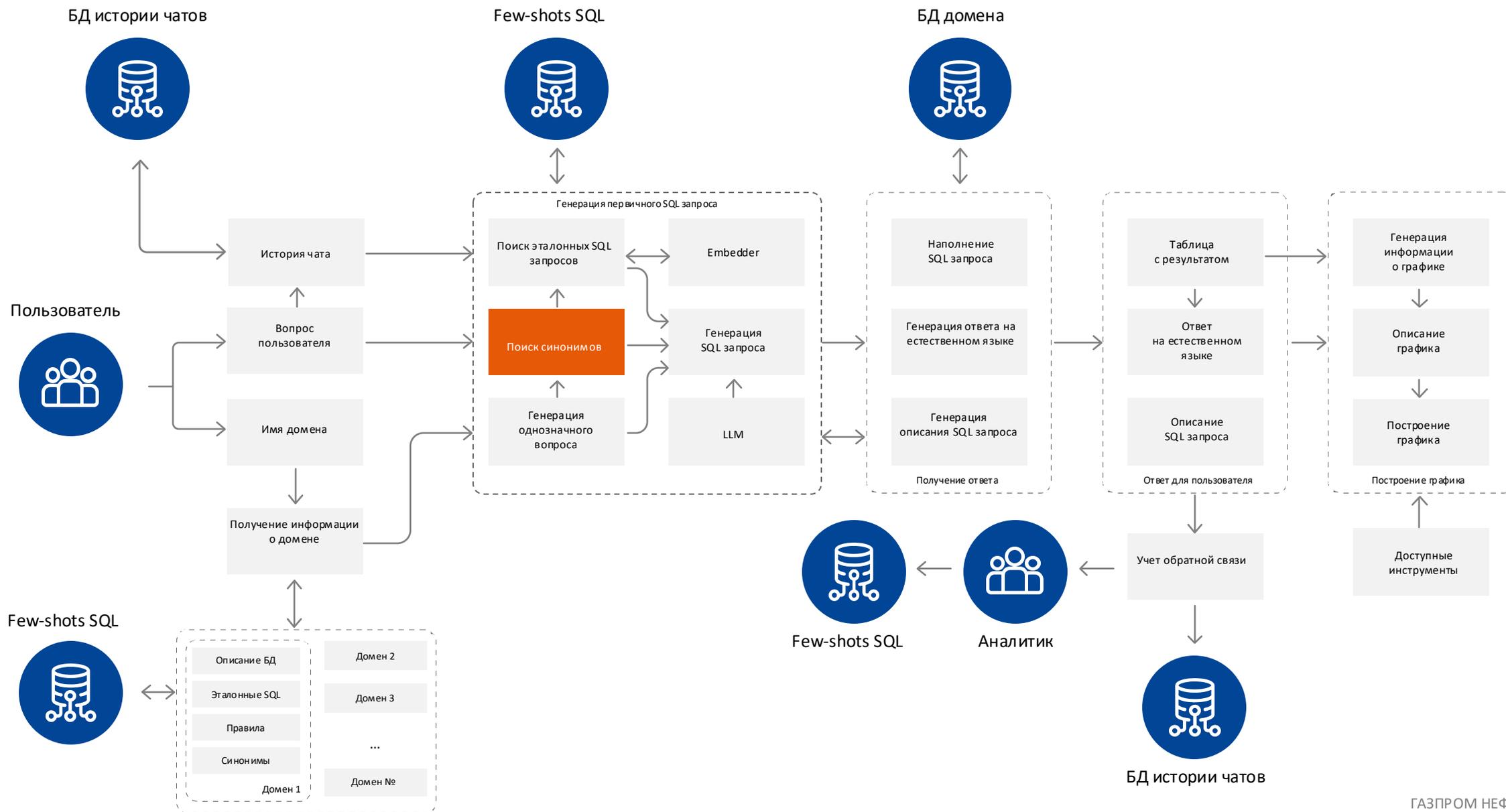


БД домена



Error

СХЕМА РЕШЕНИЯ



НАБОР ДАННЫХ. СИНОНИМЫ

ЗНАЧЕНИЕ В БД	СИНОНИМ 1	СИНОНИМ N
Системный аналитик	аналитик	
Аналитик данных	аналитик	
Автоматизированная система контролей	аск	
Когнитивный поиск	ElasticSearch	кп

СИНОНИМЫ. ПРИНЦИП ПОИСКА

ВОПРОС ПОЛЬЗОВАТЕЛЯ

Кто руководитель проекта ЦПА?

ЛЕММАТИЗАЦИЯ

кто руководитель проект цпа

ПОИСК СИНОНИМОВ

Таблица с синонимами

СИНОНИМЫ ДЛЯ ЗАПРОСА ПОЛЬЗОВАТЕЛЯ

“цпа”: ['Цифровой помощник аналитика']
“руководитель”: [
'Руководитель проекта,
'Руководитель направления',
'Менеджер продукта,
'Руководитель проекта от Заказчика'
]
“руководитель проекта”: [
'Руководитель проекта,
'Руководитель проекта от Заказчика'
]

ДОБАВЛЕНИЕ СИНОНИМОВ. ПРОМПТ

SYS_TEXT

Вы являетесь экспертом по SQL. По заданному вопросу создайте синтаксически правильный запрос SQL и верните его в ответе.

Никогда не запрашивайте все столбцы из таблицы. Вы должны запрашивать только те столбцы, которые необходимы для ответа на вопрос и те, которые участвуют в фильтрации / отборе строк.

Оберните каждое имя столбца в двойные кавычки (*) для обозначения их как ограниченных идентификаторов.

Обратите внимание, что используются только имена столбцов, которые вы видите в таблицах ниже.

Будьте внимательны, чтобы не запрашивать столбцы, которых не существует.

Также обратите внимание, какой столбец находится в какой таблице.

Правила формирования запроса:

```
{rules}
```

В запросе найдены следующие синонимы, используйте их при создании SQL фильтров:

```
{synonyms}
```

Используйте только следующие таблицы:

```
““sql  
{db_context}  
““
```

Используйте примеры запросов от пользователя и соответствующие им SQL-запросы, если посчитаете их релевантными:

```
{sql_examples}
```

ДОБАВЛЕНИЕ СИНОНИМОВ. РЕЗУЛЬТАТ

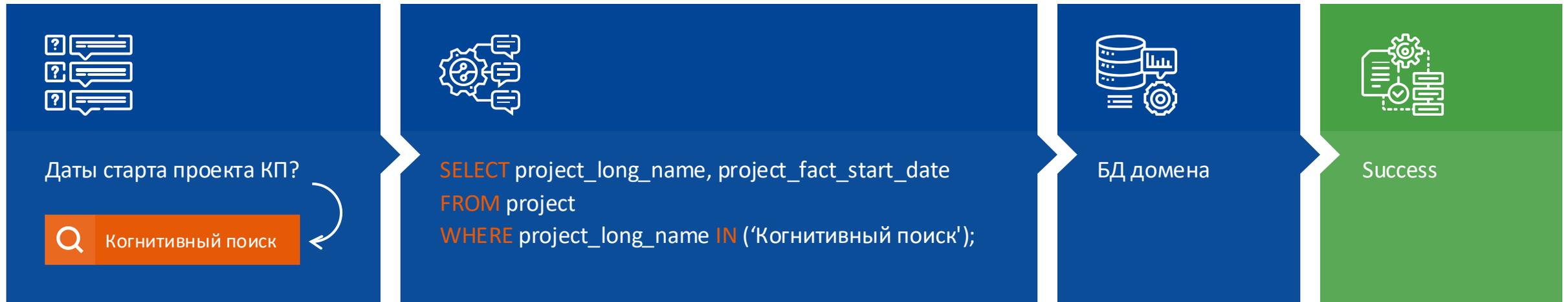
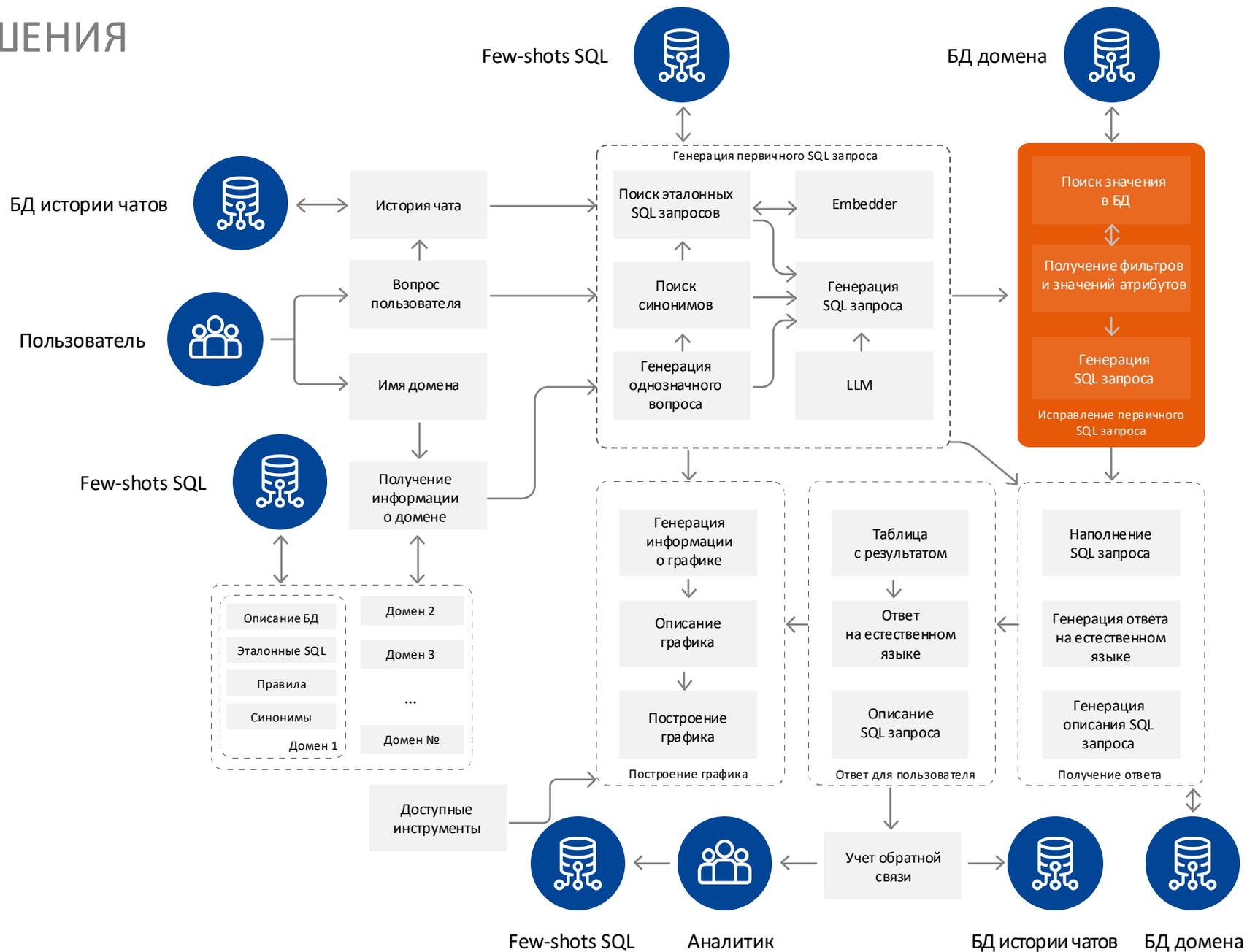
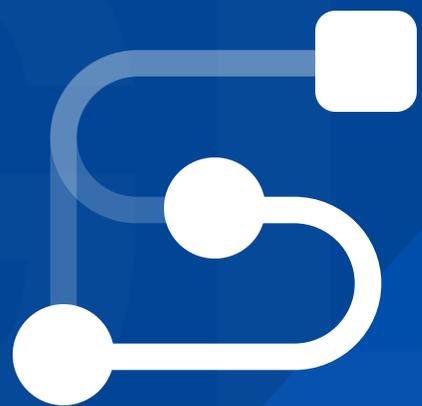


СХЕМА РЕШЕНИЯ





SQLGlot

ПОИСК ЗНАЧЕНИЙ АТТРИБУТОВ В БД

Какие проекты имеют затраты на консалтинг?

ПОИСК ЗНАЧЕНИЙ АТТРИБУТОВ В БД

Какие проекты имеют затраты на консалтинг?

```
SELECT DISTINCT
  project_num,
  project_long_name,
  cost_item_long_name
FROM
  project
WHERE
  cost_item_long_name == 'Консалтинг'
```

ПОИСК ЗНАЧЕНИЙ АТТРИБУТОВ В БД

Какие проекты имеют затраты на консалтинг?

```
SELECT DISTINCT
  project_num,
  project_long_name,
  cost_item_long_name
FROM
  project
WHERE
  cost_item_long_name == 'Консалтинг'
```



ПОИСК ЗНАЧЕНИЙ АТТРИБУТОВ В БД

Какие проекты имеют затраты на консалтинг?

```
SELECT DISTINCT
  project_num,
  project_long_name,
  cost_item_long_name
FROM
  project
WHERE
  cost_item_long_name == 'Консалтинг'
```



Атрибут: cost_item_long_name
Значение: Консалтинг

ПОИСК ЗНАЧЕНИЙ АТТРИБУТОВ В БД

Какие проекты имеют затраты на консалтинг?

```
SELECT DISTINCT
  project_num,
  project_long_name,
  cost_item_long_name
FROM
  project
WHERE
  cost_item_long_name == 'Консалтинг'
```



Атрибут: cost_item_long_name
Значение: Консалтинг



Поиск



БД домена

ПОИСК ЗНАЧЕНИЙ АТТРИБУТОВ В БД

Какие проекты имеют затраты на консалтинг?

```
SELECT DISTINCT
  project_num,
  project_long_name,
  cost_item_long_name
FROM
  project
WHERE
  cost_item_long_name == 'Консалтинг'
```



Атрибут: cost_item_long_name
Значение: Консалтинг

'Консалтинг',
'Консалтинг (Сервис)',
'Консалтинг (Архитекторы)',



Поиск



БД домена

ПОИСК ЗНАЧЕНИЙ АТТРИБУТОВ В БД

Какие проекты имеют затраты на консалтинг?

```
SELECT DISTINCT
  project_num,
  project_long_name,
  cost_item_long_name
FROM
  project
WHERE
  cost_item_long_name == 'Консалтинг'
```



```
SELECT DISTINCT
  project_num,
  project_long_name,
  cost_item_long_name
FROM
  project
WHERE
  cost_item_long_name IN ('Консалтинг', 'Консалтинг(Сервис)',
  'Консалтинг (Архитекторы)')
```

'Консалтинг',
'Консалтинг (Сервис)',
'Консалтинг (Архитекторы)',

Атрибут: cost_item_long_name
Значение: Консалтинг



Поиск



БД домена

ПОИСК ЗНАЧЕНИЙ АТТРИБУТОВ В БД. ПРОМПТ

SYS_TEXT

Вы являетесь экспертом по SQL. По заданному вопросу создайте синтаксически правильный запрос SQL и верните его в ответе.

Никогда не запрашивайте все столбцы из таблицы. Вы должны запрашивать только те столбцы, которые необходимы для ответа на вопрос и те, которые участвуют в фильтрации / отборе строк.

Оберните каждое имя столбца в двойные кавычки (*) для обозначения их как ограниченных идентификаторов.

Обратите внимание, что используются только имена столбцов, которые вы видите в таблицах ниже.

Будьте внимательны, чтобы не запрашивать столбцы, которых не существует.

Также обратите внимание, какой столбец находится в какой таблице.

Правила формирования запроса:

{rules}

В запросе найдены следующие синонимы, используйте их при создании SQL фильтров:

{synonyms}

Используйте только следующие таблицы:

““sql

{db_context}

““

Используйте примеры запросов от пользователя и соответствующие им SQL-запросы, если посчитаете их релевантными:

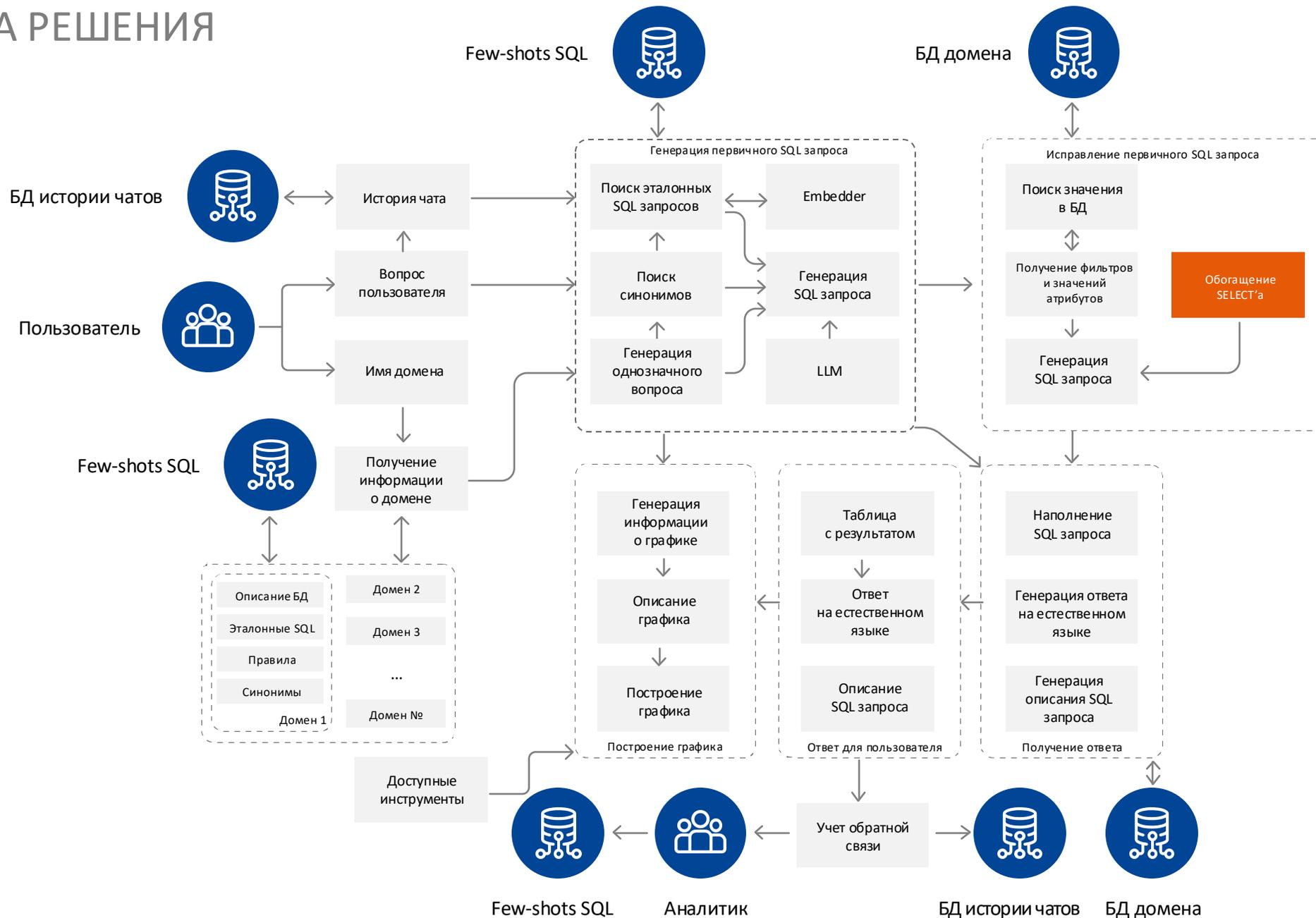
{sql_examples}

ВАЖНО: Ниже представлены значения из некоторых колонок таблиц.

Обязательно используйте ТОЛЬКО эти значения для фильтрации по указанным полям:

{db_column_values}

СХЕМА РЕШЕНИЯ



ОБОГАЩЕНИЕ SELECT'А. СХЕМА

Сколько проектов
в программе номер 1?

```
SELECT COUNT(*) AS project_count  
FROM project  
WHERE program_long_name IN ('Программа номер 1');
```



```
SELECT program_long_name, COUNT(*) AS project_count  
FROM project  
WHERE program_long_name IN ('Программа номер 1')
```

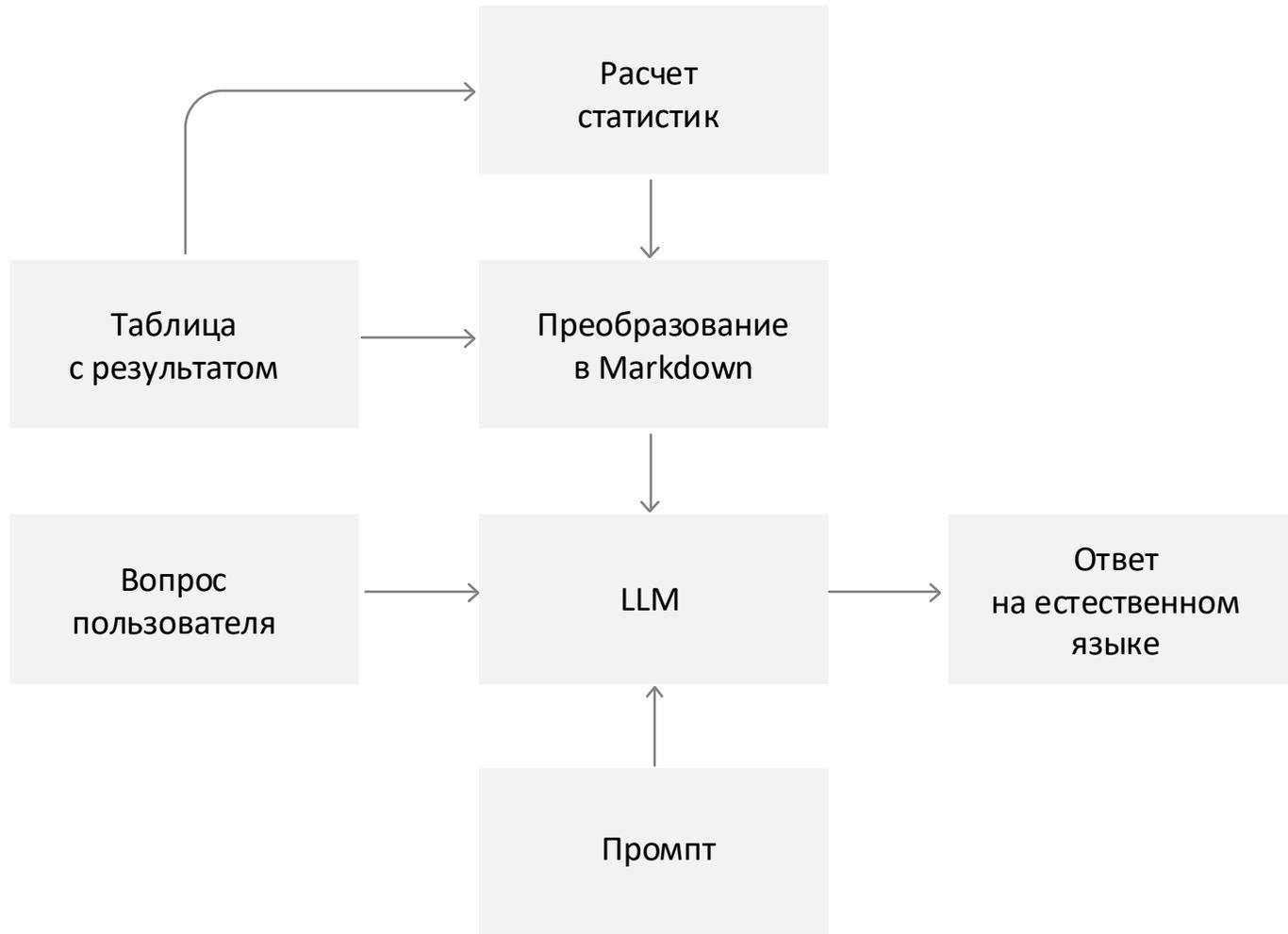
LLM

Атрибуты: program_long_name

ОБОГАЩЕНИЕ SELECT'А. РЕЗУЛЬТАТ

РЕЗУЛЬТАТЫ ИСПОЛНЕНИЯ SQL ДО ОБОГАЩЕНИЯ	РЕЗУЛЬТАТЫ ИСПОЛНЕНИЯ SQL ПОСЛЕ ОБОГАЩЕНИЯ
project_count	program_long_name project_count
5	Программа номер 1 5

ОТВЕТ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ



ОТВЕТ:

в программе «программа номер 1»
насчитывается 5 проектов

ОПИСАНИЕ ЗАПРОСА

ОПИСАНИЕ ЗАПРОСА

SQL-запрос использует одну таблицу: `project`, которая содержит информацию о проектах.

Из этой таблицы выбираются два поля: `program_long_name` - полное наименование программы.

`COUNT(*)` - количество проектов, сгруппированных по наименованию программы.

Фильтрация осуществляется по полю `program_long_name` с помощью условия `IN ('Программа номер 1')`.

Это позволяет выбрать все записи, где программа называется именно так.



ВЕРНО ЛИ СОСТАВЛЕН
SQL-ЗАПРОС?

ДОБАВЛЕНИЕ СИНОНИМОВ. РЕЗУЛЬТАТ (ПОВТОРЕНИЕ)

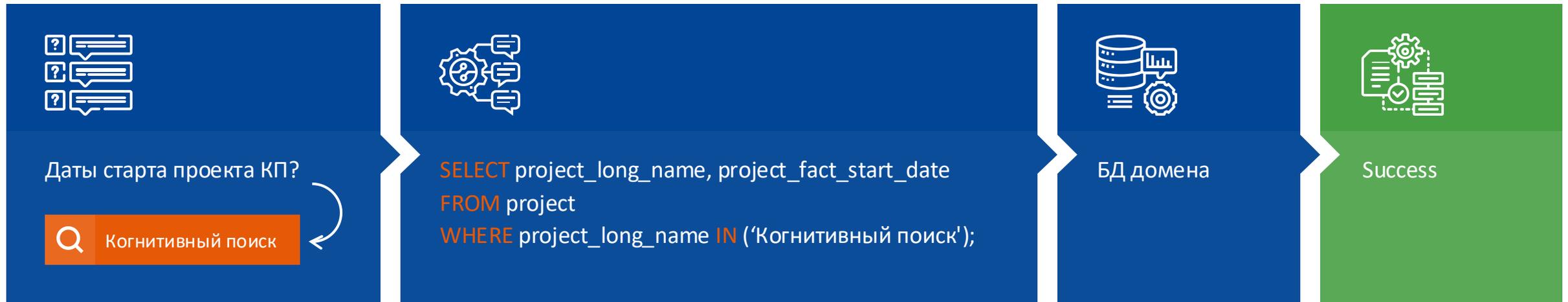
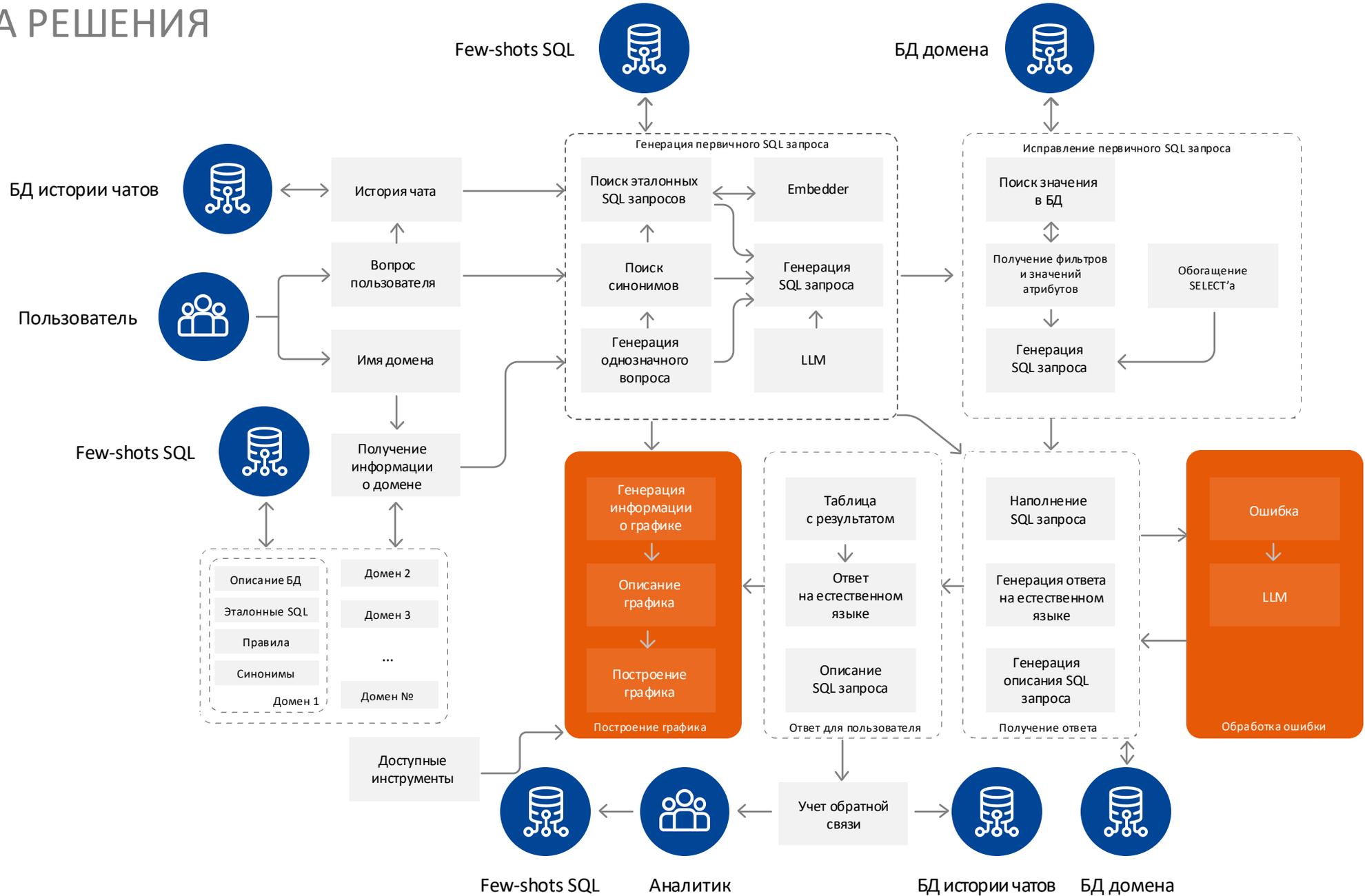
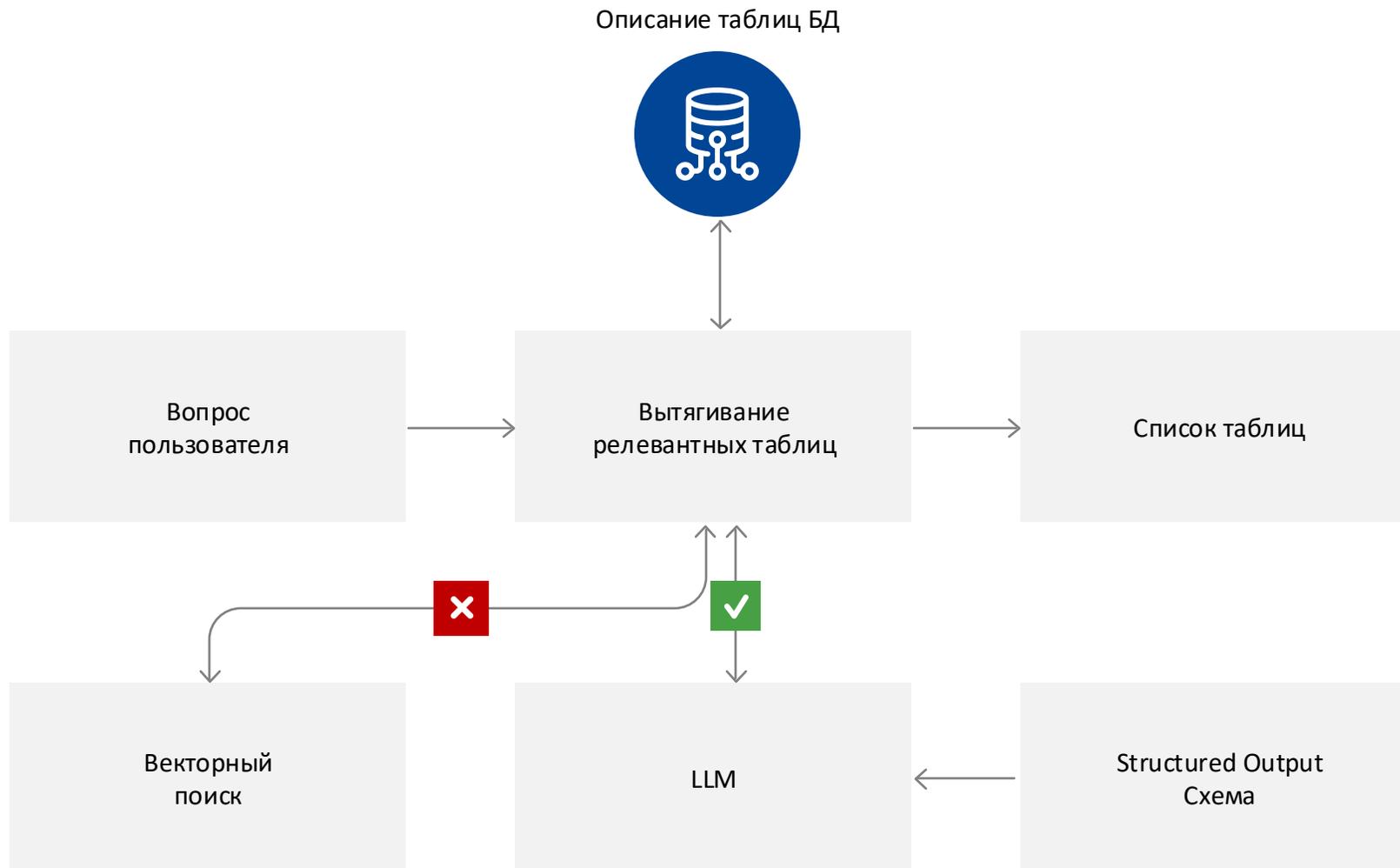


СХЕМА РЕШЕНИЯ



МАСШТАБИРОВАНИЕ СИСТЕМЫ



СРЕДА РАЗРАБОТКИ

Обезличенные данные

API (GigaChat,
YandexGPT, OpenAI)

ПРОДУКТИВНАЯ СРЕДА

Данные
продуктовой среды

Self-hosted LLM
(Ruadapt-14B)

GPU
(48/80 GB)

Inference Engine (VLLM)

ВЫВОДЫ

ВЫВОДЫ

- Классический подход к задаче Text2SQL не точен и сложно масштабируется
- Типовые решения с LLM можно взять за основу, но они плохо справляются со сложными данными
- Учесть специфику домена можно с помощью добавления синонимов и поиска по БД
- Пользовательский опыт важен, добавляем диалог, строим графики
- Большая цена ошибки – описываем пользователю как именно получили ответ



СПАСИБО ЗА ВНИМАНИЕ!
ГОТОВ ОТВЕТИТЬ НА
ВОПРОСЫ



ЕВГЕНИЙ

ВЫЛЕГЖАНИН

ГЛАВНЫЙ РАЗРАБОТЧИК И TEAM LEAD
КОМАНДЫ МАШИННОГО ОБУЧЕНИЯ (LLM)
В КОМПАНИИ ГАЗПРОМ НЕФТЬ –
ЦИФРОВЫЕ РЕШЕНИЯ