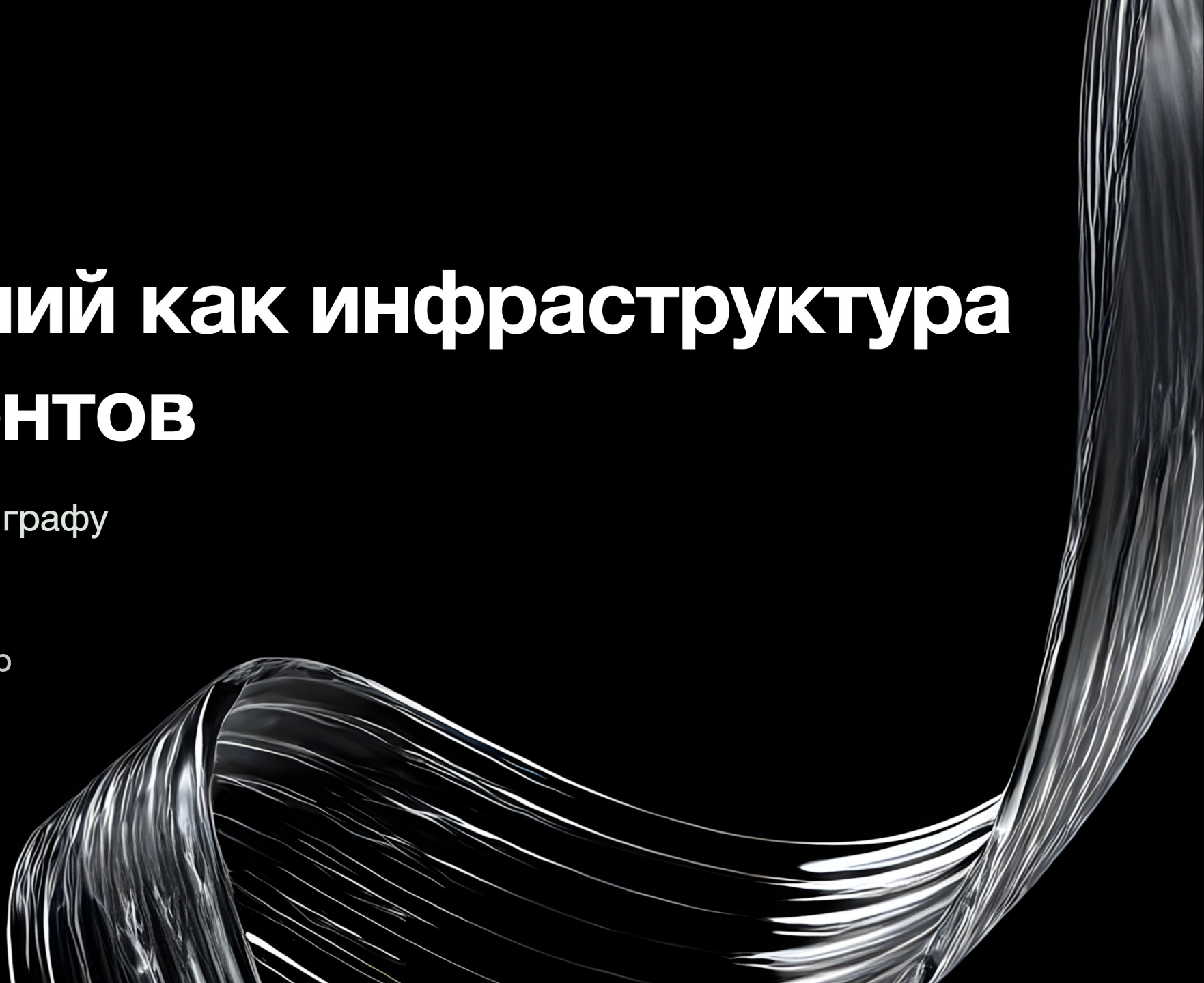


Граф знаний как инфраструктура для AI-агентов

От датасетов к единому графу

Александр Непочатых · Сбер

SmartData 2026



О себе

СБЕР

Lead Data Engineer

Как техлид, развивал платформу обработки и хранения данных для обучения фундаментальных моделей Сбера.

Как техлид и продакт, развивал граф знаний, как отдельный продукт.

В ПРОШЛОМ

Развивал хранилища данных в Авито, Делимобиле и Спортмастере.

В IT больше 10 лет.



План доклада

РАССКАЖУ

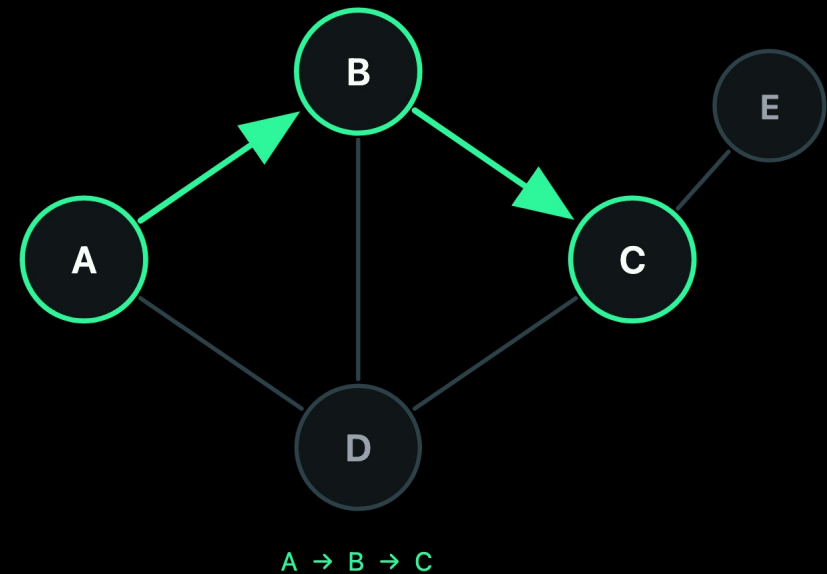
- Что такое граф
- Какую проблему мы решаем
- Почему сейчас: RAG, агенты
- Как строили текущее решение: LightRAG, технологии под капотом
- Бенчмарки и кейсы использования

ЧТО ВЫНЕСЕТЕ С ДОКЛАДА

- Когда векторного RAG недостаточно
- Как "вкатиться" в графовую тему
- Как быстро запустить графовое решение для тестирования

Что такое граф

- Граф — структура данных
- 2 сущности графа - вершины и ребра
- Вершины представляют объекты
- Ребра представляют связи между объектами
- Позволяет сделать обход графа от вершины к вершине по ребрам



Граф знаний

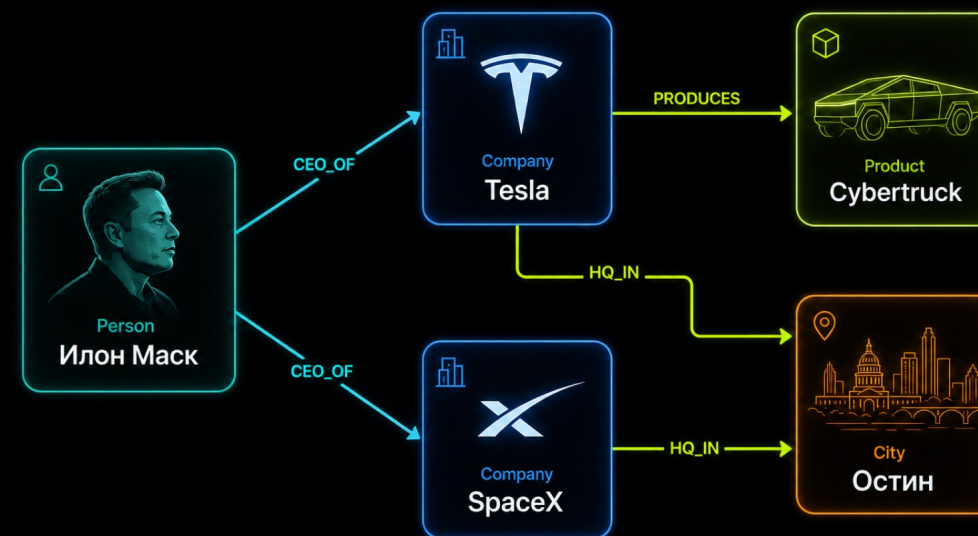
Граф знаний — граф, в котором:

- вершины обозначают сущности мира (люди, компании, продукты)
- рёбра — типизированные факты: **FOUNDED**, **CEO_OF**, **HQ_IN**

На узлах и рёбрах могут жить свойства.

Представляет собой модель предметной области

Можно обойти запросом



Где используются графы знаний



ЯНДЕКС

- Карточки объектов в Поиске
- Сущности в Картах
- Ответы Алисы



LINKEDIN

- Economic Graph: люди, компании, навыки, вакансии.
- Поиск работы и лента идут по рёбрам, не по словам в профиле.

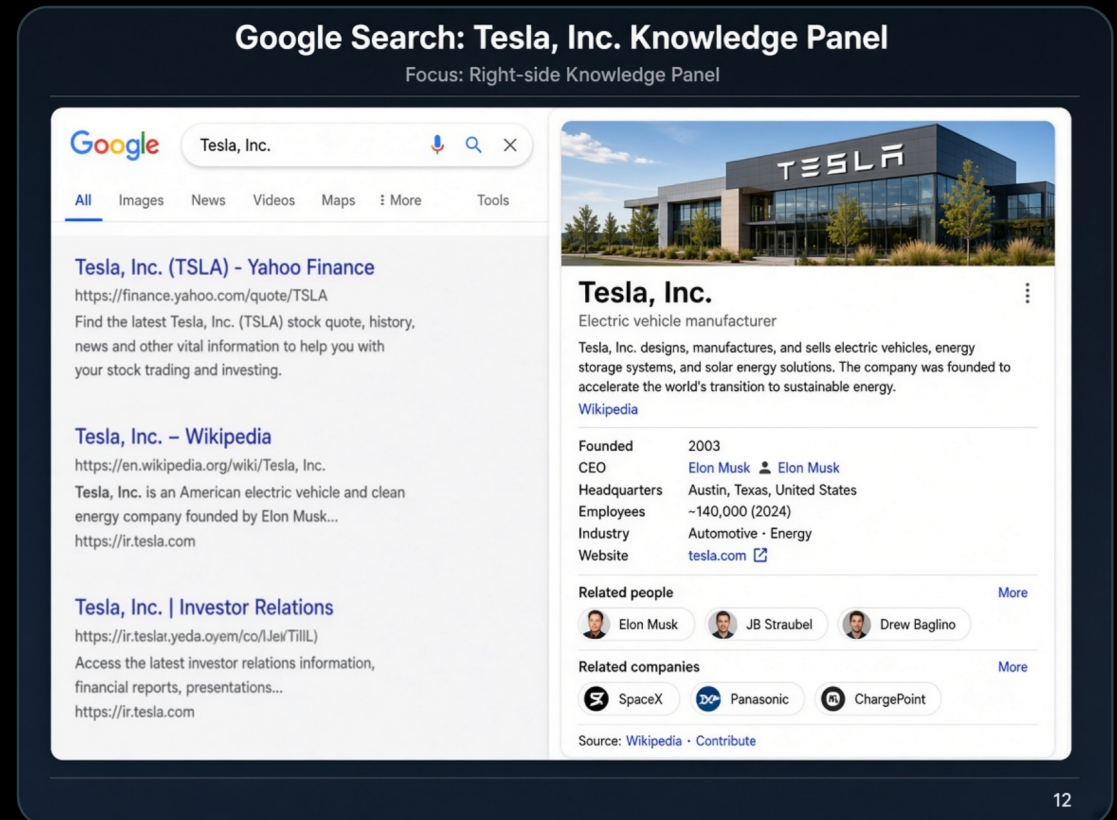


GOOGLE

- Knowledge Graph (развивается с 2012)
- Карточка в выдаче — параметры сущности

Правая часть страницы Google

- Knowledge Panel - ключевые характеристики сущности из запроса
- Принцип работы - по запросу в поисковой строке выделяется сущность, по которой отображаются ее основные свойства из графа



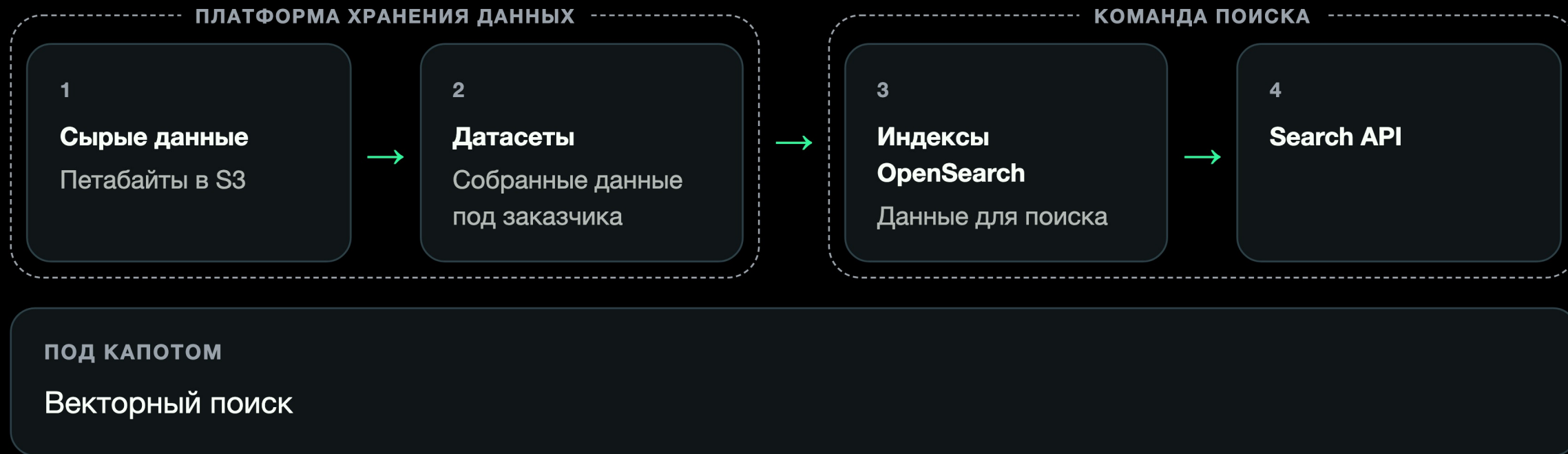
Какую проблему мы решаем

- В наличии - **70 ПБ** для обучения фундаментальных моделей
- Внутри - тексты, документы, картинки, видео, аудио
- Низкая добавочная стоимость самих данных
 - недостаточно связей, не хватает визуализации

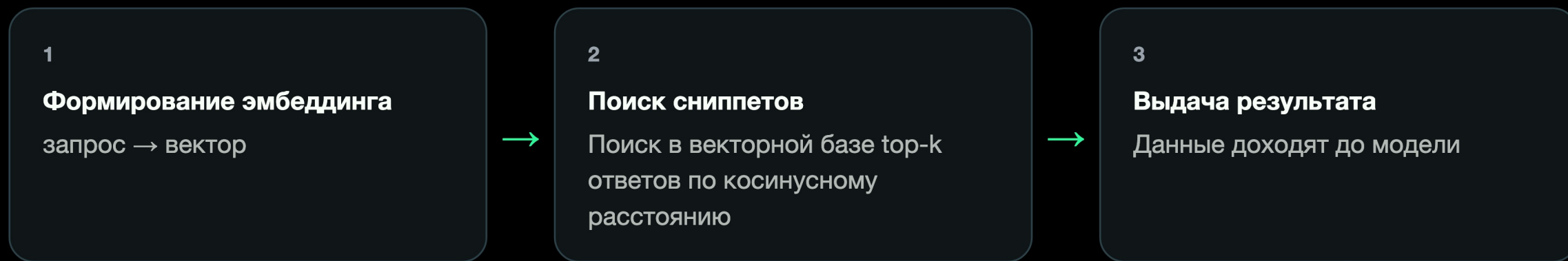


Путь потребления данных

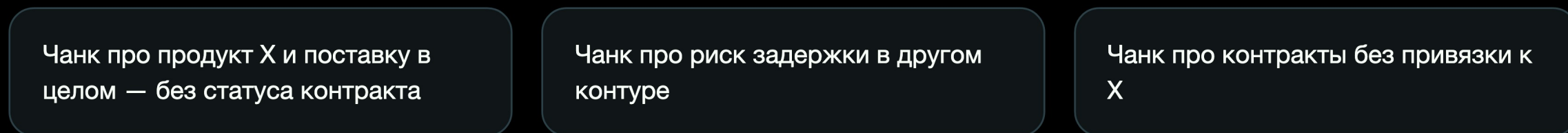
На примере векторного поиска на базе OpenSearch.



Как работает векторный поиск



Пример - Запрос: «Найди поставщиков продукта X с активным контрактом и риском задержки»



Вектор нашёл похожие формулировки. Связь «X → поставщик → активный контракт с риском» в чанках не собрана.

Гипотеза

Если есть

- данные
- потребители
- проработанные графовые подходы

то построенный граф может дать прорыв в дата-продуктах.

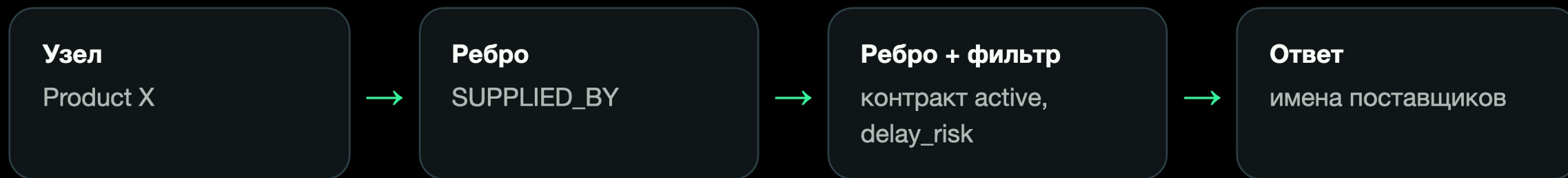
ПРОРЫВ

Качественно новое свойство информационного поиска на тех же данных, померянное на бенчмарках внутри дата-продукта

Например, цепочки фактов, которых нет в изолированных сниппетах

Как работает графовый поиск

Тот же запрос: поставщики продукта X с активным контрактом и риском задержки.



Есть одна проблема - специализированные графовые запросы

```
MATCH (p:Product {name: $x})-[:SUPPLIED_BY]->(s:Supplier)
      -[:HAS_CONTRACT]->(c:Contract {status: 'active'})
WHERE c.delay_risk = true
RETURN s.name, c.id
```

Два hop'a и фильтр по свойствам узла. Векторным top-k эту цепочку не собрать.

Векторный поиск vs графовый

	Вектор	Граф
Единица	чанк → эмбединг	сущность + типизированное ребро
Что находит	похожее по смыслу	связанное по факту
Фильтры	payload, rerank после top-k	свойства на узлах и рёбрах: контракт активен, риск есть
Multi-hop	пачка tool call, каждый — новый поиск	один MATCH по пути A→B→C
Слабое место	рвёт объект по границе чанка; «яблоко» ≈ Apple Inc.	нужен entity resolution; дороже построить
Сильное место	lookup, синонимы, нечёткое имя	соседи, ограничения, агрегации

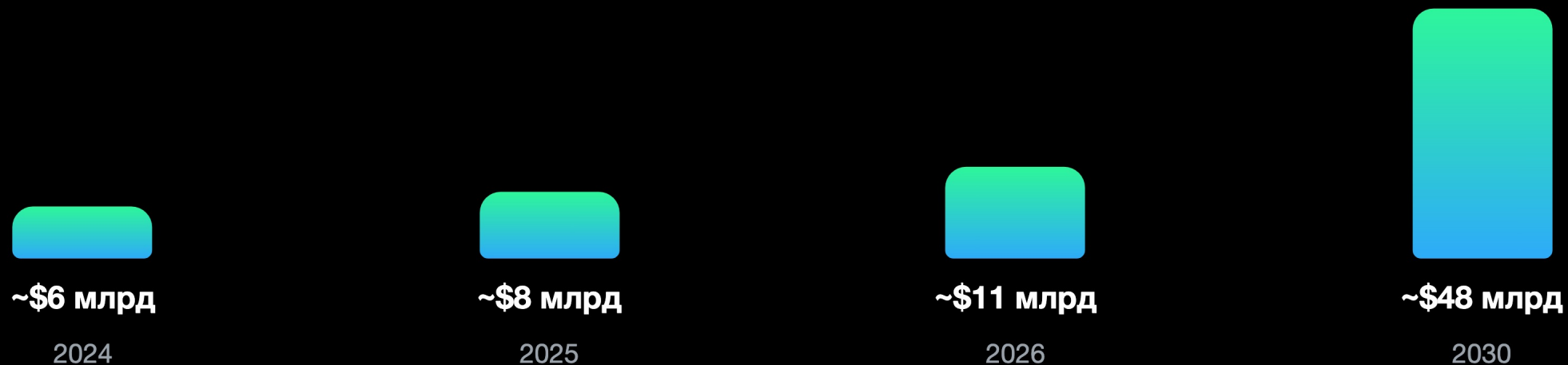
Это разные операции, не «граф вместо RAG».

**Вектор находит
похожее
Граф находит
связанное**

A high-contrast, black and white image featuring a dynamic splash of water. The water is captured in mid-motion, forming a large, circular, ring-like shape that dominates the center of the frame. The water's surface is highly reflective, with bright highlights and deep shadows that emphasize its fluid, swirling nature. The background is a solid, deep black, which makes the white and grey tones of the water stand out dramatically. The overall composition is abstract and energetic, suggesting a moment of peak action or a significant turning point.

Почему именно сейчас

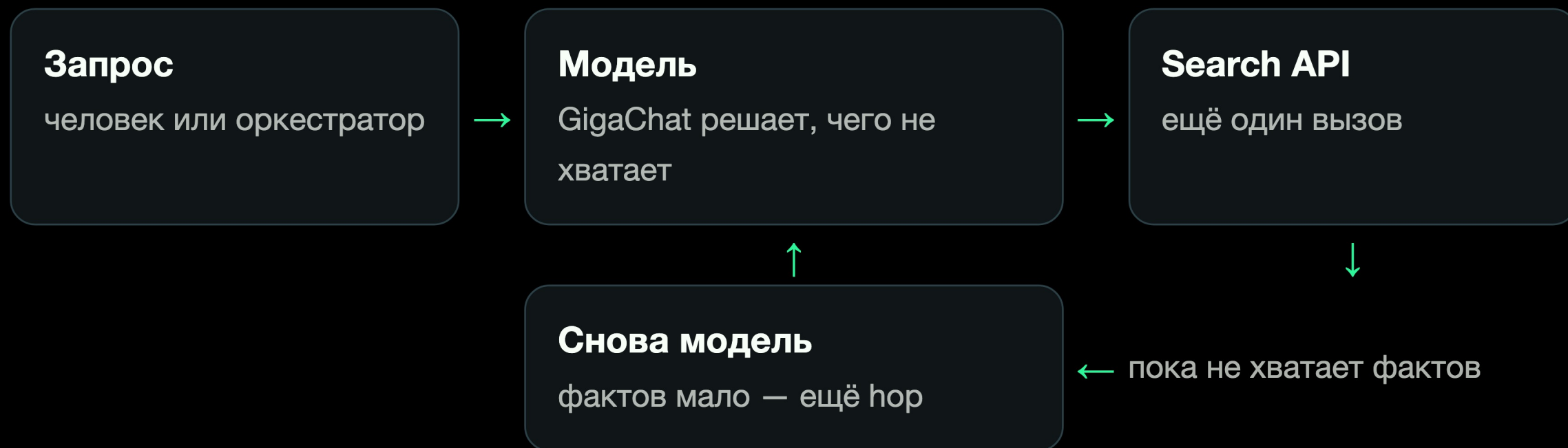
Рынок агентов растёт не линейно



2024, 2025, 2030 — BCC Research, AI Agents: Technologies, Applications and Global Markets (\$5.7 / \$8 / \$48.3 млрд)

2026 — Grand View Research, AI Agents Market (\$10.9 млрд)

Search API: кратный рост запросов за год



Один пользовательский запрос превращается в пачку вызовов Search API. Модель и повторный hop — один цикл: снова бьёт в тот же API.

Рост количества запросов в 2.5 раза за последний год.

Почему именно сейчас

ДАННЫЕ

Продукт «граф знаний» родился в команде хранилища

70 ПБ, сформированный трек потребления данных

ВЫЧИСЛЕНИЯ

GPU для построения графа (H100, A100)

Для тестов - сервисы Cloud Evolution: ML Inference, Foundation Models

СПРОС

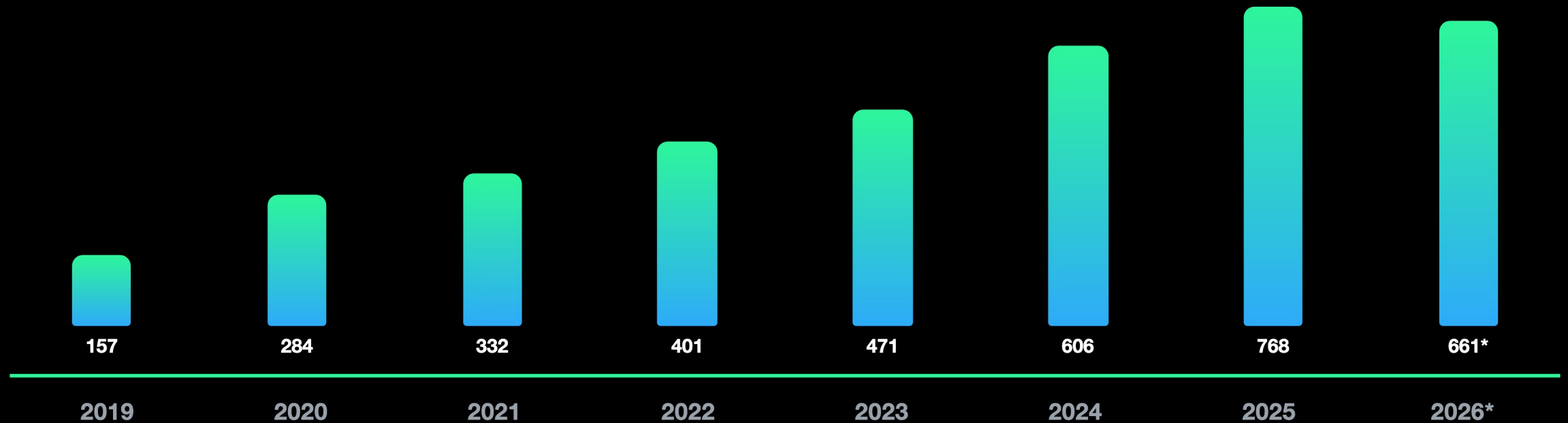
Агентам нужен структурированный контекст. Рост потребления Search API за счёт GigaChat

ПОДХОДЫ

Microsoft GraphRAG, LightRAG и облегчённые варианты. Можно не изобретать экстракцию с нуля

arXiv · knowledge graph

Статьи с «knowledge graph» в заголовке. Долгий рост темы, не вспышка одного фреймворка.



Как считали: поиск arXiv по заголовку, фраза «knowledge graph» 2026* — январь–сентябрь, поэтому столбец ниже полного 2025.

An abstract, high-contrast black and white image featuring a dynamic splash of water. The water forms a large, circular loop in the center, with several smaller droplets and ripples visible on its surface. The background is solid black, which makes the bright, reflective water stand out. The overall composition is fluid and energetic.

Как мы строили наше решение

Хранилище и технологии

Динамика развития индустрии: как быстро меняются подходы и как на это влияют LLM.

	2024	2025	2026
ХРАНИЛИЩЕ	spaCy + CommonCrawl "Ручное" построение графа	Графовый поиск на базе Wikidata, YAGO Использование внешних графов для улучшения информационного поиска	Гибридные графовые решения Доменные графы + API для агентов
ТЕХНОЛОГИИ	GraphRAG · LightRAG апрель — GraphRAG, октябрь - LightRAG	Рост популярности GraphRAG GraphRAG в тексте: 7 → 82 статьи	Зрелость стека 148 papers GraphRAG за янв–сен

Как работает spaCy

Библиотека NLP (explosion.ai): токенизация, морфология, NER, деревья зависимостей. Граф знаний сама не строит.

Илон Маск

PERSON

на машине

Tesla

PRODUCT

доехал до штаб-квартиры

SpaceX

ORG

.

- Модель размечает текст метками: PERSON, PRODUCT, ORG — это имена, не факты
- Связи рисуем правилами: PERSON рядом с PRODUCT → **DRIVES** ; PERSON рядом с ORG → **VISITED**
- На CommonCrawl это дало граф, но шумный: нет смысла отношения, только соседство меток
- Без LLM нет типа ребра «основал / работает в / поставляет» — только эвристика

Wikontic



Статья от ноября 2025

Принцип работы — граф из текста строится в онтологии Wikidata

Идея — на готовый дамп Wikidata добавлять результаты обработки датасетов хранилища

1

Формирование кандидатов-триплетов с метаданными: субъект — отношение — объект

2

Проверка онтологии — типы сущностей и допустимые связи Wikidata, иначе ребро отбрасывается или переписывается

3

Дедупликация: склеить «NYC» и «New York City»

Результат — на наших данных качество полученного графа не устроило

Microsoft GraphRAG



Разработан Microsoft Research.

Сдвиг в парадигме - как неструктурированный корпоративный текст превращают в пригодный для поиска, рассуждения и объяснения слой знаний.

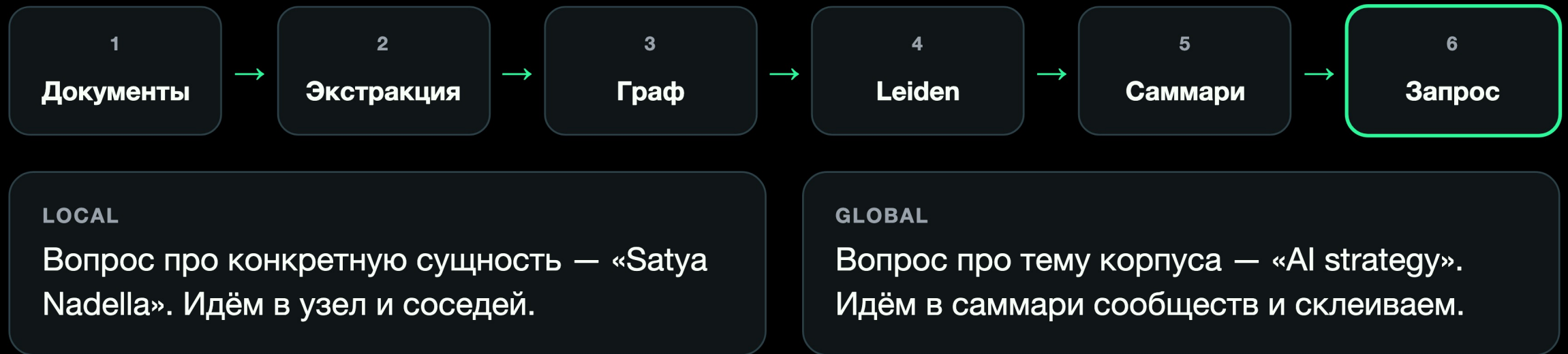
ЧТО ЭТО

LLM достаёт сущности и отношения из корпуса, складывает граф связей, режет его алгоритмом Leiden на сообщества и пишет саммари каждого сообщества.

КАКУЮ ЗАДАЧУ РЕШАЕТ

Не «кто такой X», а вопрос про весь корпус: «какие темы здесь есть, что связано с чем». Локальный поиск плюс глобальный обзор, которого у чистого векторного RAG нет.

Microsoft GraphRAG



Local нам нужен. Global через саммари — не для корпуса, который каждый день растёт.

LightRAG



Разработан University of Hong Kong, октябрь 2024

ЧТО ЭТО

Graph RAG с двухуровневым поиском: по сущностям и соседям и по темам

Граф и векторы формируются единым пайплайном

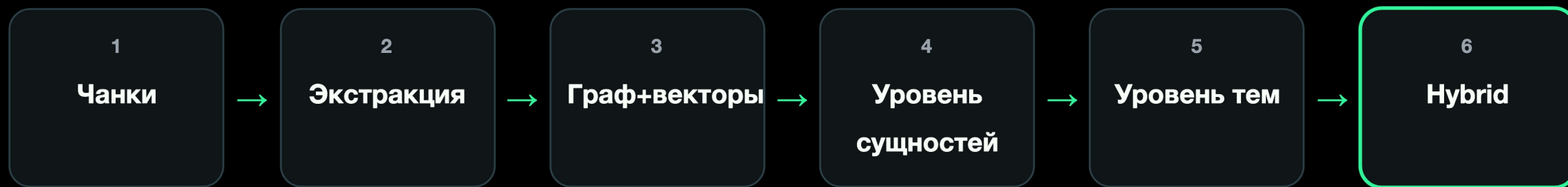
КАКУЮ ЗАДАЧУ РЕШАЕТ

Та же идея, что у Microsoft GraphRAG — ответы по связанным фактам, не только по похожим чанкам — но дешевле в построении и с живым инкрементом.

Без Leiden и саммари всех сообществ на каждый апдейт.

Основное отличие - инкрементальное наполнение графа

LightRAG



HYBRID

Склеиваем сущности низкого уровня и тематический подграф, затем rerank по релевантности и разнообразию.

MIX

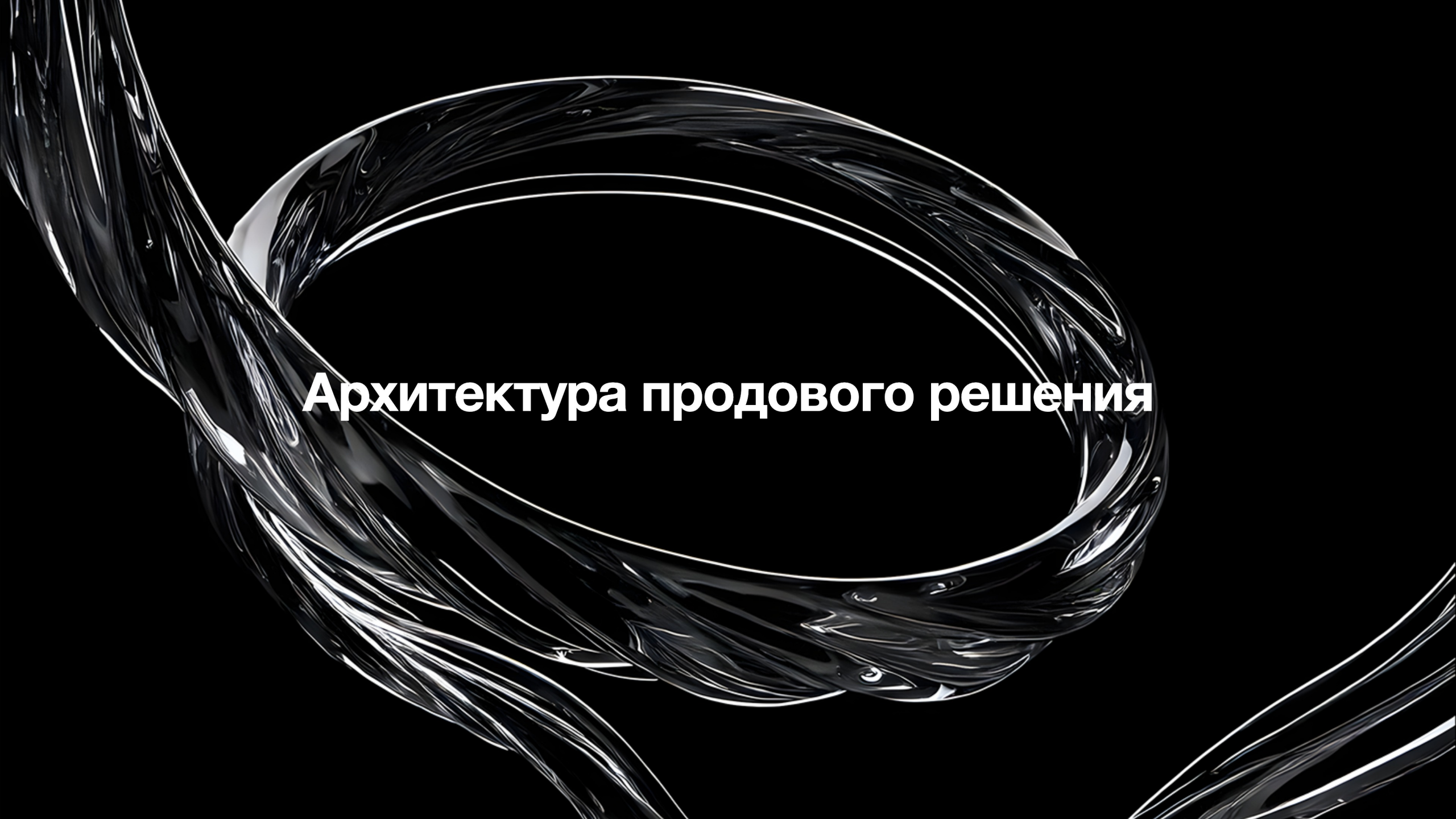
К гибриду подмешиваем обычный векторный поиск по чанкам. Агент выбирает режим, не мы зашиваем один.

GraphRAG vs LightRAG

	Microsoft GraphRAG	LightRAG
Индексация	граф → Leiden → саммари сообществ	граф + ключи двух уровней + векторы
Запрос	local по соседям / global по саммари сообществ	local / global по темам / hybrid / naive / mix
Инкремент	часто перестройка сообществ и саммари	дописали чанк, узлы и эмбеддинги
Цена на живом корпусе	высокая	приемлемая — наш выбор

Выбор инструмента: почему LightRAG

Инструмент	Экстракция	Граф	Инкремент	Комментарий
sraCy	только NER	✗	⚠	граф по правилам, низкая ценность результата
Wikontic	LLM + онтология Wikidata	✓	✓	на наших данных качество графа не устроило
GraphRAG	LLM-based	✓	✗ дорого	дорогое и долгое построение, больно на инкременте
LightRAG	LLM-based	✓	✓	Наш выбор: dual-level retrieval, дешёвые апдейты

An abstract, high-contrast black and white image featuring a dynamic splash of water. The water forms a large, swirling loop that dominates the center of the frame. The surface of the water is highly reflective, with bright highlights and deep shadows that emphasize its fluid, organic texture. The background is a solid, deep black, which makes the white highlights on the water stand out. The overall composition is energetic and modern.

Архитектура продового решения

Какие графовые базы сравнивали

Два сценария: быстрый RAG и «склад» сущностей. Шорт-лист и выбор — в статье на Хабре.

Решение	Сценарий	Почему смотрели	Итог
Apache Jena Fuseki	старт, RDF YAGO	быстро залить дампы	count(*) уезжал на десятки минут
FalkorDB	RAG	MPP in-memory, GraphRAG	доменные графы в проде
Memgraph	RAG	in-memory, быстрое чтение, Cypher	доменные графы в проде
NebulaGraph	хранилище	шарды, Tencent-scale	не взяли под агентный RAG
JanusGraph	хранилище	надстройка над Cassandra	не взяли под агентный RAG

habr.com/ru/companies/sberbank/articles/1029580

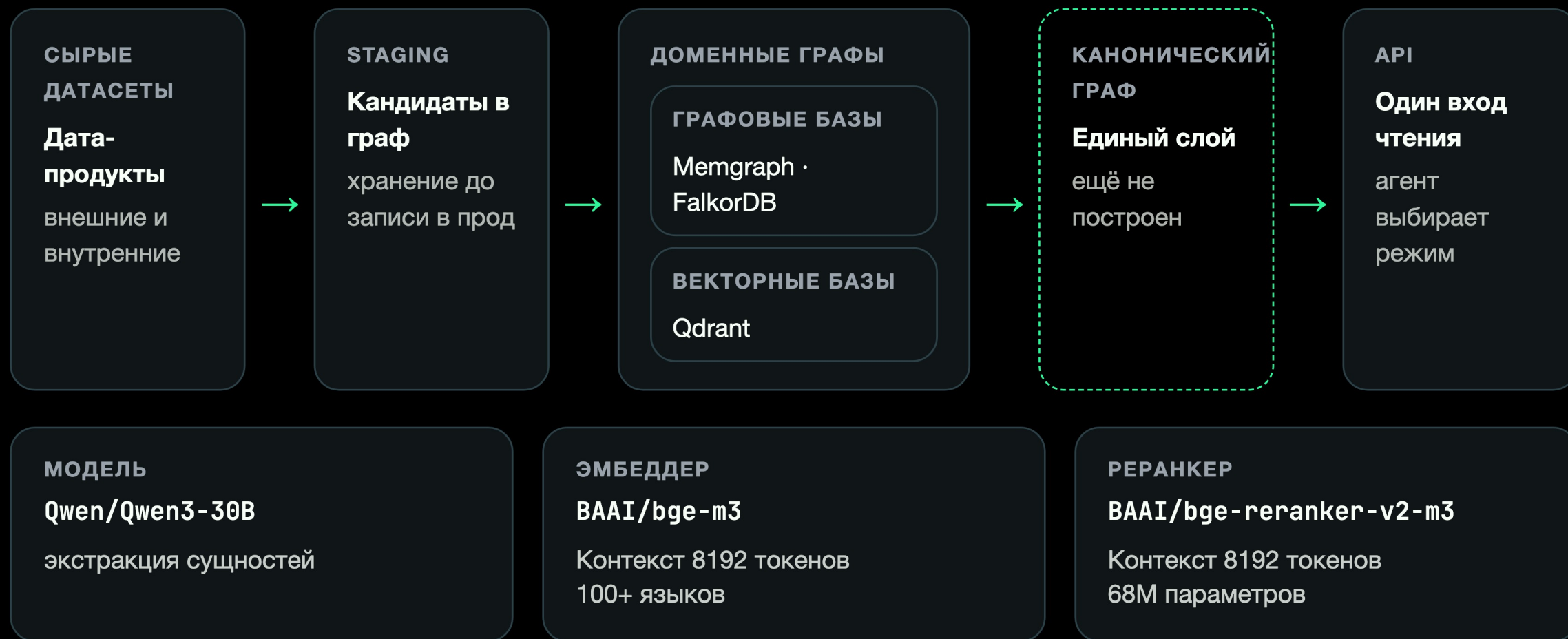


Почему рядом Qdrant, а не Milvus

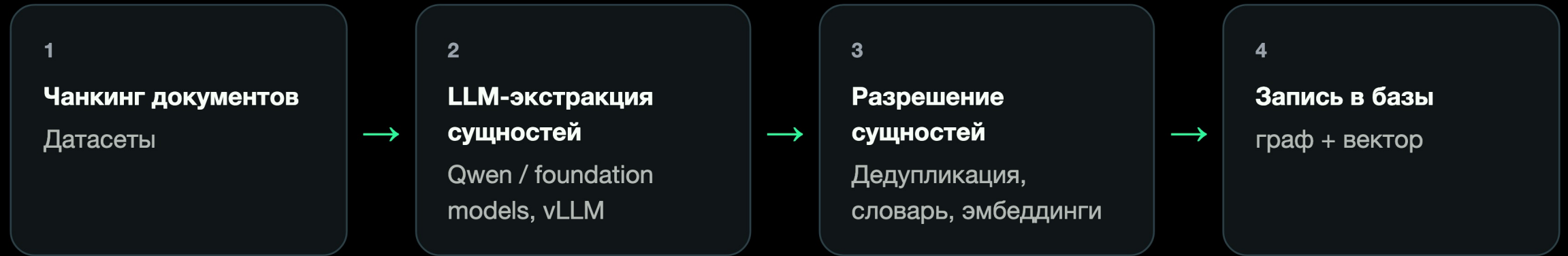
Вектор нужен отдельно от графовой базы: mix LightRAG и entity resolution берут кандидата здесь, граф подтверждает ребро.

	Qdrant	Milvus	pgvector
Стек	Rust, один сервис	etcd + MinIO + брокер	расширение Postgres
Фильтры	HNSW + payload в одном обходе	есть, тяжелее в ops	слабее на большом ANN
Для чего	self-host, метаданные: тип, домен	сотни миллионов векторов	уже живём в Postgres
Наш кейс	выбрали	избыточный кластер	не тот контур

Архитектура: от датасетов к единому графу



Как строим граф



Единый API чтения для агента

ПОИСК СУЩНОСТИ LOCAL

Конкретные сущности и соседи. «Кто контрагент этой сделки?»

ПОИСК ТЕМ GLOBAL

Тематический подграф. «Что происходит в этом домене?»

HYBRID

local + global режимы.

NAIVE

Классический векторный поиск.

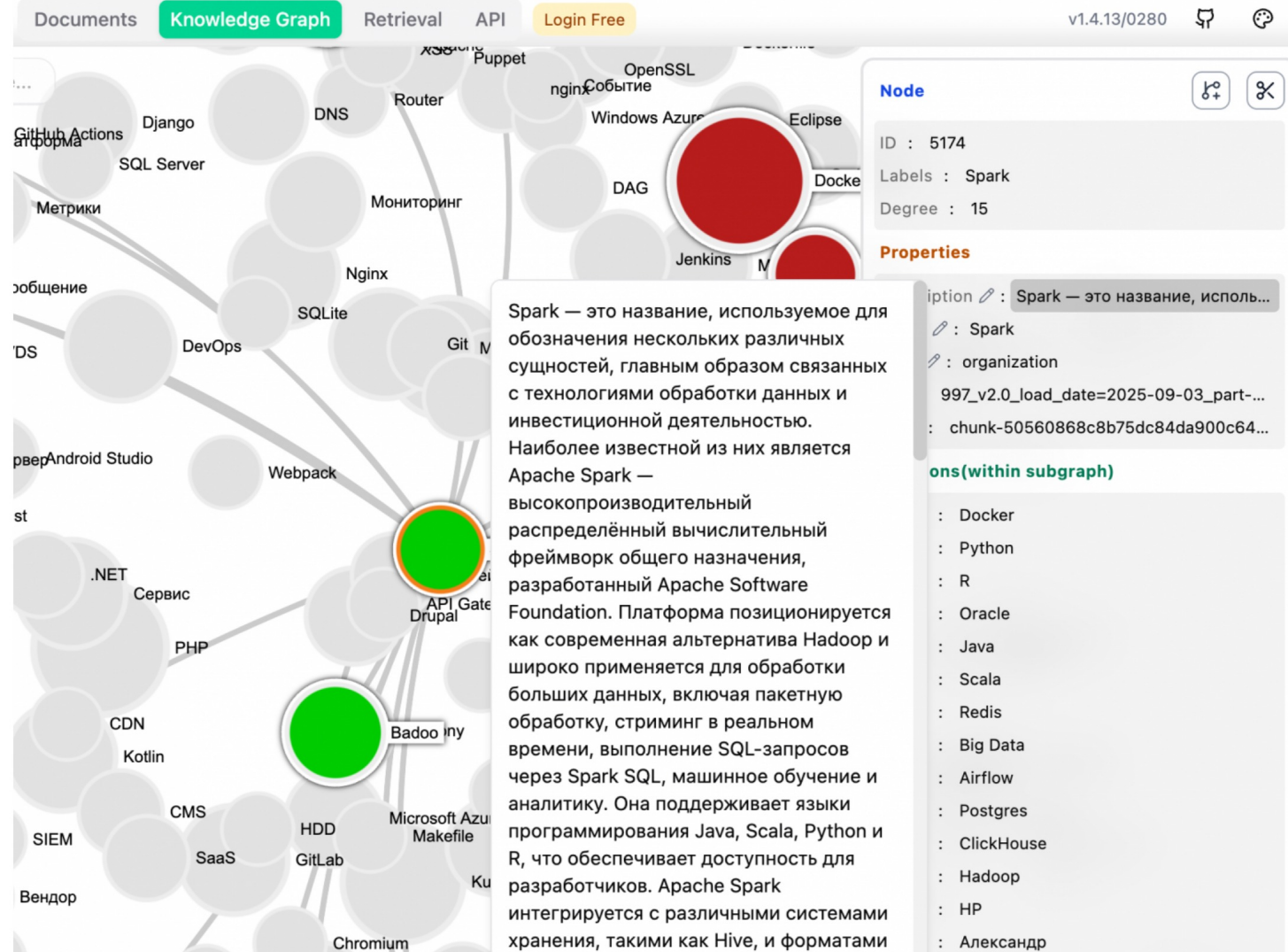
MIX

Граф + вектор. Агент решает, как достать ответ.

На что обратить внимание

- Разрешение сущностей
 - «Сбер» и «Сбербанк» - разные слова, одна сущность
 - Нужны дополнительные проверки: словарь, эмбединги
- Галлюцинации экстрактора - confidence с порогом на запись
- Рост масштаба - деградация производительности и качества на бенчмарках
- Инкремент

Пример - граф технологий



An abstract, high-contrast black and white image featuring a dynamic splash of water. The water forms a large, swirling loop that dominates the center of the frame. The surface of the water is highly reflective, with bright highlights and deep shadows that emphasize its fluid, turbulent motion. The background is a solid, deep black, which makes the white highlights on the water's surface stand out sharply. The overall composition is energetic and visually striking.

Результаты

ruSimpleQA: чем мерим поиск

Русский перевод SimpleQA — бенчмарка OpenAI. 4326 коротких фактологических вопросов, у каждого один проверяемый ответ.

ПРИМЕРЫ

Кто получил IEEE Frank Rosenblatt Award в 2010?

Michio Sugeno

В какой день альбом Carrie Underwood «Cry Pretty» получил золото RIAA?

23 октября 2018

ОСОБЕННОСТИ

- Короткий вопрос, один ответ — легко судить
- Adversarial против GPT-4: человеку просто, модели нет
- Ответ не должен меняться со временем
- Оценка: correct / incorrect / not attempted

Выбран, чтобы оценивать качество поисковой выдачи — общий язык с сервисом информационного поиска, не внутренний датасет команды графа.

Результат · ruSimpleQA

+12 п.п. на полном наборе из 4326 вопросов по сравнению с результатами Search API.

Методика оценки LLM-as-a-Judge

LightRAG верно ответил на 74% вопросов, на которых поиск молчал.

Технология работает на фактологических вопросах.

habr.com/ru/companies/sberbank/articles/1029580

Как мы улучшали поиск цифровых талантов

В базу легли открытые тексты про людей и компетенции: новости, Хабр, статьи arXiv.

НОВОСТИ

Кто где выступает, какие команды и продукты мелькают рядом.

ХАБР

Статьи, стек, проекты — рёбра «написал / работает с / в команде».

ARXIV

Публикации: темы, соавторы, область.

Из этого корпуса LightRAG собрал граф людей, навыков и проектов.

Планы на будущее

- Ускорение ответа по графу: агент не должен ждать обход
- Канонический граф поверх доменных - синергия знаний
- Ускорение построения: сейчас порядка 200 документов в час — до near real-time очень далеко

Три вывода

1 · ПОСМОТРИТЕ НА LIGHTRAG

Быстрый старт.

Удобный UI из коробки.

Большое коммьюнити.

2 · ВЕКТОР И ГРАФ — НЕ "ИЛИ", А ВМЕСТЕ

Вектор находит похожее. Граф находит связанное. Multi-hop — зона графа; простой поиск вектор закрывает.

3 · ДВА РЕЗУЛЬТАТА, КОТОРЫЕ УНОСИТЕ

+12 п.п. на ruSimpleQA и кейс поиска цифровых талантов: граф повышает доступность данных и меняет способ потребления данных.

Что почитать о нашей работе

ХРАНИЛИЩЕ

Как мы строили хранилище на 70 ПБ данных и не планируем останавливаться

habr.com/ru/companies/sberbank/articles/972078



ГРАФ ЗНАНИЙ

Как мы улучшали качество поиска с помощью графа знаний и что из этого вышло

habr.com/ru/companies/sberbank/articles/1029580



БЕНЧМАРКИНГ АГЕНТОВ

Свой инструмент для бенчмаркинга ИИ-агентов: архитектура, надёжность и интеграция с Airflow

habr.com/ru/companies/sberbank/articles/1045586



Как быстро вкатиться в тему

Поднять LightRAG — `docker compose up`

Две статьи от разработчиков

LIGHTRAG

Двухуровневый поиск: сущности и темы.
Инкремент без перестройки графа.

github.com/HKUDS/LightRAG

[arXiv 2410.05779](https://arxiv.org/abs/2410.05779)

MICROSOFT GRAPHRAG

От локального lookup к обзору корпуса.
Leiden и саммари сообществ.

github.com/microsoft/graphrag

[arXiv 2404.16130](https://arxiv.org/abs/2404.16130)

microsoft.github.io/graphrag

Спасибо за внимание!

Александр Непочатых · Сбер

Telegram

t.me/anepochatykh

SmartData 2026

