

OT TEXT-TO-SQL K TRUSTED ANALYTICS

> Строим on-prem семантический слой для AI-агентов

Игорь Дмитриев | CDO | Smart Data 2026

Давайте знакомиться

Меня зовут Игорь

- “готовлю” данные для BI, ML и AI
- выстраиваю процессы DG
- проектирую Data и ML платформы
- осознанно внедряю AI



Text to SQL и его проблемы



Бизнес-Аналитик



Почему COUNT (active_user) = 1500?

Text to SQL и его проблемы



> Сколько активных клиентов у нас?



```
SELECT COUNT(*)  
FROM users  
WHERE status = 'active'
```



Бизнес-Аналитик



Почему COUNT (active_user) = 1500?

Text to SQL и его проблемы



> Сколько активных клиентов у нас?



```
SELECT COUNT(*)  
FROM users  
WHERE status = 'active'
```



Бизнес-Аналитик



Почему COUNT (active_user) = 1500?

Финансовый отдел считает активными только тех, кто сделал транзакцию за 30 дней, их 800!

Цифры на столе у директора не бьются!

Text to SQL и его проблемы



> Сколько активных клиентов у нас?



```
SELECT COUNT(*)  
FROM users  
WHERE status = 'active'
```



LLM: от 60% до 80% точности*



Бизнес-Аналитик



Почему COUNT (active_user) = 1500?

Финансовый отдел считает активными только тех, кто сделал транзакцию за 30 дней, их 800!

Цифры на столе у директора не бьются!

* Независимые измерения на базе бенчмарка BIRD-SQL

Text to SQL и его проблемы



> Сколько активных клиентов у нас?



```
SELECT COUNT(*)  
FROM users  
WHERE status = 'active'
```



LLM: от 60% до 80% точности*



Human: ~95% точности*



Бизнес-Аналитик



Почему COUNT (active_user) = 1500?

Финансовый отдел считает активными только тех, кто сделал транзакцию за 30 дней, их 800!

Цифры на столе у директора не бьются!

* Независимые измерения на базе бенчмарка BIRD-SQL

Почему Thinking-токены не спасают отчетность

Reasoning models

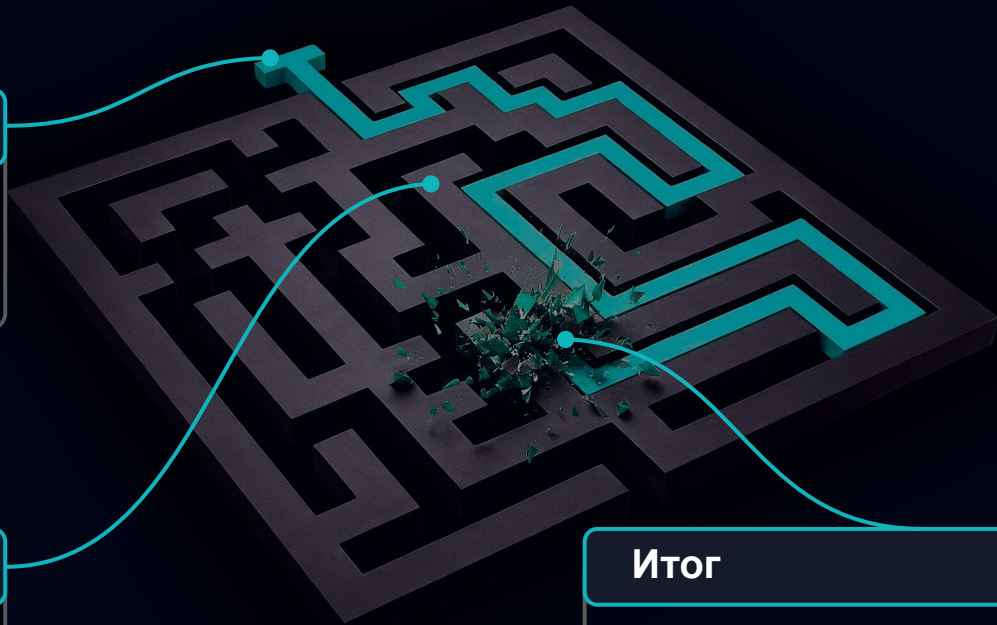
Модели вроде DeepSeek, GLM или Qwen обладают мощным аналитическим аппаратом

Умножение хаоса

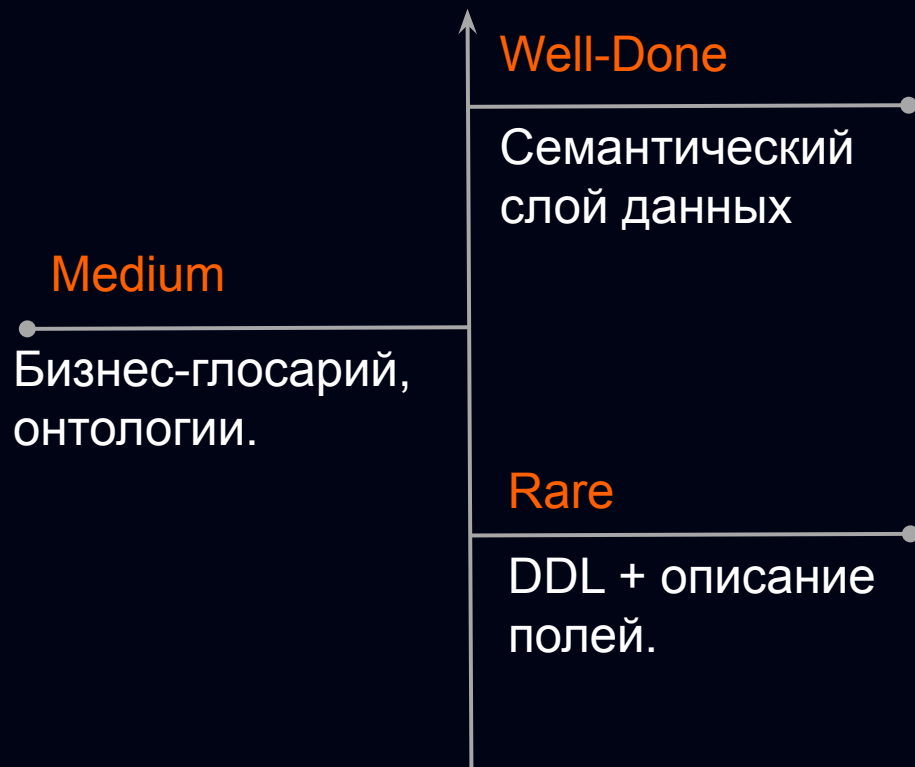
Без предметной области они используют свой интеллект для глубокого размышления над неверными предположениями

Итог

Интеллект модели не компенсирует отсутствие бизнес-контекста. Агент просто быстрее и увереннее заходит в аналитический тупик.



Уровни “прожарки” данных для AI и BI



От слепого генератора к проводнику в мир данных

Text-to-SQL

Угадывание связей таблиц
(Галлюцинации)

Точность: ~ 60%
на реальных базах

Ad-hoc
(поиск инсайтов)

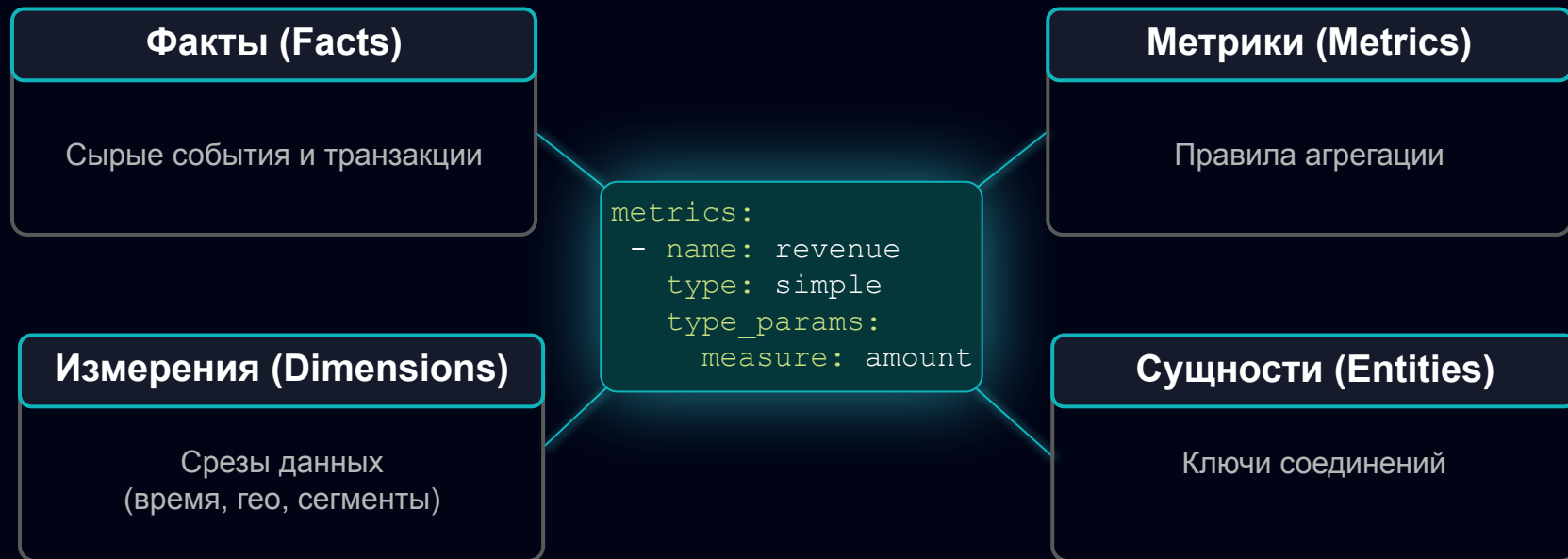
Semantic Query Engine

Детерминированный выбор
метрик из реестра

Точность: 100%
на задокументированных сущностях

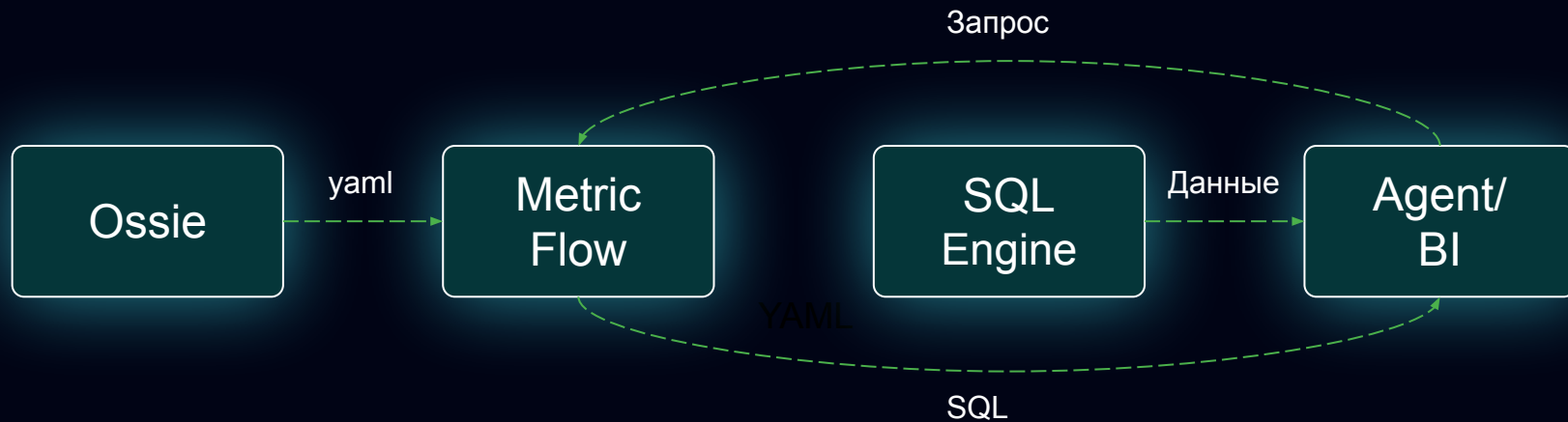
Trusted Analytics
(управленческая отчетность)

Apache Ossie (ранее Open Semantic Interchange)



Переход от логики, зашитой в проприетарный BI, к vendor-agnostic слою.
Одинаково понятно и Data-инженеру и ИИ-агенту

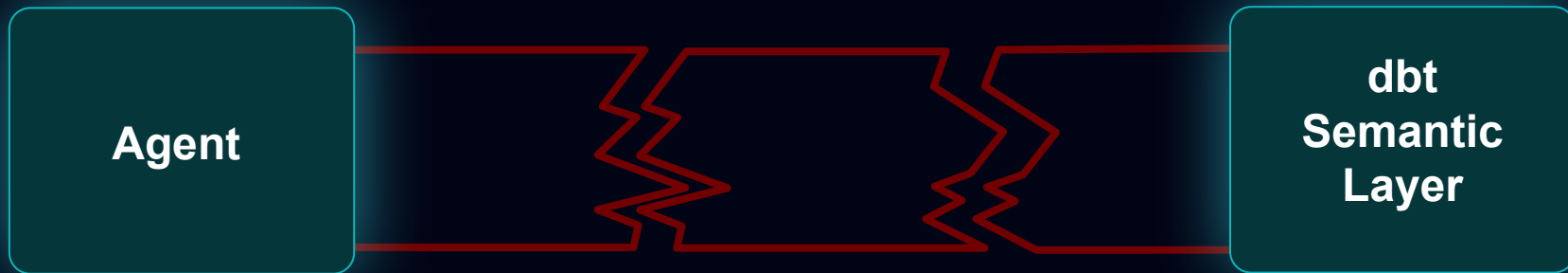
Движок вычислений семантического слоя Metric Flow



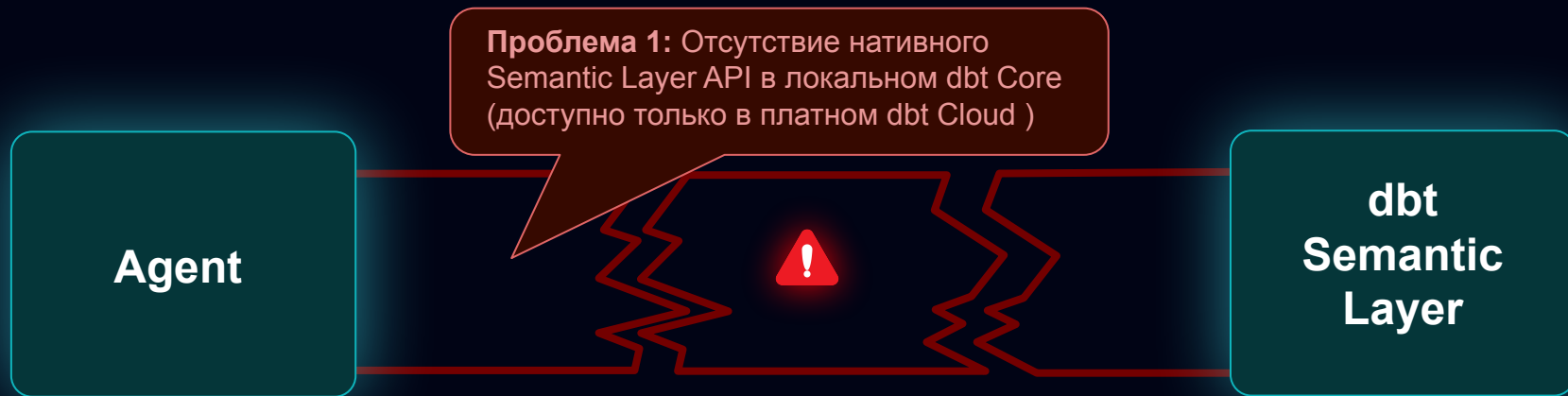
Что делает: Принимает запрос вида **Metric** + **Dimension**, строит граф зависимостей и генерирует корректный SQL под нужный диалект.

Где живет: В ядре dbt, доступен как Open Source движок.

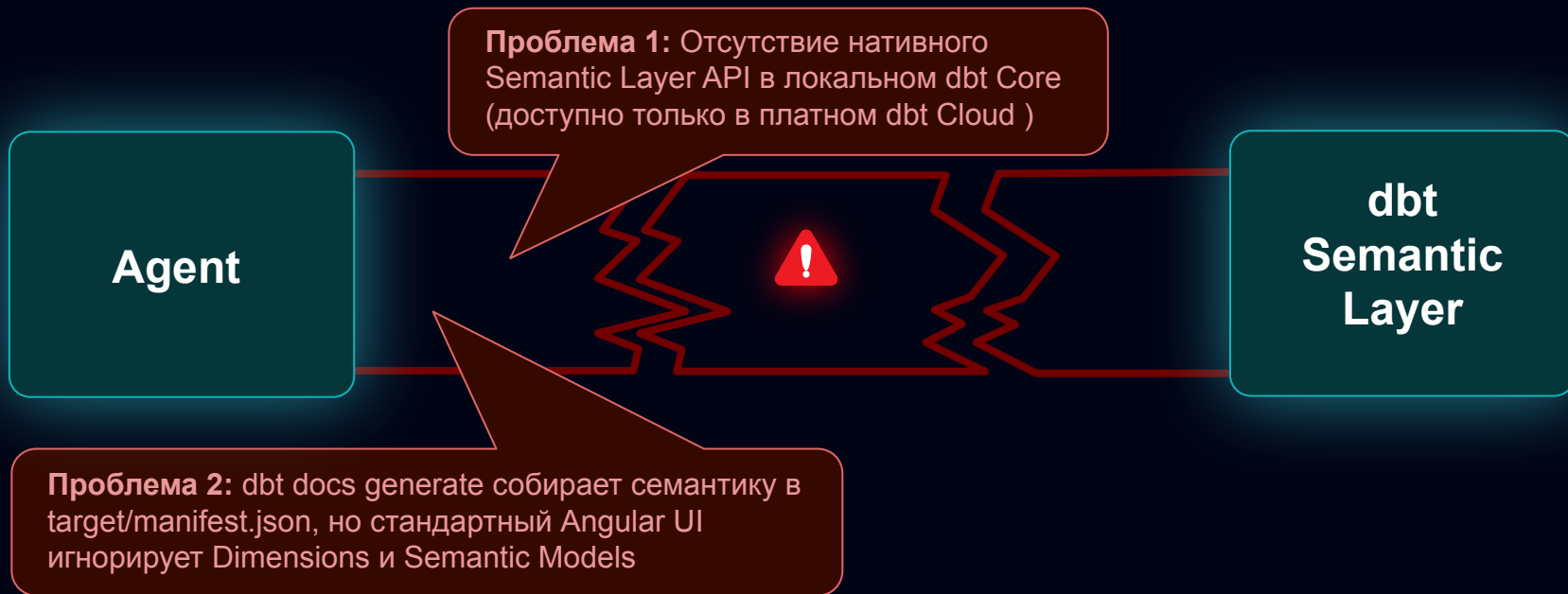
Почему это не работает из коробки



Почему это не работает из коробки



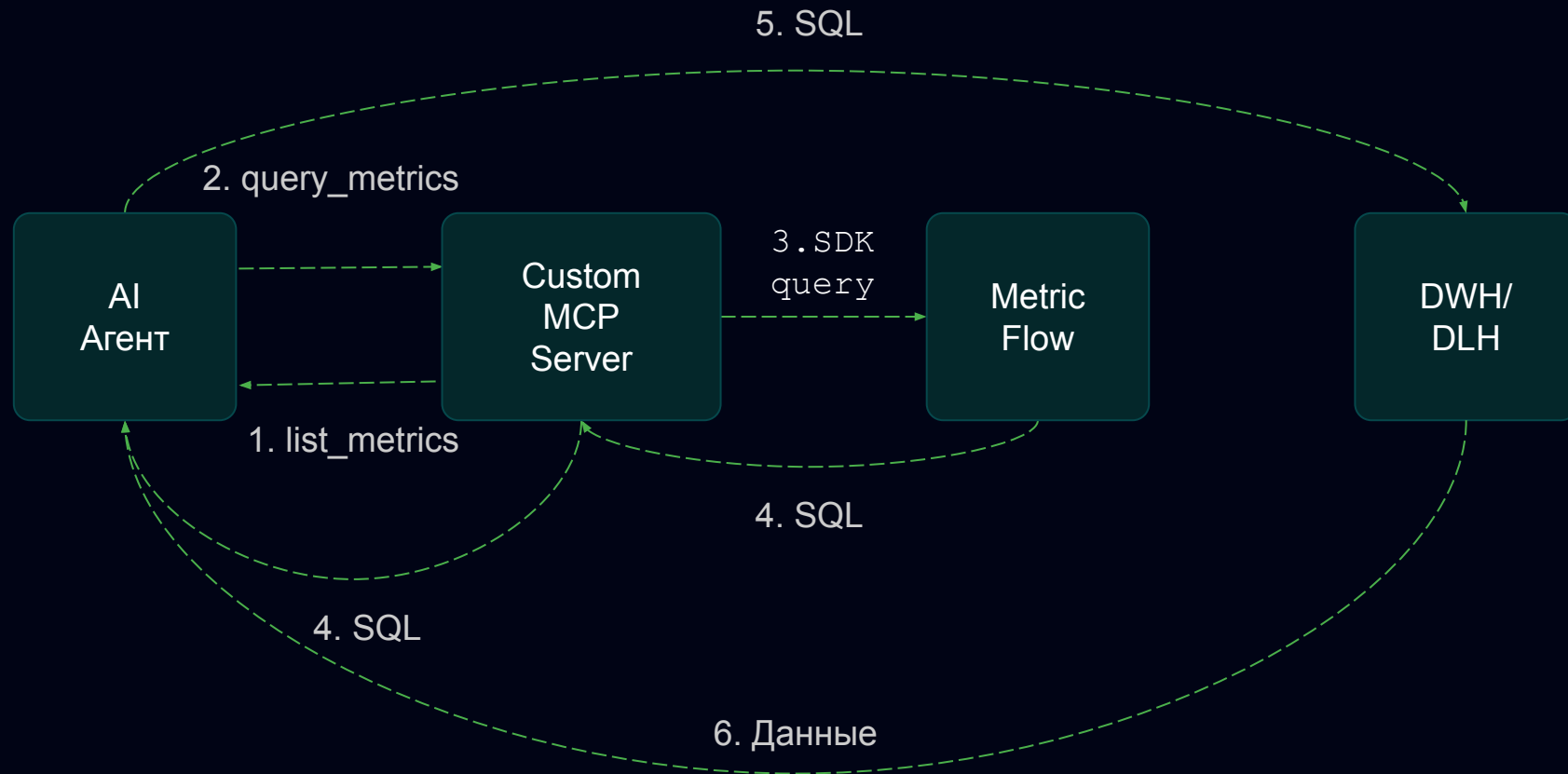
Почему это не работает из коробки



Почему это не работает из коробки



Кастомным MCP-сервером



Как не превратить YAML в цифровую помойку

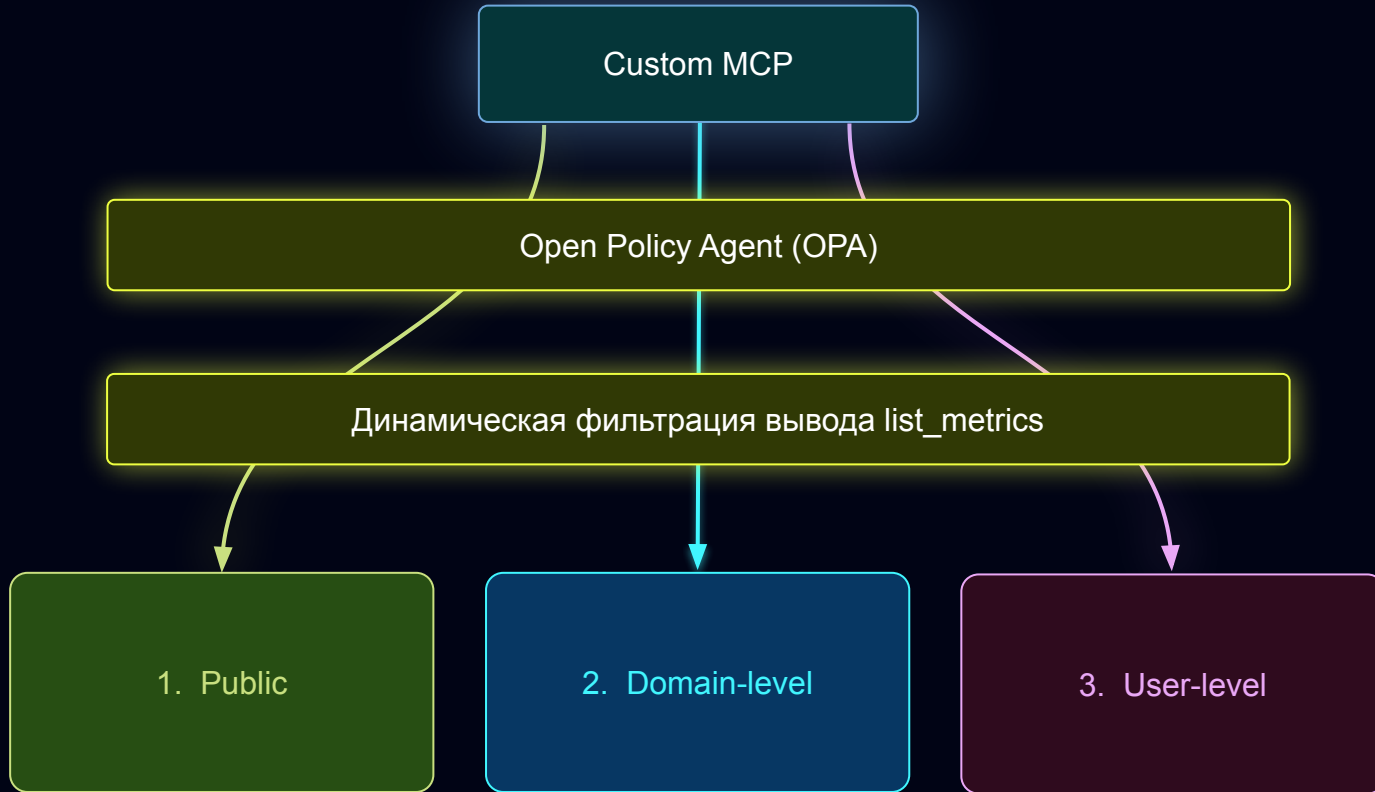


Как не превратить YAML в цифровую помойку



Внимание: Требуются Metric Owners.
Семантический слой без жестких владельцев мертв.
Инструменты не решают организационную проблему ответственности за бизнес-логику.

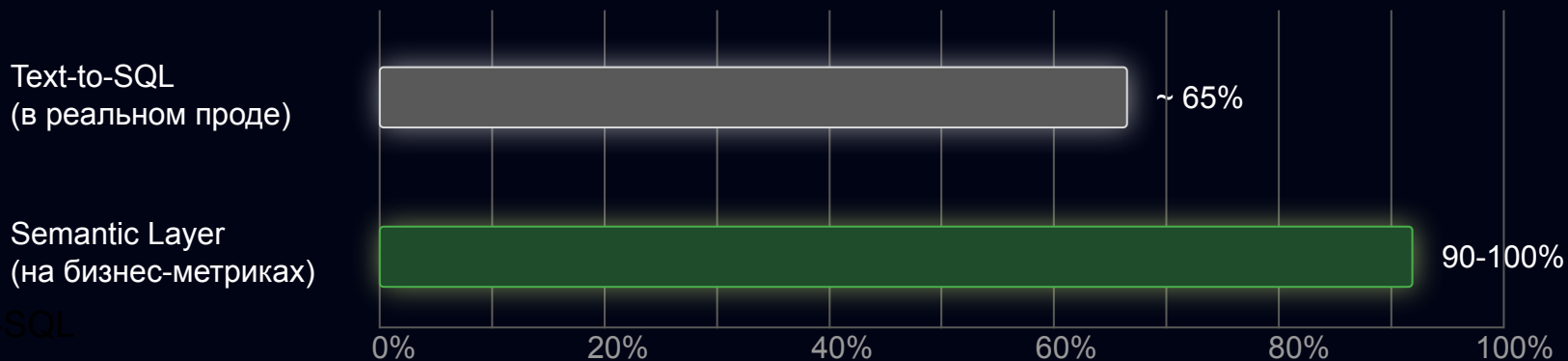
Авторизация доступа к метрикам



Гибридный подход с обратной связью



Бенчмарки



BIRD-SQL



Облачные (GPT-4) с уровнем описания Rare

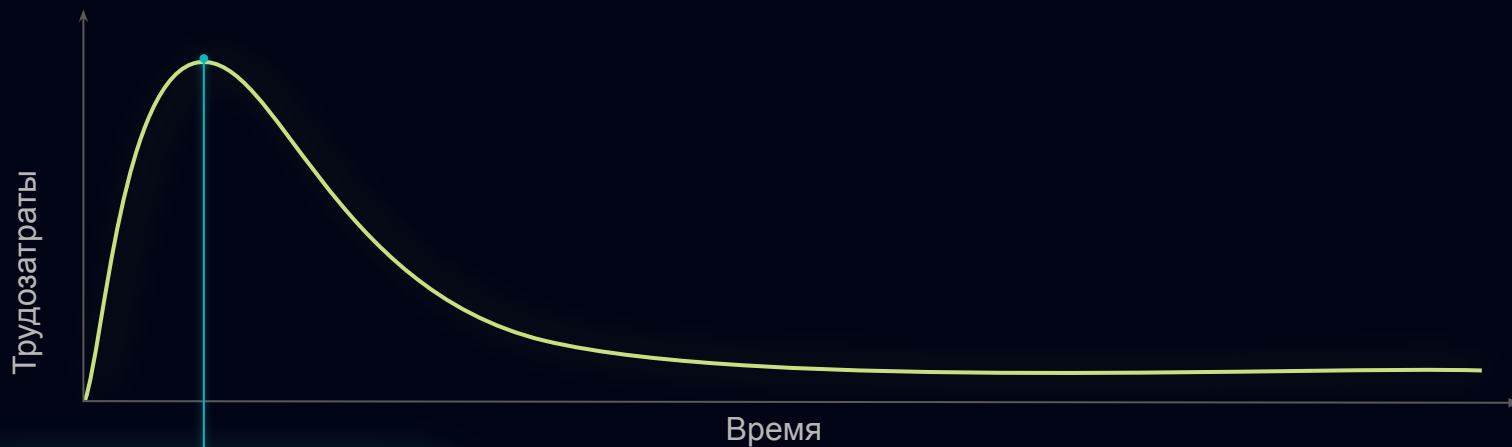
Небезопасно (утечка схем), дорого, галлюцинации в бизнес-логике



Компактные локальные (Qwen, Gemma, Mistral) + Semantic Layer

Полностью on-prem, безопасный контур, предельная точность благодаря ограничению пространства поиска

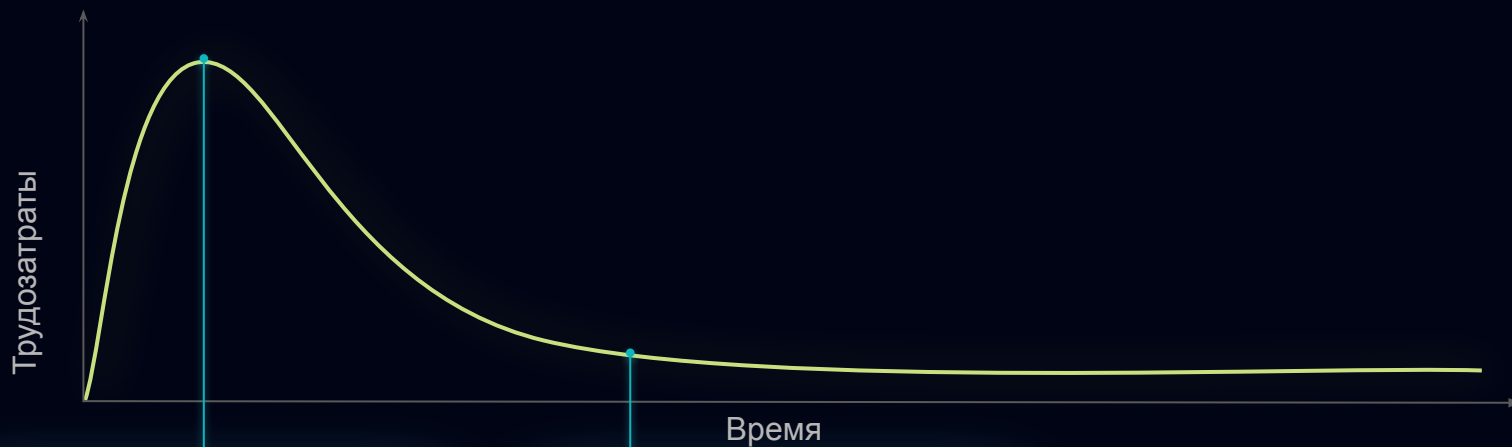
Трудозатраты MVP



Фаза 1: MVP (недели)

Описание 10-15 ключевых метрик, 1 YAML-манифест, развертывание кастомного MCP-агента. Быстрый Time-to-Market

Трудозатраты MVP



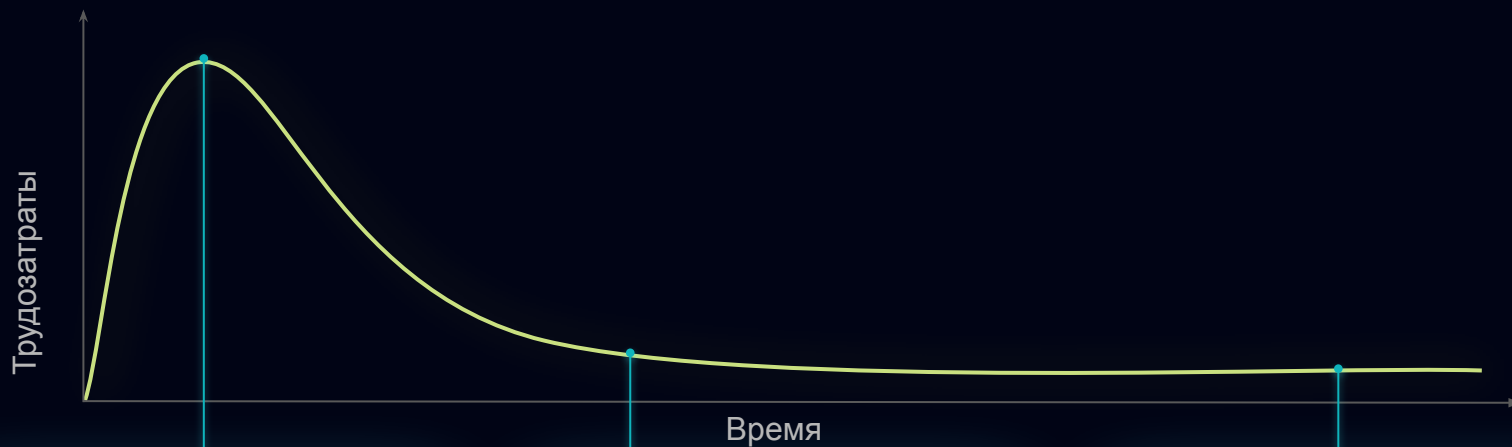
Фаза 1: MVP (недели)

Описание 10-15 ключевых метрик, 1 YAML-манифест, развертывание кастомного MCP-агента. Быстрый Time-to-Market

Фаза 2: Масштабирование (месяцы)

Описание остальных доменов данных. Главная сложность — сбор бизнес-требований, а не написание кода.

Трудозатраты MVP



Фаза 1: MVP (недели)

Описание 10-15 ключевых метрик, 1 YAML-манифест, развертывание кастомного MCP-агента. Быстрый Time-to-Market

Фаза 2: Масштабирование (месяцы)

Описание остальных доменов данных. Главная сложность — сбор бизнес-требований, а не написание кода.

Фаза 3: Плато поддержки

Основные затраты — на поддержку репозитория, code review YAML-файлов и управление изменениями

Альтернативы

	Наличие API для LLM	Analytics-as-code	Простота on-prem
OpenMetadata			
Cube.js			
Metabase			
Superset			



Takeaways

- 1 Semantic Layer - это доверие к аналитике. Его нужно было начать строить вчера. Основной потребитель AI
- 2 Есть уже открытый стандарт описания, используйте его, новые инструменты будут его поддерживать
- 3 Text To SQL тоже никуда не девается, он для ad-hoc
- 4 Начните с простого **cube.js**, и пары доменов/
- 5 Сразу думайте о governance и ЖЦ метрик.