Reverse Analytics

Как аналитики могут нанести непоправимую пользу продукту и заработать денег

Кто я

1

20+ лет в бизнес-анализе

Процессы, корпоративные системы, ERP, BI, SD, CMDB, СЭД, биллинг

2

Руководитель направления

Организационного развития в облачном провайдере (топ-3 интегратора РФ)

3

Предприниматель

В поиске работающих моделей

4

Сегодня

Студия анализа данных и стартап (LegalTech/PropTech)

Кто вы?

Разговор про вас:

Процессные аналитики

Схемы процессов, BPMN, BPMS, ARIS.

Системные аналитики

ER диаграммы, Swagger, трассировка требований.

Аналитики данных

Витрины данных, Power BI, Tableau, OLAP.

Зачем это все?

Зачем это вам:

- Убедиться в удачном выборе профессии
- Увидеть нужные скилы для BI внедрений
- Посмотреть на построение клиентской аналитики

Зачем это мне:

- Поделиться опытом смены направления аналитика
- Найти клиентов и сотрудников

Как все начиналось

1 Встроенная аналитика: где есть и где нет

Онлайн-банки и биржи активно используют BI для отображения расходов, доходов и торговых данных. Однако большинство ИТ-сервисов (вне банкинга, ритейла, телекома) не имеют качественной встроенной аналитики, предлагая только API или выгрузки.

2 Ваша возможность

Этот пробел на рынке создает отличную возможность для карьерного роста в области аналитики, предлагая уникальные ниши для внедрения BI-решений.

3 Наша история

Наш проект стартовал с запроса в небольшом сообществе от компании TeachBase, одного из топ-разработчиков корпоративных LMS в РФ.

Этап 1. Построение семантической модели предметной области и формирование метрик / визуализаций

1

Семантическая модель и метрики

Как погрузиться в контекст онлайн курсов за 1 месяц?

Знания о данных не в документах, а в головах.

Быстрее принесите результат, даже не взирая на его качество.

Почему это самое важное в проекте?

2

Сопоставление моделей

Почему делать аналитику разработчиками ~~зло~~ не оптимально?

Reverse analytics. Меппинг таблиц в метрики и события.

Ну а кто еще может перелопатить базу и понять, что где?

Что может пойти не так?

3

Построение пайплайна данных

А зачем нужны другие базы?

Когда уже звать инженеров, а когда можно и самим?

Использование management сервисов яндекса.

Гид по костылям.

На что в итоге можно замахнуться?

Этап 1: Как погрузиться в контекст онлайн курсов за 1 месяц?

Понимание предметной области

У каждой предметной области свой язык и терминология.

Определение бизнес-вопросов

Четко формулируйте, на какой вопрос пользователь хочет получить ответ и как он будет использован в дальнейшем.

Изучение сущностей и отношений

Необходимо понять бизнес-сущности и их взаимосвязи, а не просто источники данных.

Практические аспекты и автоматизация

Важно точно ответить на вопросы пользователя, а затем автоматизировать и визуализировать этот ответ.

Этап 1: Знания о данных не в документах, а в головах

Знания живут в головах экспертов
Мы тесно взаимодействовали с методологом и продактом, которые обладали всеми необходимыми знаниями. Курирование проекта CTO компании и общая вовлеченность команды заказчика оказались чрезвычайно важны.

Верификация каждого слова
Как только мы слышали новое понятие, мы всегда искали ее в интерфейсе системы, отчете или базе данных, чтобы верифицировать каждое слово с реальными данными в системе.

DevOps как союзник
Налаживание отношений с DevOps-специалистом, который обеспечивал быстрые дампы и выделял ресурсы, стало критически важным.

Люди решают судьбу проекта
Вовлеченность ключевых участников проекта, таких как CTO или тим-лид, абсолютно необходима.

Этап 1. Быстрее принесите результат, даже не взирая на его качество

Максимально быстрый первый дашборд

Построен на витрине в развернутом дампе OLTP базы PostgreSQL, за 1 месяц на одном клиенте.

Медленно, но работает

Дашборд прорисовывался 10-15 секунд, но даже так продакт и методолог начали давать обратную связь и верифицировать данные.

Первая версия всех метрик за 1,5 месяца

Это позволило сразу проверить расчеты и методики измерения на реальных данных, вместо создания макетов без данных.

Выход на конечных клиентов

После демонстрации дашборда лояльным клиентам мы получили совершенно другие, неожиданные требования.

Этап 1: Почему это самое важное в проекте?

1. **Понимание вопросов и языка клиента**

   Дашборды отвечают на вопросы людей. Важно сначала понять, на какие именно вопросы, и как эта информация будет использоваться. Начните говорить на языке клиента.

2. **Знание в головах, не в документах**

   Вся информация находится у экспертов. Создайте условия, чтобы они хотели поделиться этими знаниями с вами.

3. **Быстрое погружение в предметную область**

   Используйте доступные инструменты (ChatGPT, Perplexity, Claude) для оперативного изучения новой сферы.

4. **Ранний дашборд – быстрая обратная связь**

   Максимально быстрый первый дашборд обеспечивает ценную обратную связь. Первичное понимание часто ошибочно, и лучший способ его улучшить — это итоговая визуализация.

5. **Создание бизнес-модели сущностей**

   Ваша первая задача — построить бизнес-модели сущностей и связей предметной области.

Именно для этого нужны компетенции и опыт аналитика. Глубокое погружение в язык бизнеса — главный навык бизнес- и системного аналитика.

Этап 2. Сопоставление семантической модели и OLTP модели транзакционной базы

1 **Семантическая модель и метрики**

Как погрузиться в контекст онлайн курсов за 1 месяц?

Знания о данных не в документах, а в головах.

Быстрее принесите результат, даже не взирая на его качество.

Почему это самое важное в проекте?

2 **Сопоставление моделей**

Почему делать аналитику разработчиками (зло) не оптимально?

Reverse analytics. Меппинг таблиц в метрики и события.

Ну а кто еще может перелопатить базу и понять, что где?

Что может пойти не так?

3 **Построение пайплайна данных**

А зачем нужны другие базы?

Когда уже звать инженеров, а когда можно и самим?

Использование management сервисов яндекса.

Гид по костылям.

На что в итоге можно замахнуться?

Этап 2. Почему делать аналитику разработчиками ~~зло~~ не оптимально?

Транзакционные базы

Созданы для быстрой записи построчно и быстрого поиска отдельных строк.

Разные паттерны

Разработчики оптимизируют запись одной сущности, тогда как аналитики читают много данных для агрегаций.

Разные подходы к данным

Разработка использует CRUD и высокую нормализацию. Аналитики данных — OLAP-таблицы, денормализованные витрины и пакетную запись.

Самостоятельность аналитика

Вам придется самостоятельно разбираться в OLTP базах. Разработчики не смогут помочь с этой спецификой.

Этап 2. Reverse analytics. Меппинг таблиц в метрики и события

Сложность исходных данных

На старте проекта в Teachbase было около 200 таблиц. Это огромное количество информации, которую предстояло разобрать и систематизировать.

Глубокое погружение

Мы провели значительное время, чтобы понять, как бизнес-термины отражены в таблицах базы.

Обратный перевод

Из высоконормализованной таблиц — со всеми ее исключениями, бизнес-правилами и связями по ключам — в понятные бизнес-сущности и метрики.

Определение источника

Основная задача — однозначно определить таблицы и поля исходной транзакционной системы, которые необходимы для переноса в аналитическую базу.

Этап 2. Ну а кто еще может перелопатить базу и понять, что где?

Роль аналитика

Разработчики не углубляются в бизнес-логику, а бизнес-пользователи не всегда понимают технический язык. Аналитик — незаменимый мост между ними.

Самый долгий этап

Это самый трудоемкий и муторный этап. Системные аналитики отлично к этому подходят.

Определение сущностей

Нужно определить все таблицы, которые понадобятся в аналитике, связи между ними, правила и способы отображения бизнес-логики в системе.

Влияние на будущее

От качества проведения этого этапа зависит скорость реализации и выбор оптимальных инструментов для последующих шагов проекта.

Этап 2. Что может пойти не так?

На этом этапе могут возникнуть серьезные проблемы, требующие глубоких компетенций. Вот ключевые моменты, где могут начаться трудности:

Сложность структур данных

Уже нужно понимать структуры данных, чтобы построить эффективное аналитическое хранилище.

Различия OLTP и OLAP

Знать, чем аналитические базы данных отличаются от транзакционных и выбирать правильную

Правильные джойны

Корректно выполнять объединения таблиц, чтобы избежать кривых и не оптимальных соединений.

Витрины, измерения, метрики

Четкое разграничение и понимание витрин данных, измерений и бизнес-метрик.

Интенсивная работа с SQL

Очень много SQL для исследования дампов баз данных и извлечения нужной информации.

Ограничения LLM в SQL

LLM полезны, но их эффективность снижается из-за меньшего объема доступного SQL-кода в выборке

В конечном итоге, успех проекта во многом зависит от способности аналитика ориентироваться в этих технических сложностях, несмотря на потенциальные подводные камни.

Этап 3. Построение пайплайна данных из OLTP в широкие витрины

1 **Семантическая модель и метрики**

Как погрузиться в контекст онлайн курсов за 1 месяц?

Знания о данных не в документах, а в головах.

Быстрее принесите результат, даже не взирая на его качество.

Почему это самое важное в проекте?

2 **Сопоставление моделей**

Почему делать аналитику разработчиками ~~зло~~ не оптимально?

Reverse analytics. Меппинг таблиц в метрики и измерения.

Ну а кто еще может перелопатить базу и понять, что где?

Что может пойти не так?

3 **Построение пайплайна данных**

А зачем нужны другие базы?

Когда уже звать инженеров, а когда можно и самим?

Использование management сервисов яндекса.

Гид по костылям.

На что в итоге можно замахнуться?

Этап 3. А зачем нужны другие базы?

1. **Различные цели баз данных**
   Транзакционные базы (OLTP) оптимизированы для быстрой записи и поиска отдельных строк. Аналитические базы (OLAP) предназначены для агрегации и чтения больших объемов данных.

2. **Проблемы аналитики в OLTP**
   Попытка строить витрины данных для аналитики прямо в OLTP-базе по множеству нормализованных таблиц за длительный период неизбежно приведет к ошибкам "out of memory".

3. **Ограничения OLAP при соединениях**
   Аналитические базы плохо приспособлены для частых и сложных соединений таблиц

4. **Необходимость ETL-потоков**
   Именно здесь возникает потребность в создании потоков ETL. Данные нужно извлечь не перегрузив источник, преобразовать в нужный аналитический формат и загрузить в специализированную базу.

Этап 3. Когда уже звать инженеров, а когда можно и самим?

1. **Определите минимальную сущность**
   Найдите наименьшую атомарную бизнес-сущность в вашей предметной области

2. **Пороги объема данных**
   Ориентировочно: до 10 млн. строк может выдержать OLTP-база. От 10 до 50 млн. уже требует продуманного ETL и OLAP хранилища. Свыше 50 млн. к дата инженеру. Свыше 1 млрд. — вам нужен опытный архитектор данных.

3. **Учтите частоту обновления**
   Если данные обновляются раз в сутки, можно обойтись более простыми решениями. Если нужна аналитика в реальном времени, нужны инженеры

Этап 3. Использование management сервисов Яндекса

Полный стек управляемых сервисов
Мы использовали Yandex Transfer + Serverless функции = ETL. Managed ClickHouse как хранилище и Yandex DataLens для визуализации.

Высокая гибкость и скорость итераций
Десятки раз сносили и поднимали базы, разворачивали бэкапы, меняли трансферы и воркбуки. Все это несколько кликов

Минимум забот об инфраструктуре
Управляемые сервисы позволили нам полностью сосредоточиться на бизнес-логике и аналитике.

Мы оставили все проблемы с инфраструктурой сервису, что значительно ускорило процесс разработки и запуска аналитических решений.

Этап 3. Гид по костылям

Использование OLTP для аналитики

Дашборд на годовых данных в PostgreSQL ронял базу

Отсутствие документации и маппинга

Много времени на поиск таблиц и полей

Сложный ETL с промежуточным Postgresql

Промежуточная база PostgreSQL для ETL - добавочная точка отказа.

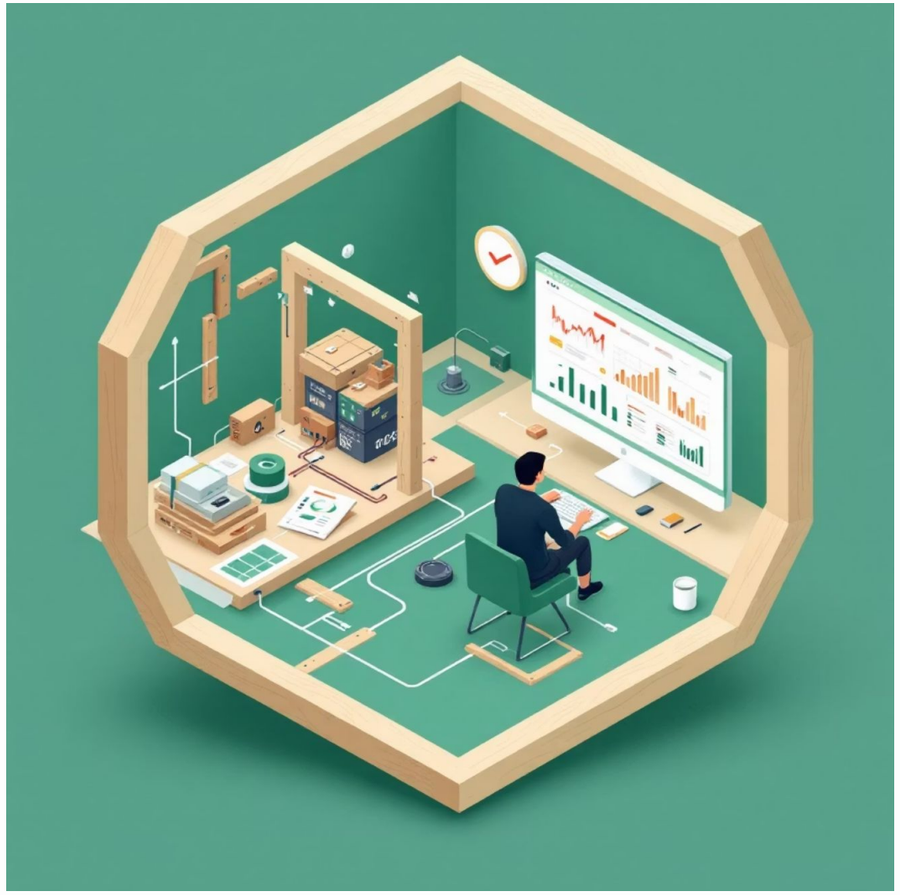
Ограничения LLM в аналитике

Тогда LLM были малоэффективны для глубокой аналитики и работы с SQL.

Лимиты времени Yandex Functions

Ограничения по времени выполнения Яндекс Функций при обработке больших объемов данных для множества клиентов приводили к сбоям.

Этап 3. На что в итоге можно замахнуться?



MVP в аналитике — это когда у клиента есть дашборд, отвечающий на вопросы пользователей. Без ошибок и, что важно, достаточно быстрый. Пользователь не видит, что там под капотом, и ему, собственно, все равно. Поэтому на первом этапе можно схитрить и накостылять процессы передачи и хранения данных.

Но у всего есть границы, и они проходят по:

- объемам строк (до 50 млн. строк);
- возможности использовать управляемые сервисы в облаке (если нет, потребуется дата-инженер и девопс);
- требованиям к частоте обновления данных (чем ближе к реалтайму, тем выше требуемые компетенции).

Что из всего этого стоит запомнить?

1

Результат с правильной командой и инструментами

При наличии реальной потребности клиента, мотивированной команды (даже из двух аналитиков) и облачных инструментов, возможно:

- За 6 месяцев создать и продать аналитический продукт.
- Предложить клиентам приносящий ценность продукт, за который они готовы платить.
- Получить полноценное конкурентное преимущество на рынке.

2

Значение софт-скиллов

Чем лучше выстроены отношения и коммуникация в проектной команде, тем быстрее и эффективнее достигаются результаты.

3

Когда пора звать подмогу

"Более 50 млн. строк", необходимости "real time сбора данных" или "Никакого облака, быстро на виртуалках поднимем!". Это явные сигналы к привлечению дополнительных специалистов.

Потенциал бизнес- и системных аналитиков

Вы можете создавать полноценные аналитические продукты, как внутренние, так и для клиентов. Хотя потребуется активное освоение новых инструментов. Базовые принципы мышления уже заложены, что открывает отличные возможности для карьерного роста.

Попробуйте сами!

Отсканируйте QR-код, чтобы перейти к интерактивному демо дашборду, разработанному на основе описанного проекта. Он наглядно демонстрирует результаты нашей работы и возможности аналитики.

Спасибо за проект! https://teachbase.ru/

Контакты

Если у вас возникли вопросы, предложения или вы хотите обсудить возможности сотрудничества, пожалуйста, свяжитесь со мной.

Телефон

+7 (909) 909-39-59

Наш сайт

www.smart3asy.com

Мой TG

Отсканируйте QR-код, чтобы связаться со мной еще быстрее