Mage: волшебный инструмент оркестрации

Валентин Пановский,

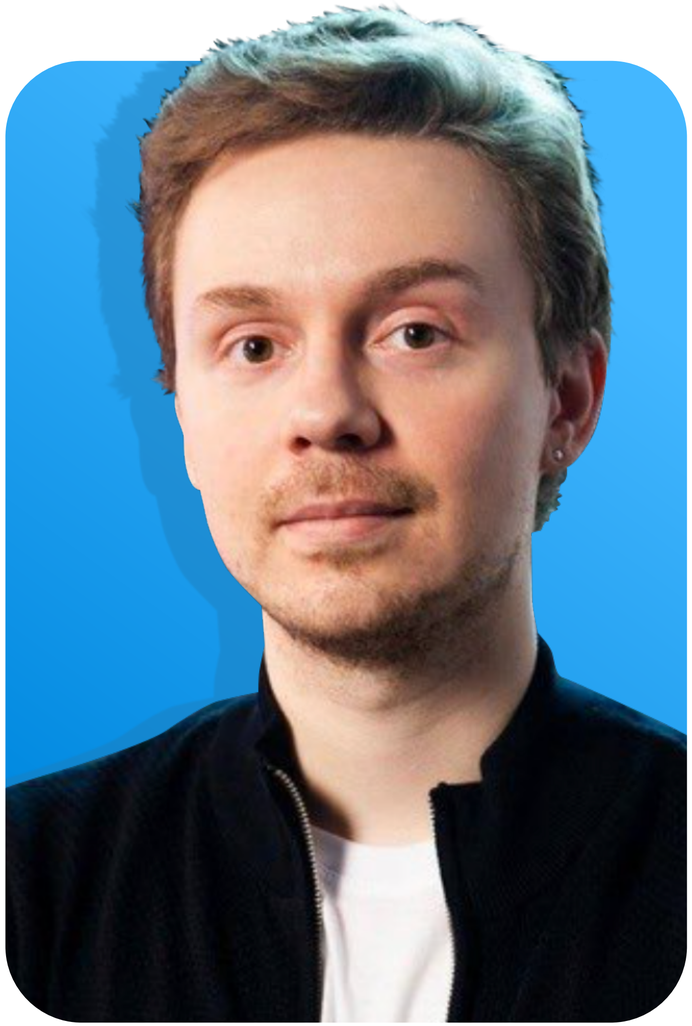
Head of Data Platform @ WB, ex-CDO @ more.tv

Валентин Пановский

- Head of Data Platform маркетплейса WB (ex-CDO видеосервиса more.tv, ex-CDO образовательной платформы Skillbox)

- Кандидат физико-математических наук, доцент факультета «Информационные технологии и прикладная математика» МАИ (НИУ),

Более десяти лет суммарного опыта работы и ведения научных и коммерческих проектов, включая оцифровку бизнеса с последующим переходом на DDD-стратегию управления, построение и развитие алгоритмов машинного обучения для решения прикладных задач в разных сферах (от E-commerce до экологии и EdTech)



О чём сегодня расскажу?

1. Что такое more.tv и какие задачи выполняет департамент аналитики и управления данных в стриминговом сервисе
2. Как выглядел legacy контур для BI системы?
3. Целевое решение для переезда
4. ~~Опыт эксплуатации Dagster и сравнение с Airflow~~ Личный опыт и взгляд
5. Что такое Mage и почему он может быть полезен?

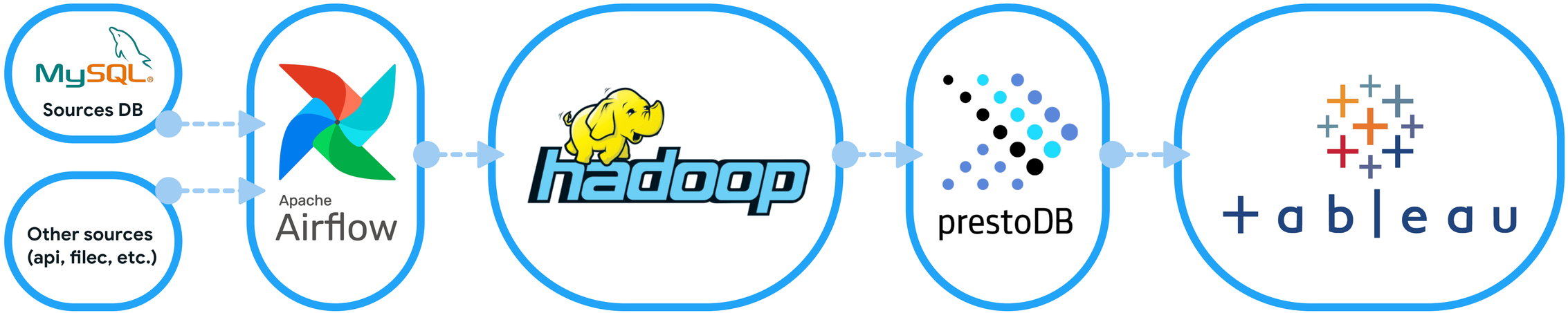
Аналитика и управление данными в more.tv

- **Инженерный блок**
  - обеспечение работы DWH
  - загрузка и обработка данных (ETL/ELT)
  - развёртывание и поддержка внутренних сервисов
- **Аналитический блок**
  - продуктовая и маркетинговая аналитика (сопровождение feature-команд, AB-тесты, валидация гипотез, оценка эффективности performance кампаний)
  - отчётность и дашборды
  - ML-команда (RecSys & Uplift)

Disclaimer

- Сравнительно небольшой объём ETL процессов (до ~100 штук)
- Общий объём обрабатываемых данных < 100 ТБ
- Количество пользователей системы ±10 специалистов
- На момент переезда не было бизнес-процессов, завязанных на получение данных в Real Time
- Использовались внешние решения для рекомендательной системы

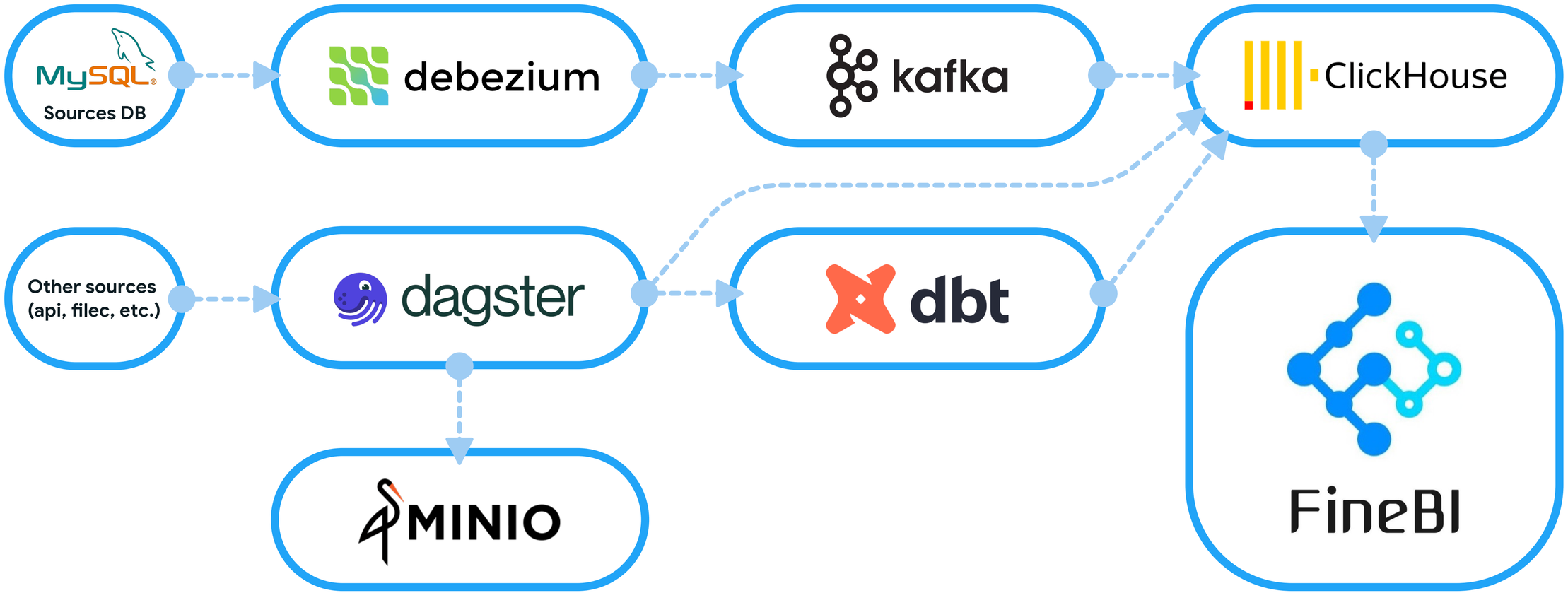
Как выглядел legacy контур для BI системы?



Что не устраивало?

- Ручное управление VPS в рамках выделенного пула Bare Metal серверов и отсутствие возможности оперативно проводить деплой сервисов
- Неравномерная нагрузка на БД backend сервисов во время загрузки данных в DWH
- Используемые ML решения были внедрены как SaaS => риски отключения по внешним обстоятельствам
- Несоответствие используемого стека навыкам инженеров (сложилось ввиду обновления команды)
- Не было возможности закупить или продлить лицензии Tableau

Целевое решение для переезда



Целевое решение для переезда

Реализация процесса загрузки данных с источника через CDC протокол средствами Debezium не только делает нагрузку на БД более равномерной, но также дополнительно даёт возможность в режиме (near) realtime реагировать на события на платформе

Целевое решение для переезда

Внедрение DBT позволяет реорганизовать нагрузку между командами инженеров и аналитиков и сделать их более независимыми (аналитики сами могут писать скрипты для создания сущностей на разных уровнях хранения данных и автоматически получать результаты, когда runner’ы gitlab’а раскатывают их в DBT)

Дополнительно использование тестов внутри DBT позволяет начать управлять качеством данных на уровне базовых проверок консистентности

Целевое решение для переезда

Для снижения рисков, связанных с использованием внешних ML-сервисов, было принято решение разработать и запустить собственные технологии; в рамках реализации этого решения был выбран достаточно классический стек: **Jupyter** для прототипирования, **mlflow** для запуска и мониторинга экспериментов, **dvc** для версионирования и **kubernetes** для деплоя сервисов в production

В качестве основного ядра DWH — **Clickhouse** (ориентировались в том числе на bus factor и наличие компетенций у смежных команд)

Целевое решение для переезда

Dagster в качестве основной ETL/ELT системы

MinIO для передачи не realtime данных во внешние сервисы для последующего использования (например, сегменты для CRM системы в рамках триггерных цепочек и «полок» с тайтлами)

FineBI как целевое решения для служб отчётности

Личный взгляд и опыт

Личный взгляд и опыт

- **Что было на старте?**
  - нестабильно работающий контур работы с данными (в истории наблюдались падения, на восстановление после которых уходили недели)
  - низкий технологический bus factor

- **Почему в итоге у нас получилось?**
  - смогли договориться с заказчиками на временный холд на поток задач, которые объективно не являются критическими для ведения бизнеса
  - выбрали решение, не требующее переписывания огромного количества legacy кода, т. е. начали from a scratch
  - «заряженная» команда, открытая к изучению нового

А причём тут Mage? O_o

Предпосылки к тестированию

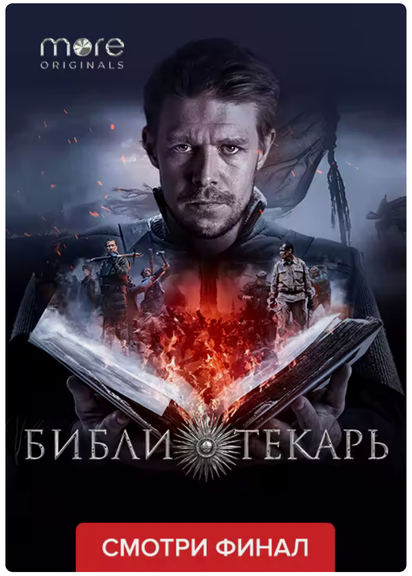
А может ли аналогичное желание облегчить разработку ETL процессов быть успешным в разрезе пайплайна подготовки ML-моделей?

- **Disclaimer**
  - нет realtime моделей
  - подготавливать данные от результатов применения моделей и передавать их во внешние production системы можно batch’ами

Как может выглядеть?

А давайте посмотрим на «живом» примере по подготовки индивидуальных рекомендательных полок для пользователей

Рекомендуем тебе ›



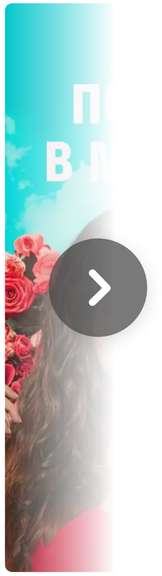
Библиотекарь

Великолепный век

Фишер

Молодёжка

Ветреный



Постучі

Заключение

Несмотря на наличие индустриальных стандартов в каждом из направлений полезно знакомиться с технологиями, чтобы расширять кругозор и быть в курсе имеющихся возможностей и альтернатив. В совокупности это позволит более комплексно смотреть на решаемую задачу и оптимальным образом выбирать используемый инструментарий.

В нашем кейсе это привело к оптимизации ETL-процессов и выделению отдельного инструмента для запуска несложных ML-моделей. Последнее облегчило работу со стажёрами, ускорив процесс их обучения и погружения в доменную область.

Вопросы

@PANOVSKIY_V