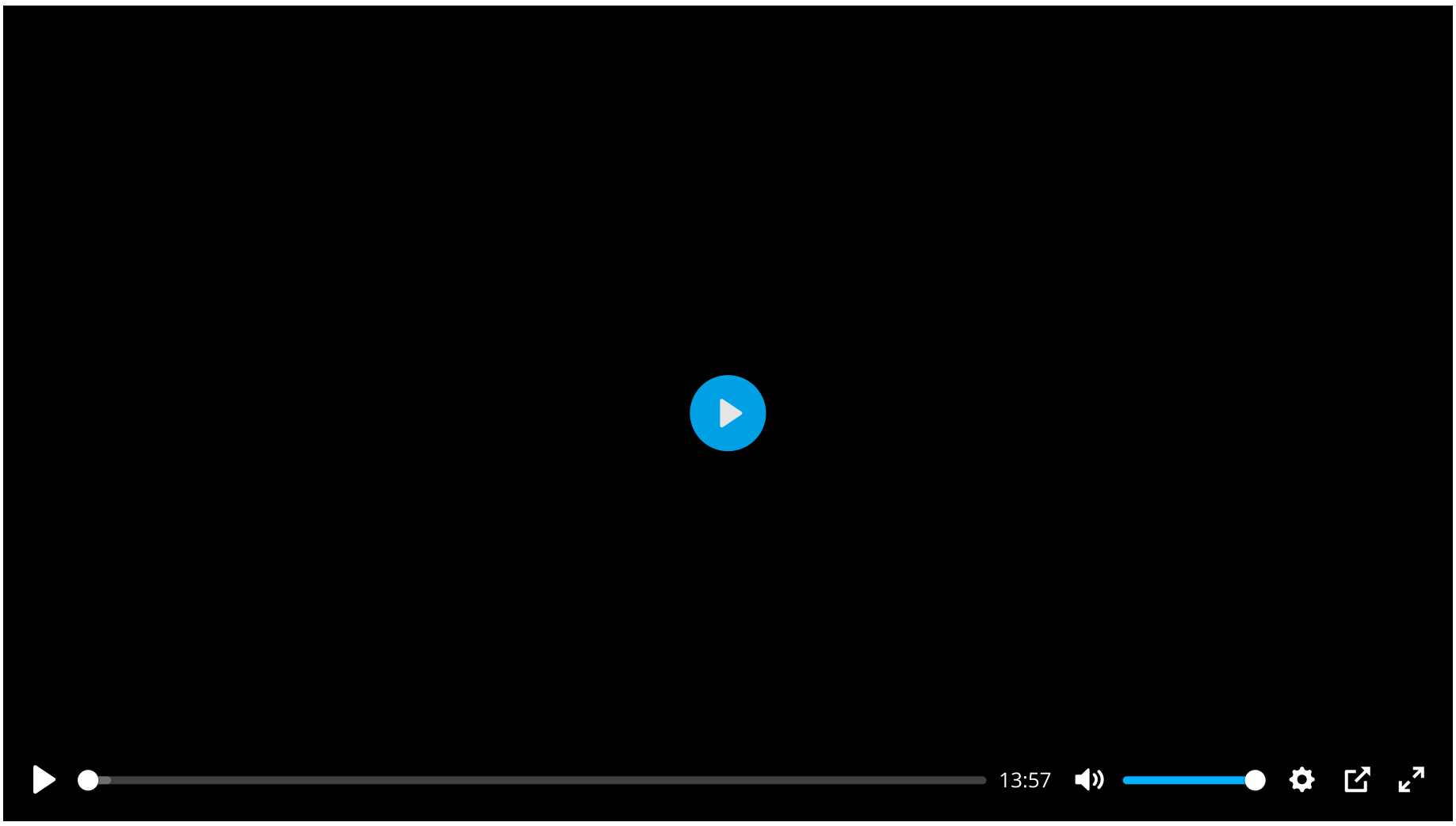
7.3 Пайплайны. Оркестрация данных. Обзор ETL-инструментов



Текстовая расшифровка видео:

ПАЙПЛАЙНЫ

План:

- ETL / ELT и пайплайны (конвейеры данных);
- Как сделать свой;
- Критерии выбора;
- Чтобы все не пошло не так;
- Варианты запуска;
- Интеграция данных.

ETL / ELT и пайплайны (конвейеры данных)

Слова «Пайплайн» и «Конвейер» достаточно близки, поэтому можно использовать слово «Конвейер» вместо «Пайплайн». К этой взаимозаменяемости относятся и ETL/ELT.

Пайплайн – более широкий термин, охватывающий разные стадии процессинга и преобразования данных (**ETL** и **ELT** могут быть его **подмножествами**).

Например, пайплайн может реализовывать **стадии обучения и/или инференса ML-модели**.

Стадии ETL могут быть как пакетные, так и потоковые (в зависимости от ситуации). Как в случае **пакетной (batch)** обработки, так и в случае **потоковой (streaming)** обработки, стадии **Extract**, **Transform** и **Load** могут быть реализованы как **шаги (таски)** пайплайна.

В результате выполнения пайплайна получается какой-то **артефакт** – данные **в конечном хранилище**, обученная **ML-модель** или же, например, **отчет** в каком-то BI-дэшборде.

Часто можно встретить слово «Сценарий». **Сценарий** – еще более общий термин, чем пайплайн. Однако с точки зрения бизнеса – это сценарий.

Как сделать свой

Вопрос: как переложить данные?

Ответ: можно взять инструменты, такие как Python, Spark, написать на них решение. Другой вариант – взять готовое решение в том же Airflow или других инструментах (например, Терабайт). К разным источникам есть уже готовые коннекторы.

Часто получается что-то среднее: мы не пишем все с нуля, процесс по перекладыванию не реализуем, однако часто реализуем какую-то часть/логику, которую было бы сложнее автоматизировать.

Критерии выбора

Критерии выбора для инструмента:

- **Connectivity (возможность подключения)**

У инструмента должны быть коннекторы для ваших источников и целевых систем. Это справедливо как для визуального инструмента, так и для языка программирования.

- **Scalability (масштабируемость)**

Возможность параллелить обработку потоков данных.

- **Extensibility (расширяемость)**

Возможность расширять функциональность и добавлять плагины.

Чтобы все не пошло не так

- **Security (безопасность)**

Возможность интеграции с имеющимися системами аутентификации и управления секретами, ролевой контроль доступа.

Так, например, к вам могут прийти безопасники из соседнего отдела и потребовать, чтобы все реализуемые вами процессы по перекладыванию данных, отправляли логи в CM-систему для дальнейшего мониторинга.

- **Disaster Recovery (восстановление после сбоев)**

Возможность с минимальными потерями восстановить работоспособность в случае нарушения работы.

Варианты запуска

Вопрос: Каким образом мы можем брать данные из источника и куда-то перекладывать?

Ответ: мы можем это делать **по расписанию** (когда, например, раз в час запускаем какой-то процесс, выгребающий данные и заливающий их куда-либо); часто это делается при помощи механизма «cron» или других инструментов со встроенной возможностью использовать периодичные таски.

Альтернативный вариант – отслеживание изменений в источнике напрямую. Можно использовать механизм «Change Data Capture» (для отлова и пересылки изменений), если же речь идет о создании файла, который сразу же должен попасть в хранилище, можно использовать механизм «inotify».

Интеграция данных

Часто используется терминология «Интеграция данных».

Интеграция данных подразумевает под собой наличие множества источников данных (например, из разных отделов предприятия), которые имеют разные наборы полей, которые по-разному унифицируются.

Частая задача – настроить систему, которая будет собирать данные из разрозненных систем внутри предприятия.

Разные отделы часто используют разные решения для работы с данными и/или строят **Data Silo** (изолированные куски данных, лежащие по разным отделам).

В идеальном случае перекладывание данных между разными системами для последующего анализа должно быть максимально простой операцией.

Как вам урок? 😠 😐 🔥

Изучил, далее >

Слёрм ©

+7 (495) 248-05-80

Лицензия №ДЛ-1368 от 22.08.2019

Политика конфиденциальности

Публичная оферта