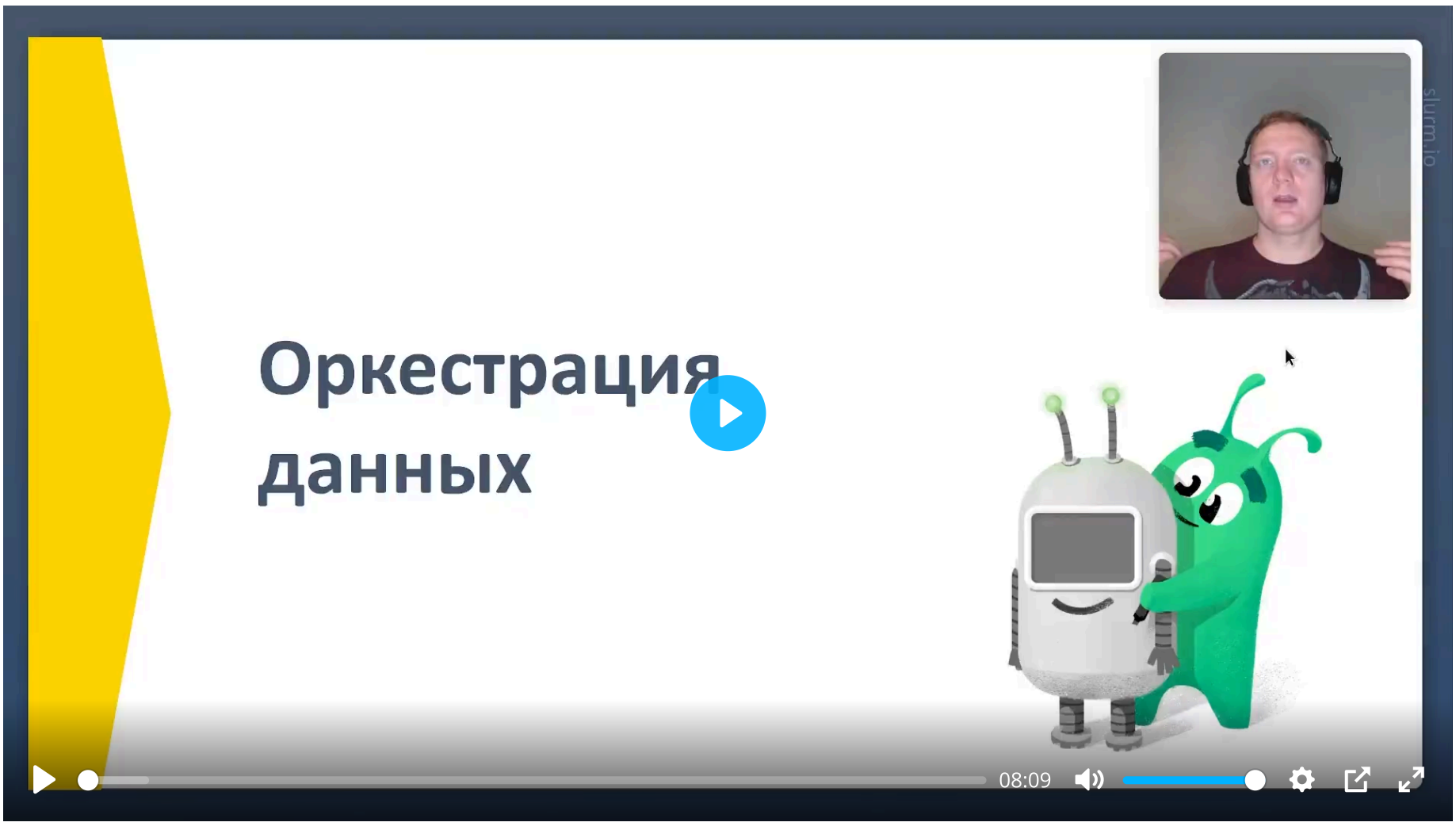
7.3 Пайплайны. Оркестрация данных. Обзор ETL-инструментов



Текстовая расшифровка видео:

ОРКЕСТРАЦИЯ ДАННЫХ

План:

- Что такое «Оркестрация данных»;
- Популярные решения оркестрации данных;
- У вас облако и лень писать самому;
- Все-таки пишем сами;
- Так, все-таки, ETL или оркестратор данных.

Что такое «Оркестрация данных»

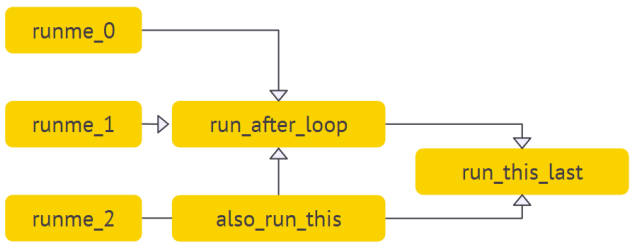
Оркестрация данных – это общее название процесса по перекладыванию данных, состоящего из разных хитрых стадий, которые могут этот процесс составлять.

Мы можем собирать данные из разных источников, агрегировать их, интегрировать, обрабатывать, преобразовывать и т.д.

Между разными этапами пайплайна могут быть **сложные зависимости**. Кроме того, может потребоваться **передача артефактов** из предыдущего этапа в следующий.

Также необходимо **отслеживать прогресс** выполнения **направленного ациклического графа (DAG)** задач и **сигнализировать об ошибках**.

Английский термин «**Workflow management**» часто переводят на русский язык как «**Оркестрация**», а в применении к задачам DE – «**Оркестрация данных**».



Популярные решения оркестрации данных

Инструментов, которые могут строить DAG’и, достаточно много:

- Apache Airflow;
- Prefect;
- Cloud dataflow;
- Dagster;
- Beam;
- Apache NiFi.

В рамках учебного курса мы познакомимся и научимся работать с инструментами Airflow и NiFi.

У вас облако и лень писать самому

Существуют разные облачные инструменты:

- Airbyte;
- Fivetran (на сегодняшний день очень популярный);
- Meltano (на сегодняшний день очень популярный);
- Matillion.

Fivetran, Matillion в первую очередь работают в клауде и именно там реализуют оркестрацию.

Airbyte, Meltano тоже современные Open-source-инструменты (немного конкурируют между собой) для того, чтобы делать оркестрацию локально.

Эти решения относятся к так называемому «**Modern Data Stack**». Это современные стильные инструменты, к которым присматриваются и используют многие западные компании.

Все-таки пишем сами

Очевидный выбор:

- Python;
- Java;
- Scala (предпочтительнее в случае Spark);
- Flink;
- Groovy (специфический язык программирования; его используют в сочетании с Apache NiFi, так как порой необходимо реализовывать разные преобразования и т.д.).

Так, все-таки, ETL или оркестратор данных

- **Задача оркестратора** – запускать стадии в DAG’е.
- **Задача ETL** – перекладывать данные.

Иногда и то, и другое встречается в одном инструменте.

Один из самых популярных инструментов – **Apache Airflow**. Обычно он заставляет **писать много кода** вручную, но можно вместо этого попробовать **интегрироваться** с другими внешними инструментами, например, **Airbyte** или **Meltano**.

На **внедрение** инструмента может уйти довольно **много времени**, но в итоге оно окупится, когда предоставит вам **возможность оркестрировать** данные, затратив минимум усилий.

Как вам урок? 😠 😐 🔥

Изучил, далее >