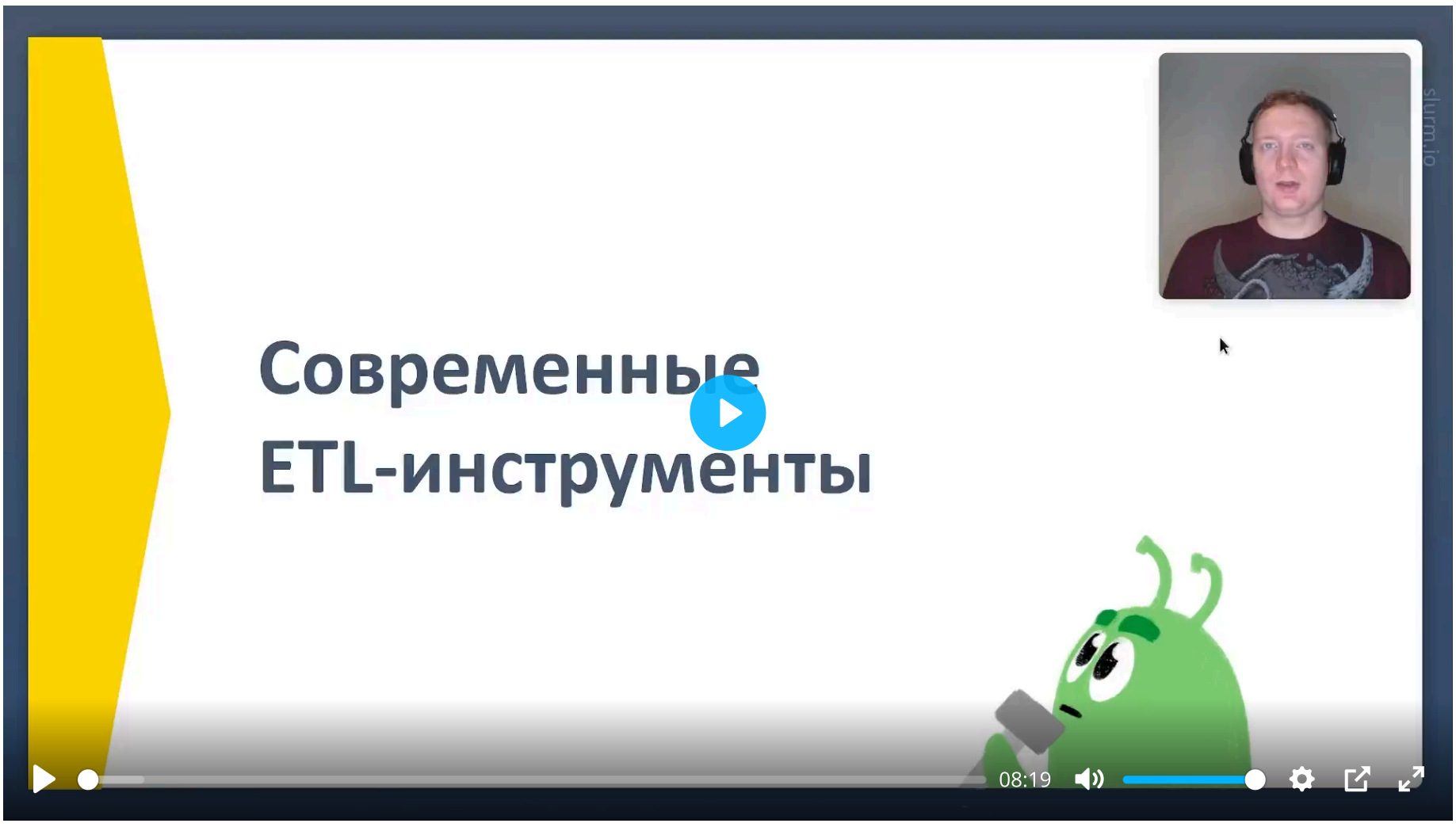
7.3 Пайплайны. Оркестрация данных. Обзор ETL-инструментов



Текстовая расшифровка видео:

СОВРЕМЕННЫЕ ETL-ИНСТРУМЕНТЫ

План:

- Мечты менеджеров;
- Квадрат Gartner’a;
- Talend Open Studio;
- AWS Glue;
- Apache Nifi;
- Apache Airflow и Python;
- «Убийцы» Apache Airflow.

Мечты менеджеров

Современные инструменты ETL корпоративного уровня обычно включают следующие функции:

- Автоматизация: полностью автоматизированные конвейеры;
- Drag'n'drop интерфейс: правила «0-code» и dataflows (взаимосвязи между стадиями);
- Поддержка сложных вычислений;
- Безопасность и соответствие требованиям: шифрование данных и соответствие требованиям HIPAA и GDPR.

Квадрат Gartner’a

Существует компания «Gartner», которая периодически публикует Magic Quadrant, публикует набор инструментов, в том числе по перекладыванию данных. Данный квадрат графически описывает ситуацию на рынке, позволяющую оценить возможности продуктов и их производителей (лидерство, претенденты и т.д.):

Figure 1: Magic Quadrant for Data Integration Tools

Source: Gartner (August 2022)

Например, на данном изображении в топе:

- Инструменты от Oracle;
- Informatica;
- SAP и т.д.

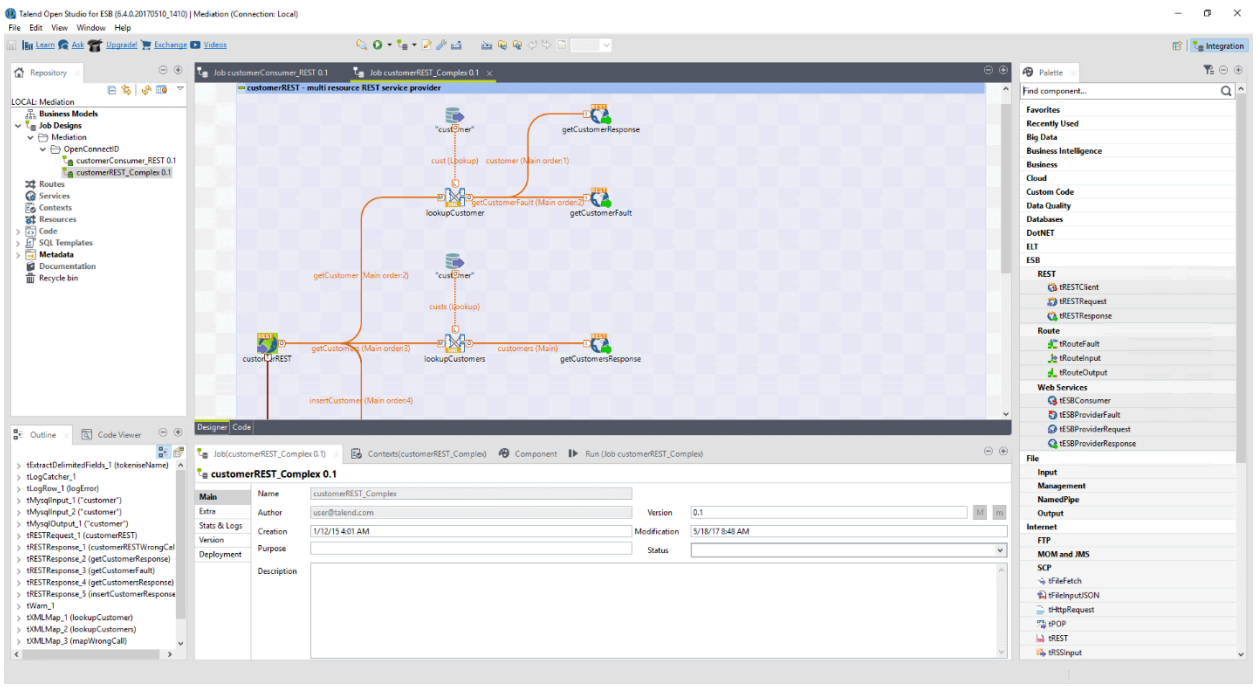
Talend Open Studio

Talend Open Studio – это иностранная компания, иностранный инструмент, иностранная оркестрация. Talend поддерживает почти все, о чем было сказано нами ранее.

Основные характеристики:

- Поддерживает системы класса BigData, Data Warehousing;
- Поддерживает профилирование;
- Имеет функции для совместной работы, мониторинга и планирования;
- Графический интерфейс drag'n'drop для создания конвейеров ETL;
- Автоматически генерирует код Java;
- Интегрируется со многими хранилищами данных;
- Открытый исходный код.

Выглядит следующим образом:



Имеется графический интерфейс.

AWS Glue

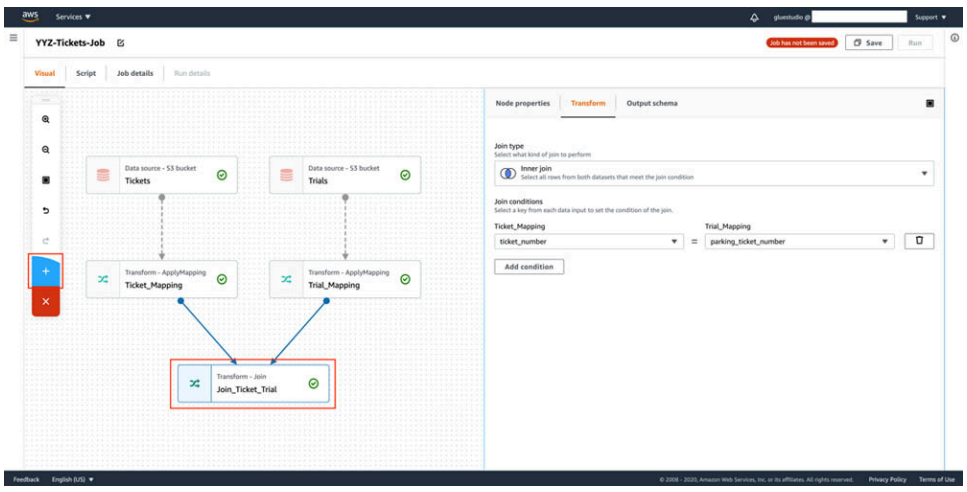
AWS Glue – это ETL-сервис, который упрощает перекладывание, подготовку и простое преобразование данных.

Основные характеристики:

- Предлагает различные схемы для хранения данных;
- Позволяет создавать задания ETL из консоли AWS;
- Можно вставлять куски кода на Python и SQL.

Применима лишь в случае, если ваша инфраструктура в клауде (например, в Амазоне).

Выглядит следующим образом:



Имеются стадии, стрелочки взаимосвязей между стадиями. Также мы можем подправлять код каждой стадии в ячейках и организовать ETL.

Apache Nifi

Apache Nifi – это Open-source ETL/ELT-инструмент.

Основные характеристики:

- Умеет работать со множеством систем, причем не только класса BigData и Data Warehousing;
- Работа с конкретной СУБД реализуется за счет добавления соответствующего JDBC драйвера;
- API для написания своего модуля в качестве дополнительного приемника или преобразователя данных;
- Графический веб-интерфейс для создания DataFlow;
- Метрики и мониторинг доступны по REST API.

Apache Airflow и Python

Apache Airflow и Python – универсально настраиваемые платформы с открытым исходным кодом.

Основные характеристики:

- Позволяют создавать, планировать и отслеживать воркфлоу;
- Имеют массу плагинов и интеграций, например, с Airbyte и Meltano;
- Масштабируемы до BigData;
- Интегрируются с облачными платформами;
- Пока что «нормально» не умеют в декларативные DAG’и.

«Убийцы» Apache Airflow

На сегодняшний день существуют три проекта, заявляющих о себе как об альтернативах Airflow, которые по каким-то параметрам лучше:

- **Prefect**

Основные характеристики:

- Минималистичный;
- Имеет весьма ограниченные возможности по безопасности и контролю над воркфлоу;
- Кодовая база проще и чище;
- Конфигурируется YAML’ом.

- **Dagster**

Основные характеристики:

- Создан бывшими разработчиками Airflow;
- Подходит для небольших команд;
- Лучше заточен под облака;
- Мало что умеет за пределами простого запуска задач.

- **Mage (появился сравнительно недавно)**

Напомним: суперновое не всегда лучше. Airflow может решать более широкий круг задач.

Для объективного сравнения советуем посмотреть разные видеоролики с конференции «SmartData», где спикеры подробно рассказывают о том, как работают данные проекты, а также проводят сравнения между ними и Airflow.

Как вам урок? 😠 😐 🔥

Изучил, далее >