СЛЁРМ Обучение для инженеров и технических лидеров

Дата-инженер

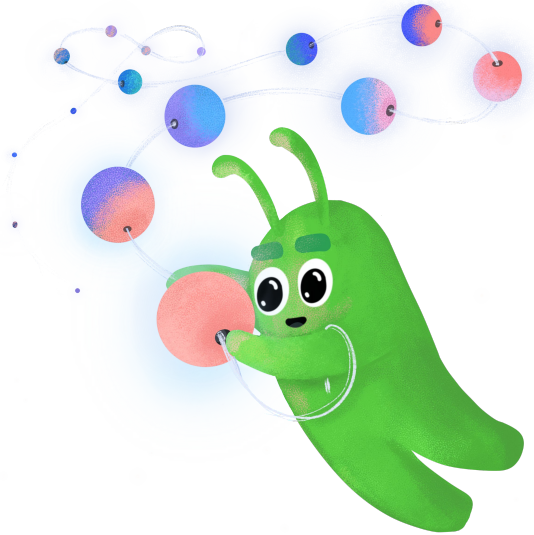
Пайплайны. Оркестрация данных. Обзор ETL-инструментов

Николай Марков

Цели урока. Что вы узнаете:

1. Что означают слова «пайплайн», «воркфлоу», «конвейер» и прочие профессиональные термины
2. Как устроена оркестрация данных
3. Критерии выбора инструмента оркестрации
4. Пару слов о существующих рыночных решениях

Пайплайны

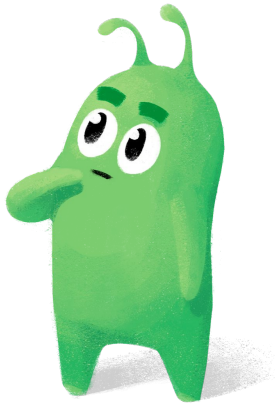


ETL / ELT и пайплайны (конвейеры данных)

01 Часто термины **ETL / ELT** и конвейеры данных (**пайплайны**, **data pipelines**) используют взаимозаменяемо

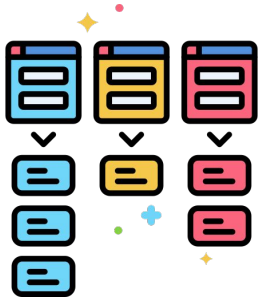
02 **Пайплайн** – более широкий термин, охватывающий разные стадии процессинга и преобразования данных (**ETL** и **ELT** могут быть его **подмножествами)**

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

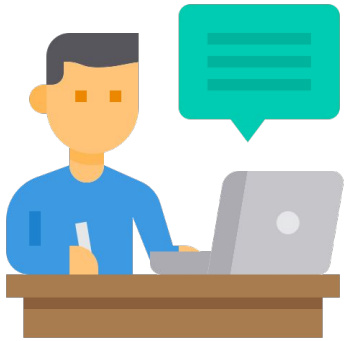


ETL / ELT и пайплайны (конвейеры данных)

- 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 используется еще одно слово — «сценарии»

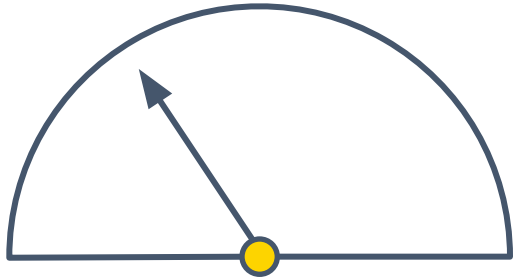


Как сделать свой?



Написать самостоятельно

Взять готовый



Критерии выбора

Connectivity (возможность подключения)
У инструмента должны быть коннекторы для ваших источников и целевых систем

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

Disaster Recovery (восстановление после сбоев)
Возможность с минимальными потерями восстановить работоспособность в случае нарушения работы

Варианты запуска

Отслеживание изменений в источнике (ключевые слова – inotify, Change Data Capture)

По расписанию (ключевое слово – cron)

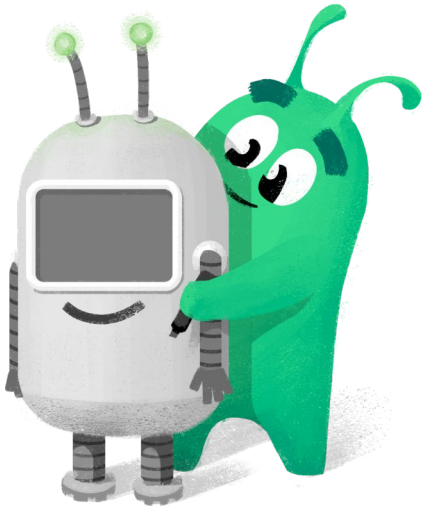
Интеграция данных

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

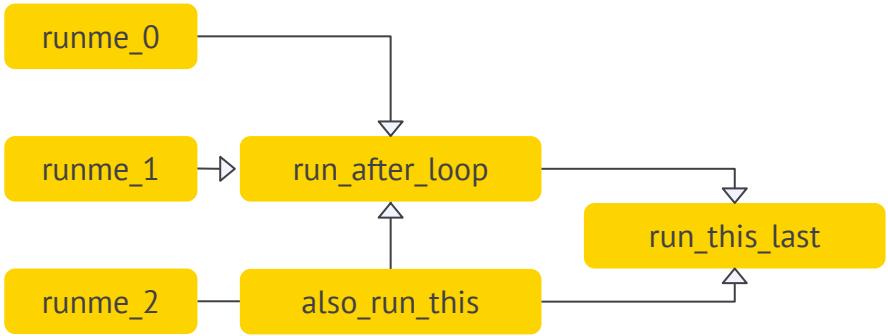
В идеальном случае перекладывание данных между разными системами для последующего анализа должно быть максимально простой операцией

Оркестрация данных



Что такое Оркестрация данных?

- Между разными этапами пайплайна могут быть **сложные зависимости**
- Кроме того, может потребоваться **передавать артефакты** из предыдущего этапа в следующий
- Также необходимо **отслеживать прогресс** выполнения **направленного ациклического графа (DAG)** задач и **сигнализировать об ошибках**
- Английский термин **«workflow management»** часто на русский переводят как **«оркестрация»**, а в применении к задачам DE — **«оркестрация данных»**



Популярные решения оркестрации данных

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

У вас облако и лень писать самому?

AIRBYTE

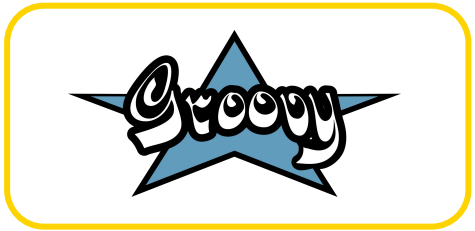
Fivetran

meltano

MATILLION

Эти решения тоже относятся к так называемому Modern Data Stack

Все-таки пишем сами



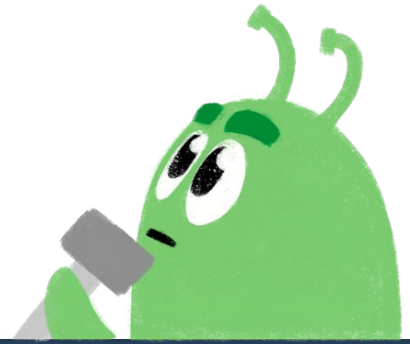
Так, все-таки, ETL или оркестратор данных?

1. Задача **оркестратора – запускать стадии** в DAG’е. Задача **ETL – перекладывать** данные. Иногда и то, и другое встречается в одном инструменте

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

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

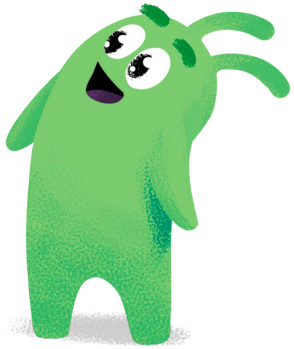
Современные ETL-инструменты



Мечты менеджеров

Современные инструменты ETL корпоративного уровня обычно включают следующие функции:

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



Квадрат Gartner’a

Figure 1: Magic Quadrant for Data Integration Tools

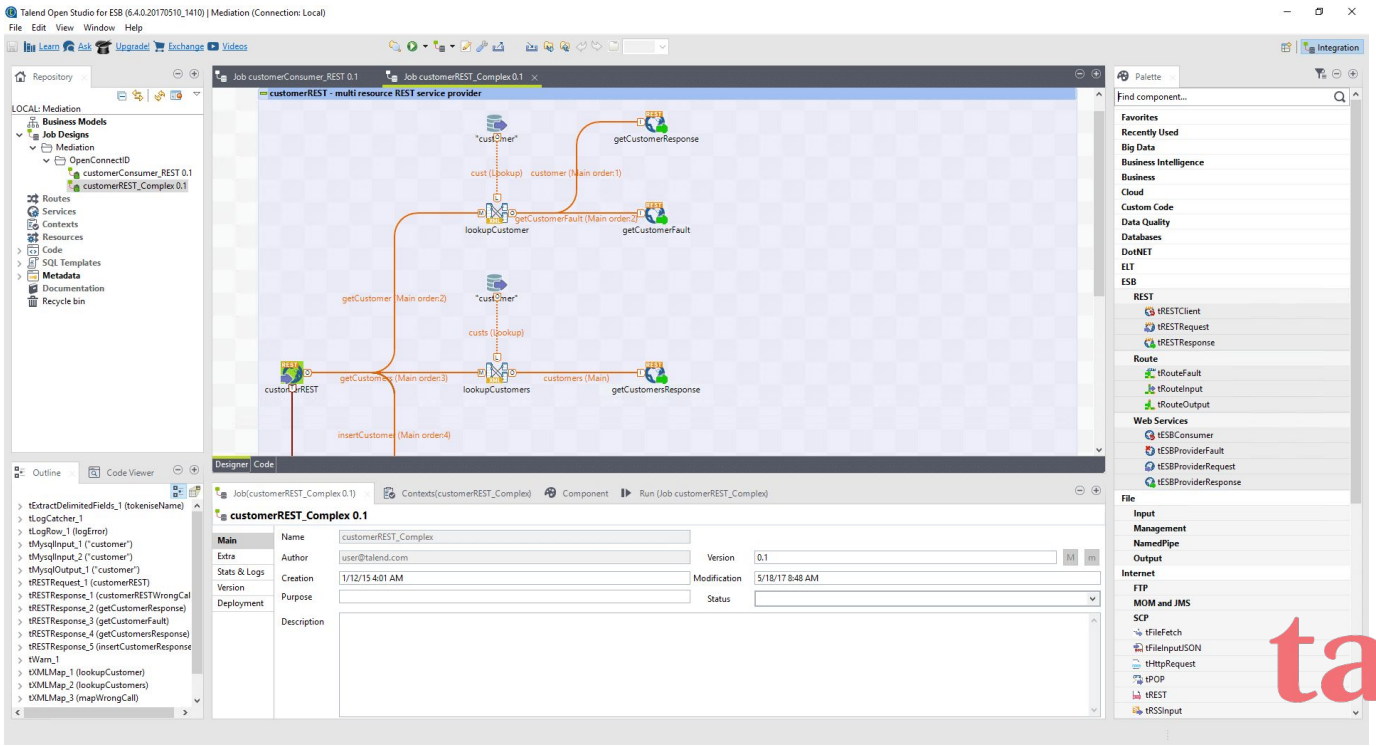
Source: Gartner (August 2022)

Talend Open Studio

- 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

talend

Talend Open Studio

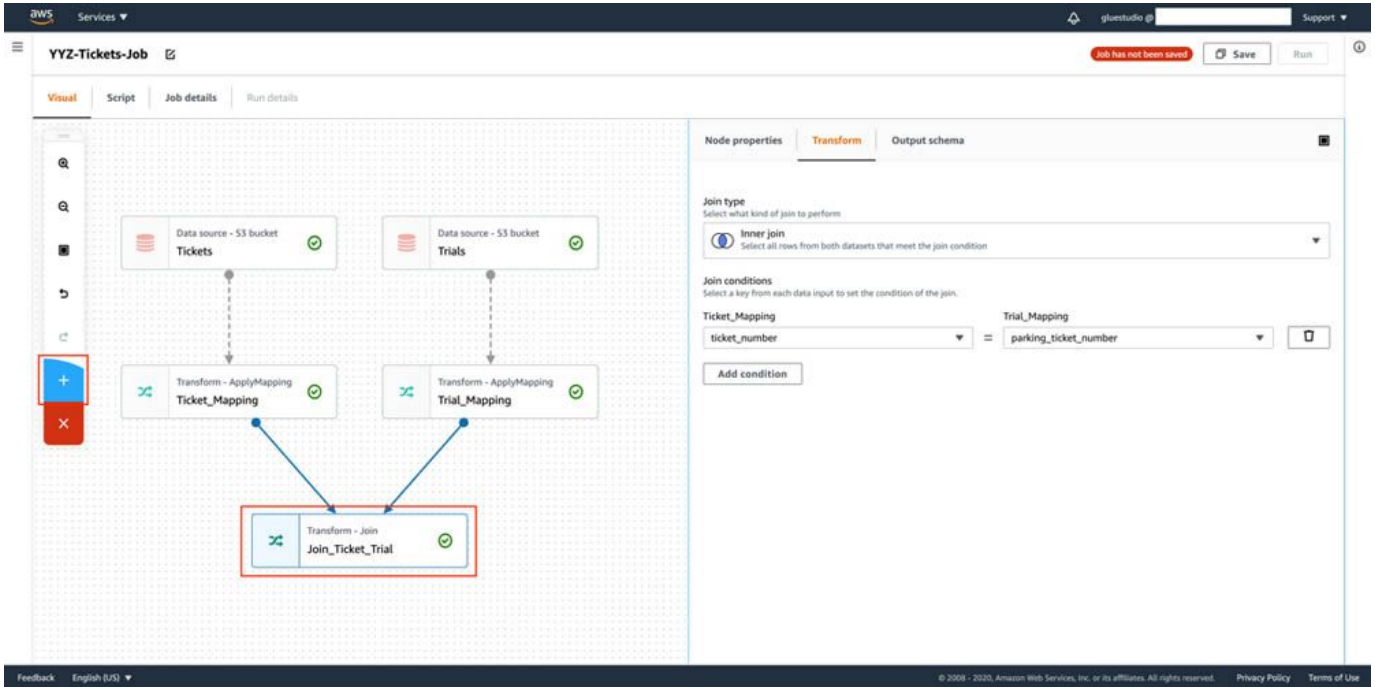


talend

AWS Glue

- Сервис ETL, упрощающий подготовку данных для аналитики
- Предлагает различные схемы для хранения данных
- Позволяет создавать задания ETL из консоли AWS
- Можно вставлять куски кода на Python и SQL

AWS Glue



Apache Nifi

- Open-source ETL/ELT-инструмент
- 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

- Универсально настраиваемая платформа с открытым исходным кодом
- Позволяет создавать, планировать и отслеживать воркфлоу
- Имеет массу плагинов и интеграций, например, с Airbyte и Meltano
- Масштабируема до BigData
- Интегрируется с облачными платформами
- Пока что «нормально» не умеет в декларативные DAG’и

«Убийцы» Apache Airflow

- Довольно минималистичен
- Имеет весьма ограниченные возможности по безопасности и контролю над воркфлоу
- Кодовая база сильно проще и чище
- Конфигурируется YAML'ом

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

Также недавно появился

Итоги. О чем поговорили:

1. Пайплайн охватывает разные стадии процессинга и преобразования данных, ETL/ELT — подмножества
2. Что такое оркестрация данных и чем она отличается от ETL
3. На какие критерии стоит смотреть при выборе инструмента
4. Какие популярные решения существуют сейчас на рынке

Спасибо
за внимание!