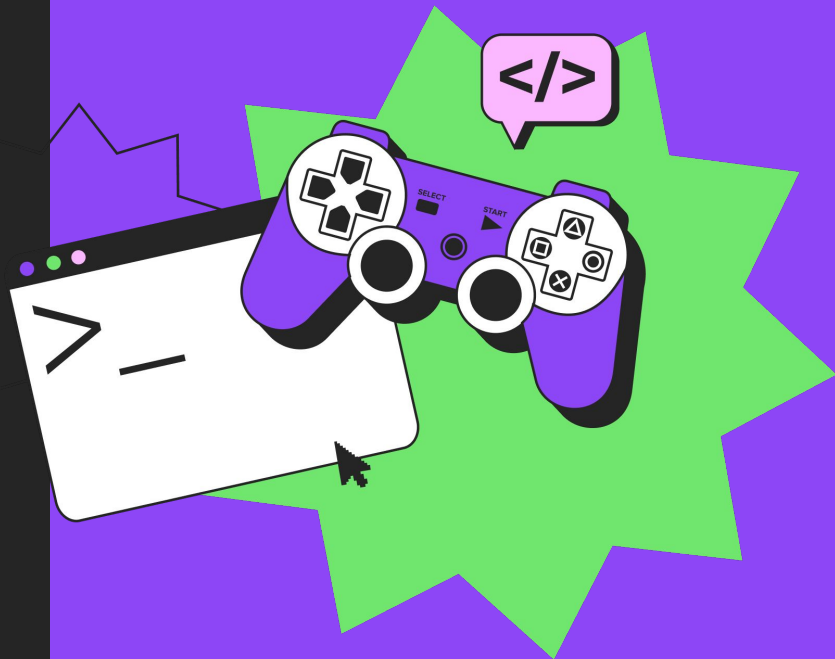


Обзор возможностей Airflow, установка и настройка

Урок 5










План курса (вертикальный)

- 1 Модели данных и нормализация таблиц. Схема "звезда".
Прошедший урок
- 2 Введение в подготовку данных для аналитиков. Таблицы фактов и таблицы измерений.
Прошедший урок
- 3 Получение денормализованных таблиц из нормализованных.
Прошедший урок
- 4 Партиционирование данных.
Сегодняшний урок

- 5 Обзор возможностей Airflow, установка и настройка.
Будущий урок
- 6 Операторы в Airflow и их применение для ETL.
Будущий урок
- 7 Построение пайплайнов и визуализация потоков данных в Airflow.
Будущий урок
- 8 Специфика применения ETL в различных предметных сферах
Будущий урок



Что будет на уроке сегодня

-  Зачем нужна система управления рабочим процессом
-  Что такое Apache Airflow
-  Принципы Airflow
-  Преимущества Airflow
-  Apache Airflow установка и настройка



Викторина



3 самые большие проблемы для дата-инженеров?

1. Обслуживание конвейера, стратегия обработки данных, интеграция различных систем
2. Маппинг данных, проектирование пайплайна, масштабируемость
3. Автоматизация, дублирование данных, ошибки в данных



3 самые большие проблемы для дата-инженеров?

1. Обслуживание конвейера, стратегия обработки данных, интеграция различных систем
2. Маппинг данных, проектирование пайплайна, масштабируемость
3. Автоматизация, дублирование данных, ошибки в данных



Для чего система управления рабочим процессом (выберите подходящие варианты)?

1. Автоматизация
2. Визуализация
3. Увеличение производительности
4. Контроль качества данных
5. Мониторинг
6. Масштабируемость



Для чего система управления рабочим процессом (выберите подходящие варианты)?

1. Автоматизация
2. Визуализация
3. Увеличение производительности
4. Контроль качества данных
5. Мониторинг
6. Масштабируемость



Что такое DAG?

1. Пайплайн для Airflow
2. Направленный ациклический граф описывающий пайплайн



Что такое DAG?

1. Пайплайн для Airflow
2. Направленный ациклический граф описывающий пайплайн



Что задачи в Airflow?

1. Операторы
2. События в рабочем процессе
3. Будущие указания для Airflow
4. Все варианты верны



Что задачи в Airflow?

1. Операторы
2. События в рабочем процессе
3. Будущие указания для Airflow
4. Все варианты верны



Что такое датчики или сенсоры в Airflow?

1. Операторы которые ждут наступления определенных событий
2. Операторы для хранения специфической информации о состоянии пайплайна
3. Все варианты верны



Что такое датчики или сенсоры в Airflow?

1. **Операторы которые ждут наступления определенных событий**
2. Операторы для хранения специфической информации о состоянии пайплайна
3. Все варианты верны



Как операторы связаны с задачами?

1. Операторы не связаны с задачами
2. Операторы описывают задачи
3. Операторы позволяют выполнять различные типы задач в зависимости от функциональности



Как операторы связаны с задачами?

1. Операторы не связаны с задачами
2. Операторы описывают задачи
3. Операторы позволяют выполнять различные типы задач в зависимости от функциональности



Вопросы?

Вопросы?



Вопросы?





Практика



Установка Airflow

Чтобы начать работу с Apache Airflow нам необходимо развернуть его на нашем устройстве. Существует несколько способов сделать это:

1. Установка из исходников
2. Установка с использованием PyPI
3. Установка с помощью docker images



Задание 1

Используя материалы лекции установите Apache Airflow в чистый virtualenv с использованием constraint файла. Убедитесь что по адресу localhost:8080 открывается список Dag сценариев



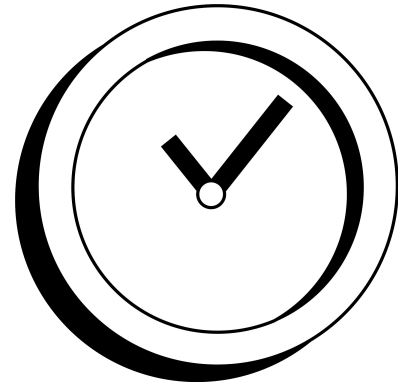
40 минут



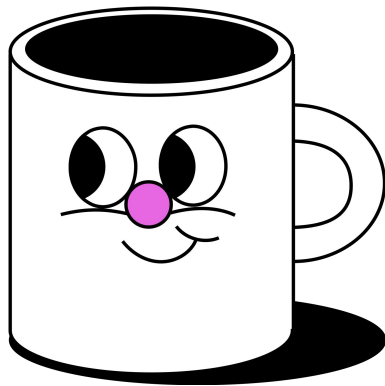
Задание 1

Используя материалы лекции установите Apache Airflow в чистый virtualenv с использованием constraint файла. Убедитесь что по адресу localhost:8080 открывается список Dag сценариев

<<40:00->>



Перерыв



<<5:00->>



Задание 2

Создайте нового пользователя с правами админа и авторизуйтесь с его помощью в интерфейсе airflow

A large teal circle containing the text '15 минут' in white, indicating a 15-minute time limit for the task.

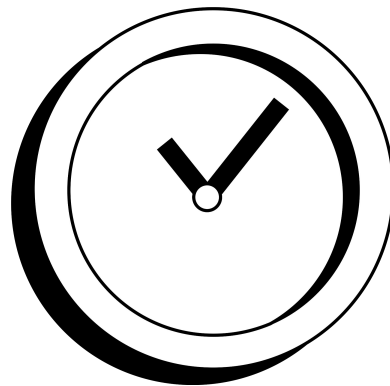
15 минут



Задание 2

Создайте нового пользователя с правами админа и авторизуйтесь с его помощью в интерфейсе airflow

<<15:00->>





Задание 3

Посмотрите список существующих пайплайнов. Запустите некоторые из них, посмотрите на результат выполнения. Откройте логи выполнения пайплайнов. Отправьте в чат скриншот из логов, говорящий об успешном завершении пайплайна

A large teal circle on the right side of the slide, containing the text '15 минут' in white.

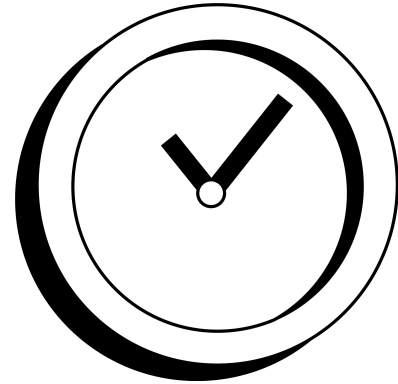
15 минут



Задание 3

Посмотрите список существующих пайплайнов. Запустите некоторые из них, посмотрите на результат выполнения. Откройте логи выполнения пайплайнов. Отправьте в чат скриншот из логов, говорящий об успешном завершении пайплайна

<<15:00->>





Вопросы?

Вопросы?



Вопросы?





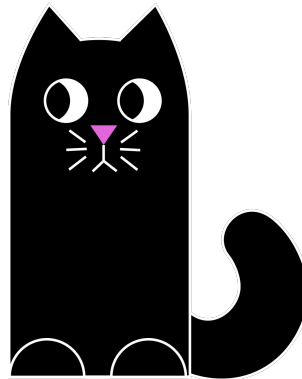
Домашнее задание



Домашнее задание

На основе сайта yandex.ru:

- Определите, на каком протоколе работает сайт.
- Проанализируйте структуру страницы сайта
- Внесите не менее 10 изменений на страницу с помощью инструмента разработчика и представьте скриншоты было/стало.
- Создайте прототип низкой детализации (дополнительное задание, если на семинаре дошли до задания №8)





Спасибо за внимание

