

Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение высшего
образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

Проектный практикум по разработке ETL-решений

Вебинар 21-03-2025

Практическая работа

Выполнила: Сачкова Г.Г (st_98)., группа: АДЭУ-211

Преподаватель:

Москва

2025

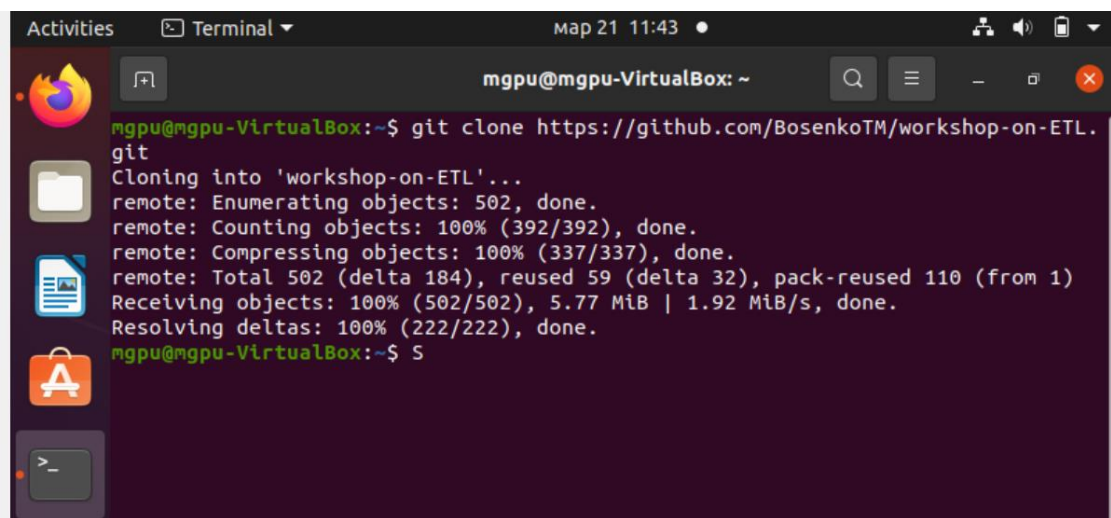
Задачи:

1. Развернуть ВМ ubuntu_mgpu.ova в VirtualBox.
2. Клонировать на ПК задание бизнес-кейс Umbrella в домашний каталог ВМ.
3. Запустить контейнер 01_umbrella.py с кейсом, изучить и описать основные элементы интерфейса Apache Airflow.
4. Спроектировать верхнеуровневую архитектуру

11 вариант. Получить прогноз в Лос-Анджелесе на 5 дней. Отсортировать по температуре (по убыванию). Сохранить топ-3 самых жарких дня.

Ход работы:

В первую очередь была развернута виртуальная машина и клонировано задание в домашний репозиторий (рисунок 1).



```
mgpu@mgpu-VirtualBox: ~  
$ git clone https://github.com/BosenkoTM/workshop-on-ETL.  
git  
Cloning into 'workshop-on-ETL'...  
remote: Enumerating objects: 502, done.  
remote: Counting objects: 100% (392/392), done.  
remote: Compressing objects: 100% (337/337), done.  
remote: Total 502 (delta 184), reused 59 (delta 32), pack-reused 110 (from 1)  
Receiving objects: 100% (502/502), 5.77 MiB | 1.92 MiB/s, done.  
Resolving deltas: 100% (222/222), done.  
mgpu@mgpu-VirtualBox: ~ $
```

Далее был изучен интерфейс Apache Airflow, описание основных элементов изложено ниже:

- DAG (Directed Acyclic Graph). Это структура, состоящая из объектов (узлов), которые связаны между собой. DAG описывает логику выполнения задач: какие должны быть выполнены, в каком порядке и как часто.
- Задача (Task). Описывает, что делать, например, выборку данных, анализ, запуск других систем. Каждая задача — это экземпляр оператора с определёнными параметрами.
- Оператор (Operator). Класс Python, который определяет, что нужно сделать в рамках задачи. Есть операторы для выполнения скриптов Bash, кода Python, SQL-запросов.
- Веб-сервер. Предоставляет пользовательский интерфейс для мониторинга, управления и запуска задач. Через веб-интерфейс пользователи могут просматривать список задач, проверять их статус и управлять расписанием выполнения.

- База метаданных. Хранит информацию о задачах, их статусе, зависимостях и истории выполнения

17 ✓
colab

```
[64] from google.colab import files
      uploaded = files.upload()
```

Выбрать файлы Число файлов: 2

- joined_data (3).csv(text/csv) - 127 bytes, last modified: 21.03.2025 - 100% done
- ml_model (1).pkl(n/a) - 926 bytes, last modified: 21.03.2025 - 100% done

Saving joined_data (3).csv to joined_data (3) (1).csv
Saving ml_model (1).pkl to ml_model (1) (3).pkl

2 ✓
colab

```
[22] !pip install dill
```

Requirement already satisfied: dill in /usr/local/lib/python3.11/dist-packages (0.3.9)

0 ✓
colab

```
[32] import joblib
      model = joblib.load("ml_model (1).pkl")
      import pandas as pd
      print(model.predict(pd.DataFrame({'temperature': [0, 5, 10, 15, 20, 25]})))
```

```
[ 86.15384615  57.30769231  28.46153846  -0.38461538 -29.23076923
 -58.07692308]
```

0 ✓
colab

```
[62] df = pd.read_csv('joined_data (3).csv')
```

df

	date	temperature	sales
0	2025-03-21	16.3	10
1	2025-03-22	14.7	15
2	2025-03-23	15.9	20
3	2025-03-24	16.3	25
4	2025-03-25	16.1	30

Далее: [Посмотреть рекомендованные графики](#) [New interactive sheet](#)

Рисунок 5 – Просмотр файла с предсказаниями

Далее по индивидуальному заданию надо отсортировать по убыванию температуры, а также вывести топ 3 жарких дня (рисунок 6).

Сортировка по убыванию температуры

```
[65] sorted_by_temperature = df.sort_values(by='temperature', ascending=False)
sorted_by_temperature
```

	date	temperature	sales
0	2025-03-21	16.3	10
3	2025-03-24	16.3	25
4	2025-03-25	16.1	30
2	2025-03-23	15.9	20
1	2025-03-22	14.7	15

Далее: [Посмотреть рекомендованные графики](#) [New interactive sheet](#)

Топ 3 жарких дня

```
sorted_by_temperature.head(3)
```

	date	temperature	sales
0	2025-03-21	16.3	10
3	2025-03-24	16.3	25
4	2025-03-25	16.1	30

Далее: [Посмотреть рекомендованные графики](#) [New interactive sheet](#)

Рисунок 6 – Сортировка по убыванию температуры и топ 3 жарких дня

Затем была построена архитектура кейса, которая представлена на рисунке 7.

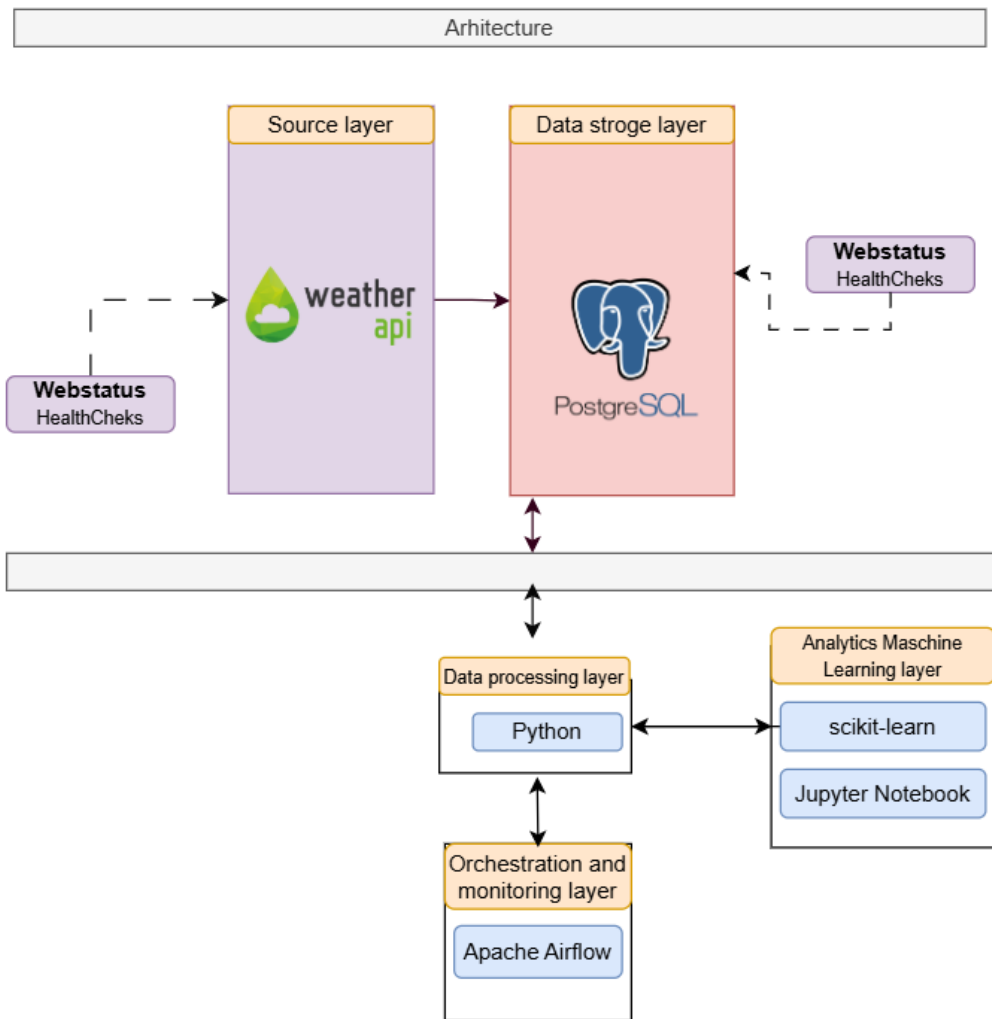


Рисунок 7 – Архитектура кейса

Выводы:

1. Получен прогноз в Лос-Анджелесе на 5 дней
2. Отсортирован по температуре (по убыванию)
3. Сохранены топ-3 самых жарких дня
4. Построена архитектура кейса
5. Описаны основные элементы Apache Airflow