

Модуль
Python для сетевого
программирования и
автоматизации



Домашняя работа № 3. Python для ETL (итоговый проект)

Python для продвинутых
специалистов

Домашнее работа № 3

Цель работы: научиться создавать потоки работ (pipeline) ETL.

В итоговом проекте мы объединим результаты, полученные в предыдущих домашних работах, и построим единый поток работ ETL.

Задание

Разработайте поток работ для Apache AirFlow, который должен включать следующие этапы:

1. Загрузка архива Единого государственного реестр юридических лиц в формате JSON по ссылке <https://ofdata.ru/open-data/download/egrul.json.zip>.
2. Выбор из реестра компаний, занимающихся деятельностью в сфере телекоммуникаций (Группировка ОКВЭД 61).
3. Поиск на сайте hh.ru вакансий middle python developer от компаний, занимающихся деятельностью в сфере телекоммуникаций. Допускается использовать парсинг Web-страниц или REST API.
4. Составление списка ТОП-10 наиболее востребованных ключевых навыков middle python developer в компаниях, занимающихся деятельностью в сфере телекоммуникаций.

Для сдачи задания пришлите исходные коды программы на Python.