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

Мошенина Елена Дмитриевна БД-241м

Программные средства сбора, консолидации и аналитики данных

**Вариант 14**

**Лабораторная работа 3. Программные средства консолидации данных из различных источников с использованием Python и Apache Airflow**

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Руководитель дисциплины:

Босенко Т.М., доцент департамента

информатики, управления и технологий,

доктор экономических наук

Москва

2025

**Цель работы:** освоить практические навыки проектирования и автоматизации ETL-процессов (Extract, Transform, Load) с использованием Apache Airflow. научиться создавать конвейеры данных (DAG), которые извлекают информацию из разнородных источников, выполняют их консолидацию и трансформацию с помощью Python, и загружают результат в целевую базу данных с отправкой уведомлений.

**Оборудование и ПО:**

* система контейнеризации Docker и Docker Compose.
* Apache Airflow (разворачивается в Docker).
* база данных SQLite.
* Python 3.x с библиотеками pandas, openpyxl.
* email-сервис для настройки уведомлений (например, MailHog,

входящий в сборку).

**Порядок выполнения работы**

1. **Подготовка окружения:**
   * клонируйте репозиторий проекта git clone https://github.com/BosenkoTM/DCCAS.git и перейдите в каталог **business\_case\_umbrella**.
   * запустите все сервисы (Airflow, Postgres, MailHog) с помощью команды **docker compose up -d**.
   * убедитесь в доступе к веб-интерфейсу Airflow по адресу

[**http://localhost:8080**.](http://localhost:8080/)

1. **Анализ бизнес-кейса и проектирование DAG:**
   * выберите ваш вариант задания из таблицы ниже. каждое задание представляет собой бизнес-кейс, требующий автоматизации сбора и обработки данных.
   * спроектируйте логику вашего DAG (Directed Acyclic Graph): определите последовательность задач (tasks), их зависимости и итоговый результат.
2. **Разработка DAG:**
   * в папке **dags** создайте Python-файл для вашего DAG.
   * реализуйте **Extract**: напишите Python-функции для чтения данных из трех источников (CSV, Excel, JSON).
   * реализуйте **Transform**: напишите Python-функцию, которая принимает данные из предыдущего шага, выполняет их консолидацию, очистку, обогащение и аналитические расчеты согласно вашему заданию, используя библиотеку pandas.
   * реализуйте **Load**: напишите Python-функцию для сохранения обработанных данных в базу данных SQLite.
   * определите **задачи (Operators)** в вашем DAG, связав их с разработанными Python-функциями (PythonOperator).
   * настройте **уведомления:** добавьте в конец DAG EmailOperator для отправки отчета об успешном выполнении на тестовый email.
   * установите зависимости между задачами (>>, <<).
3. **Тестирование и запуск:**
   * поместите исходные файлы с данными в папку dags/data.
   * в веб-интерфейсе Airflow активируйте ваш DAG и запустите его выполнение вручную.
   * отследите выполнение всех задач, проверьте логи в случае ошибок.
   * убедитесь, что данные корректно загрузились в SQLite (можно проверить с помощью утилиты sqlite3 внутри контейнера или написав отдельный скрипт).
   * проверьте получение email-уведомления в интерфейсе MailHog

([**http://localhost:8025**).](http://localhost:8025/)

1. **Подготовка отчета и исходного кода:**
   * подготовьте электронный отчет согласно требованиям.
   * опубликуйте ваш исходный код (файл DAG и вспомогательные скрипты) в публичном Git-репозитории.

**Варианты заданий: бизнес-кейсы для автоматизации**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Файл 1 (CSV)** | **Файл 2 (Excel)** | **Файл 3 (JSON)** | **Аналитическая задача DAG** |
| **14** | **Пользователи приложения:** user\_id, registration\_date | **Сессии:** user\_id, session\_duration  \_minutes | **Покупки в приложении:** user\_id, purchase\_amount | рассчитать среднее время сессии и средний чек для пользователей, зарегистрированн ых в последнем месяце. |

**Основная часть:**

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Содержимое, созданное искусственным интеллектом, может быть неверным.

Остальные также посмотрим:

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Содержимое, созданное искусственным интеллектом, может быть неверным.

Достигнутые цели:

- Успешно реализован ETL-процесс анализа поведения пользователей мобильного приложения

- Освоены практические навыки работы с Apache Airflow для оркестрации данных

- Автоматизирован расчет ключевых метрик: среднее время сессии и средний чек

Технические результаты:

- Спроектирован и настроен DAG с 6 задачами (Extract, Transform, Load, Report, Notify)

- Реализована обработка данных из трех источников: CSV, Excel, JSON

- Настроена автоматическая отправка email-отчетов через MailHog

- Обеспечено сохранение результатов в SQLite базу данных

Полученные навыки:

- Работа с Docker Compose для развертывания Airflow

- Создание и настройка DAG в Apache Airflow

- Обработка данных с помощью Pandas

- Настройка email-уведомлений и мониторинг процессов

- Работа с разноформатными данными

Практическая ценность:

Лабораторная работа продемонстрировала полный цикл создания production-ready ETL-решения для анализа пользовательского поведения, готового к интеграции с реальными системами мобильной аналитики.