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**.
2. **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.
3. **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 (>>, <<).

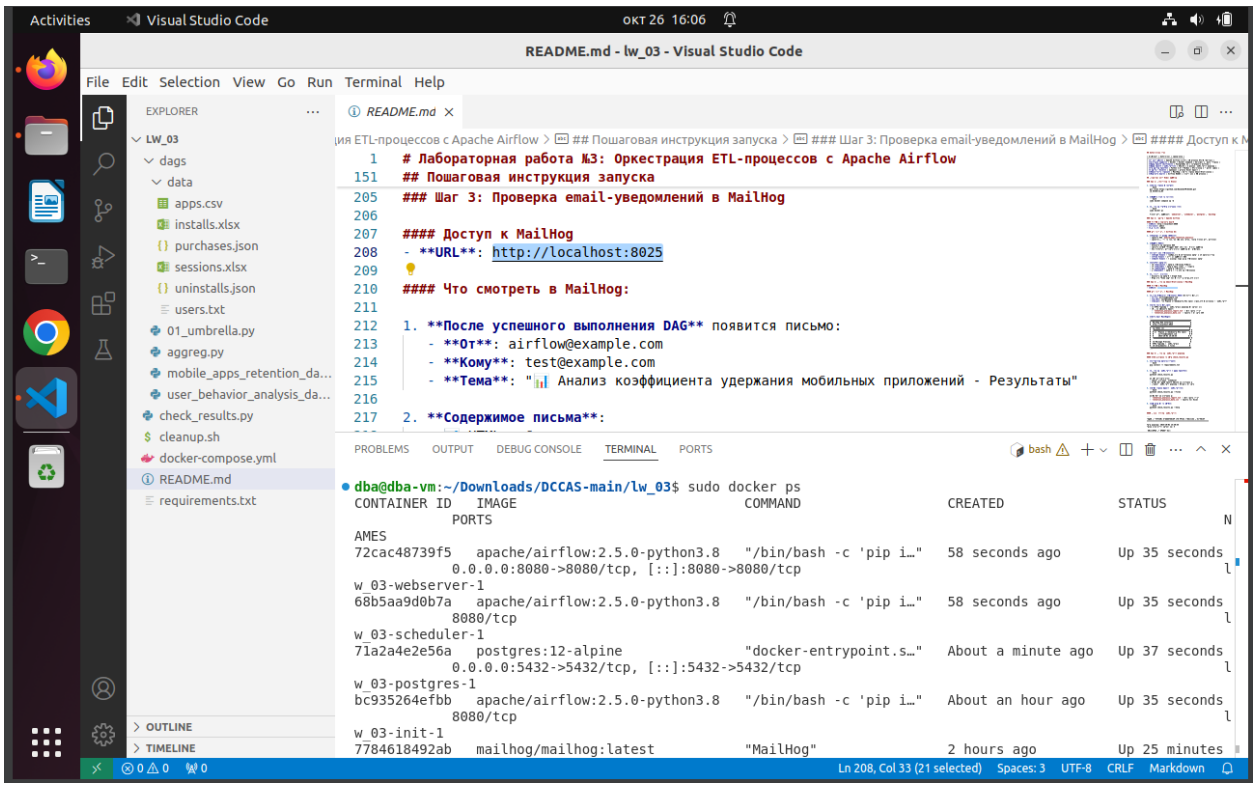
4. **Тестирование и запуск:**
- поместите исходные файлы с данными в папку dags/data.
- в веб-интерфейсе Airflow активируйте ваш DAG и запустите его выполнение вручную.
- отследите выполнение всех задач, проверьте логи в случае ошибок.
- убедитесь, что данные корректно загрузились в SQLite (можно проверить с помощью утилиты sqlite3 внутри контейнера или написав отдельный скрипт).
- проверьте получение email-уведомления в интерфейсе MailHog (**http://localhost:8025**).

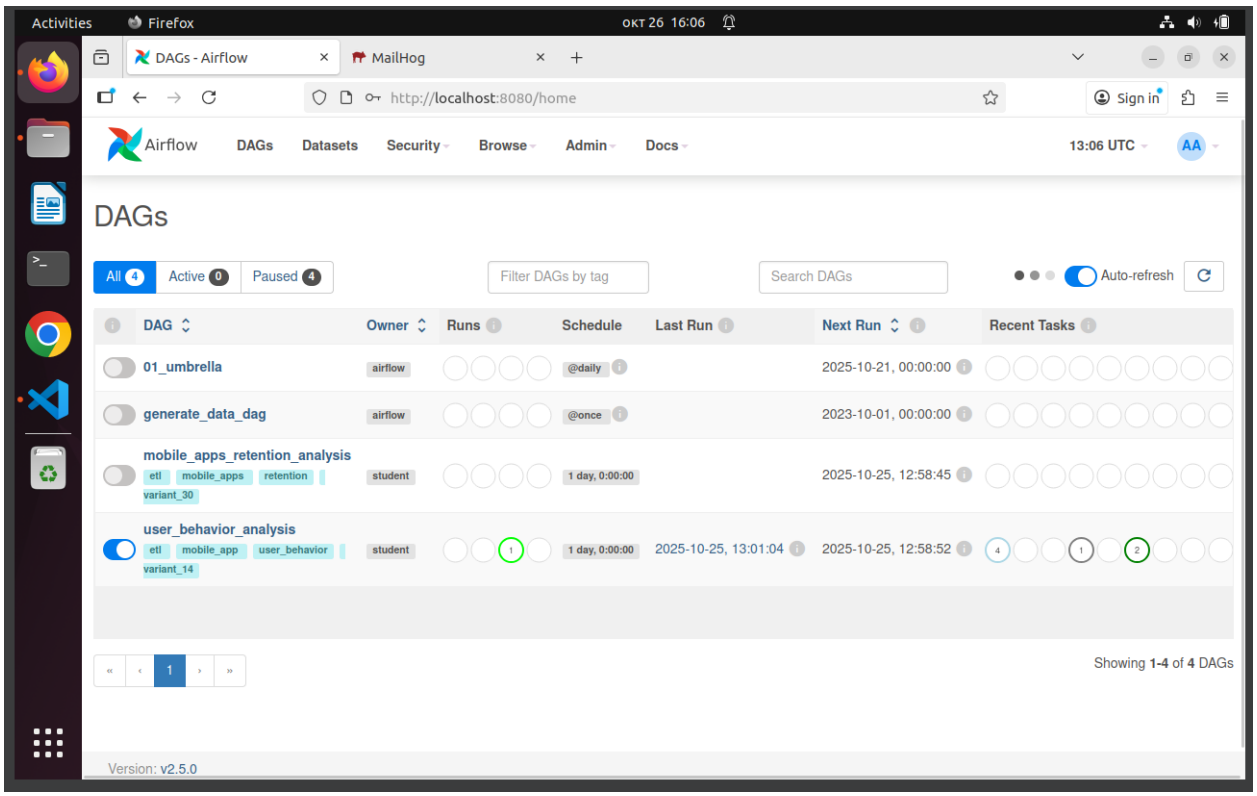
5. **Подготовка отчета и исходного кода:**
- подготовьте электронный отчет согласно требованиям.
- опубликуйте ваш исходный код (файл DAG и вспомогательные скрипты) в публичном Git-репозитории.

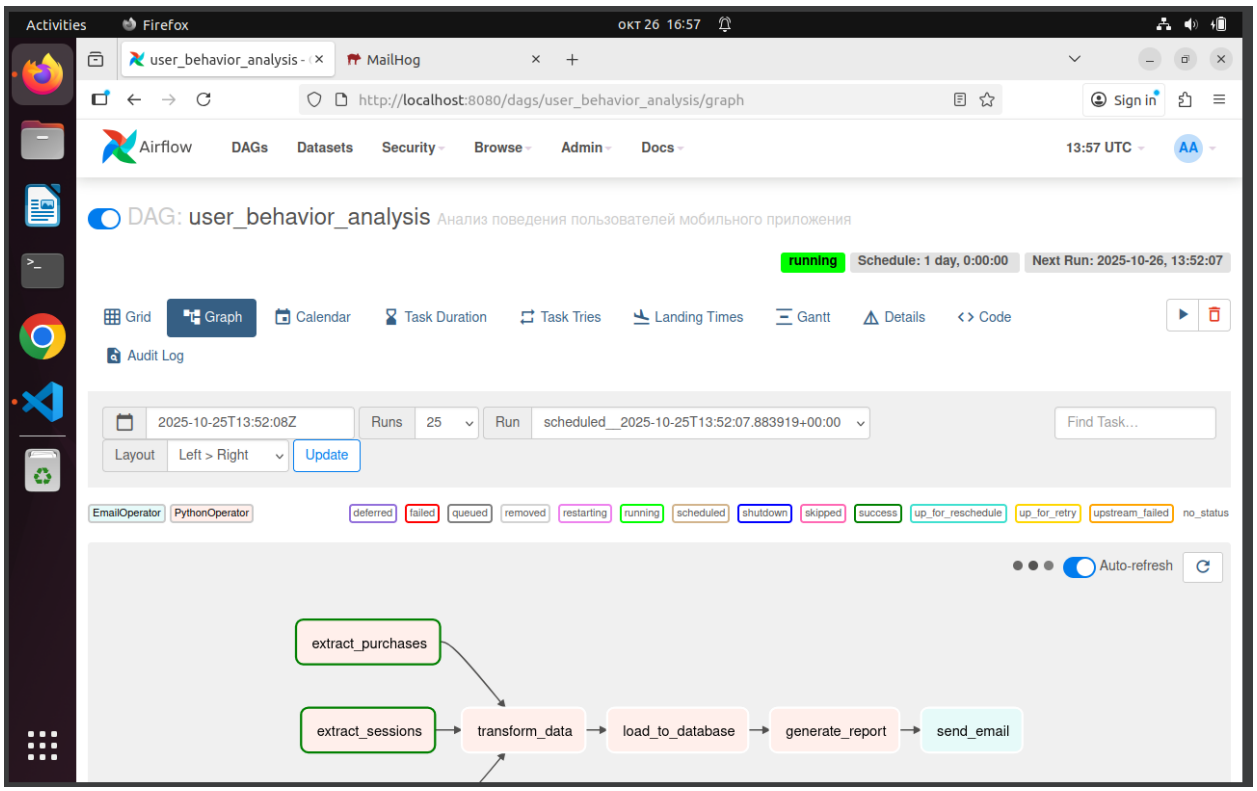
Варианты заданий: бизнес-кейсы для автоматизации

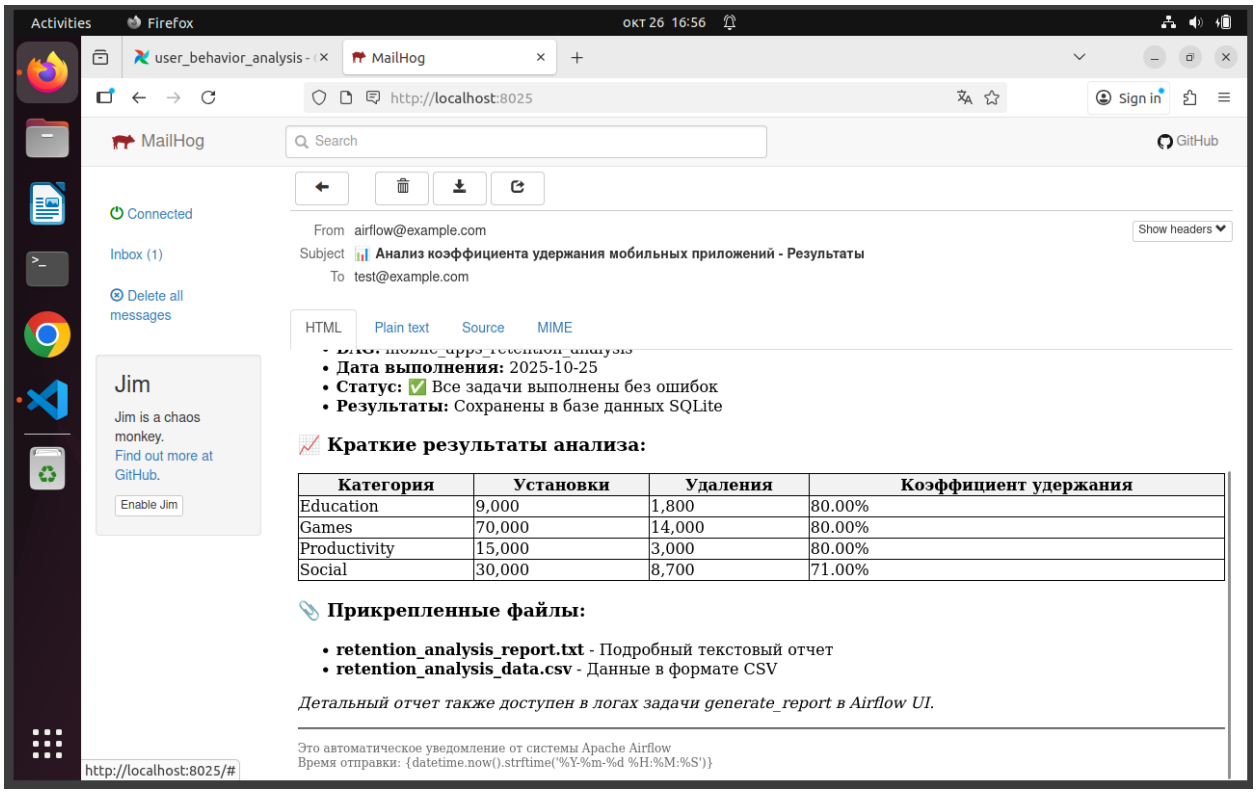
№	Файл 1 (CSV)	Файл 2 (Excel)	Файл 3 (JSON)	Аналитическая задача DAG
14	**Пользователи приложения:** user_id, registration_date	**Сессии:** user_id, session_duration_minutes	**Покупки в приложении:** user_id, purchase_amount	рассчитать среднее время сессии и средний чек для пользователей, зарегистрированных в последнем месяце.

Основная часть:

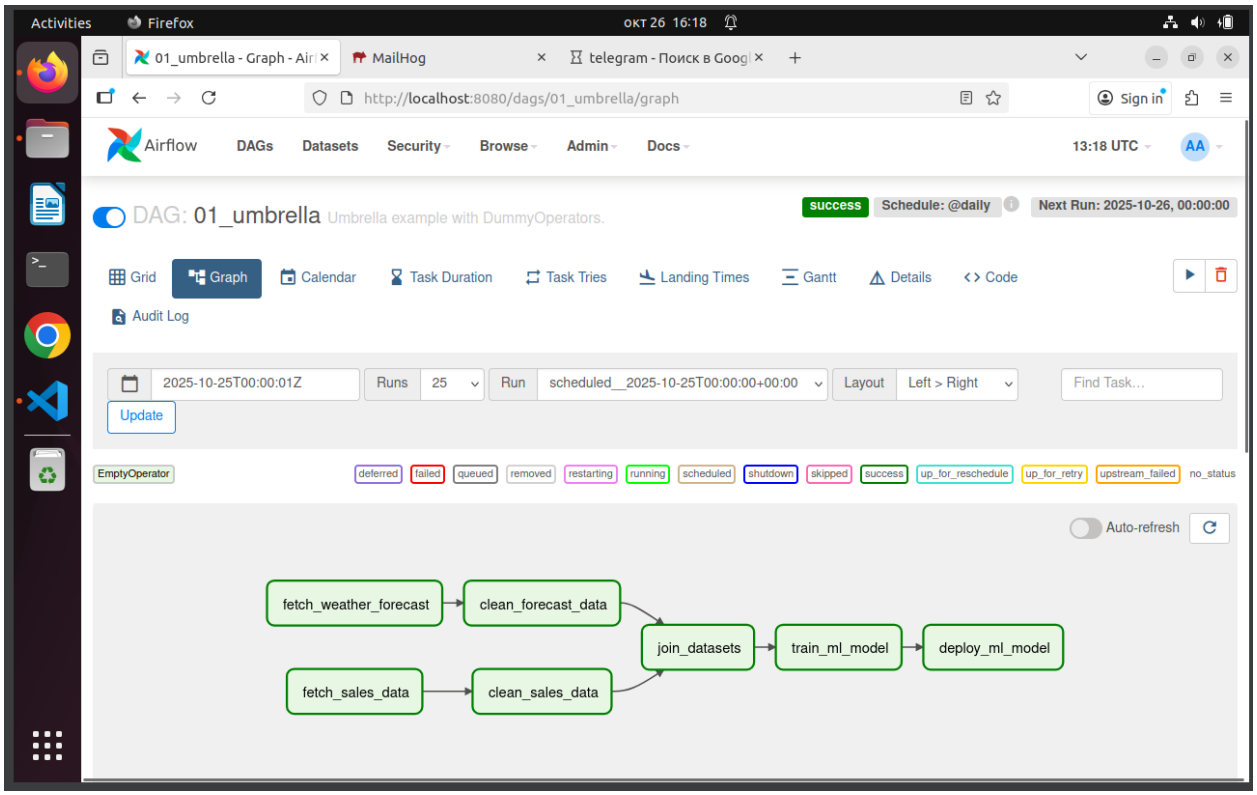

Activities
Visual Studio Code
окт 26 16:06
README.md - lw_03 - Visual Studio Code
File Edit Selection View Go Run Terminal Help
EXPLORER
README.md
LW_03
dags
data
apps.csv
installs.xlsx
purchases.json
sessions.xlsx
uninstalls.json
users.txt
01_umbrella.py
aggreg.py
mobile_apps_retention_da...
user_behavior_analysis_da...
check_results.py
cleanup.sh
docker-compose.yml
README.md
requirements.txt
OUTLINE
TIMELINE
Лабораторная работа №3: Оркестрация ETL-процессов с Apache Airflow
Пошаговая инструкция запуска
Шаг 3: Проверка email-уведомлений в MailHog
Доступ к MailHog
- **URL**: http://localhost:8025
Что смотреть в MailHog:
1. **После успешного выполнения DAG** появится письмо:
- **От**: airflow@example.com
- **Кому**: test@example.com
- **Тема**: "📊 Анализ коэффициента удержания мобильных приложений - Результаты"
2. **Содержимое письма**:
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
bash
dba@dba-vm:~/Downloads/DCCAS-main/lw_03$ sudo docker ps
CONTAINER ID IMAGE COMMAND CREATED STATUS PORTS NAMES
72cac48739f5 apache/airflow:2.5.0-python3.8 "/bin/bash -c 'pip i…" 58 seconds ago Up 35 seconds 0.0.0.0:8080->8080/tcp, [::]:8080->8080/tcp w_03-webserver-1
68b5aa9d0b7a apache/airflow:2.5.0-python3.8 "/bin/bash -c 'pip i…" 58 seconds ago Up 35 seconds 8080/tcp w_03-scheduler-1
71a2a4e2e56a postgres:12-alpine "docker-entrypoint.s…" About a minute ago Up 37 seconds 0.0.0.0:5432->5432/tcp, [::]:5432->5432/tcp w_03-postgres-1
bc935264efbb apache/airflow:2.5.0-python3.8 "/bin/bash -c 'pip i…" About an hour ago Up 35 seconds 8080/tcp w_03-init-1
7784618492ab mailhog/mailhog:latest "MailHog" 2 hours ago Up 25 minutes
Ln 208, Col 33 (21 selected) Spaces: 3 UTF-8 CRLF Markdown

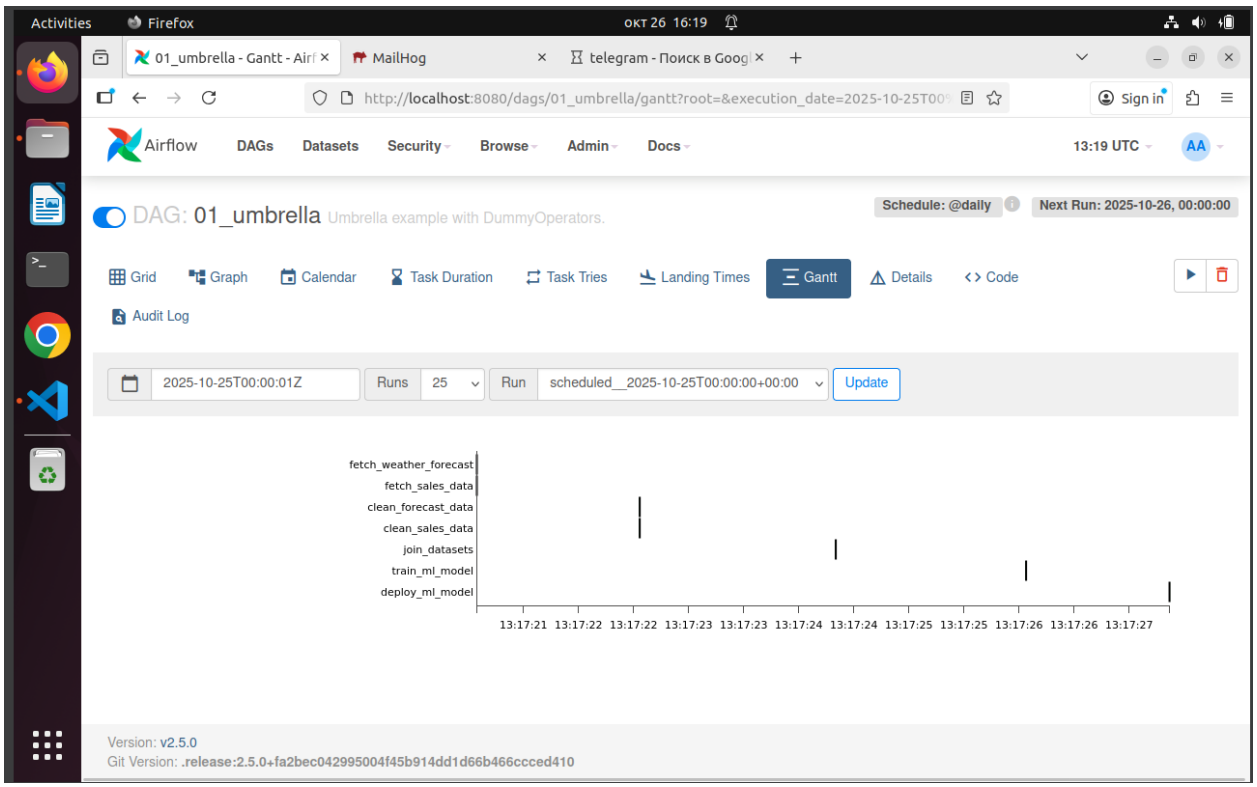

Activities
Firefox
окт 26 16:06
DAGs - Airflow
MailHog
http://localhost:8080/home
Sign in
Airflow
DAGs
Datasets
Security
Browse
Admin
Docs
13:06 UTC
AA
DAGs
All 4
Active 0
Paused 4
Filter DAGs by tag
Search DAGs
Auto-refresh
DAG
Owner
Runs
Schedule
Last Run
Next Run
Recent Tasks
01_umbrella
airflow
@daily
2025-10-21, 00:00:00
generate_data_dag
airflow
@once
2023-10-01, 00:00:00
mobile_apps_retention_analysis
etl
mobile_apps
retention
variant_30
student
1 day, 0:00:00
2025-10-25, 12:58:45
user_behavior_analysis
etl
mobile_app
user_behavior
variant_14
student
1 day, 0:00:00
2025-10-25, 13:01:04
2025-10-25, 12:58:52
Showing 1-4 of 4 DAGs
Version: v2.5.0

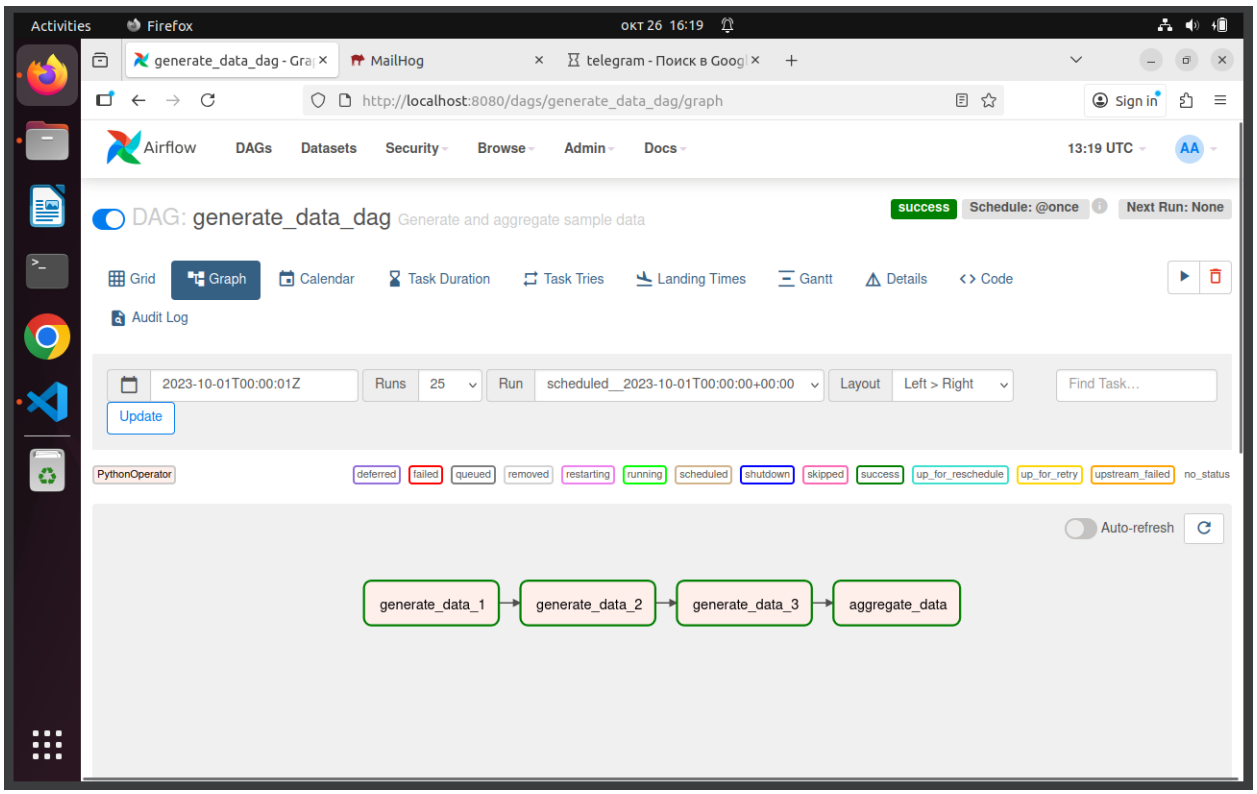

Activities
Firefox
окт 26 16:57
user_behavior_analysis -
MailHog
http://localhost:8080/dags/user_behavior_analysis/graph
Sign in
Airflow
DAGs
Datasets
Security
Browse
Admin
Docs
13:57 UTC
AA
DAG: user_behavior_analysis
Анализ поведения пользователей мобильного приложения
running
Schedule: 1 day, 0:00:00
Next Run: 2025-10-26, 13:52:07
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
2025-10-25T13:52:08Z
Runs
25
Run
scheduled__2025-10-25T13:52:07.883919+00:00
Find Task...
Layout
Left > Right
Update
EmailOperator
PythonOperator
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
extract_purchases
extract_sessions
transform_data
load_to_database
generate_report
send_email

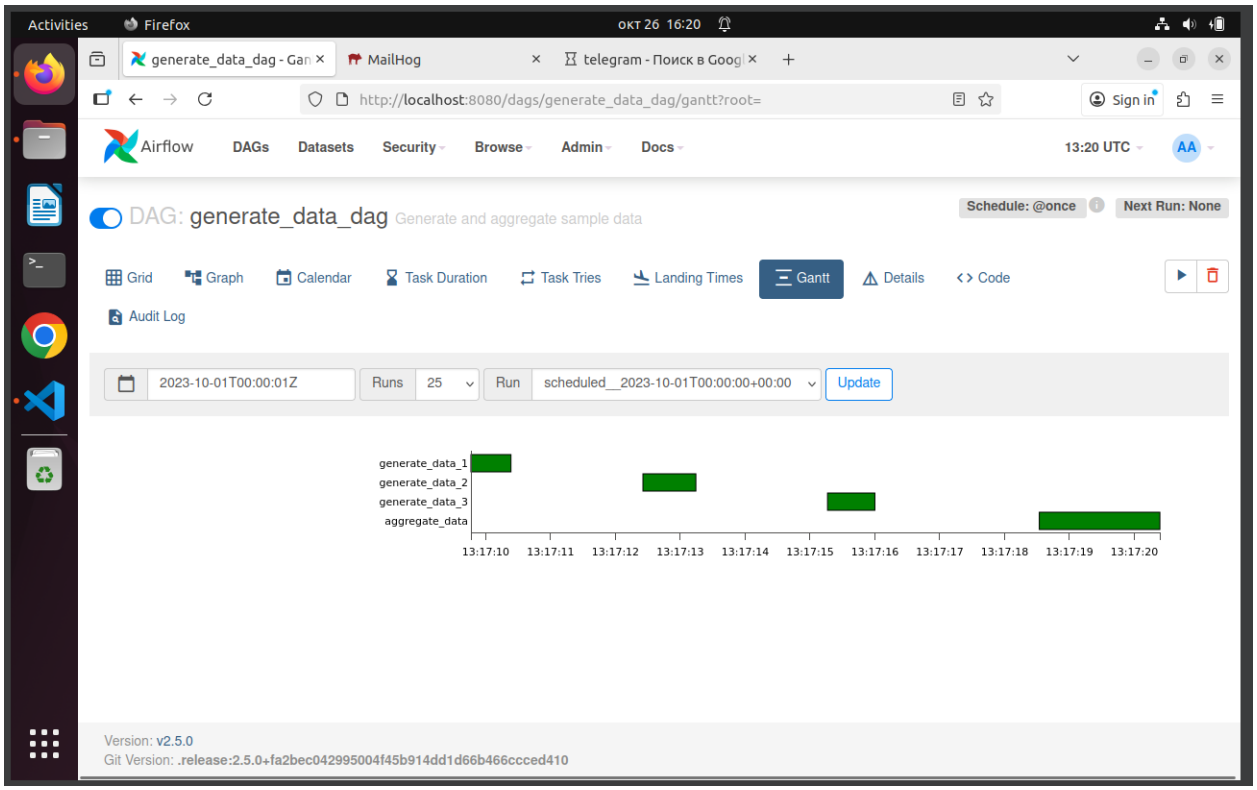

Activities
Firefox
окт 26 16:56
user_behavior_analysis -
MailHog
http://localhost:8025
Sign in
MailHog
Search
GitHub
Connected
Inbox (1)
Delete all messages
Jim
Jim is a chaos monkey.
Find out more at GitHub.
Enable Jim
From airflow@example.com
Subject Анализ коэффициента удержания мобильных приложений - Результаты
To test@example.com
Show headers
HTML
Plain text
Source
MIME
Дата выполнения: 2025-10-25
Статус: ✅ Все задачи выполнены без ошибок
Результаты: Сохранены в базе данных SQLite
Краткие результаты анализа:
Категория | Установки | Удаления | Коэффициент удержания
Education | 9,000 | 1,800 | 80.00%
Games | 70,000 | 14,000 | 80.00%
Productivity | 15,000 | 3,000 | 80.00%
Social | 30,000 | 8,700 | 71.00%
Прикрепленные файлы:
retention_analysis_report.txt - Подробный текстовый отчет
retention_analysis_data.csv - Данные в формате CSV
Детальный отчет также доступен в логах задачи generate_report в Airflow UI.
Это автоматическое уведомление от системы Apache Airflow
Время отправки: {datetime.now().strftime('%Y-%m-%d %H:%M:%S')}
http://localhost:8025/#

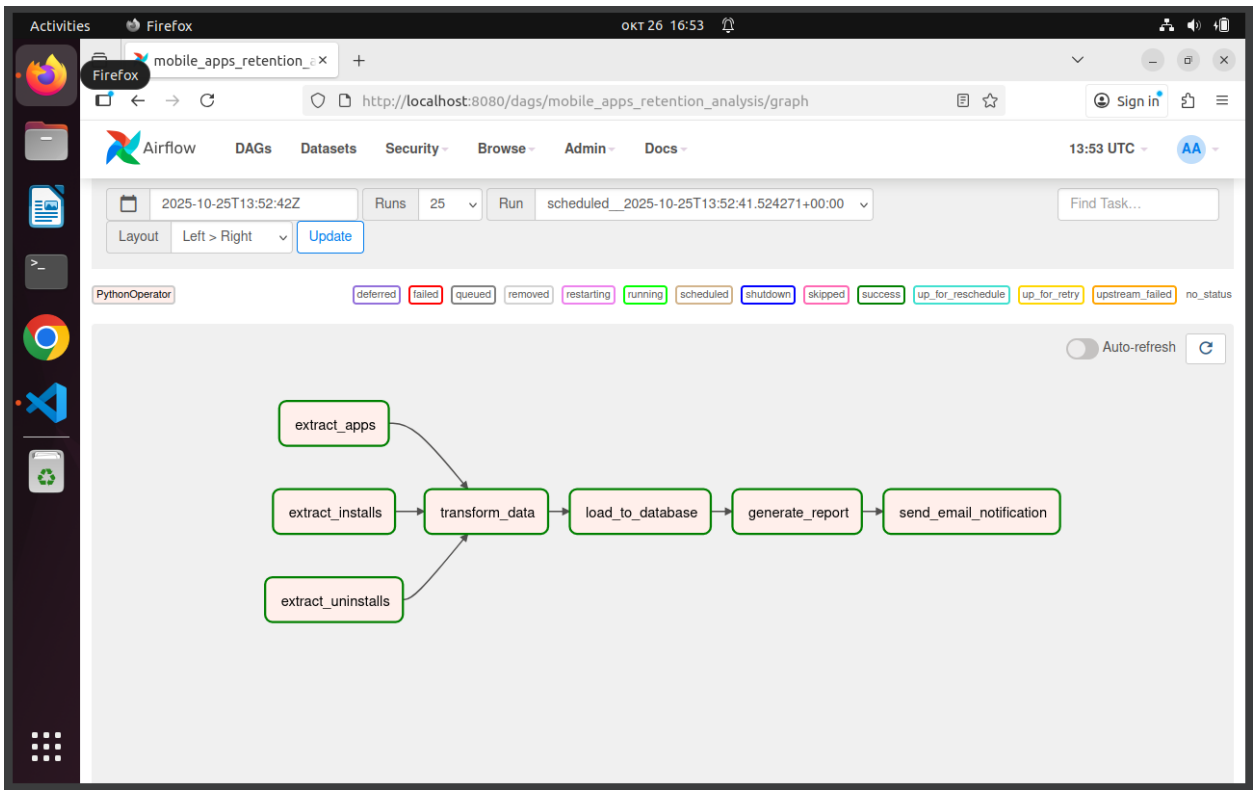
Остальные также посмотрим:

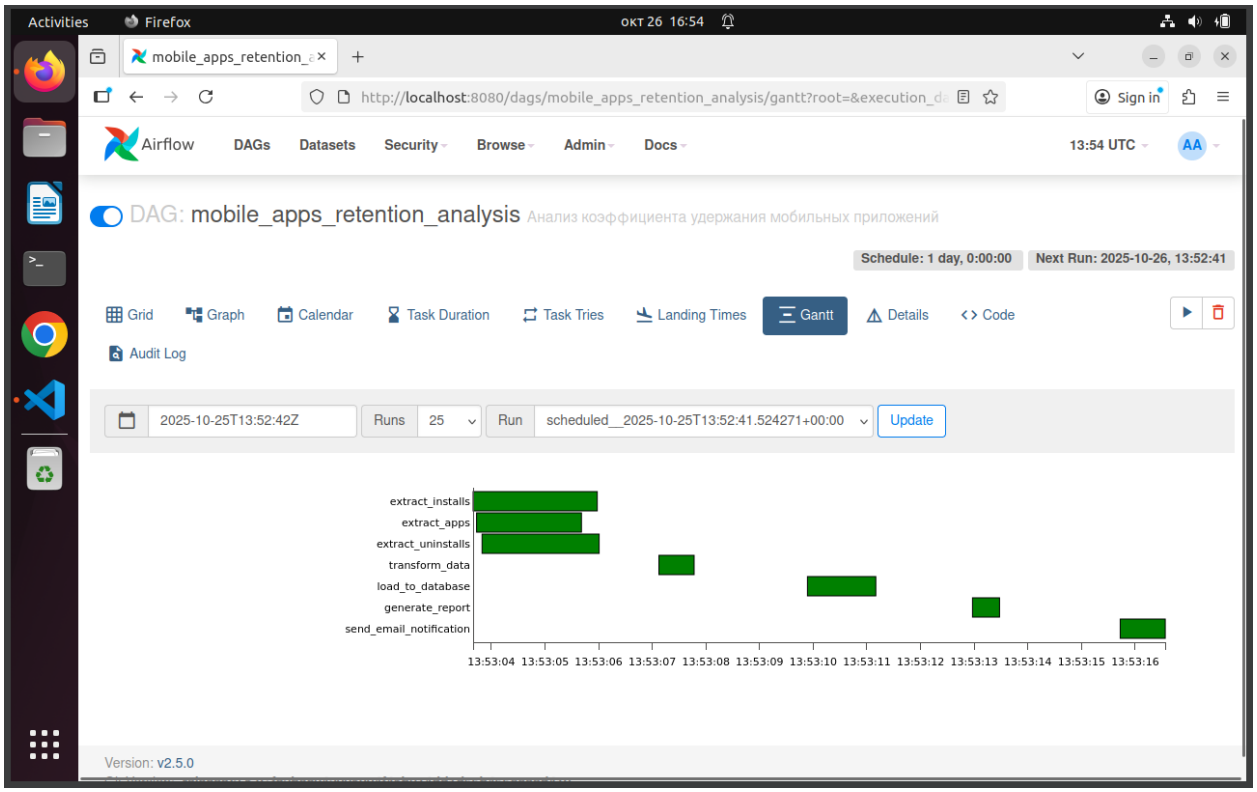


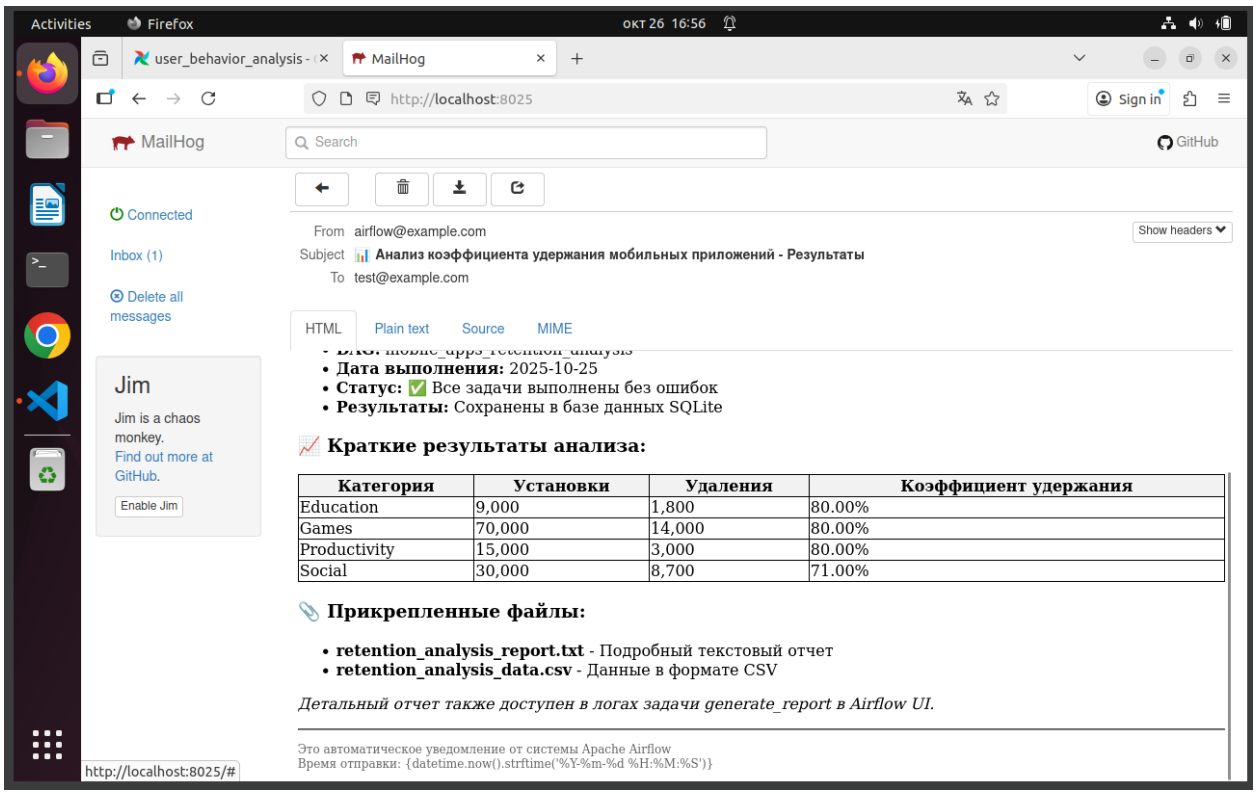



Activities
Firefox
окт 26 16:19
generate_data_dag - Gra
MailHog
telegram - Поиск в Goog
http://localhost:8080/dags/generate_data_dag/graph
Sign in
Airflow
DAGs
Datasets
Security
Browse
Admin
Docs
13:19 UTC
AA
DAG: generate_data_dag
Generate and aggregate sample data
success
Schedule: @once
Next Run: None
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
2023-10-01T00:00:01Z
Runs
25
Run
scheduled__2023-10-01T00:00:00+00:00
Layout
Left > Right
Find Task...
Update
PythonOperator
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
generate_data_1
generate_data_2
generate_data_3
aggregate_data


Activities
Firefox
окт 26 16:20
generate_data_dag - Gan
MailHog
telegram - Поиск в Goog
http://localhost:8080/dags/generate_data_dag/gantt?root=
Sign in
Airflow
DAGs
Datasets
Security
Browse
Admin
Docs
13:20 UTC
AA
DAG: generate_data_dag
Generate and aggregate sample data
Schedule: @once
Next Run: None
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
2023-10-01T00:00:01Z
Runs
25
Run
scheduled__2023-10-01T00:00:00+00:00
Update
generate_data_1
generate_data_2
generate_data_3
aggregate_data
13:17:10
13:17:11
13:17:12
13:17:13
13:17:14
13:17:15
13:17:16
13:17:17
13:17:18
13:17:19
13:17:20
Version: v2.5.0
Git Version: .release:2.5.0+fa2bec042995004f45b914dd1d66b466ccced410


Activities
Firefox
окт 26 16:53
mobile_apps_retention_
http://localhost:8080/dags/mobile_apps_retention_analysis/graph
Sign in
Airflow
DAGs
Datasets
Security
Browse
Admin
Docs
13:53 UTC
AA
2025-10-25T13:52:42Z
Runs
25
Run
scheduled__2025-10-25T13:52:41.524271+00:00
Find Task...
Layout
Left > Right
Update
PythonOperator
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
extract_apps
extract_installs
transform_data
load_to_database
generate_report
send_email_notification
extract_uninstalls


Activities
Firefox
окт 26 16:54
mobile_apps_retention_
http://localhost:8080/dags/mobile_apps_retention_analysis/gantt?root=&execution_d
Sign in
Airflow
DAGs
Datasets
Security
Browse
Admin
Docs
13:54 UTC
AA
DAG: mobile_apps_retention_analysis
Анализ коэффициента удержания мобильных приложений
Schedule: 1 day, 0:00:00
Next Run: 2025-10-26, 13:52:41
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
2025-10-25T13:52:42Z
Runs
25
Run
scheduled__2025-10-25T13:52:41.524271+00:00
Update
extract_installs
extract_apps
extract_uninstalls
transform_data
load_to_database
generate_report
send_email_notification
13:53:04
13:53:05
13:53:06
13:53:07
13:53:08
13:53:09
13:53:10
13:53:11
13:53:12
13:53:13
13:53:14
13:53:15
13:53:16
Version: v2.5.0



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.