

Отчёт о выполнении домашнего задания к семинару номер семь

Файл d7dag.py

```
import os
import logging
import pandas as pd
from airflow.operators.bash import BashOperator
from sqlalchemy import create_engine
from sqlalchemy.exc import SQLAlchemyError
from airflow import DAG
from airflow.operators.python import PythonOperator
from datetime import datetime, timedelta
import pendulum

# Настройки логирования
logging.basicConfig(level=logging.INFO)
logger = logging.getLogger(__name__)

default_args = {
    'owner': 'Kostia',
    'depends_on_past': False,
    'start_date': pendulum.datetime(2024, 6, 1, tz='Europe/Moscow'),
    'email': ['dom@dom.ru'],
    'email_on_failure': False,
    'email_on_retry': False,
    'retries': 0,
    'retry_delay': timedelta(minutes=5)
}

dag2 = DAG('Kostia002',
            default_args=default_args,
            description="seminar_7",
            catchup=False,
            schedule_interval='0 8 * * *')

def percent(**kwargs):
    files = ['/opt/airflow/dags/d4_1.xlsx', '/opt/airflow/dags/d4_2.xlsx',
            '/opt/airflow/dags/d4_3.xlsx']

    # Настройки подключения к MySQL
    MYSQL_USER = "airflow"
    MYSQL_PASSWORD = "airflow"
    MYSQL_HOST = "mysql-db"
    MYSQL_DB = "spark"

    # Подключение к MySQL через SQLAlchemy с таймаутом
    con =
    create_engine(f"mysql+pymysql://{MYSQL_USER}:{MYSQL_PASSWORD}@{MYSQL_HOST}:3306/{MYSQL_DB}",
                  connect_args={'connect_timeout': 10})

    for file in files:
        if not os.path.exists(file):
            logger.error(f"❌ Файл {file} не найден!")
            raise FileNotFoundError(f"Файл {file} не найден!")

        logger.info(f"📁 Обработка файла {file}")
        df = pd.read_excel(file, engine='openpyxl')

        required_columns = ['Платеж по основному долгу', 'Платеж по процентам']
        for col in required_columns:
            if col not in df.columns:
                logger.error(f"❌ В файле {file} отсутствует колонка {col}!")
                raise KeyError(f"Ошибка: в файле {file} отсутствует колонка {col}!")

        df['долг'] = df['Платеж по основному долгу'].cumsum().round(2)
        df['проценты'] = df['Платеж по процентам'].cumsum().round(2)
```

```

# Проверка существования таблицы перед заменой
try:
    table_exists = con.dialect.has_table(con.connect(), "credit",
schema="spark")
    if not table_exists or os.path.basename(file) == "d4_1.xlsx":
        df.to_sql('credit', con, schema='spark', if_exists='replace',
index=False, chunksize=500)
    else:
        df.to_sql('credit', con, schema='spark', if_exists='append',
index=False, chunksize=500)
except SQLAlchemyError as e:
    logger.error(f"Ошибка при записи в БД: {e}")
    raise


# Устанавливаем Python-библиотеки
task1 = BashOperator(
    task_id='pip_install',
    bash_command="""
pip install --no-cache-dir cryptography pandas pymysql sqlalchemy openpyxl
pendulum
""",
    dag=dag2
)

# Этот DAG автоматизирует обработку платежных данных из Excel-файлов
task2 = PythonOperator(
    task_id='python3',
    dag=dag2,
    python_callable=percent
)

task1 >> task2 # Устанавливаем порядок выполнения

```

Лог выполнения

 DAGs Cluster Activity Datasets Security Browse Admin Docs 19:29 UTC AU

Triggered Kostia002 with new Run ID manual__2025-02-19T19:24:24.979846+00:00, it should start any moment now.

DAG: Kostia002 seminar_7 Schedule: 0 8 * * * Next Run ID: 2025-02-19, 05:00:00 UTC

19.02.2025 19:24:25 All Run Types All Run States Clear Filters Auto-refresh 25

Press **SHIFT** + **/** for Shortcuts

Task: Kostia002 > 2025-02-19, 19:24:24 UTC python3

Clear task Mark state as... Filter DAG by task

Details Graph Gantt Code Event Log **Logs** Xcom Task Duration

All Levels All File Sources Wrap Download See More

66f2542a6351

Log message source details

[2025-02-19, 19:24:29 UTC] (local_task_job_runner.py:123) Pre task execution logs

[2025-02-19, 19:24:29 UTC] (taskinstance.py:2613) INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: Kostia002.python3 manual__2025-02-19T19:24:24.979846+00:00 [queued]>

[2025-02-19, 19:24:29 UTC] (taskinstance.py:2613) INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: Kostia002.python3 manual__2025-02-19T19:24:24.979846+00:00 [queued]>

[2025-02-19, 19:24:29 UTC] (taskinstance.py:2866) INFO - Starting attempt 1 of 1

[2025-02-19, 19:24:29 UTC] (taskinstance.py:2869) INFO - Executing <Task(PythonOperator): python3> on 2025-02-19 19:24:24.979846+00:00

[2025-02-19, 19:24:29 UTC] (standard_task_runner.py:184) INFO - Running: ['***', 'tasks', 'run', 'Kostia002', 'python3', 'manual__2025-02-19T19:24:24.979846+00:00', '--job-id', '29', '--raw', '--subdir', 'DAGS_F

[2025-02-19, 19:24:29 UTC] (standard_task_runner.py:185) INFO - Job 29: Subtask python3

[2025-02-19, 19:24:29 UTC] (logging_mixin.py:106) WARNING - /home/eee/locallib/python3.12/site-packages/xxx/task/task_runner/standard_task_runner.py:70 DeprecationWarning: This process (pid:1878) is multi-thre

[2025-02-19, 19:24:29 UTC] (standard_task_runner.py:72) INFO - Started process 1871 to run task

[2025-02-19, 19:24:29 UTC] (task_command.py:467) INFO - Running <TaskInstance: Kostia002.python3 manual__2025-02-19T19:24:24.979846+00:00 [running]> on host 66f2542a6351

[2025-02-19, 19:24:29 UTC] (taskinstance.py:3132) INFO - Exporting env vars: AIRFLOW_CTX_DAG_EMAIL="dom@dom.ru" AIRFLOW_CTX_DAG_OWNER="Kostia" AIRFLOW_CTX_DAG_ID="Kostia002" AIRFLOW_CTX_TASK_ID="python3" AIRFLOW

[2025-02-19, 19:24:29 UTC] (taskinstance.py:731) Log group end

[2025-02-19, 19:24:29 UTC] (dtdag.py:52) INFO - Ошибка файла /opt/xxx/dags/d4_1.xlsx

[2025-02-19, 19:24:29 UTC] (dtdag.py:52) INFO - Ошибка файла /opt/xxx/dags/d4_2.xlsx

[2025-02-19, 19:24:29 UTC] (dtdag.py:52) INFO - Ошибка файла /opt/xxx/dags/d4_3.xlsx

[2025-02-19, 19:24:29 UTC] (python.py:248) INFO - Done. Returned value was: None

[2025-02-19, 19:24:29 UTC] (taskinstance.py:348) Post task execution logs

[2025-02-19, 19:24:29 UTC] (taskinstance.py:352) INFO - Marking task as SUCCESS. dag_id=Kostia002, task_id=python3, run_id=manual__2025-02-19T19:24:24.979846+00:00, execution_date=20250219T192424, start_date=202

[2025-02-19, 19:24:29 UTC] (local_task_job_runner.py:266) INFO - Task exited with return code 0

[2025-02-19, 19:24:29 UTC] (taskinstance.py:3895) INFO - 0 downstream tasks scheduled from follow-on schedule check

[2025-02-19, 19:24:29 UTC] (local_task_job_runner.py:245) Log group end

Project

Airflow_MySQL_Spark ~/Programs/

airflow

logs

plugins

airflow.cfg

mysql-init

spark

env

docker-compose.yaml

docker-compose ---.xonix.yaml

README.md

External Libraries

Scratches and Consoles

WHERE

ORDER BY

	Период	И	Месяц	Сумма платежа	Платеж по основному долгу	Платеж
1		360	1 2023-11-01 00:00:00	86689.04	3655.71	
2		360	2 2023-12-01 00:00:00	86689.04	3688	
3		360	3 2024-01-01 00:00:00	86689.04	3720.58	
4		360	4 2024-02-01 00:00:00	86689.04	3753.44	
5		360	5 2024-03-01 00:00:00	86689.04	3786.6	
6		360	6 2024-04-01 00:00:00	86689.04	3820.04	
7		360	7 2024-05-01 00:00:00	86689.04	3853.79	
8		360	8 2024-06-01 00:00:00	86689.04	3887.83	
9		360	9 2024-07-01 00:00:00	86689.04	3922.17	
10		360	10 2024-08-01 00:00:00	86689.04	3956.82	
11		360	11 2024-09-01 00:00:00	86689.04	3991.77	
12		360	12 2024-10-01 00:00:00	86689.04	4027.03	
13		360	13 2024-11-01 00:00:00	86689.04	4062.6	
14		360	14 2024-12-01 00:00:00	86689.04	4098.49	
15		360	15 2025-01-01 00:00:00	86689.04	4134.69	
16		360	16 2025-02-01 00:00:00	86689.04	4171.22	
17		360	17 2025-03-01 00:00:00	86689.04	4208.06	
18		360	18 2025-04-01 00:00:00	86689.04	4245.23	
19		360	19 2025-05-01 00:00:00	86689.04	4282.73	
20		360	20 2025-06-01 00:00:00	86689.04	4320.56	
21		360	21 2025-07-01 00:00:00	86689.04	4358.73	
22		360	22 2025-08-01 00:00:00	86689.04	4397.23	
23		360	23 2025-09-01 00:00:00	86689.04	4436.07	
24		360	24 2025-10-01 00:00:00	86689.04	4475.26	
25		360	25 2025-11-01 00:00:00	86689.04	4514.79	
26		360	26 2025-12-01 00:00:00	86689.04	4554.67	

Database

@localhost 4

schemas 4

airflow

information_schema

performance_schema

spark

tables 2

credit

tasket4b

views

routes

events

virtual views

Server Objects

Database > @localhost > schemas > spark > tables > credit

SUM: Not enough values

Remote Python 3.12.8 Doc...e (airflow-webserver) (2)