**PROBLEM:**

**🡪DOWNLOAD TWO CSV FILES TO LOCAL AND STORE THEM IN ANY DATABASE.**

**🡪MERGE TWO TABLES IN DATABASE**

**🡪DELETE THE DOWNLOADED FILE FROM LOCAL**

**🡪MERGE TWO TABLES IN THE DATABASE BASED ON SOME CRITERIA**

**🡪DISPLAY THE MERGED TABLE BACK TO LOCAL**

**“PERFORM THE TASK USING AIRFLOW”**

**SOLUTION:**

* **STEP1: task.py**

\_\_\_\_import datetime

from airflow import DAG

from airflow.operators.bash\_operator import BashOperator

from airflow.operators.python\_operator import PythonOperator

import csv

import psycopg2

import airflow

import requests

import os

#def print\_world():

# print('world')

def LoadFile\_1():

file1\_url = "http://insight.dev.schoolwires.com/HelpAssets/C2Assets/C2Files/C2MassAssignUsersSample\_.csv"

input1 = requests.get(file1\_url, stream = True)

output1 = open("/home/baadmin/airflow\_ishi/dags/input/file1.csv",'a')

for line in input1:

content = ''.join(chr(x) for x in line)

c=content.rstrip()

print(c, file=output1)

def LoadFile\_2():

file2\_url = "http://insight.dev.schoolwires.com/HelpAssets/C2Assets/C2Files/C2ImportFamRelSample.csv"

input2 = requests.get(file2\_url, stream = True)

output2 = open("/home/baadmin/airflow\_ishi/dags/input/file2.csv",'a')

for line in input2:

content = ''.join(chr(x) for x in line)

c=content.rstrip()

print(c, file=output2)

def ConnectDb\_1():

conn = psycopg2.connect("host='localhost' port='5432' dbname='sangeedb' user='postgres' password='Password!234'")

cur = conn.cursor()

f=open("/home/baadmin/airflow\_ishi/dags/input/file1.csv", 'r')

#copy\_stmt = """COPY {} FROM STDIN WITH CSV {} QUOTE AS '"' """.format('csvtable1',('"UserCode"','"GroupCode"'))

cur.copy\_from(f, 'csvtable1', sep=',',columns=('"UserCode"','"GroupCode"'))

cur.execute(""" Delete from csvtable1 where "UserCode" = 'UserCode' """)

#cur.copy\_expert("COPY csvtable1 TO STDOUT WITH CSV HEADER", sys.stdout)

cur = conn.cursor()

conn.commit()

conn.close()

def ConnectDb\_2():

conn = psycopg2.connect("host='localhost' port='5432' dbname='sangeedb' user='postgres' password='Password!234'")

cur = conn.cursor()

f=open("/home/baadmin/airflow\_ishi/dags/input/file2.csv", 'r')

cur.copy\_from(f, 'csvtable2', sep=',',columns=('"Parent Identifier"','"Student Identifier"'))

cur.execute(""" Delete from csvtable2 where "Parent Identifier" = 'Parent Identifier' """)

#cur.copy\_expert("COPY csvtable2 TO STDOUT WITH CSV HEADER", sys.stdout)

cur = conn.cursor()

conn.commit()

conn.close()

def RemoveCsvFile():

os.remove("/home/baadmin/airflow\_ishi/dags/input/file1.csv")

os.remove("/home/baadmin/airflow\_ishi/dags/input/file2.csv")

def MergeTable():

conn = psycopg2.connect("host='localhost' port='5432' dbname='sangeedb' user='postgres' password='Password!234'")

cur = conn.cursor()

f=open("/home/baadmin/december6/input/merge.csv", 'w')

cur.execute("""copy mergetable to '/home/baadmin/airflow\_ishi/dags/input/merge.csv' delimiter ',' csv header""")

#cur.copy\_to(f, mergetable, sep='\t')

cur = conn.cursor()

conn.commit()

conn.close()

default\_args = {

'owner': 'ishi\_aa',

'start\_date': datetime.datetime(2018, 12, 9),

'retries': 1,

'retry\_delay': datetime.timedelta(minutes=1),

}

with DAG('1\_FileToDb',

default\_args=default\_args,

) as dag:

csvfile1 = PythonOperator(task\_id='csvfile1',

python\_callable= LoadFile\_1)

csvfile2 = PythonOperator(task\_id='csvfile2',

python\_callable= LoadFile\_2)

db1 = PythonOperator(task\_id='connectdb1',

python\_callable= ConnectDb\_1)

db2 = PythonOperator(task\_id='connectdb2',

python\_callable= ConnectDb\_2)

remove= PythonOperator(task\_id='removecsv',

python\_callable= RemoveCsvFile)

merge = PythonOperator(task\_id='merge',

python\_callable= MergeTable)

csvfile1 >> csvfile2 >> db1 >> db2 >> remove >> merge

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* **STEP2: postgres**

su sudo postgres

psql

\c sangee db

create table csvtable1(“UserCode” varchar(20),”GroupCode” varcahr2(20));

sangeedb=# select \* from csvtable1;

**UserCode | GroupCode**

**----------+-----------**

**(0 rows)**

sangeedb=# select \* from csvtable2;

**Parent Identifier | Student Identifier**

**-------------------+--------------------**

**(0 rows)**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**To run the code : python task.py**

sangeedb=# select \* from csvtable1;

**UserCode | GroupCode**

**----------+-----------**

**1051 | 555**

**1150 | 555**

**1152 | 555**

**1154 | 555**

**(4 rows)**

sangeedb=# select \* from csvtable2;

**Parent Identifier | Student Identifier**

**-------------------+--------------------**

**1001 | 1002**

**1010 | 1020**

**(2 rows)**

sangeedb=# create table mergetable as(select \* from csvtable1,csvtable2);

sangeedb=# select \* from mergetable;

**UserCode | GroupCode | Parent Identifier | Student Identifier**

**----------+-----------+-------------------+--------------------**

**1051 | 555 | 1001 | 1002**

**1051 | 555 | 1010 | 1020**

**1150 | 555 | 1001 | 1002**

**1150 | 555 | 1010 | 1020**

**1152 | 555 | 1001 | 1002**

**1152 | 555 | 1010 | 1020**

**1154 | 555 | 1001 | 1002**

**1154 | 555 | 1010 | 1020**

**(8 rows)**

STEPS

1. LOAD first csv file from http to local (**file1.csv)**
2. Load second csv file from http to local (**file2.csv)**
3. Connect to postgres db. Send the first csv file to db.(it will be saved as table) (**csvtable1)**
4. Connect to postgres db. Send the second csv file to db.(it will be saved as table) (**csvtable2)**
5. Remove the files from local (**file1.csv,file2.csv)**
6. Merge the table based on some criteria (**mergetable)**
7. Send the table to the local (**merge.csv)**