import os

from datetime import datetime

from airflow.decorators import dag, task

default\_args = {

  'owner': 'DWH',

  'depends\_on\_past': False,

  'retries': 0,

  'pg\_schema': 'project',

  'pg\_connection\_id': 'dwh\_pg'

}

INPUT\_FOLDER = os.path.join(os.getenv('DATA\_FOLDER'), 'input')

SQL\_FOLDER = os.path.join(os.getenv('DATA\_FOLDER'), 'sql')

NEO4J\_FOLDER = os.path.join(os.getenv('DATA\_FOLDER'), 'neo4j')

# define custom function to load data into Neo4j

def load\_into\_neo4j(uri, user, password, file\_path):

  driver = GraphDatabase.driver(uri, auth=(user, password))

  with driver.session() as session:

    with session.begin\_transaction() as tx:

      query = open(file\_path, 'r').read()

      tx.run(query)

  driver.close()

@dag(

  default\_args=default\_args,

  description='ETL pipeline for Kaggle data',

  schedule\_interval='\* \* \* \* \*',  # execute every minute

  start\_date=datetime(2023, 10, 1, 0, 0, 0),

  catchup=False,

  concurrency=1,

  max\_active\_runs=1,

  template\_searchpath=[os.getenv('DATA\_FOLDER'), SQL\_FOLDER],

)

def etl\_submissions():

  from airflow.operators.python import ShortCircuitOperator

  from airflow.providers.postgres.operators.postgres import PostgresOperator

  from airflow.utils.state import State

  from etl import process\_authors, process\_submissions, process\_citations

  from transform import generate\_authors\_csv #---------

  @task(

    trigger\_rule='all\_done'

  )

  def complete\_chunk(\*\*kwargs):

    tasks = kwargs["ti"].get\_dagrun().get\_task\_instances()

    failed\_tasks = [ti for ti in tasks if ti.state == State.FAILED]

    chunk\_id = kwargs["ti"].xcom\_pull(task\_ids='extract\_chunk', key='return\_value')[0][0]

    chunk\_status = 'failed' if len(failed\_tasks) > 0 else 'processed'

    print(f"Marking chunk {chunk\_id} as {chunk\_status}")

    pg\_operator = PostgresOperator(

      task\_id='complete\_chunk',

      postgres\_conn\_id=default\_args["pg\_connection\_id"],

      sql="update project.chunk\_queue set status = %(chunk\_status)s where id = %(chunk\_id)s",

      parameters={

        'chunk\_id': chunk\_id,

        'chunk\_status': chunk\_status

      }

    )

    return pg\_operator.execute({})

  @task()

  def load\_into\_neo4j\_task():

    uri = os.getenv('NEO4J\_URI', 'if\_missing\_uri')

    user = os.getenv('NEO4J\_USER', 'if\_missing\_user')

    password = os.getenv('NEO4J\_PASSWORD', 'if\_missing\_password')

    file\_path = os.path.join(NEO4J\_FOLDER, 'neo4j\_query.cql')

    load\_into\_neo4j(uri, user, password, file\_path)

  def chunk\_in\_xcom(\*\*kwargs):

    chunk = kwargs["ti"].xcom\_pull(task\_ids='extract\_chunk', key='return\_value')

    return len(chunk) > 0

  extract\_chunk\_task = PostgresOperator(

    task\_id='extract\_chunk',

    postgres\_conn\_id=default\_args["pg\_connection\_id"],

    sql="select \* from project.chunk\_queue where status = 'pending' or status = 'failed' order by id limit 1"

  )

  check\_for\_chunks\_to\_process = ShortCircuitOperator(

    task\_id="check\_for\_chunks\_to\_process",

    python\_callable=chunk\_in\_xcom,

  )

  process\_authors\_task = process\_authors()

  process\_submissions\_task = process\_submissions()

  process\_citations\_task = process\_citations()

  neo4j\_task = load\_into\_neo4j\_task()   #---------

  complete\_chunk\_task = complete\_chunk()

  extract\_chunk\_task >> check\_for\_chunks\_to\_process >> process\_authors\_task >> process\_submissions\_task >> process\_citations\_task >> complete\_chunk\_task

etl\_submissions()