# import the libraries

from datetime import timedelta

# The DAG object; we'll need this to instantiate a DAG

from airflow import DAG

# Operators; we need this to write tasks!

from airflow.operators.bash\_operator import BashOperator

# This makes scheduling easy

from airflow.utils.dates import days\_ago

#defining DAG arguments

# You can override them on a per-task basis during operator initialization

default\_args = {

    'owner': 'Ram Param',

    'start\_date': days\_ago(0),

    'email': ['ramesh@somemail.com'],

    'email\_on\_failure': False,

    'email\_on\_retry': False,

    'retries': 0,

    'retry\_delay': timedelta(seconds=2),

}

# defining the DAG

# define the DAG

dag = DAG(

    'etl\_airflow',

    default\_args=default\_args,

    description='ETL pipeline text change',

    schedule\_interval=timedelta(seconds=1),

)

# define the tasks

# define the first task         download task

'''

download = BashOperator(

    task\_id='download',

    bash\_command='wget "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Apache%20Airflow/Build%20a%20DAG%20using%20Airflow/web-server-access-log.txt"',

    dag=dag,

)

'''

# define the second task        extract task

extract = BashOperator(

    task\_id='extract',

    bash\_command='cut -d "#" -f 1,4 /home/project/airflow/dags/web-server-access-log.txt > /home/project/airflow/dags/extracted-data\_etl.txt',

    dag=dag,

)

# define the third task         transform task

transform = BashOperator(

    task\_id='transform',

    bash\_command='tr [:lower:] [:upper:] < /home/project/airflow/dags/extracted-data\_etl.txt > /home/project/airflow/dags/transformed-data\_etl.csv',

    dag=dag,

)

# define the fourth task        load task

load = BashOperator(

    task\_id='load',

    bash\_command='zip /home/project/airflow/dags/log.zip /home/project/airflow/dags/transformed-data\_etl.csv',

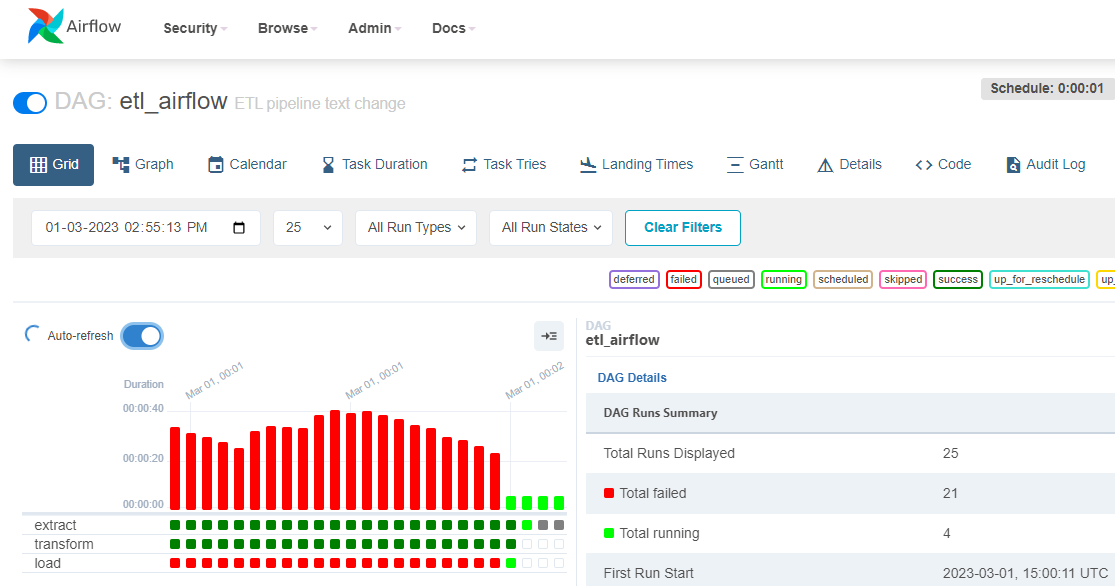
    dag=dag,

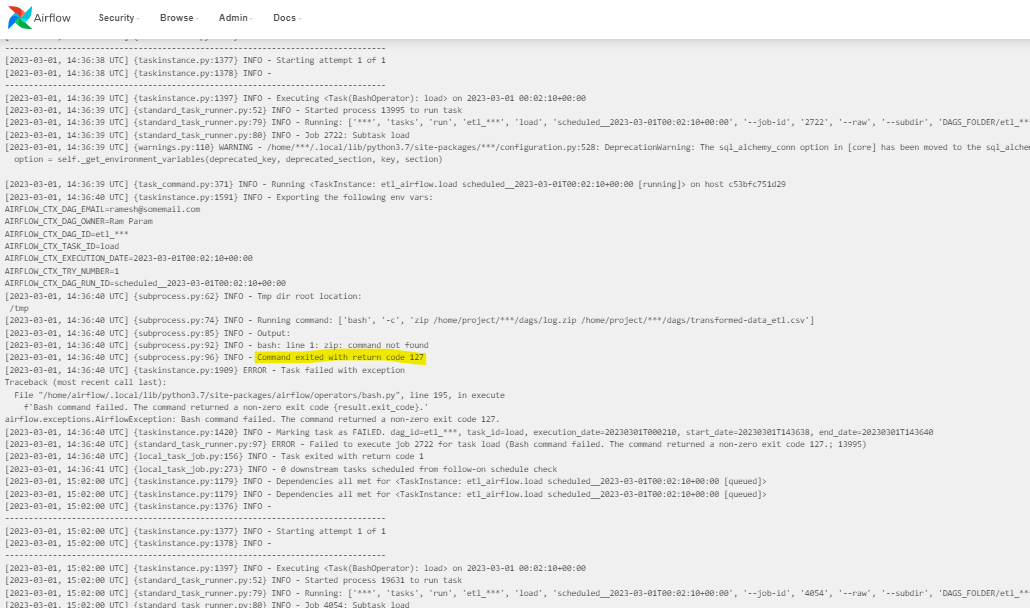
)

# task pipeline

extract >> transform >> load

this time the 3rs task is failing again and again :-





Below is the log in text for last task :-

\*\*\* Reading local file: /home/project/airflow/logs/dag\_id=etl\_airflow/run\_id=scheduled\_\_2023-03-01T00:02:10+00:00/task\_id=load/attempt=1.log

[2023-03-01, 14:36:38 UTC] {taskinstance.py:1179} INFO - Dependencies all met for <TaskInstance: etl\_airflow.load scheduled\_\_2023-03-01T00:02:10+00:00 [queued]>

[2023-03-01, 14:36:38 UTC] {taskinstance.py:1179} INFO - Dependencies all met for <TaskInstance: etl\_airflow.load scheduled\_\_2023-03-01T00:02:10+00:00 [queued]>

[2023-03-01, 14:36:38 UTC] {taskinstance.py:1376} INFO -

--------------------------------------------------------------------------------

[2023-03-01, 14:36:38 UTC] {taskinstance.py:1377} INFO - Starting attempt 1 of 1

[2023-03-01, 14:36:38 UTC] {taskinstance.py:1378} INFO -

--------------------------------------------------------------------------------

[2023-03-01, 14:36:39 UTC] {taskinstance.py:1397} INFO - Executing <Task(BashOperator): load> on 2023-03-01 00:02:10+00:00

[2023-03-01, 14:36:39 UTC] {standard\_task\_runner.py:52} INFO - Started process 13995 to run task

[2023-03-01, 14:36:39 UTC] {standard\_task\_runner.py:79} INFO - Running: ['\*\*\*', 'tasks', 'run', 'etl\_\*\*\*', 'load', 'scheduled\_\_2023-03-01T00:02:10+00:00', '--job-id', '2722', '--raw', '--subdir', 'DAGS\_FOLDER/etl\_\*\*\*.py', '--cfg-path', '/tmp/tmpy12bzk5o', '--error-file', '/tmp/tmpvhkxgpmm']

[2023-03-01, 14:36:39 UTC] {standard\_task\_runner.py:80} INFO - Job 2722: Subtask load

[2023-03-01, 14:36:39 UTC] {warnings.py:110} WARNING - /home/\*\*\*/.local/lib/python3.7/site-packages/\*\*\*/configuration.py:528: DeprecationWarning: The sql\_alchemy\_conn option in [core] has been moved to the sql\_alchemy\_conn option in [database] - the old setting has been used, but please update your config.

option = self.\_get\_environment\_variables(deprecated\_key, deprecated\_section, key, section)

[2023-03-01, 14:36:39 UTC] {task\_command.py:371} INFO - Running <TaskInstance: etl\_airflow.load scheduled\_\_2023-03-01T00:02:10+00:00 [running]> on host c53bfc751d29

[2023-03-01, 14:36:40 UTC] {taskinstance.py:1591} INFO - Exporting the following env vars:

AIRFLOW\_CTX\_DAG\_EMAIL=ramesh@somemail.com

AIRFLOW\_CTX\_DAG\_OWNER=Ram Param

AIRFLOW\_CTX\_DAG\_ID=etl\_\*\*\*

AIRFLOW\_CTX\_TASK\_ID=load

AIRFLOW\_CTX\_EXECUTION\_DATE=2023-03-01T00:02:10+00:00

AIRFLOW\_CTX\_TRY\_NUMBER=1

AIRFLOW\_CTX\_DAG\_RUN\_ID=scheduled\_\_2023-03-01T00:02:10+00:00

[2023-03-01, 14:36:40 UTC] {subprocess.py:62} INFO - Tmp dir root location:

/tmp

[2023-03-01, 14:36:40 UTC] {subprocess.py:74} INFO - Running command: ['bash', '-c', 'zip /home/project/\*\*\*/dags/log.zip /home/project/\*\*\*/dags/transformed-data\_etl.csv']

[2023-03-01, 14:36:40 UTC] {subprocess.py:85} INFO - Output:

[2023-03-01, 14:36:40 UTC] {subprocess.py:92} INFO - bash: line 1: zip: command not found

[2023-03-01, 14:36:40 UTC] {subprocess.py:96} INFO - Command exited with return code 127

[2023-03-01, 14:36:40 UTC] {taskinstance.py:1909} ERROR - Task failed with exception

Traceback (most recent call last):

File "/home/airflow/.local/lib/python3.7/site-packages/airflow/operators/bash.py", line 195, in execute

f'Bash command failed. The command returned a non-zero exit code {result.exit\_code}.'

airflow.exceptions.AirflowException: Bash command failed. The command returned a non-zero exit code 127.

[2023-03-01, 14:36:40 UTC] {taskinstance.py:1420} INFO - Marking task as FAILED. dag\_id=etl\_\*\*\*, task\_id=load, execution\_date=20230301T000210, start\_date=20230301T143638, end\_date=20230301T143640

[2023-03-01, 14:36:40 UTC] {standard\_task\_runner.py:97} ERROR - Failed to execute job 2722 for task load (Bash command failed. The command returned a non-zero exit code 127.; 13995)

[2023-03-01, 14:36:40 UTC] {local\_task\_job.py:156} INFO - Task exited with return code 1

[2023-03-01, 14:36:41 UTC] {local\_task\_job.py:273} INFO - 0 downstream tasks scheduled from follow-on schedule check

[2023-03-01, 15:02:00 UTC] {taskinstance.py:1179} INFO - Dependencies all met for <TaskInstance: etl\_airflow.load scheduled\_\_2023-03-01T00:02:10+00:00 [queued]>

[2023-03-01, 15:02:00 UTC] {taskinstance.py:1179} INFO - Dependencies all met for <TaskInstance: etl\_airflow.load scheduled\_\_2023-03-01T00:02:10+00:00 [queued]>

[2023-03-01, 15:02:00 UTC] {taskinstance.py:1376} INFO -

--------------------------------------------------------------------------------

[2023-03-01, 15:02:00 UTC] {taskinstance.py:1377} INFO - Starting attempt 1 of 1

[2023-03-01, 15:02:00 UTC] {taskinstance.py:1378} INFO -

--------------------------------------------------------------------------------

[2023-03-01, 15:02:00 UTC] {taskinstance.py:1397} INFO - Executing <Task(BashOperator): load> on 2023-03-01 00:02:10+00:00

[2023-03-01, 15:02:00 UTC] {standard\_task\_runner.py:52} INFO - Started process 19631 to run task

[2023-03-01, 15:02:00 UTC] {standard\_task\_runner.py:79} INFO - Running: ['\*\*\*', 'tasks', 'run', 'etl\_\*\*\*', 'load', 'scheduled\_\_2023-03-01T00:02:10+00:00', '--job-id', '4054', '--raw', '--subdir', 'DAGS\_FOLDER/etl\_\*\*\*.py', '--cfg-path', '/tmp/tmp1zlyn6gz', '--error-file', '/tmp/tmpb1uzqh7w']

[2023-03-01, 15:02:00 UTC] {standard\_task\_runner.py:80} INFO - Job 4054: Subtask load

[2023-03-01, 15:02:00 UTC] {warnings.py:110} WARNING - /home/\*\*\*/.local/lib/python3.7/site-packages/\*\*\*/configuration.py:528: DeprecationWarning: The sql\_alchemy\_conn option in [core] has been moved to the sql\_alchemy\_conn option in [database] - the old setting has been used, but please update your config.

option = self.\_get\_environment\_variables(deprecated\_key, deprecated\_section, key, section)

[2023-03-01, 15:02:00 UTC] {task\_command.py:371} INFO - Running <TaskInstance: etl\_airflow.load scheduled\_\_2023-03-01T00:02:10+00:00 [running]> on host c53bfc751d29

[2023-03-01, 15:02:01 UTC] {taskinstance.py:1591} INFO - Exporting the following env vars:

AIRFLOW\_CTX\_DAG\_EMAIL=ramesh@somemail.com

AIRFLOW\_CTX\_DAG\_OWNER=Ram Param

AIRFLOW\_CTX\_DAG\_ID=etl\_\*\*\*

AIRFLOW\_CTX\_TASK\_ID=load

AIRFLOW\_CTX\_EXECUTION\_DATE=2023-03-01T00:02:10+00:00

AIRFLOW\_CTX\_TRY\_NUMBER=1

AIRFLOW\_CTX\_DAG\_RUN\_ID=scheduled\_\_2023-03-01T00:02:10+00:00

[2023-03-01, 15:02:01 UTC] {subprocess.py:62} INFO - Tmp dir root location:

/tmp

[2023-03-01, 15:02:01 UTC] {subprocess.py:74} INFO - Running command: ['bash', '-c', 'zip /home/project/\*\*\*/dags/log.zip /home/project/\*\*\*/dags/transformed-data\_etl.csv']

[2023-03-01, 15:02:01 UTC] {subprocess.py:85} INFO - Output:

[2023-03-01, 15:02:01 UTC] {subprocess.py:92} INFO - bash: line 1: zip: command not found

[2023-03-01, 15:02:01 UTC] {subprocess.py:96} INFO - Command exited with return code 127

[2023-03-01, 15:02:01 UTC] {taskinstance.py:1909} ERROR - Task failed with exception

Traceback (most recent call last):

File "/home/airflow/.local/lib/python3.7/site-packages/airflow/operators/bash.py", line 195, in execute

f'Bash command failed. The command returned a non-zero exit code {result.exit\_code}.'

airflow.exceptions.AirflowException: Bash command failed. The command returned a non-zero exit code 127.

[2023-03-01, 15:02:01 UTC] {taskinstance.py:1420} INFO - Marking task as FAILED. dag\_id=etl\_\*\*\*, task\_id=load, execution\_date=20230301T000210, start\_date=20230301T150200, end\_date=20230301T150201

[2023-03-01, 15:02:01 UTC] {standard\_task\_runner.py:97} ERROR - Failed to execute job 4054 for task load (Bash command failed. The command returned a non-zero exit code 127.; 19631)

[2023-03-01, 15:02:01 UTC] {local\_task\_job.py:156} INFO - Task exited with return code 1

[2023-03-01, 15:02:01 UTC] {local\_task\_job.py:273} INFO - 0 downstream tasks scheduled from follow-on schedule check