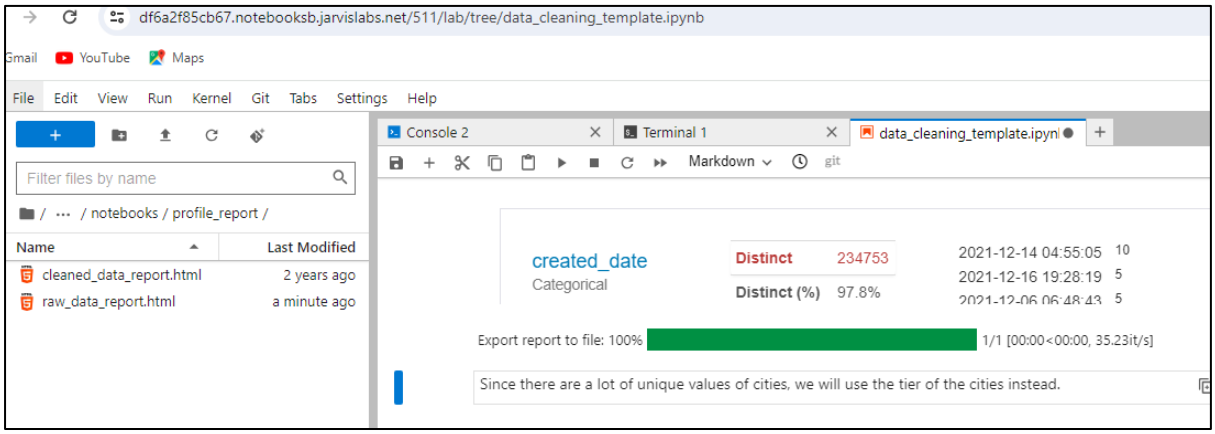
Data Loading

Unzip the Assignment file in Console:

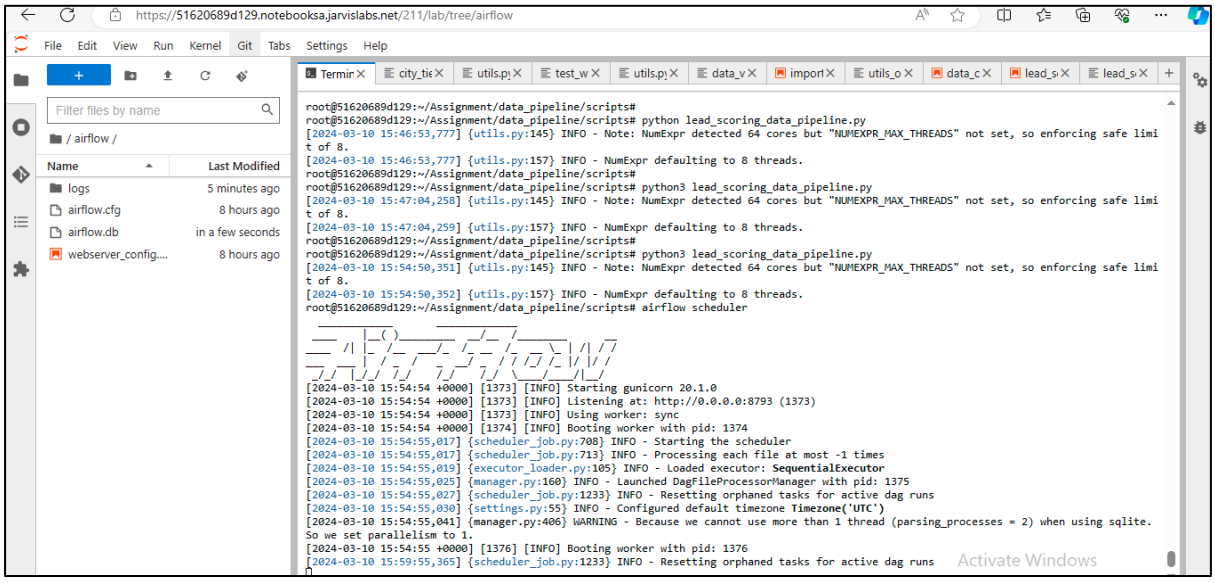
```
root@df6a2f85cb67:/# pwd
/
root@df6a2f85cb67:/# cd /home
root@df6a2f85cb67:~# unzip Assignment.zip
Archive:  Assignment.zip
   creating: Assignment/01_data_pipeline/
  inflating: Assignment/01_data_pipeline/INSTRUCTIONS.txt
   creating: Assignment/01_data_pipeline/notebooks/
   creating: Assignment/01_data_pipeline/notebooks/Data/
  inflating: Assignment/01_data_pipeline/notebooks/Data/cleaned_data.csv
```

Creation of profile report:



Data Pipeline

Starting Airflow in command line:



AirFlow process:

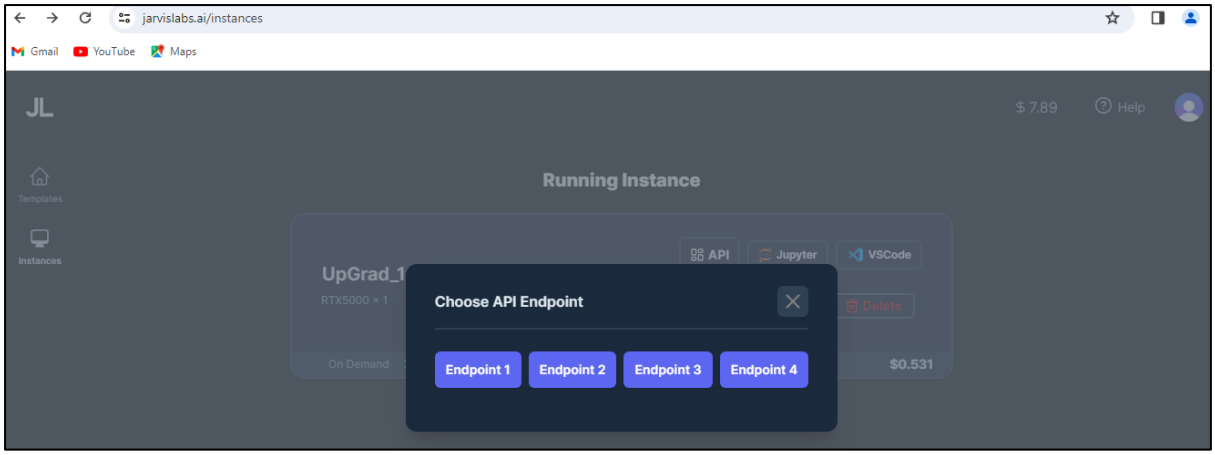
```
root@3b6b5b51627a:~/Assignment/01_data_pipeline/scripts# airflow db init
DB: sqlite:////home/airflow/airflow.db
[2024-03-17 04:26:30,094] {db.py:1462} INFO - Creating tables
INFO  [alembic.runtime.migration] Context impl SQLiteImpl.
INFO  [alembic.runtime.migration] Will assume non-transactional DDL.
WARNI [airflow.models.crypto] empty cryptography key - values will not be stored encrypted.
Initialization done
root@3b6b5b51627a:~/Assignment/01_data_pipeline/scripts#
```

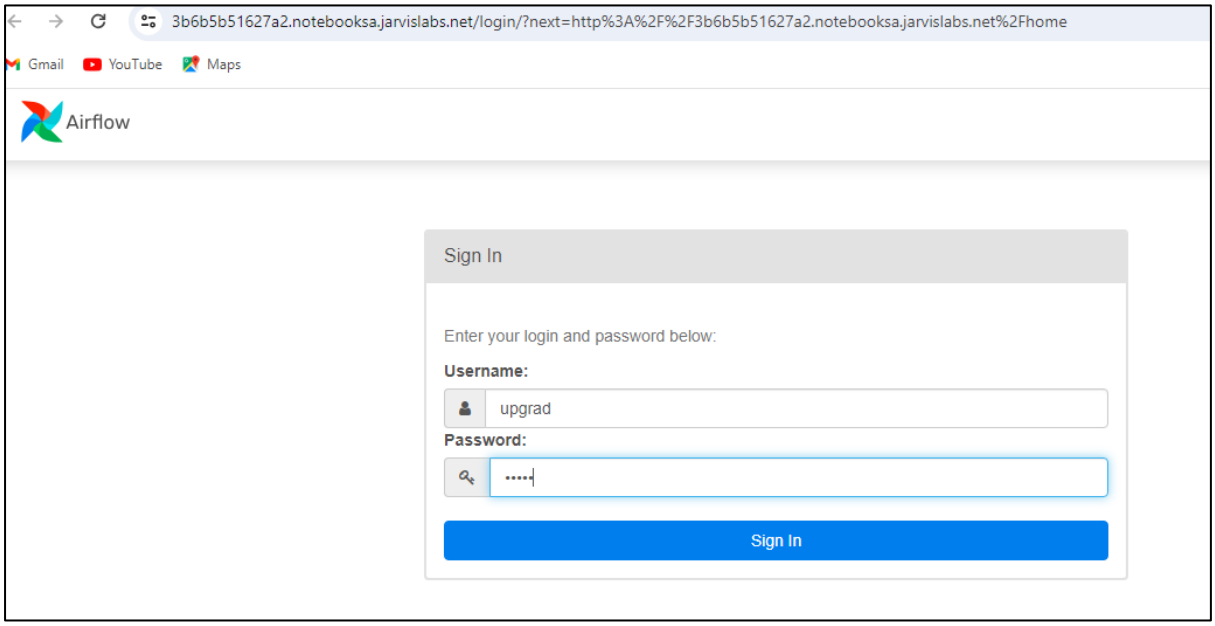
```
root@3b6b5b51627a:~/Assignment/01_data_pipeline/scripts# airflow users create --username upgrad --firstname upgrad --lastname
upgrad --role Admin --email shindesagarm@yahoo.co.in --password admin
[2024-03-17 04:33:55,567] {manager.py:817} WARNING - No user yet created, use flask fab command to do it.
[2024-03-17 04:33:55,690] {manager.py:508} INFO - Created Permission View: menu access on List Users
[2024-03-17 04:33:55,705] {manager.py:568} INFO - Added Permission menu access on List Users to role Admin
[2024-03-17 04:33:55,730] {manager.py:508} INFO - Created Permission View: menu access on Security
```

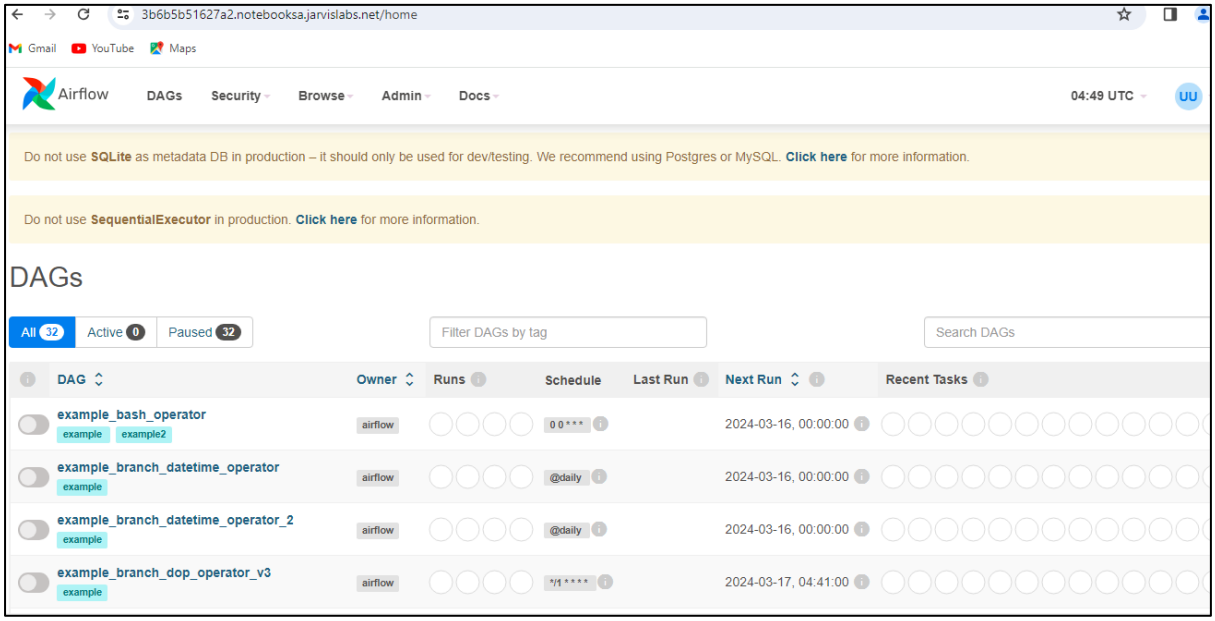
```
[2024-03-17 04:33:57,091] {manager.py:568} INFO - Added Permission can create on XComs to role Admin
[2024-03-17 04:33:58,331] {manager.py:213} INFO - Added user upgrad
User "upgrad" created with role "Admin"
root@3b6b5b51627a:~/Assignment/01_data_pipeline/scripts#
```

```
root@3b6b5b51627a:~/Assignment/01_data_pipeline/scripts# airflow webserver
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /    _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
Running the Gunicorn Server with:
Workers: 4 sync
Host: 0.0.0.0:6007
Timeout: 120
Logfiles: - -
Access Logformat:
=================================================================
[2024-03-17 04:36:48 +0000] [822] [INFO] Starting gunicorn 20.1.0
[2024-03-17 04:36:48 +0000] [822] [INFO] Listening at: http://0.0.0.0:6007 (822)
[2024-03-17 04:36:48 +0000] [822] [INFO] Using worker: sync
[2024-03-17 04:36:48 +0000] [824] [INFO] Booting worker with pid: 824
[2024-03-17 04:36:48 +0000] [825] [INFO] Booting worker with pid: 825
[2024-03-17 04:36:48 +0000] [826] [INFO] Booting worker with pid: 826
[2024-03-17 04:36:48 +0000] [827] [INFO] Booting worker with pid: 827
```

```
root@3b6b5b51627a:~# airflow scheduler
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /    _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
[2024-03-17 04:38:55 +0000] [860] [INFO] Starting gunicorn 20.1.0
[2024-03-17 04:38:55 +0000] [860] [INFO] Listening at: http://0.0.0.0:8793 (860)
[2024-03-17 04:38:55 +0000] [860] [INFO] Using worker: sync
[2024-03-17 04:38:55,335] {scheduler_job.py:708} INFO - Starting the scheduler
[2024-03-17 04:38:55,335] {scheduler_job.py:713} INFO - Processing each file at most -1 times
[2024-03-17 04:38:55 +0000] [861] [INFO] Booting worker with pid: 861
[2024-03-17 04:38:55,337] {executor_loader.py:105} INFO - Loaded executor: SequentialExecutor
[2024-03-17 04:38:55,343] {manager.py:160} INFO - Launched DagFileProcessorManager with pid: 862
[2024-03-17 04:38:55,344] {scheduler_job.py:1233} INFO - Resetting orphaned tasks for active dag runs
[2024-03-17 04:38:55,347] {settings.py:55} INFO - Configured default timezone Timezone('UTC')
[2024-03-17 04:38:55,359] {manager.py:406} WARNING - Because we cannot use more than 1 thread (parsing_processes = 2) when us
ing sqlite. So we set parallelism to 1.
[2024-03-17 04:38:55 +0000] [864] [INFO] Booting worker with pid: 864
```


jarvislabs.ai/instances
Gmail
YouTube
Maps
JL
$ 7.89
Help
Templates
Instances
Running Instance
API
Jupyter
VSCode
UpGrad_1
RTX5000 × 1
Delete
On Demand
$0.531
Choose API Endpoint
Endpoint 1
Endpoint 2
Endpoint 3
Endpoint 4


3b6b5b51627a2.notebooksa.jarvislabs.net/login/?next=http%3A%2F%2F3b6b5b51627a2.notebooksa.jarvislabs.net%2Fhome
Gmail
YouTube
Maps
Airflow
Sign In
Enter your login and password below:
Username:
upgrad
Password:
Sign In


3b6b5b51627a2.notebooksa.jarvislabs.net/home
Gmail
YouTube
Maps
Airflow
DAGs
Security
Browse
Admin
Docs
04:49 UTC
UU
Do not use SQLite as metadata DB in production – it should only be used for dev/testing. We recommend using Postgres or MySQL. Click here for more information.
Do not use SequentialExecutor in production. Click here for more information.
DAGs
All 32
Active 0
Paused 32
Filter DAGs by tag
Search DAGs
DAG
Owner
Runs
Schedule
Last Run
Next Run
Recent Tasks
example_bash_operator
example
example2
airflow
0 0 * * *
2024-03-16, 00:00:00
example_branch_datetime_operator
example
airflow
@daily
2024-03-16, 00:00:00
example_branch_datetime_operator_2
example
airflow
@daily
2024-03-16, 00:00:00
example_branch_dop_operator_v3
example
airflow
*/1 * * * *
2024-03-17, 04:41:00

Training Pipeline:

```
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline#
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline# cp /home/Assignment/training_pipeline/scripts/lead_scoring_t
raining_pipeline.py .
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline# cp /home/Assignment/training_pipeline/scripts/constants.py .
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline#
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline# cp /home/Assignment/training_pipeline/scripts/utils.py .
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline#
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline# ls -lrt
total 16
drwxr-xr-x 2 root root   34 Mar 17 06:47 __pycache__
drwxr-xr-x 3 root root   69 Mar 17 08:02 data
-rw-r--r-- 1 root root 3010 Mar 17 08:06 lead_scoring_training_pipeline.py
-rw-r--r-- 1 root root 2258 Mar 17 08:07 constants.py
-rw-r--r-- 1 root root 6379 Mar 17 08:07 utils.py
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline#
```

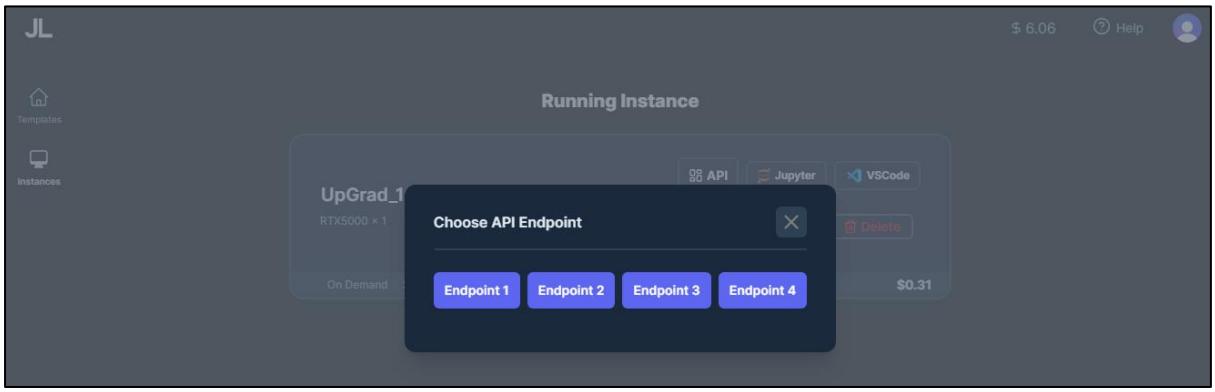
Airflow server:

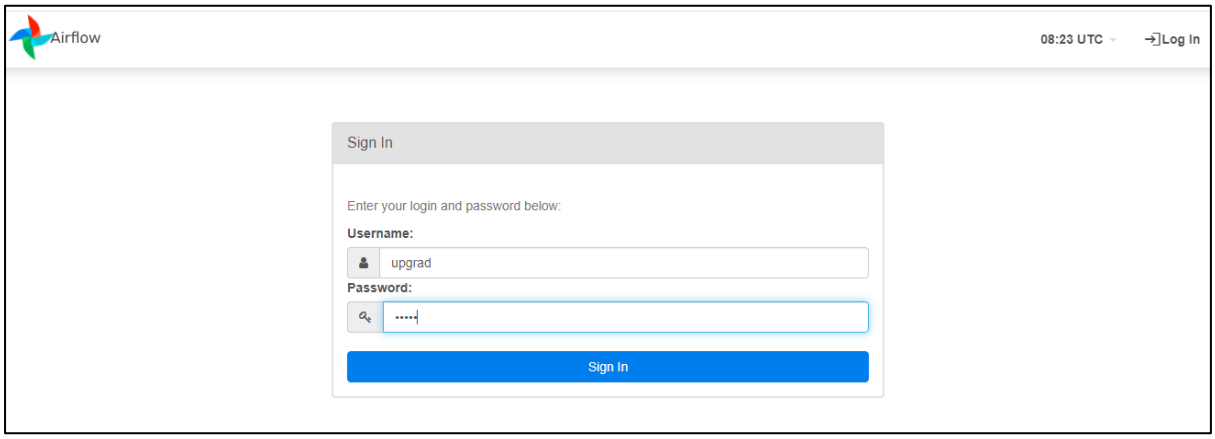
```
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline# airflow db init
DB: sqlite:////home/airflow/airflow.db
[2024-03-17 08:13:59,576] {db.py:1462} INFO - Creating tables
INFO  [alembic.runtime.migration] Context impl SQLiteImpl.
INFO  [alembic.runtime.migration] Will assume non-transactional DDL.
WARNI [airflow.models.crypto] empty cryptography key - values will not be stored encrypted.
Initialization done
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline#
```

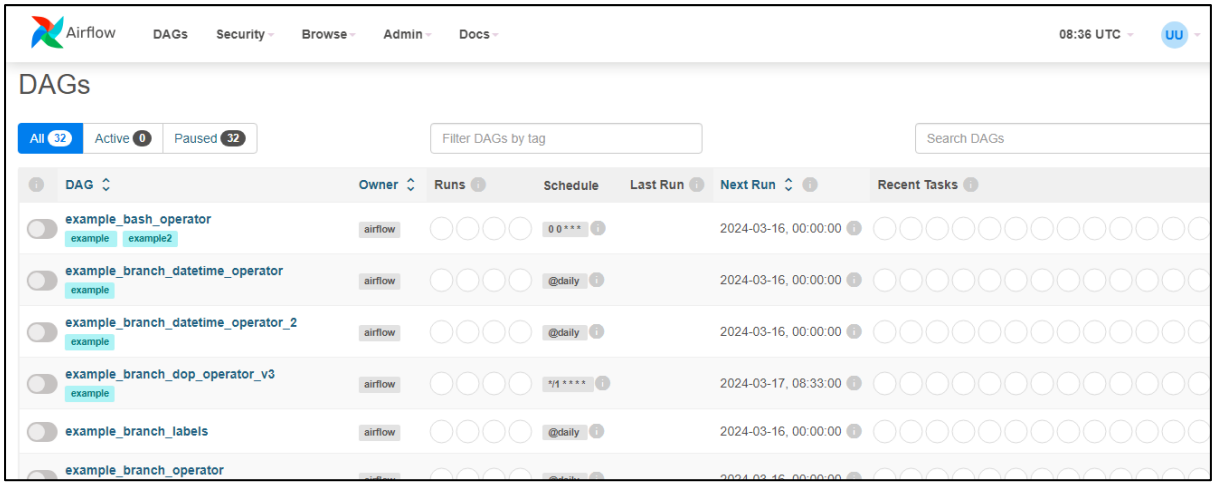
```
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline# airflow users create --username upgrad --firstname upgrad --
lastname upgrad --role Admin --email shindesagarm@yahoo.co.in --password admin
upgrad already exist in the db
```

```
root@bd079ac964cf:~/airflow/dags/Lead_scoring_training_pipeline# airflow webserver
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /   _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
Running the Gunicorn Server with:
Workers: 4 sync
Host: 0.0.0.0:6007
Timeout: 120
Logfiles: - -
Access Logformat:
==================================================================
[2024-03-17 08:19:16 +0000] [1010] [INFO] Starting gunicorn 20.1.0
[2024-03-17 08:19:17 +0000] [1010] [INFO] Listening at: http://0.0.0.0:6007 (1010)
[2024-03-17 08:19:17 +0000] [1010] [INFO] Using worker: sync
[2024-03-17 08:19:17 +0000] [1012] [INFO] Booting worker with pid: 1012
[2024-03-17 08:19:17 +0000] [1013] [INFO] Booting worker with pid: 1013
[2024-03-17 08:19:17 +0000] [1014] [INFO] Booting worker with pid: 1014
[2024-03-17 08:19:17 +0000] [1015] [INFO] Booting worker with pid: 1015
```

```
root@bd079ac964cf:~# airflow scheduler
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /   _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
[2024-03-17 08:20:10 +0000] [1048] [INFO] Starting gunicorn 20.1.0
[2024-03-17 08:20:10 +0000] [1048] [INFO] Listening at: http://0.0.0.0:8793 (1048)
[2024-03-17 08:20:10 +0000] [1048] [INFO] Using worker: sync
[2024-03-17 08:20:10 +0000] [1049] [INFO] Booting worker with pid: 1049
[2024-03-17 08:20:10,116] {scheduler_job.py:708} INFO - Starting the scheduler
[2024-03-17 08:20:10,116] {scheduler_job.py:713} INFO - Processing each file at most -1 times
[2024-03-17 08:20:10,118] {executor_loader.py:105} INFO - Loaded executor: SequentialExecutor
[2024-03-17 08:20:10,123] {manager.py:160} INFO - Launched DagFileProcessorManager with pid: 1050
[2024-03-17 08:20:10,124] {scheduler_job.py:1233} INFO - Resetting orphaned tasks for active dag runs
[2024-03-17 08:20:10,128] {settings.py:55} INFO - Configured default timezone Timezone('UTC')
[2024-03-17 08:20:10 +0000] [1051] [INFO] Booting worker with pid: 1051
[2024-03-17 08:20:10,133] {scheduler_job.py:1256} INFO - Marked 1 SchedulerJob instances as failed
[2024-03-17 08:20:10,139] {manager.py:406} WARNING - Because we cannot use more than 1 thread (parsing_processes = 2) when us
ing sqlite. So we set parallelism to 1.
```





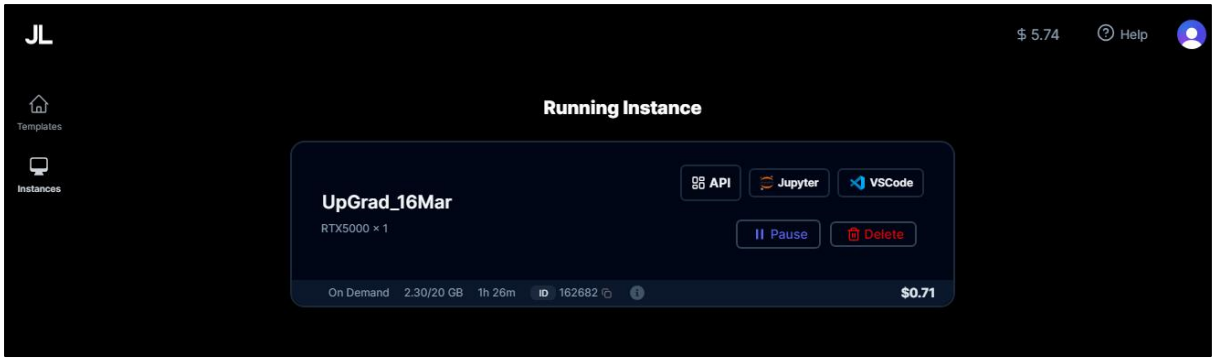


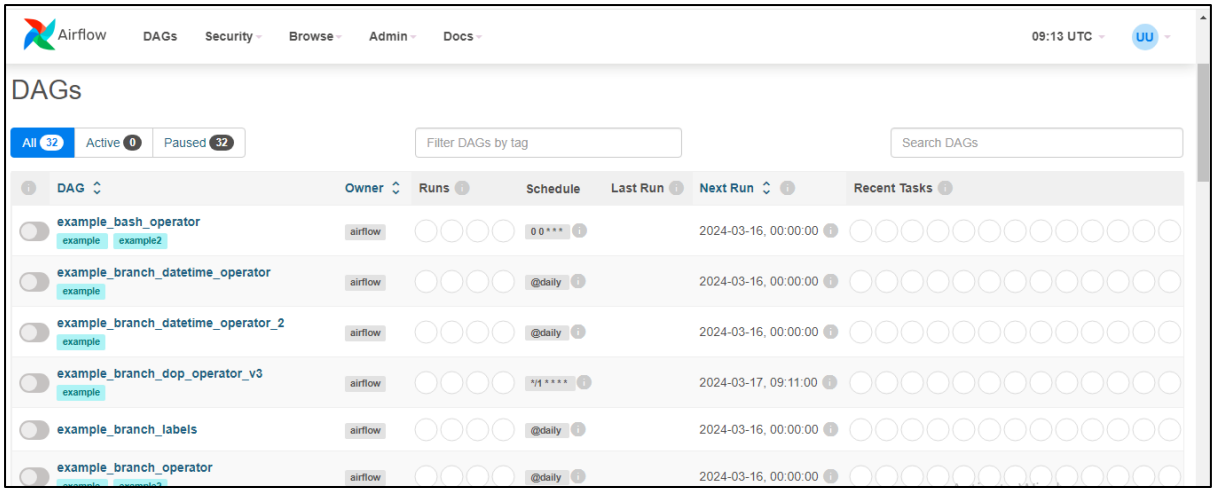
Inference Pipeline:

```
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts# cp utils.py /home/airflow/dags/Lead_scoring_inference_pipeline/
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts#
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts# cp constants.py /home/airflow/dags/Lead_scoring_inference_pipeline/
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts#
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts# cp lead_scoring_inference_pipeline.py /home/airflow/dags/Lead_scoring_infer
ence_pipeline/
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts#
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts# ls -lrt
total 16
-rw-r--r-- 1 root root 2390 Mar 17 08:54 constants.py
-rw-r--r-- 1 root root 6414 Mar 17 08:57 utils.py
-rw-r--r-- 1 root root 2641 Mar 17 09:00 lead_scoring_inference_pipeline.py
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts#
root@bd079ac964cf:~/Assignment/03_inference_pipeline/scripts# cd /home/airflow/dags/Lead_scoring_inference_pipeline/
root@bd079ac964cf:~/airflow/dags/Lead_scoring_inference_pipeline# ls -lrt
total 16
-rw-r--r-- 1 root root 6414 Mar 17 09:02 utils.py
-rw-r--r-- 1 root root 2390 Mar 17 09:03 constants.py
-rw-r--r-- 1 root root 2641 Mar 17 09:03 lead_scoring_inference_pipeline.py
root@bd079ac964cf:~/airflow/dags/Lead_scoring_inference_pipeline#
```

```
[2024-03-17 09:07:41,699] {db.py:1462} INFO - Creating tables
INFO  [alembic.runtime.migration] Context impl SQLiteImpl.
INFO  [alembic.runtime.migration] Will assume non-transactional DDL.
WARNI [airflow.models.crypto] empty cryptography key - values will not be stored encrypted.
Initialization done
root@bd079ac964cf:~/airflow/dags/Lead_scoring_inference_pipeline# airflow users create --username upgrad --firstname upgrad --lastname up
grad --role Admin --email shindesagarm@yahoo.co.in --password admin
upgrad already exist in the db
root@bd079ac964cf:~/airflow/dags/Lead_scoring_inference_pipeline# airflow webserver
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /   _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
Traceback (most recent call last):
  File "/opt/conda/bin/airflow", line 11, in <module>
    sys.exit(main())
  File "/opt/conda/lib/python3.8/site-packages/airflow/__main__.py", line 38, in main
    args.func(args)
  File "/opt/conda/lib/python3.8/site-packages/airflow/cli/cli_parser.py", line 51, in command
    return func(*args, **kwargs)
  File "/opt/conda/lib/python3.8/site-packages/airflow/utils/cli.py", line 99, in wrapper
    return f(*args, **kwargs)
  File "/opt/conda/lib/python3.8/site-packages/airflow/cli/commands/webserver_command.py", line 363, in webserver
    check_if_pidfile_process_is_running(pid_file=pid_file, process_name="webserver")
  File "/opt/conda/lib/python3.8/site-packages/airflow/utils/process_utils.py", line 300, in check_if_pidfile_process_is_running
    raise AirflowException(f"The {process_name} is already running under PID {pid}.")
airflow.exceptions.AirflowException: The webserver is already running under PID 1010.
```

```
root@bd079ac964cf:~# airflow scheduler
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /   _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
[2024-03-17 09:09:48 +0000] [8097] [INFO] Starting gunicorn 20.1.0
[2024-03-17 09:09:48 +0000] [8097] [ERROR] Connection in use: ('0.0.0.0', 8793)
[2024-03-17 09:09:48 +0000] [8097] [ERROR] Retrying in 1 second.
[2024-03-17 09:09:48,180] {scheduler_job.py:708} INFO - Starting the scheduler
[2024-03-17 09:09:48,180] {scheduler_job.py:713} INFO - Processing each file at most -1 times
[2024-03-17 09:09:48,182] {executor_loader.py:105} INFO - Loaded executor: SequentialExecutor
[2024-03-17 09:09:48,188] {manager.py:160} INFO - Launched DagFileProcessorManager with pid: 8098
[2024-03-17 09:09:48,189] {scheduler_job.py:1233} INFO - Resetting orphaned tasks for active dag runs
[2024-03-17 09:09:48,193] {settings.py:55} INFO - Configured default timezone Timezone('UTC')
[2024-03-17 09:09:48,204] {manager.py:406} WARNING - Because we cannot use more than 1 thread (parsing_processes = 2) when using sqlite.
So we set parallelism to 1.
```




Airflow
DAGs
Security
Browse
Admin
Docs
09:13 UTC
UU
DAGs
All 32
Active 0
Paused 32
Filter DAGs by tag
Search DAGs
DAG
Owner
Runs
Schedule
Last Run
Next Run
Recent Tasks
example_bash_operator
example
example2
airflow
0 0 * * *
2024-03-16, 00:00:00
example_branch_datetime_operator
example
airflow
@daily
2024-03-16, 00:00:00
example_branch_datetime_operator_2
example
airflow
@daily
2024-03-16, 00:00:00
example_branch_dop_operator_v3
example
airflow
*/1 * * * *
2024-03-17, 09:11:00
example_branch_labels
airflow
@daily
2024-03-16, 00:00:00
example_branch_operator
airflow
@daily
2024-03-16, 00:00:00