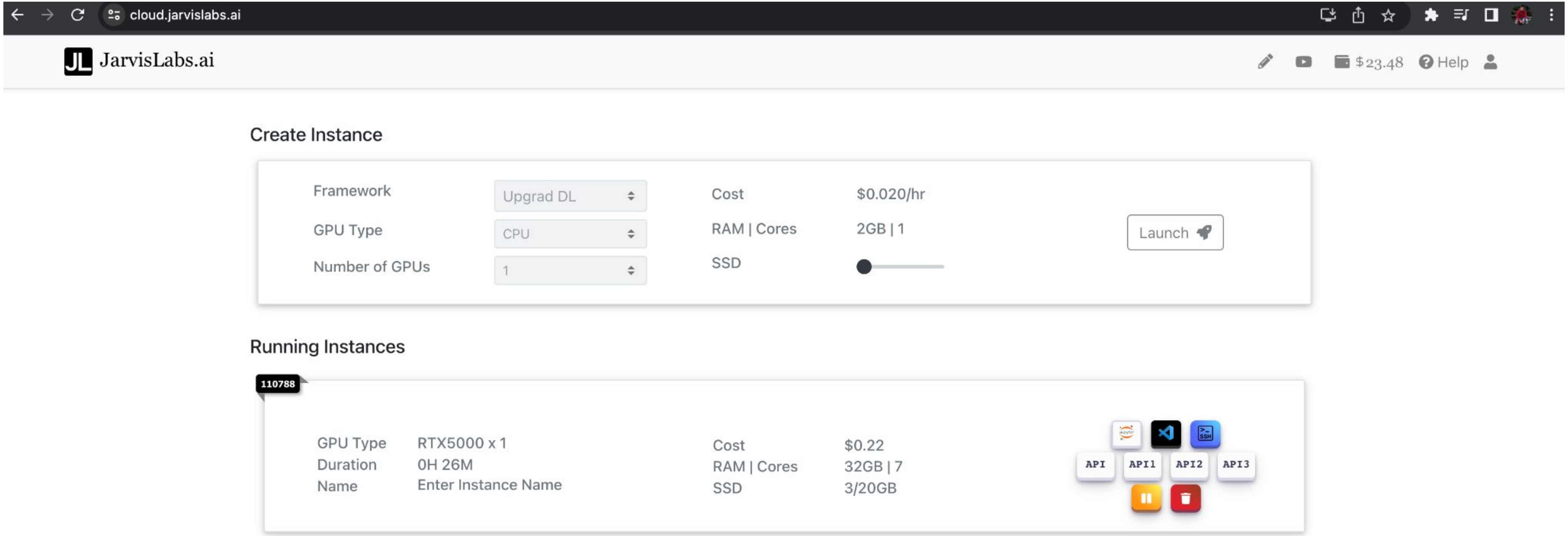
MLOps CodePro Case Study

PRAVIN TAWADE



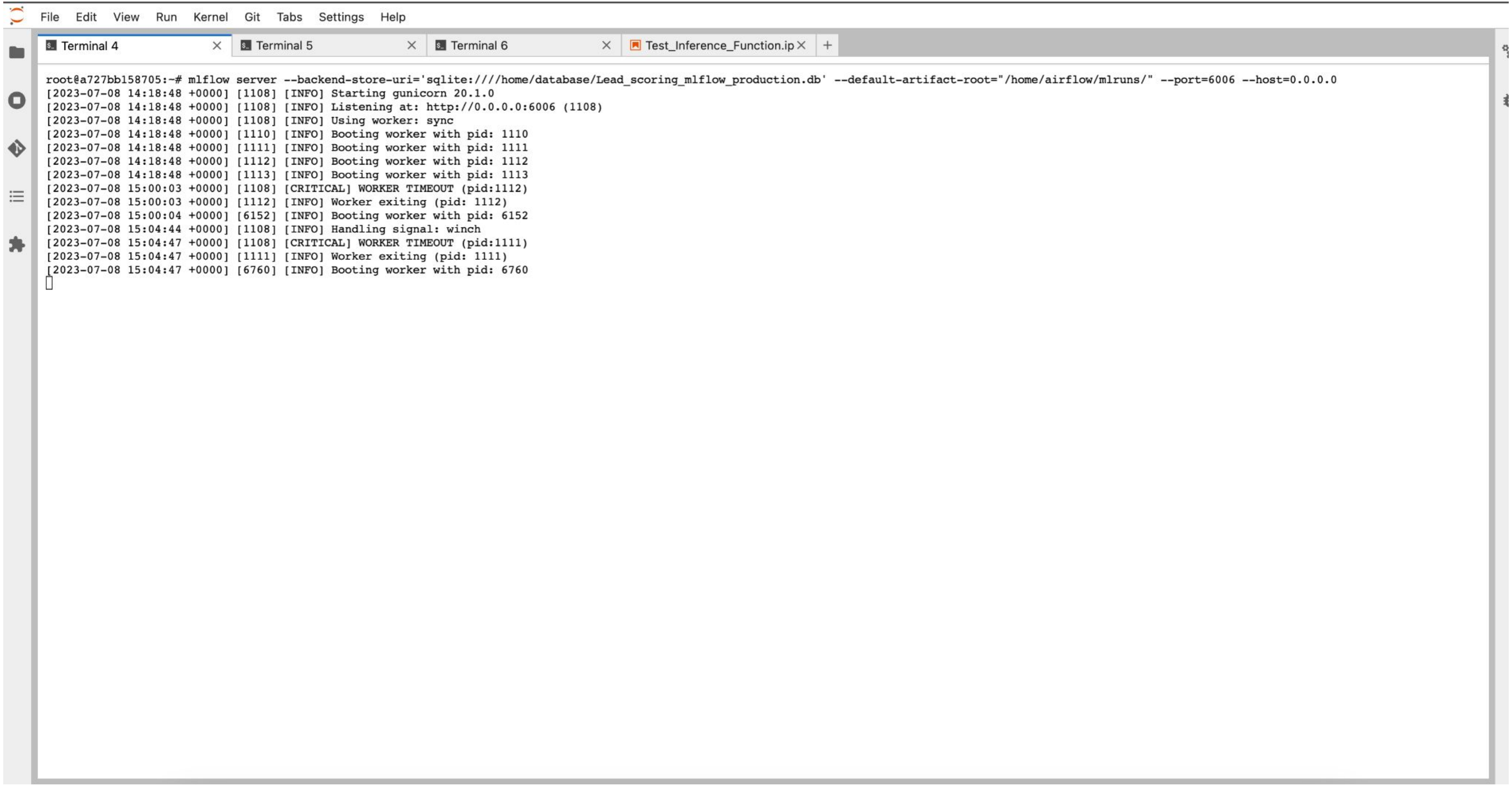
Create Instance

Framework	Upgrad DL	Cost	$0.020/hr
GPU Type	CPU	RAM | Cores	2GB | 1
Number of GPUs	1	SSD	

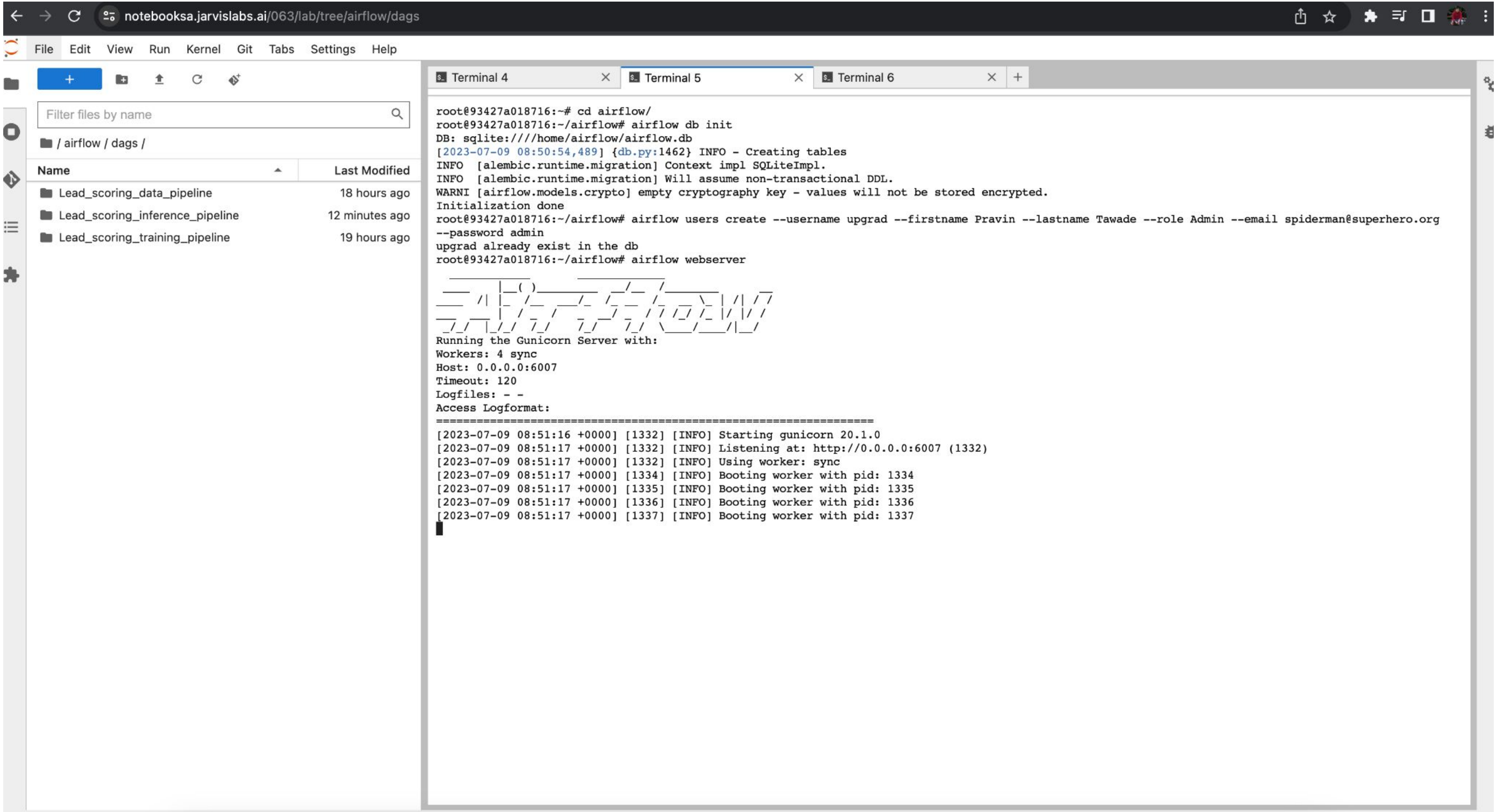
Running Instances

110788

GPU Type	RTX5000 x 1	Cost	$0.22
Duration	0H 26M	RAM | Cores	32GB | 7
Name	Enter Instance Name	SSD	3/20GB

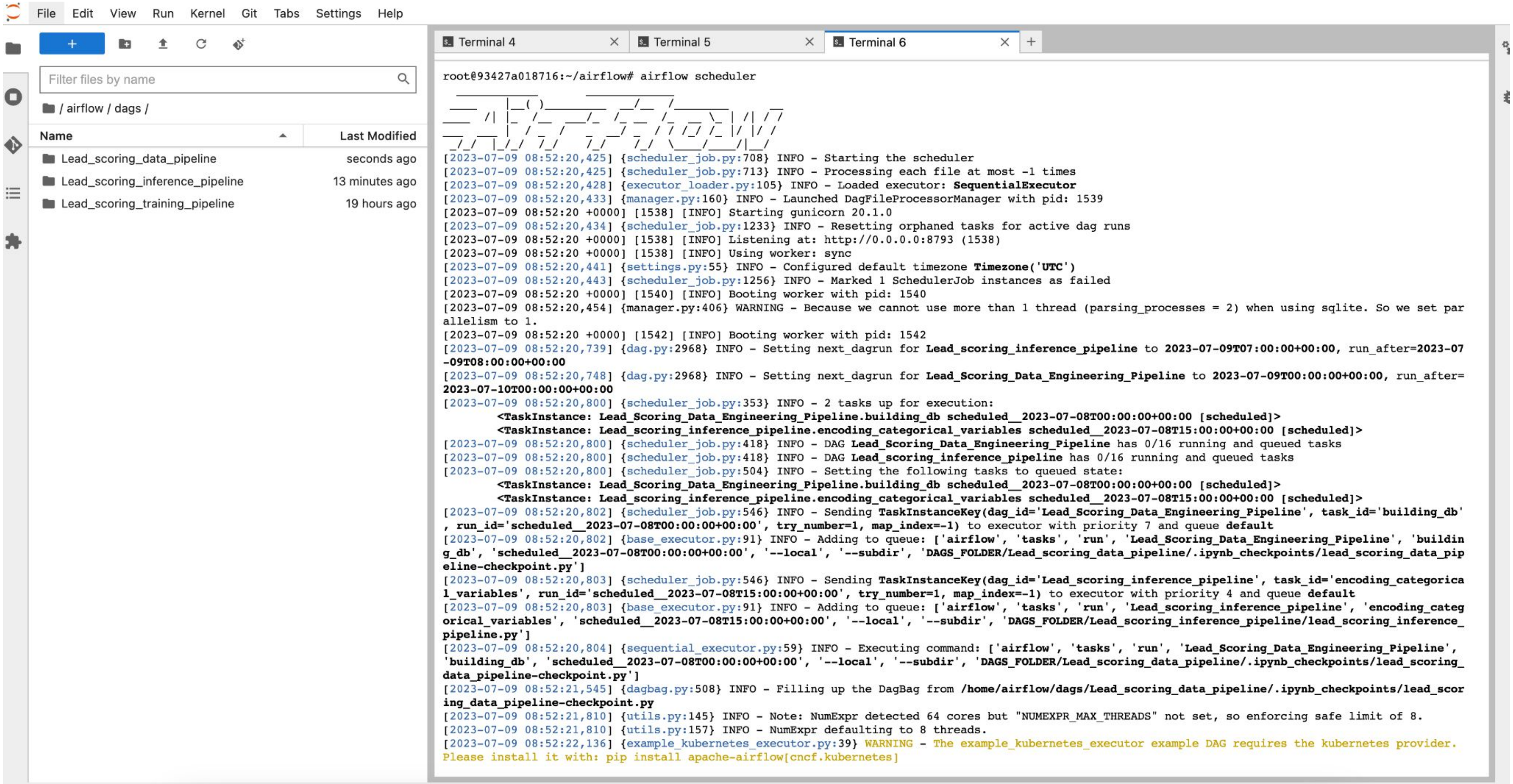


```
root@a727bb158705:~# mlflow server --backend-store-uri='sqlite:////home/database/Lead_scoring_mlflow_production.db' --default-artifact-root="/home/airflow/mlruns/" --port=6006 --host=0.0.0.0
[2023-07-08 14:18:48 +0000] [1108] [INFO] Starting gunicorn 20.1.0
[2023-07-08 14:18:48 +0000] [1108] [INFO] Listening at: http://0.0.0.0:6006 (1108)
[2023-07-08 14:18:48 +0000] [1108] [INFO] Using worker: sync
[2023-07-08 14:18:48 +0000] [1110] [INFO] Booting worker with pid: 1110
[2023-07-08 14:18:48 +0000] [1111] [INFO] Booting worker with pid: 1111
[2023-07-08 14:18:48 +0000] [1112] [INFO] Booting worker with pid: 1112
[2023-07-08 14:18:48 +0000] [1113] [INFO] Booting worker with pid: 1113
[2023-07-08 15:00:03 +0000] [1108] [CRITICAL] WORKER TIMEOUT (pid:1112)
[2023-07-08 15:00:03 +0000] [1112] [INFO] Worker exiting (pid: 1112)
[2023-07-08 15:00:04 +0000] [6152] [INFO] Booting worker with pid: 6152
[2023-07-08 15:04:44 +0000] [1108] [INFO] Handling signal: winch
[2023-07-08 15:04:47 +0000] [1108] [CRITICAL] WORKER TIMEOUT (pid:1111)
[2023-07-08 15:04:47 +0000] [1111] [INFO] Worker exiting (pid: 1111)
[2023-07-08 15:04:47 +0000] [6760] [INFO] Booting worker with pid: 6760
```

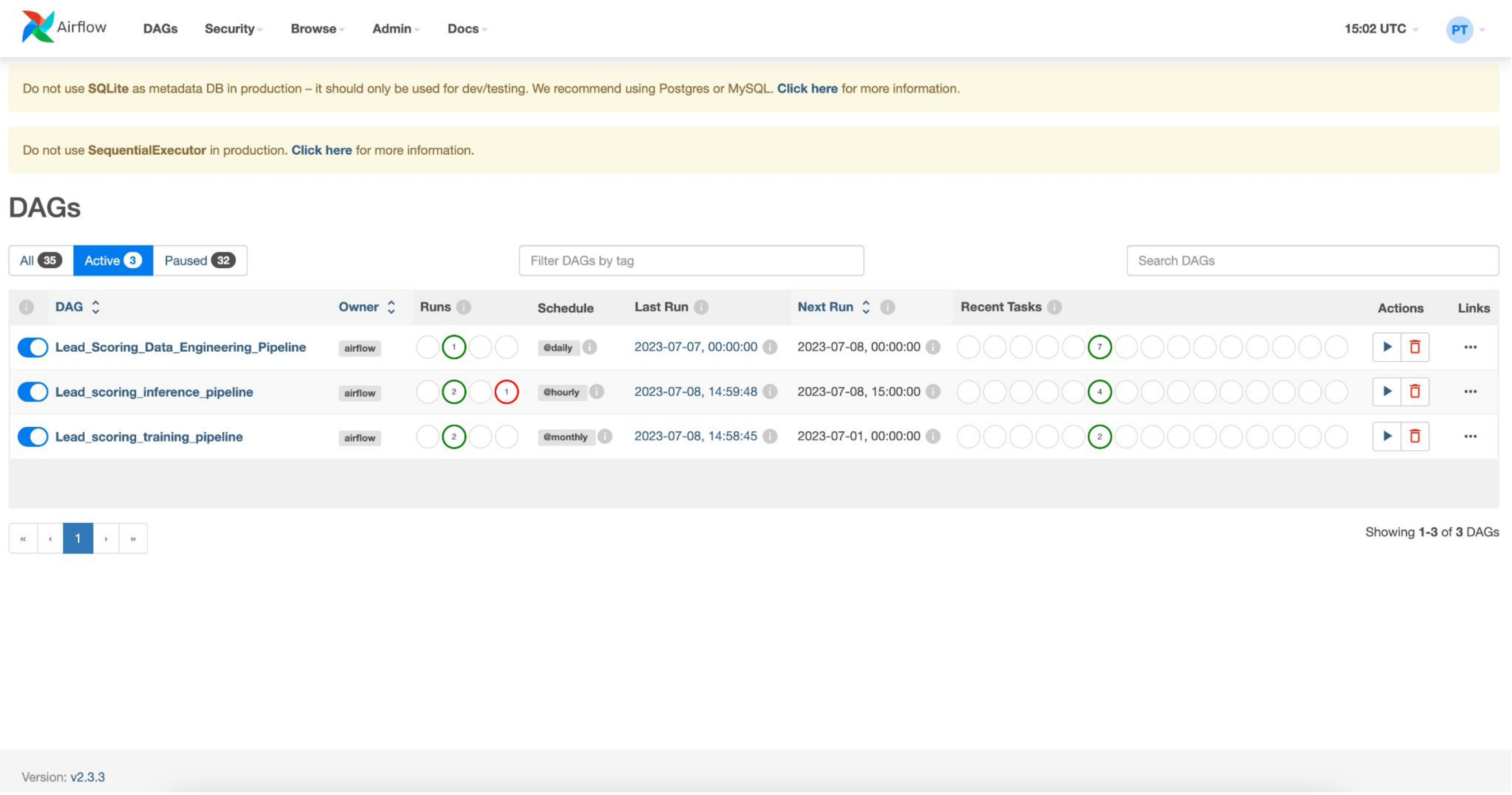


Name	Last Modified
Lead_scoring_data_pipeline	18 hours ago
Lead_scoring_inference_pipeline	12 minutes ago
Lead_scoring_training_pipeline	19 hours ago

```
root@93427a018716:~# cd airflow/
root@93427a018716:~/airflow# airflow db init
DB: sqlite:////home/airflow/airflow.db
[2023-07-09 08:50:54,489] {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@93427a018716:~/airflow# airflow users create --username upgrad --firstname Pravin --lastname Tawade --role Admin --email spiderman@superhero.org
--password admin
upgrad already exist in the db
root@93427a018716:~/airflow# airflow webserver
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /   _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
Running the Gunicorn Server with:
Workers: 4 sync
Host: 0.0.0.0:6007
Timeout: 120
Logfiles: - -
Access Logformat:
=================================================================
[2023-07-09 08:51:16 +0000] [1332] [INFO] Starting gunicorn 20.1.0
[2023-07-09 08:51:17 +0000] [1332] [INFO] Listening at: http://0.0.0.0:6007 (1332)
[2023-07-09 08:51:17 +0000] [1332] [INFO] Using worker: sync
[2023-07-09 08:51:17 +0000] [1334] [INFO] Booting worker with pid: 1334
[2023-07-09 08:51:17 +0000] [1335] [INFO] Booting worker with pid: 1335
[2023-07-09 08:51:17 +0000] [1336] [INFO] Booting worker with pid: 1336
[2023-07-09 08:51:17 +0000] [1337] [INFO] Booting worker with pid: 1337
```



```
root@93427a018716:~/airflow# airflow scheduler
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /   _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
[2023-07-09 08:52:20,425] {scheduler_job.py:708} INFO - Starting the scheduler
[2023-07-09 08:52:20,425] {scheduler_job.py:713} INFO - Processing each file at most -1 times
[2023-07-09 08:52:20,428] {executor_loader.py:105} INFO - Loaded executor: SequentialExecutor
[2023-07-09 08:52:20,433] {manager.py:160} INFO - Launched DagFileProcessorManager with pid: 1539
[2023-07-09 08:52:20 +0000] [1538] [INFO] Starting gunicorn 20.1.0
[2023-07-09 08:52:20,434] {scheduler_job.py:1233} INFO - Resetting orphaned tasks for active dag runs
[2023-07-09 08:52:20 +0000] [1538] [INFO] Listening at: http://0.0.0.0:8793 (1538)
[2023-07-09 08:52:20 +0000] [1538] [INFO] Using worker: sync
[2023-07-09 08:52:20,441] {settings.py:55} INFO - Configured default timezone Timezone('UTC')
[2023-07-09 08:52:20,443] {scheduler_job.py:1256} INFO - Marked 1 SchedulerJob instances as failed
[2023-07-09 08:52:20 +0000] [1540] [INFO] Booting worker with pid: 1540
[2023-07-09 08:52:20,454] {manager.py:406} WARNING - Because we cannot use more than 1 thread (parsing_processes = 2) when using sqlite. So we set par
allelism to 1.
[2023-07-09 08:52:20 +0000] [1542] [INFO] Booting worker with pid: 1542
[2023-07-09 08:52:20,739] {dag.py:2968} INFO - Setting next_dagrun for Lead_scoring_inference_pipeline to 2023-07-09T07:00:00+00:00, run_after=2023-07
-09T08:00:00+00:00
[2023-07-09 08:52:20,748] {dag.py:2968} INFO - Setting next_dagrun for Lead_Scoring_Data_Engineering_Pipeline to 2023-07-09T00:00:00+00:00, run_after=
2023-07-10T00:00:00+00:00
[2023-07-09 08:52:20,800] {scheduler_job.py:353} INFO - 2 tasks up for execution:
        <TaskInstance: Lead_Scoring_Data_Engineering_Pipeline.building_db scheduled__2023-07-08T00:00:00+00:00 [scheduled]>
        <TaskInstance: Lead_scoring_inference_pipeline.encoding_categorical_variables scheduled__2023-07-08T15:00:00+00:00 [scheduled]>
[2023-07-09 08:52:20,800] {scheduler_job.py:418} INFO - DAG Lead_Scoring_Data_Engineering_Pipeline has 0/16 running and queued tasks
[2023-07-09 08:52:20,800] {scheduler_job.py:418} INFO - DAG Lead_scoring_inference_pipeline has 0/16 running and queued tasks
[2023-07-09 08:52:20,800] {scheduler_job.py:504} INFO - Setting the following tasks to queued state:
        <TaskInstance: Lead_Scoring_Data_Engineering_Pipeline.building_db scheduled__2023-07-08T00:00:00+00:00 [scheduled]>
        <TaskInstance: Lead_scoring_inference_pipeline.encoding_categorical_variables scheduled__2023-07-08T15:00:00+00:00 [scheduled]>
[2023-07-09 08:52:20,802] {scheduler_job.py:546} INFO - Sending TaskInstanceKey(dag_id='Lead_Scoring_Data_Engineering_Pipeline', task_id='building_db'
, run_id='scheduled__2023-07-08T00:00:00+00:00', try_number=1, map_index=-1) to executor with priority 7 and queue default
[2023-07-09 08:52:20,802] {base_executor.py:91} INFO - Adding to queue: ['airflow', 'tasks', 'run', 'Lead_Scoring_Data_Engineering_Pipeline', 'buildin
g_db', 'scheduled__2023-07-08T00:00:00+00:00', '--local', '--subdir', 'DAGS_FOLDER/Lead_scoring_data_pipeline/.ipynb_checkpoints/lead_scoring_data_pip
eline-checkpoint.py']
[2023-07-09 08:52:20,803] {scheduler_job.py:546} INFO - Sending TaskInstanceKey(dag_id='Lead_scoring_inference_pipeline', task_id='encoding_categorica
l_variables', run_id='scheduled__2023-07-08T15:00:00+00:00', try_number=1, map_index=-1) to executor with priority 4 and queue default
[2023-07-09 08:52:20,803] {base_executor.py:91} INFO - Adding to queue: ['airflow', 'tasks', 'run', 'Lead_scoring_inference_pipeline', 'encoding_categ
orical_variables', 'scheduled__2023-07-08T15:00:00+00:00', '--local', '--subdir', 'DAGS_FOLDER/Lead_scoring_inference_pipeline/lead_scoring_inference_
pipeline.py']
[2023-07-09 08:52:20,804] {sequential_executor.py:59} INFO - Executing command: ['airflow', 'tasks', 'run', 'Lead_Scoring_Data_Engineering_Pipeline',
'building_db', 'scheduled__2023-07-08T00:00:00+00:00', '--local', '--subdir', 'DAGS_FOLDER/Lead_scoring_data_pipeline/.ipynb_checkpoints/lead_scoring_
data_pipeline-checkpoint.py']
[2023-07-09 08:52:21,545] {dagbag.py:508} INFO - Filling up the DagBag from /home/airflow/dags/Lead_scoring_data_pipeline/.ipynb_checkpoints/lead_scor
ing_data_pipeline-checkpoint.py
[2023-07-09 08:52:21,810] {utils.py:145} INFO - Note: NumExpr detected 64 cores but "NUMEXPR_MAX_THREADS" not set, so enforcing safe limit of 8.
[2023-07-09 08:52:21,810] {utils.py:157} INFO - NumExpr defaulting to 8 threads.
[2023-07-09 08:52:22,136] {example_kubernetes_executor.py:39} WARNING - The example_kubernetes_executor example DAG requires the kubernetes provider.
Please install it with: pip install apache-airflow[cncf.kubernetes]
```



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.

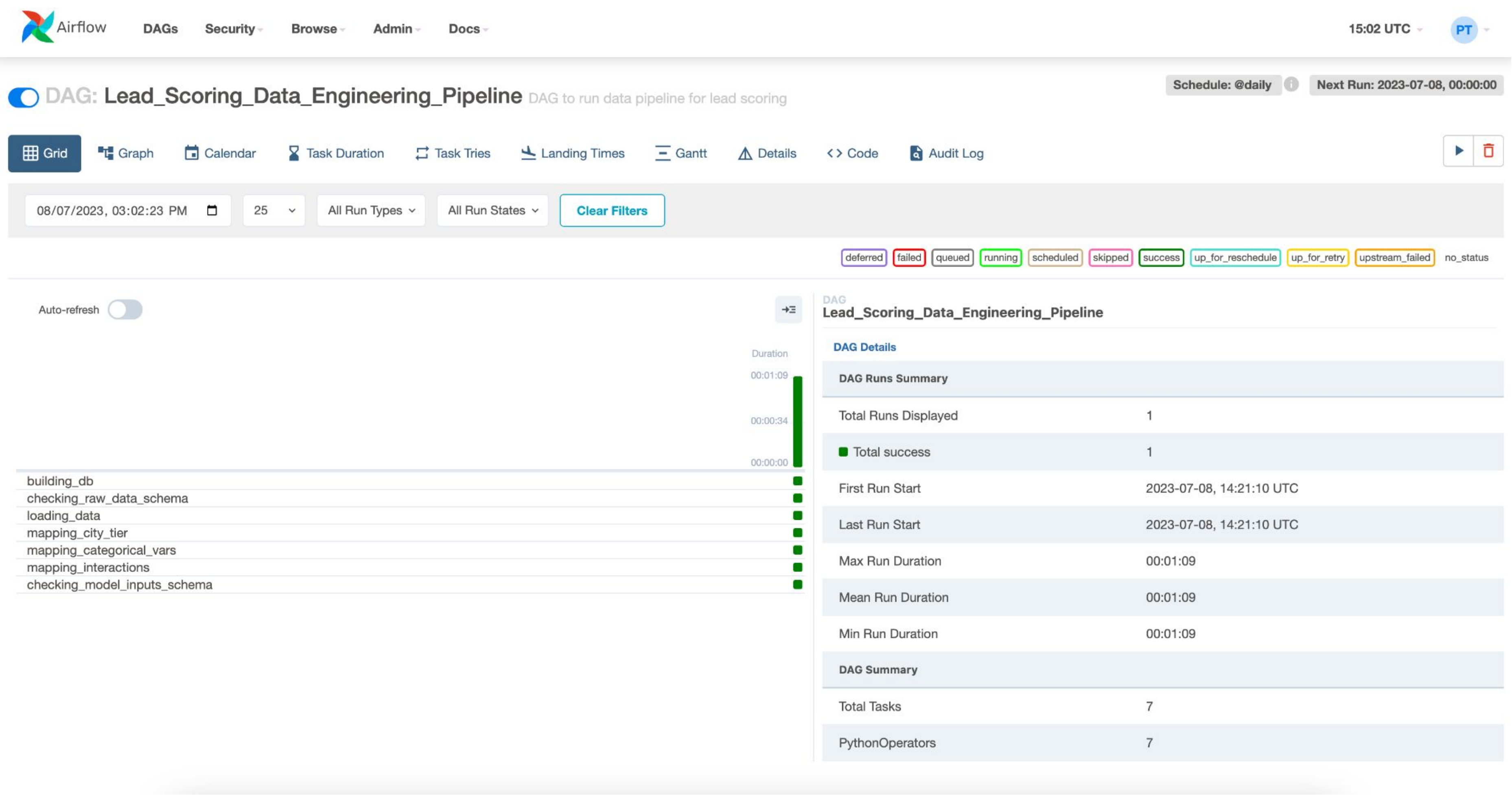
Do not use **SequentialExecutor** in production. **Click here** for more information.

DAGs

	DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
●	Lead_Scoring_Data_Engineering_Pipeline	airflow	1	@daily	2023-07-07, 00:00:00	2023-07-08, 00:00:00	7	▶ 🗑	•••
●	Lead_scoring_inference_pipeline	airflow	2 1	@hourly	2023-07-08, 14:59:48	2023-07-08, 15:00:00	4	▶ 🗑	•••
●	Lead_scoring_training_pipeline	airflow	2	@monthly	2023-07-08, 14:58:45	2023-07-01, 00:00:00	2	▶ 🗑	•••

Showing **1-3** of **3** DAGs

Version: v2.3.3



DAG: Lead_Scoring_Data_Engineering_Pipeline DAG to run data pipeline for lead scoring

Schedule: @daily

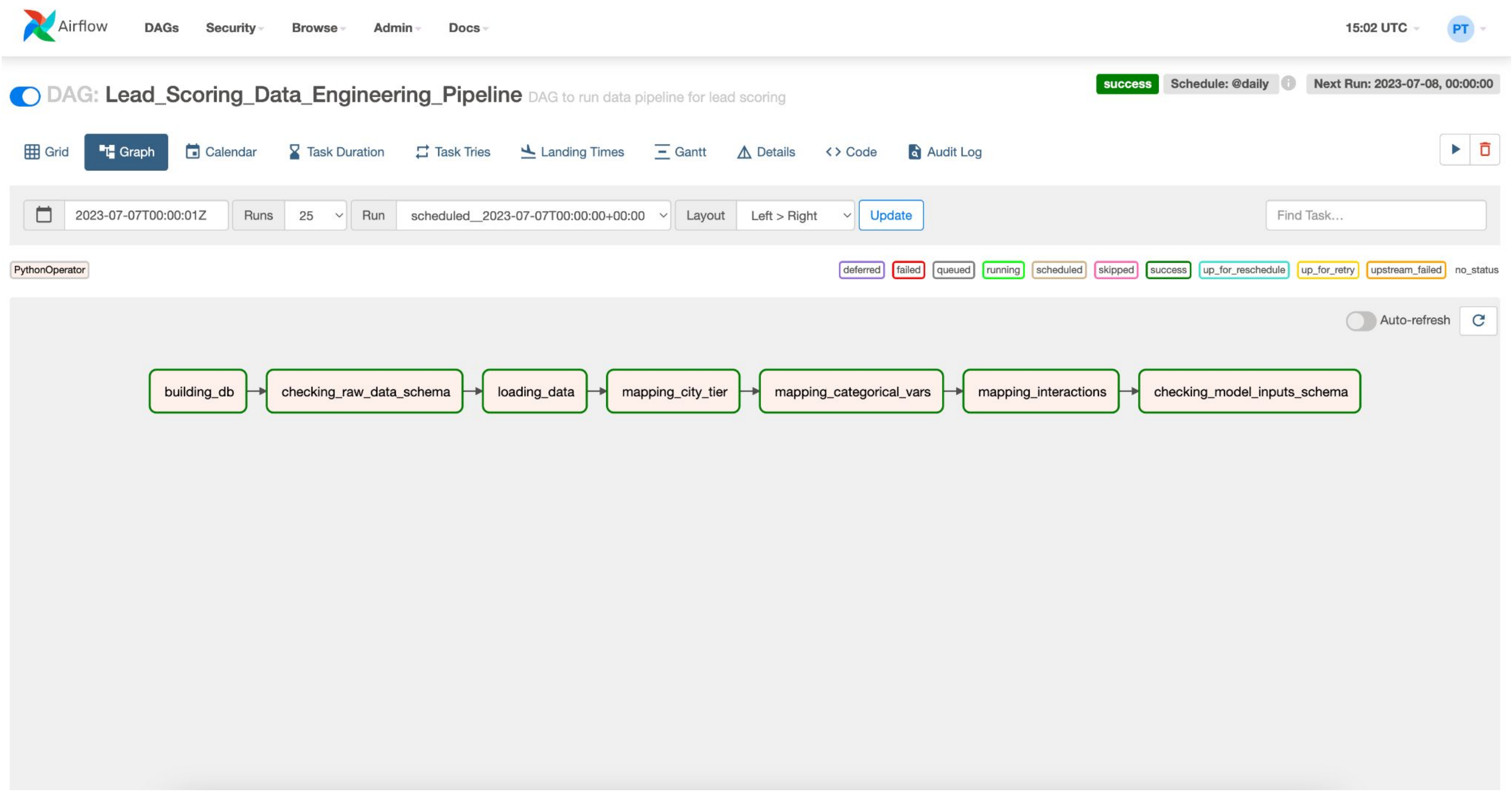
Next Run: 2023-07-08, 00:00:00

building_db
checking_raw_data_schema
loading_data
mapping_city_tier
mapping_categorical_vars
mapping_interactions
checking_model_inputs_schema

DAG
Lead_Scoring_Data_Engineering_Pipeline

DAG Details

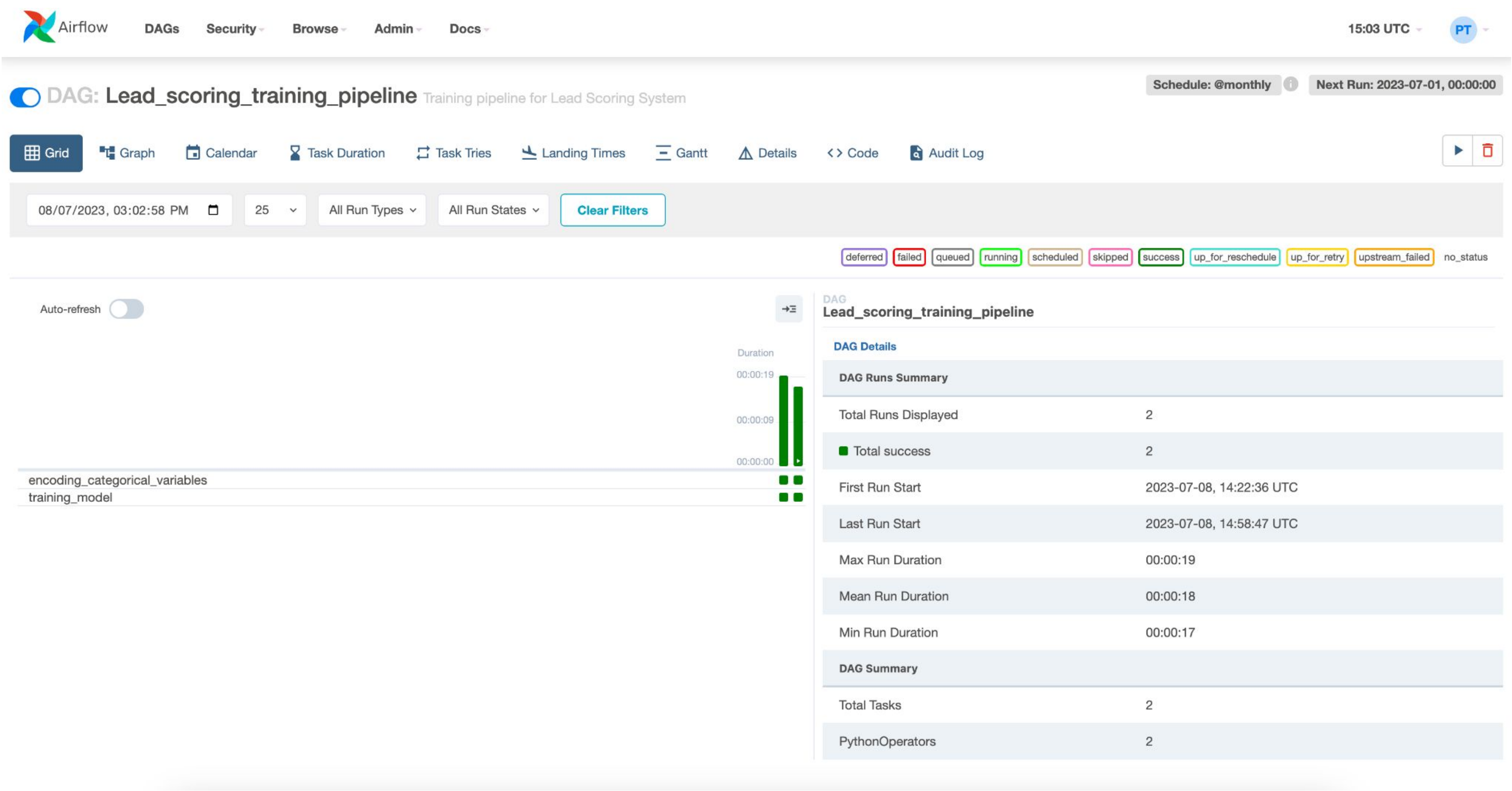
DAG Runs Summary	
Total Runs Displayed	1
🟩 Total success	1
First Run Start	2023-07-08, 14:21:10 UTC
Last Run Start	2023-07-08, 14:21:10 UTC
Max Run Duration	00:01:09
Mean Run Duration	00:01:09
Min Run Duration	00:01:09
DAG Summary	
Total Tasks	7
PythonOperators	7



DAG: Lead_Scoring_Data_Engineering_Pipeline DAG to run data pipeline for lead scoring

success Schedule: @daily Next Run: 2023-07-08, 00:00:00

2023-07-07T00:00:01Z Runs 25 Run scheduled__2023-07-07T00:00:00+00:00 Layout Left > Right Update

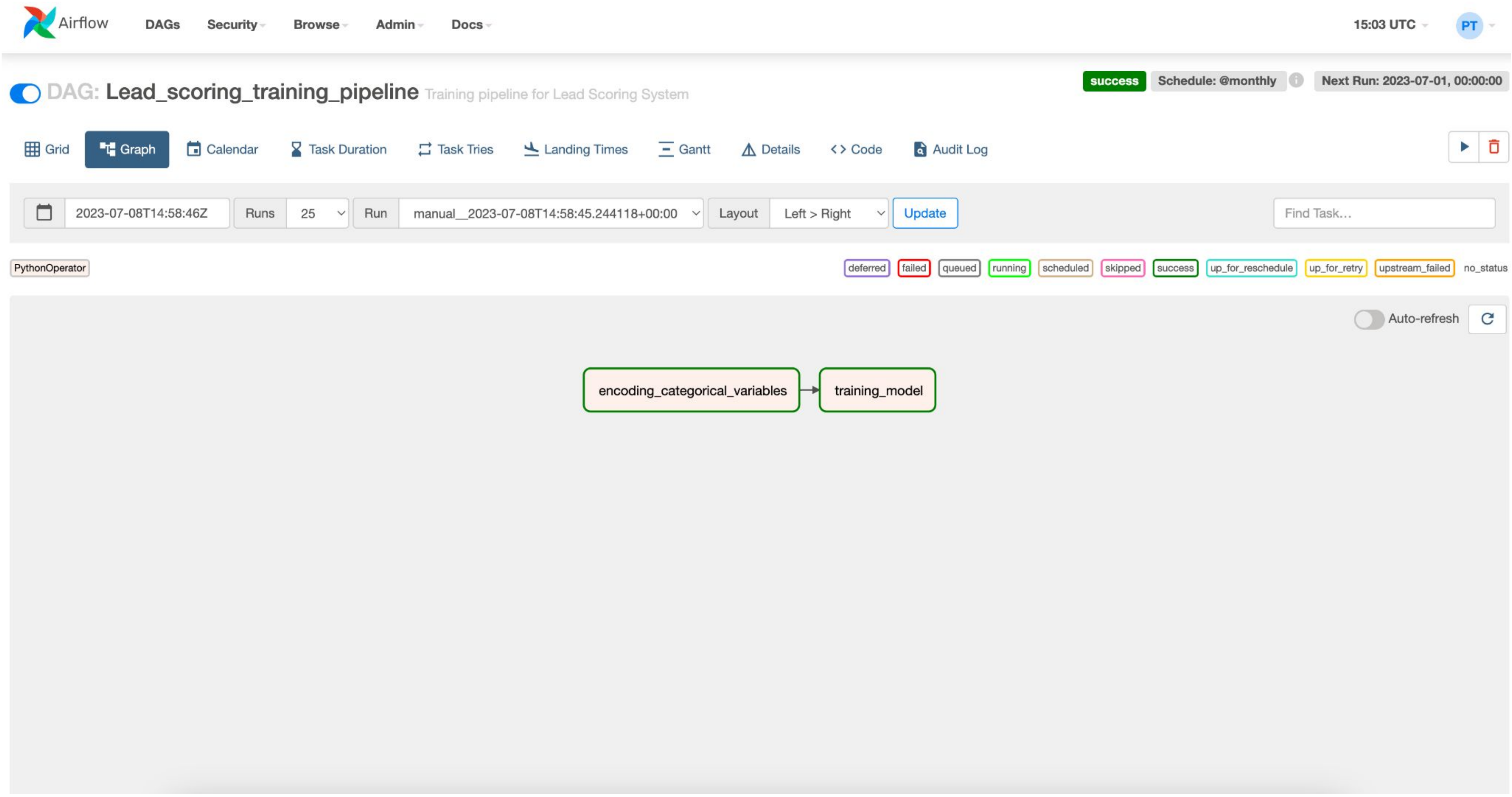


DAG

Lead_scoring_training_pipeline

DAG Details

DAG Runs Summary	
Total Runs Displayed	2
🟩 Total success	2
First Run Start	2023-07-08, 14:22:36 UTC
Last Run Start	2023-07-08, 14:58:47 UTC
Max Run Duration	00:00:19
Mean Run Duration	00:00:18
Min Run Duration	00:00:17
DAG Summary	
Total Tasks	2
PythonOperators	2

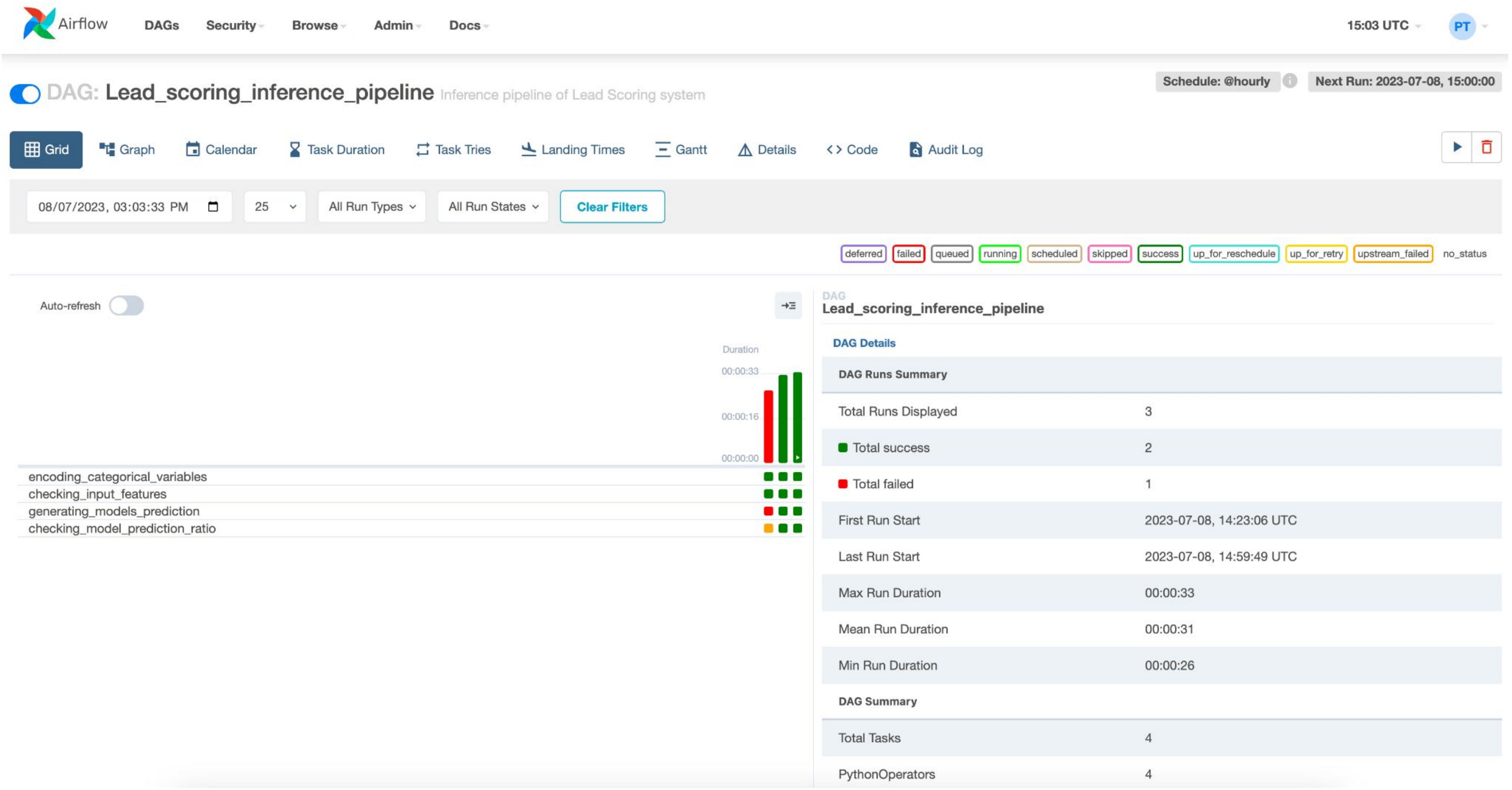


DAG: Lead_scoring_training_pipeline Training pipeline for Lead Scoring System

success Schedule: @monthly Next Run: 2023-07-01, 00:00:00

2023-07-08T14:58:46Z Runs 25 Run manual__2023-07-08T14:58:45.244118+00:00 Layout Left > Right

encoding_categorical_variables → training_model



DAG: Lead_scoring_inference_pipeline Inference pipeline of Lead Scoring system

Schedule: @hourly

Next Run: 2023-07-08, 15:00:00

08/07/2023, 03:03:33 PM

25

All Run Types

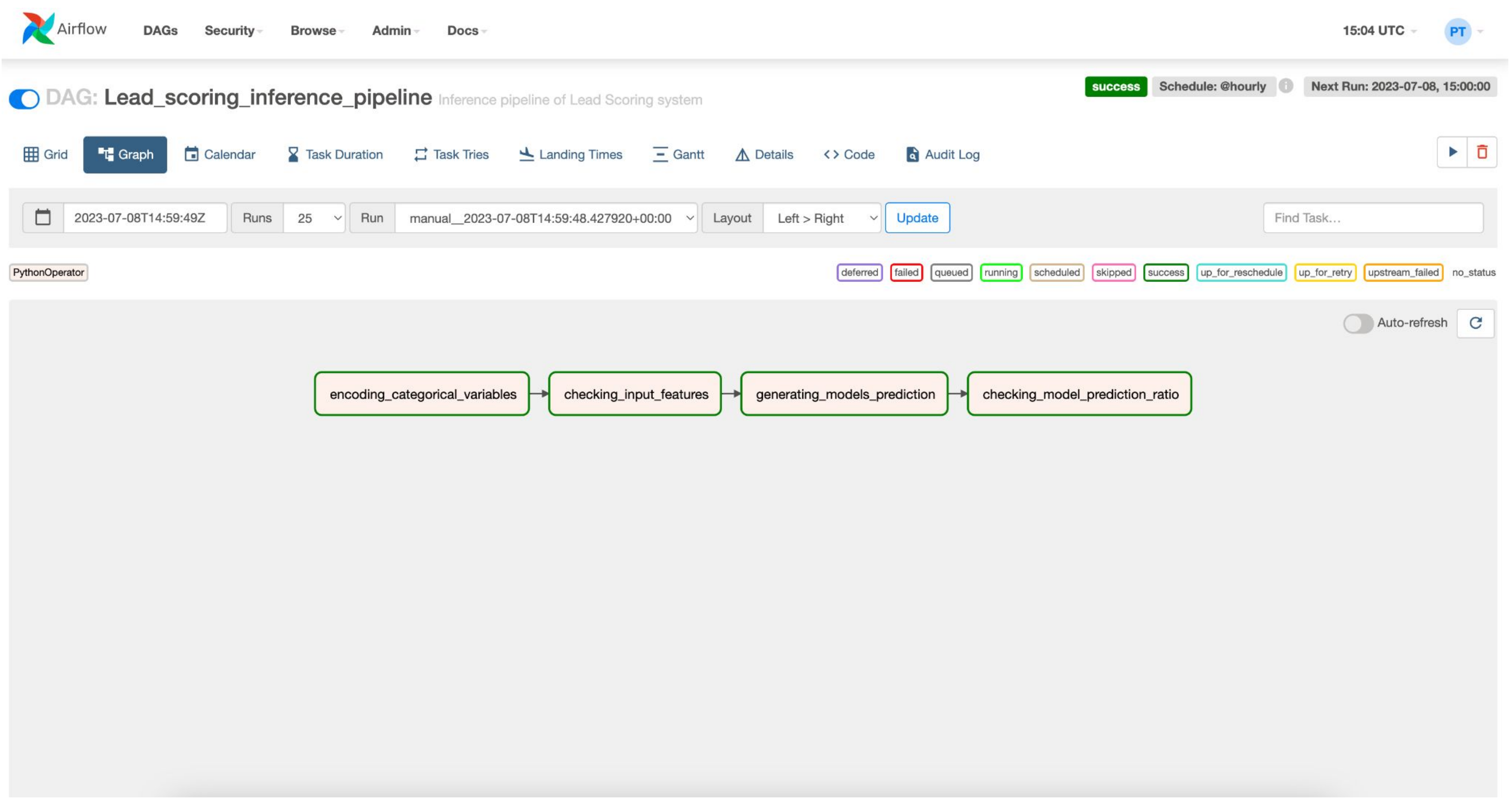
All Run States

encoding_categorical_variables	🟩	🟩	🟩
checking_input_features	🟩	🟩	🟩
generating_models_prediction	🟥	🟩	🟩
checking_model_prediction_ratio	🟧	🟩	🟩

DAG

Lead_scoring_inference_pipeline

DAG Runs Summary	
Total Runs Displayed	3
🟩 Total success	2
🟥 Total failed	1
First Run Start	2023-07-08, 14:23:06 UTC
Last Run Start	2023-07-08, 14:59:49 UTC
Max Run Duration	00:00:33
Mean Run Duration	00:00:31
Min Run Duration	00:00:26
DAG Summary	
Total Tasks	4
PythonOperators	4

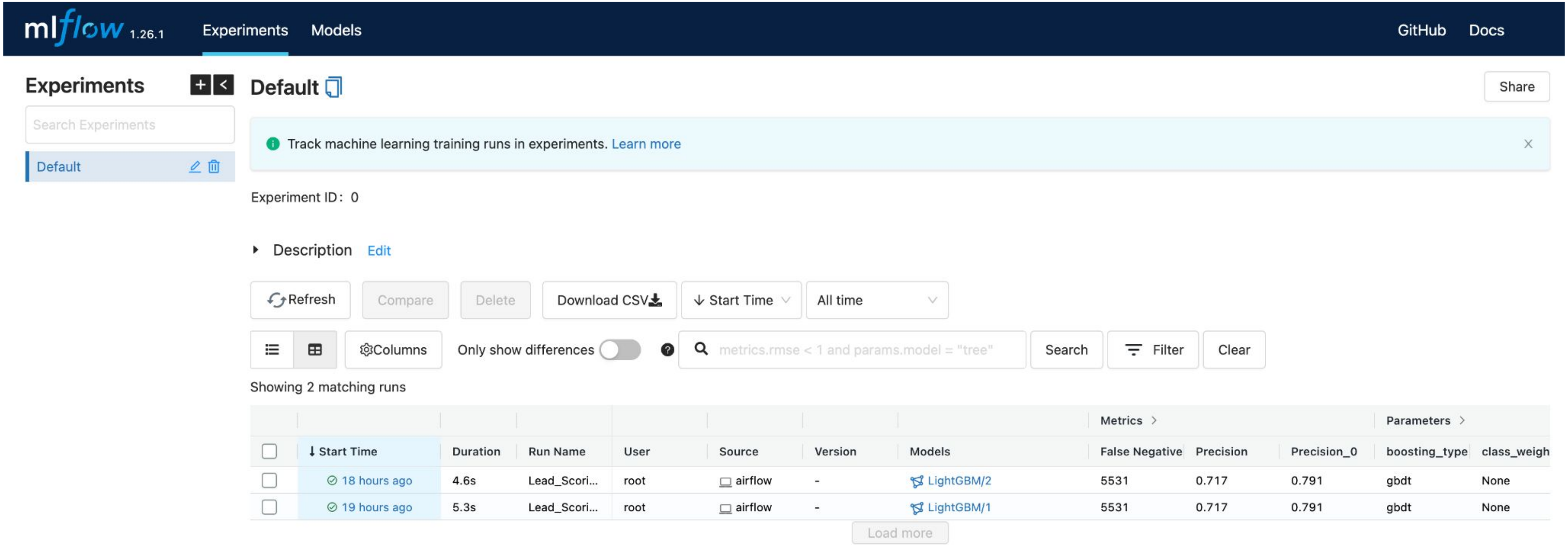


DAG: Lead_scoring_inference_pipeline Inference pipeline of Lead Scoring system

success Schedule: @hourly Next Run: 2023-07-08, 15:00:00

2023-07-08T14:59:49Z Runs 25 Run manual__2023-07-08T14:59:48.427920+00:00 Layout Left > Right

encoding_categorical_variables → checking_input_features → generating_models_prediction → checking_model_prediction_ratio



Default

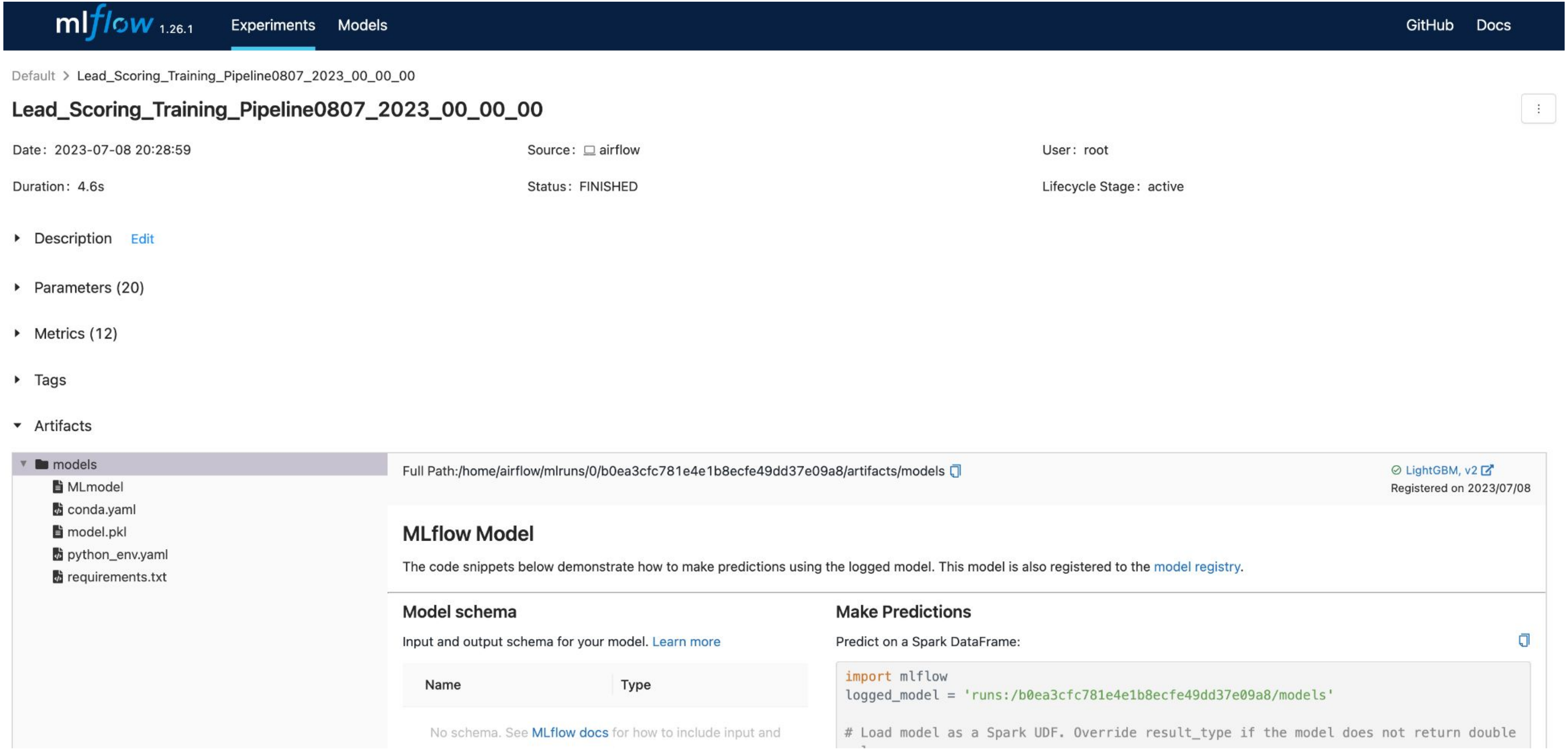
Track machine learning training runs in experiments. Learn more

Experiment ID: 0

Description Edit

Showing 2 matching runs

								Metrics			Parameters	
□	↓ Start Time	Duration	Run Name	User	Source	Version	Models	False Negative	Precision	Precision_0	boosting_type	class_weigh
□	18 hours ago	4.6s	Lead_Scori...	root	airflow	-	LightGBM/2	5531	0.717	0.791	gbdt	None
□	19 hours ago	5.3s	Lead_Scori...	root	airflow	-	LightGBM/1	5531	0.717	0.791	gbdt	None



Lead_Scoring_Training_Pipeline0807_2023_00_00_00

Date: 2023-07-08 20:28:59

Source: airflow

User: root

Duration: 4.6s

Status: FINISHED

Lifecycle Stage: active

- models
  - MLmodel
  - conda.yaml
  - model.pkl
  - python_env.yaml
  - requirements.txt

Full Path:/home/airflow/mlruns/0/b0ea3cfc781e4e1b8ecfe49dd37e09a8/artifacts/models

LightGBM, v2
Registered on 2023/07/08

MLflow Model

The code snippets below demonstrate how to make predictions using the logged model. This model is also registered to the model registry.

Model schema

Input and output schema for your model. Learn more

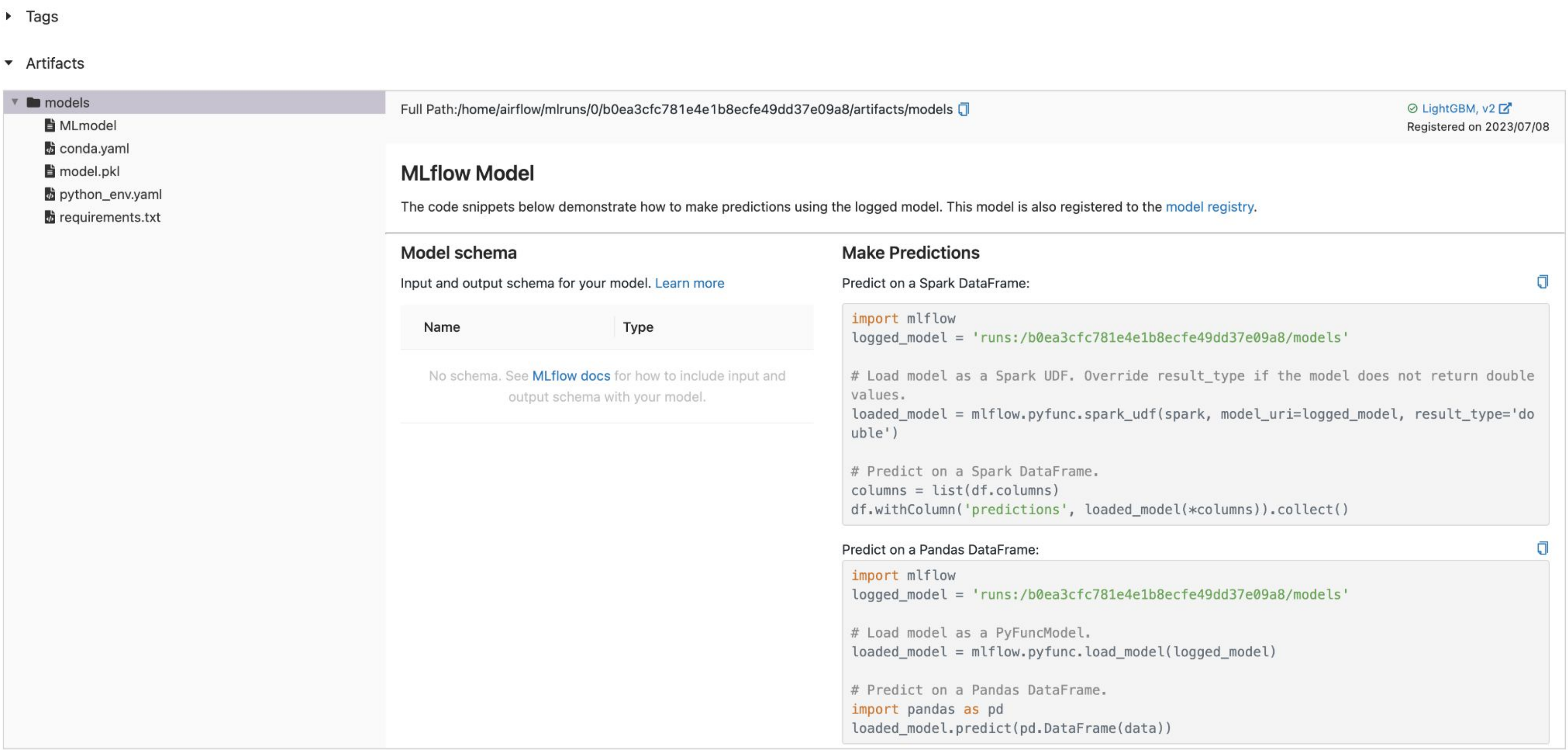
Name	Type

No schema. See MLflow docs for how to include input and

Make Predictions

Predict on a Spark DataFrame:

```
import mlflow
logged_model = 'runs:/b0ea3cfc781e4e1b8ecfe49dd37e09a8/models'

# Load model as a Spark UDF. Override result_type if the model does not return double
```



▸ Tags

▾ Artifacts

- ▾ models
  - MLmodel
  - conda.yaml
  - model.pkl
  - python_env.yaml
  - requirements.txt

Full Path:/home/airflow/mlruns/0/b0ea3cfc781e4e1b8ecfe49dd37e09a8/artifacts/models

LightGBM, v2

Registered on 2023/07/08

MLflow Model

The code snippets below demonstrate how to make predictions using the logged model. This model is also registered to the model registry.

Model schema

Input and output schema for your model. Learn more

Name	Type

No schema. See MLflow docs for how to include input and output schema with your model.

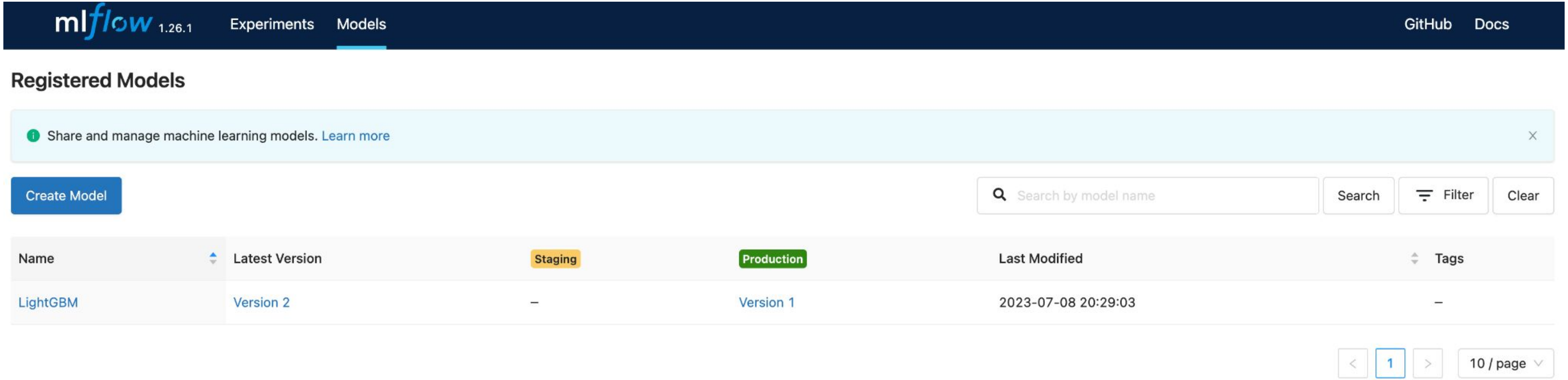
Make Predictions

Predict on a Spark DataFrame:

```
import mlflow
logged_model = 'runs:/b0ea3cfc781e4e1b8ecfe49dd37e09a8/models'

# Load model as a Spark UDF. Override result_type if the model does not return double values.
loaded_model = mlflow.pyfunc.spark_udf(spark, model_uri=logged_model, result_type='double')

# Predict on a Spark DataFrame.
columns = list(df.columns)
df.withColumn('predictions', loaded_model(*columns)).collect()
```

Predict on a Pandas DataFrame:

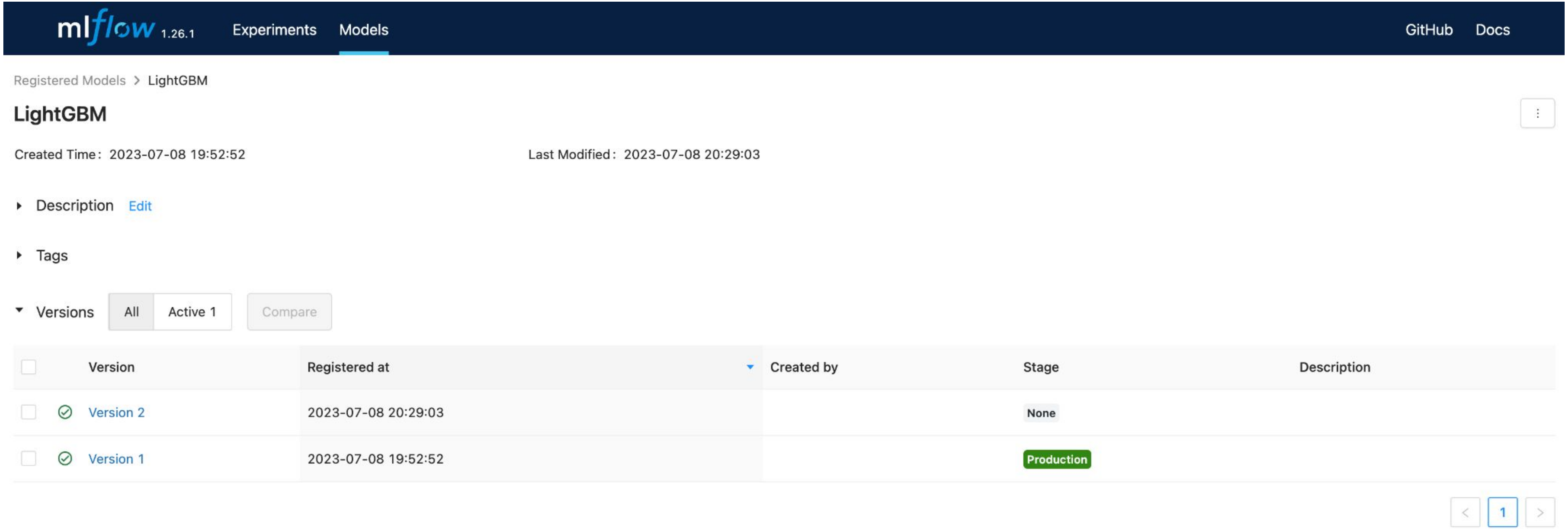
```
import mlflow
logged_model = 'runs:/b0ea3cfc781e4e1b8ecfe49dd37e09a8/models'

# Load model as a PyFuncModel.
loaded_model = mlflow.pyfunc.load_model(logged_model)

# Predict on a Pandas DataFrame.
import pandas as pd
loaded_model.predict(pd.DataFrame(data))
```



Registered Models

Share and manage machine learning models. Learn more

Name	Latest Version	Staging	Production	Last Modified	Tags
LightGBM	Version 2	–	Version 1	2023-07-08 20:29:03	–



LightGBM

Created Time: 2023-07-08 19:52:52

Last Modified: 2023-07-08 20:29:03

□	Version	Registered at	Created by	Stage	Description
□	✓ Version 2	2023-07-08 20:29:03		None	
□	✓ Version 1	2023-07-08 19:52:52		Production	

Thank you