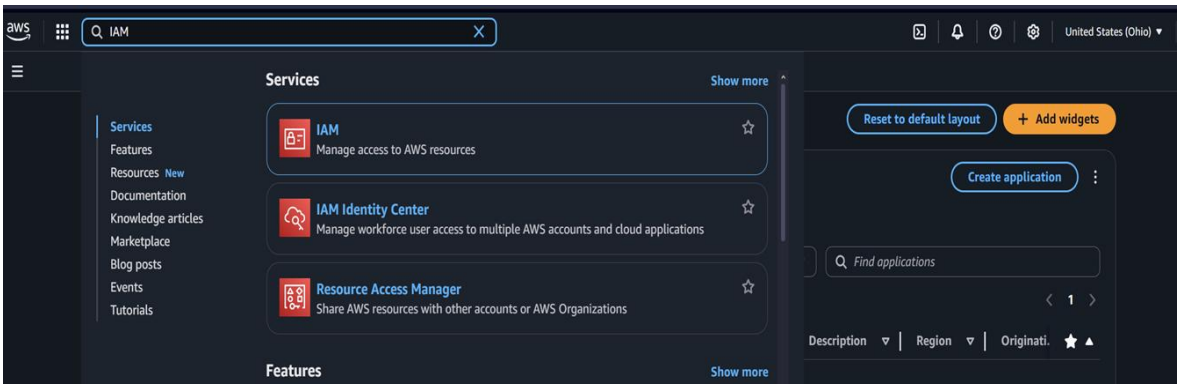
PROJECT – END TO END ML BUSINESS PIPELINE

Create an AWS Account and set up IAM Roles.

Go to https://aws.amazon.com and sign up.

Set up IAM Roles



The IAM Roles helps in granting permission to entities that interact with AWS resources. Before creating the roles, a user profile must be created, with access key and secret access key. Policies are created for each user or group.

Install AWS-CLI on windows or Linux to interact with AWS services using commands in your command-line shell. This can be done using: https://docs.aws.amazon.com/cli/latest/userguide/getting-started-install.html

After downloading, open the CMD or PowerShell:

aws onfigure

and enter the access key ID, secret access key, region and output format.

S3 bucket was created using

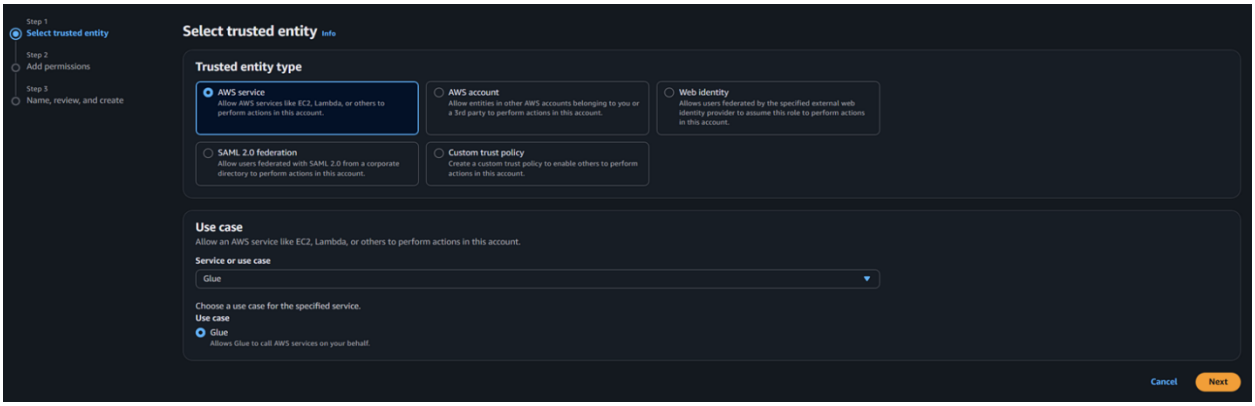
Aws s3 mb s3://folder-name

Extra folders were created: raw, scripts, temp, production, transformation, athena-results.

Some Python packages are installed using CMD.

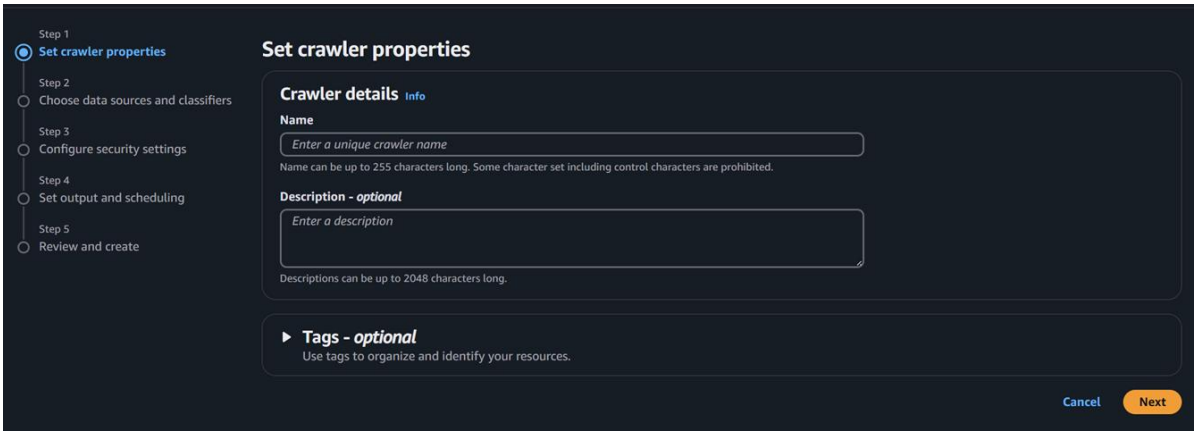
The python files ran in the CMD: ingest_fmp.py and ingest_mysql.py is ran and the data saved at their respective folders in the S3 bucket.

An IAM Role is created for Glue which is used to manage data integration. It is used to automate modern data pipelines with built in ETL processes.

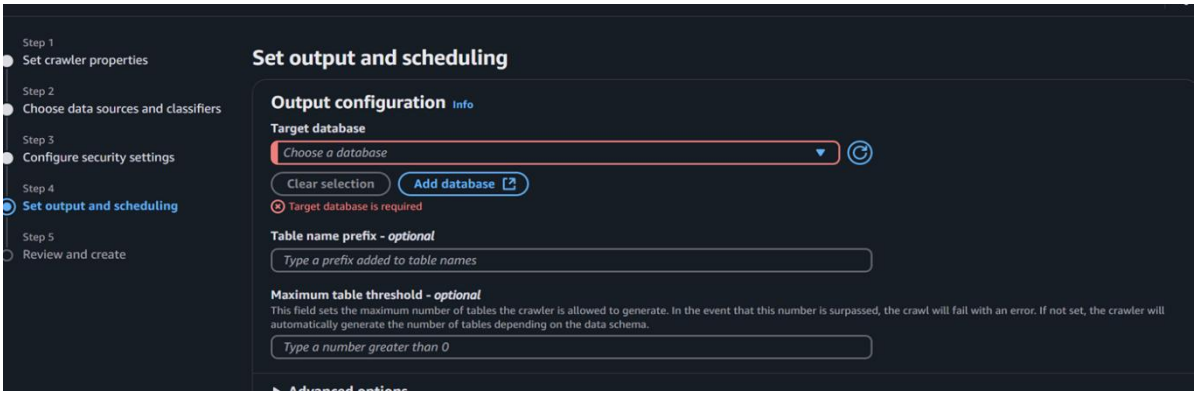


The role is given AmazonS3FullAccess and AWSGlueServices permissioons.

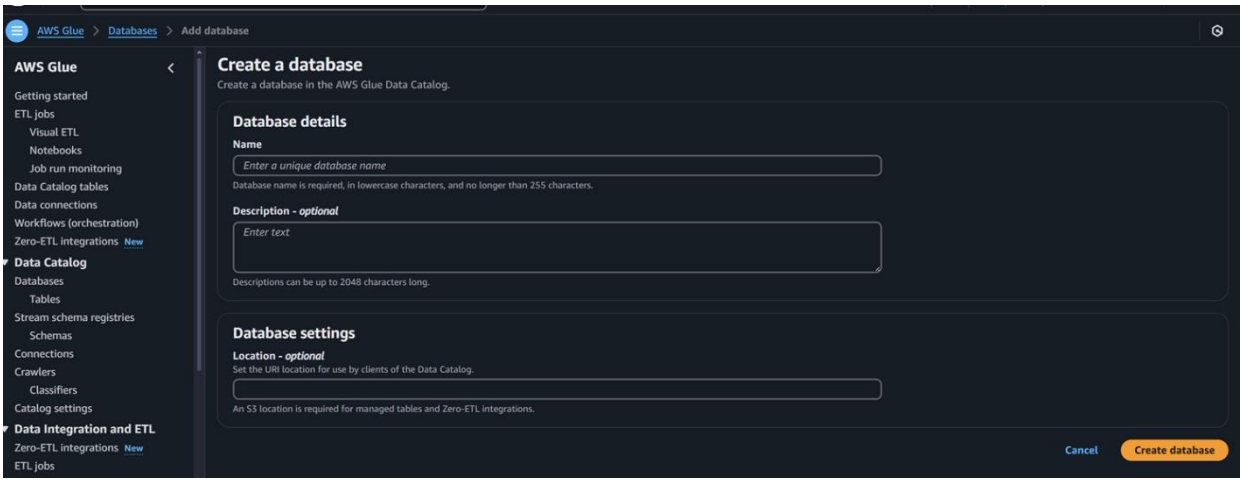
After that a Crawler and a database are created.



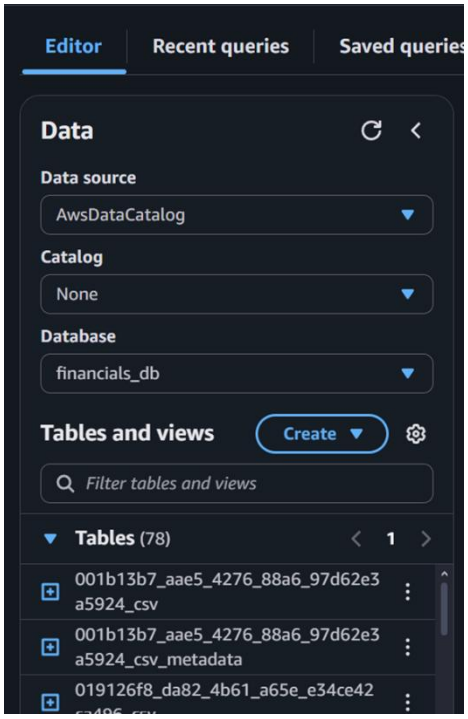
The database can be created using the Crawler or with Athena



Clicking the add database goes straight to creating a database



OR



In Athena, go to the Athena editor and type

Create DATABASE

Files: transform_financial_data.py, join_financial_summary.py and data_wrangler.py are run via glue jobs which normalizes the raw fmp data saving it as a parquete file and joins transformed income statements of stock prices. The data_wrangler preprocesses and enriches the income data by adding new KPIs. The data is saved as a csv file, which is used by PowerBI for visualization and for modelling using machine learning.

Connect to Power BI to Athena

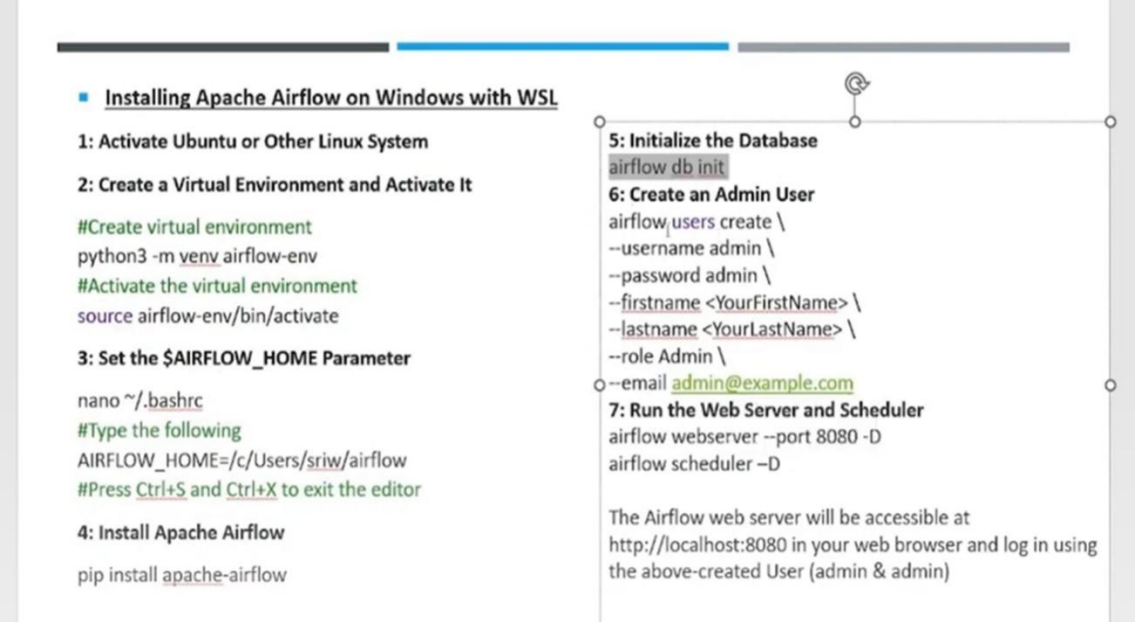
Athena is connected to PowerBI using odbc driver. The KPIs created were used to create visualizations.

company	year	revenue_old	netincome_old	netprofitmargin	grossprofit	revenue	netincome	costandexpenses	costofgoodsandservices	netproft
NFLX	2023	$567.81	$56.54	99.58	$55,565,114.00	$56,781,000	$56,540,000	$2,026,476.67	$1,215,886.00	$45,382,572.
TSLA	2023	$305.68	$78.39	256.44	$30,127,208.50	$30,568,000	$78,390,000	$734,652.50	$440,791.50	$24,941,191.
NVDA	2023	$235.02	$40.79	173.56	$23,298,867.51	$23,502,000	$40,790,000	$338,554.14	$203,132.49	$19,491,129.
AMZN	2023	$318.72	$76.35	239.55	$31,468,709.19	$31,872,000	$76,350,000	$672,151.36	$403,290.81	$26,123,766.
NVDA	2023	$296.5	$16.1	54.30	$29,468,966.11	$29,650,000	$16,100,000	$301,723.16	$181,033.89	$24,749,512.
NVDA	2023	$292.26	$28.08	96.08	$28,867,250.08	$29,226,000	$28,080,000	$597,916.54	$358,749.92	$24,018,215.
GOOG	2023	$320.46	$62.25	194.25	$31,172,338.91	$32,046,000	$62,250,000	$1,456,101.82	$873,661.09	$25,204,114.
NFLX	2023	$429.15	$71.87	167.47	$42,664,467.57	$42,915,000	$71,870,000	$417,554.05	$250,532.43	$35,905,503.
TSLA	2023	$118.39	$26.3	222.15	$11,529,260.14	$11,839,000	$26,300,000	$516,233.10	$309,739.86	$9,344,164.
BABA	2023	$553.84	$52.82	95.37	$55,125,859.59	$55,384,000	$52,820,000	$430,234.01	$258,140.41	$46,407,515.
ORCL	2023	$399.13	$69.02	172.93	$39,391,463.75	$39,913,000	$69,020,000	$869,227.08	$521,536.25	$32,684,258.
GOOG	2023	$241.38	$32.18	133.32	$23,563,472.09	$24,138,000	$32,180,000	$957,546.52	$574,527.91	$19,211,975.
TSLA	2023	$497.46	$93.29	187.53	$49,156,901.26	$49,746,000	$93,290,000	$981,831.23	$589,098.74	$40,867,999.
NVDA	2023	$467.99	$47.22	100.90	$45,448,044.56	$46,799,000	$47,220,000	$2,251,592.41	$1,350,955.44	$36,638,070.
GOOG	2023	$570.88	$120.99	211.94	$56,427,075.26	$57,088,000	$120,990,000	$1,101,541.23	$660,924.74	$46,999,308.
MSFT	2023	$562.96	$69.33	132.96	$55,080,878.21	$56,296,000	$69,330,000	$2,025,202.99	$1,215,121.79	$44,972,705.

See the PowerBI Business Report file.

Airflow and Automation

An environment was created for the airflow and the ran on the webserver

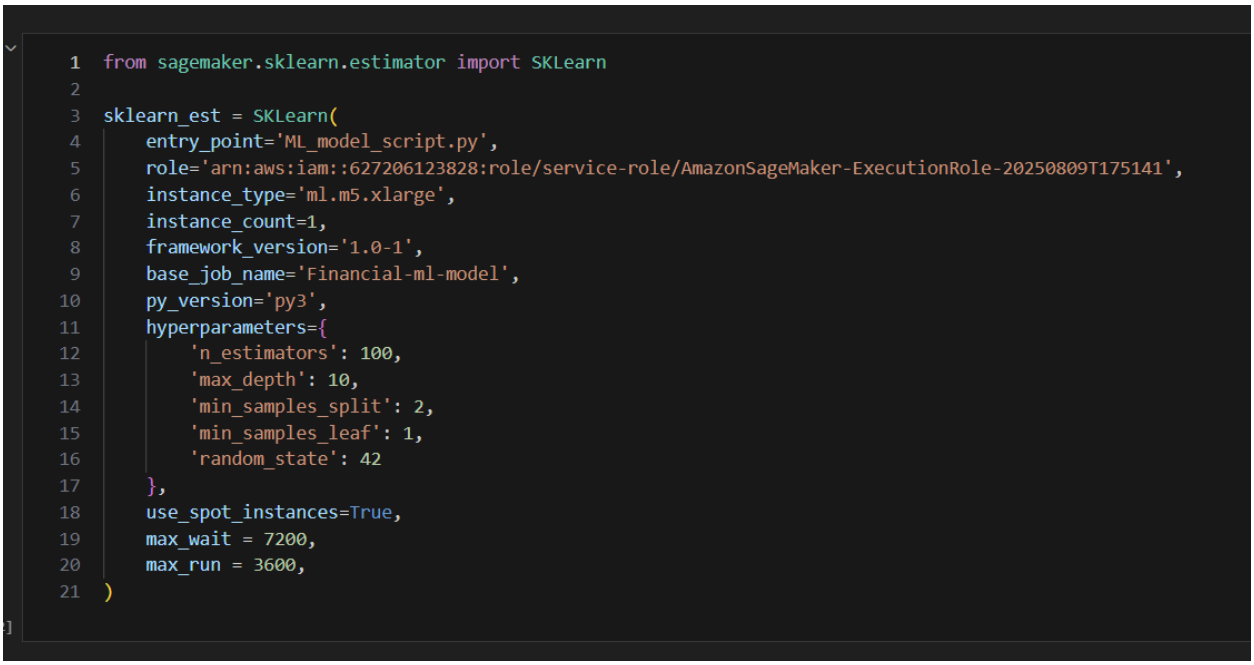


A folder called dag is created, which has files: api_ingestion_dag.py, financial_pipeline_dag.py

The dag files contain specific tasks that must scheduled.

ML Training

The data saved as a csv file is modelled using a machine learning technique. AWS Sagemaker was used for this.



```python
from sagemaker.sklearn.estimator import SKLearn

sklearn_est = SKLearn(
    entry_point='ML_model_script.py',
    role='arn:aws:iam::627206123828:role/service-role/AmazonSageMaker-ExecutionRole-20250809T175141',
    instance_type='ml.m5.xlarge',
    instance_count=1,
    framework_version='1.0-1',
    base_job_name='Financial-ml-model',
    py_version='py3',
    hyperparameters={
        'n_estimators': 100,
        'max_depth': 10,
        'min_samples_split': 2,
        'min_samples_leaf': 1,
        'random_state': 42
    },
    use_spot_instances=True,
    max_wait = 7200,
    max_run = 3600,
)
```