Chapter 5

Week 5

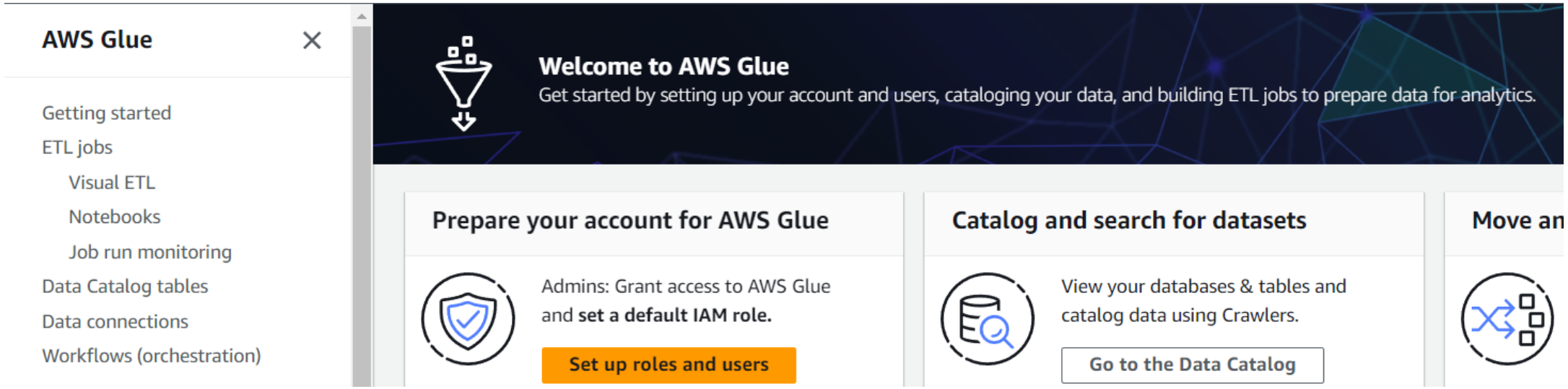
Data Engineer

Trainer: Balazs Balogh

CUBIX
INSTITUTE OF TECHNOLOGY

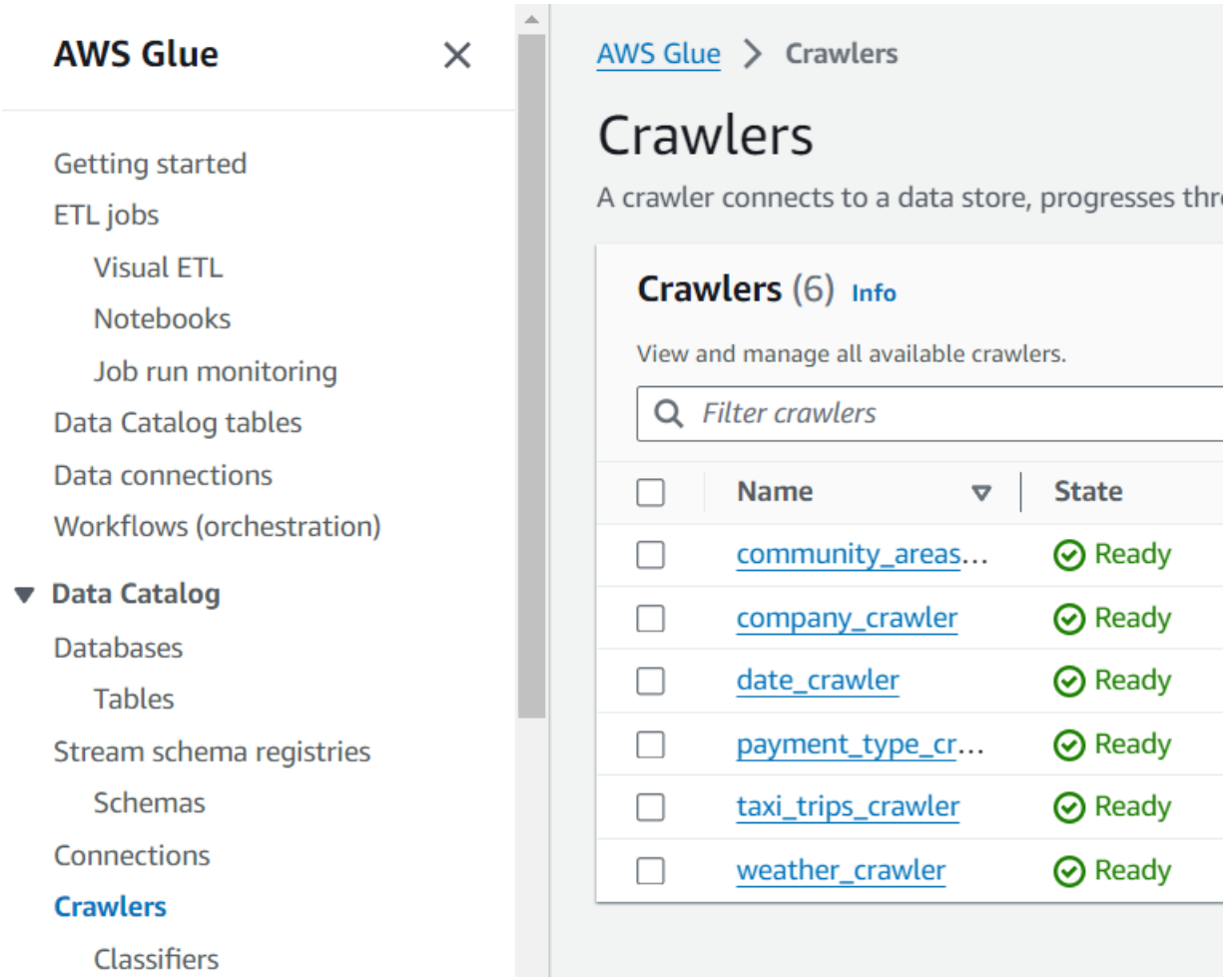
AWS Glue

- AWS Glue is a serverless data integration service that makes it easy for analytics users to discover, prepare, move, and integrate data from multiple sources.



AWS Glue

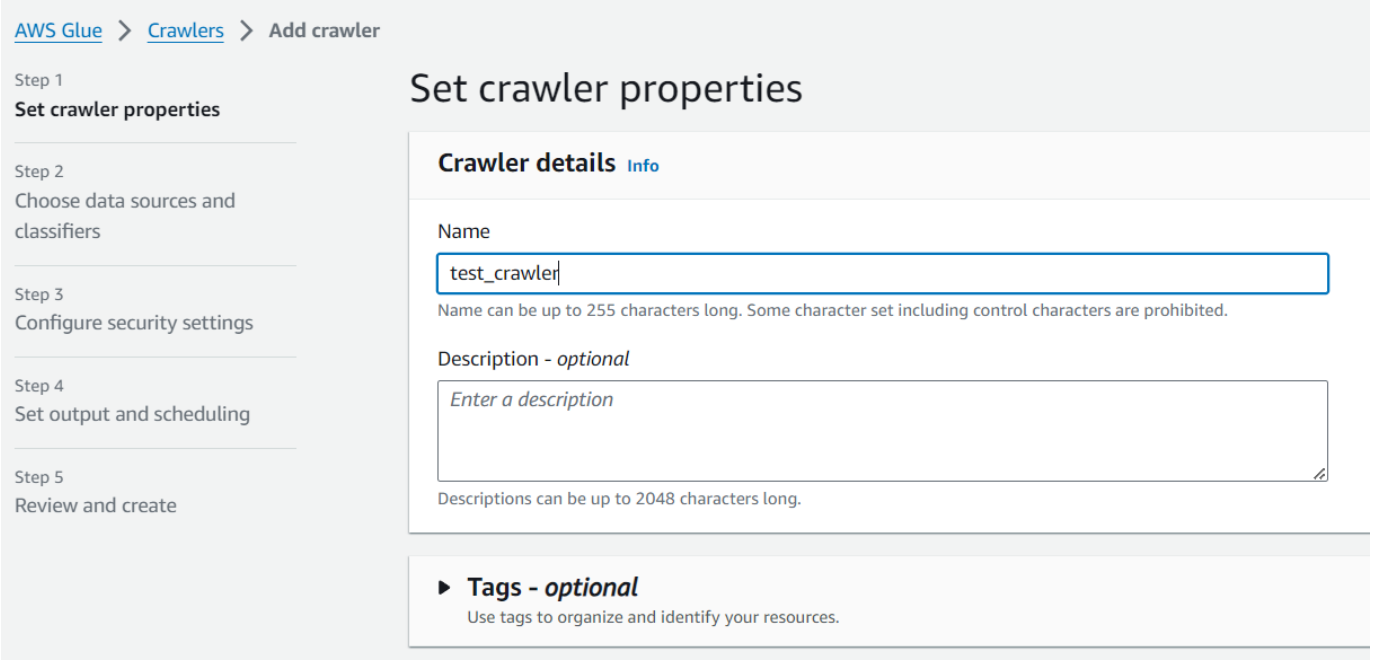
- We can create ETL jobs, Databases, Crawlers and so on.
- We use the databases and crawlers from the offerings.
- Crawlers are used to populate AWS Glue Data Catalog tables. It goes through the data, and infer it's schema, to make it queryable.



CUBIX
INSTITUTE OF TECHNOLOGY

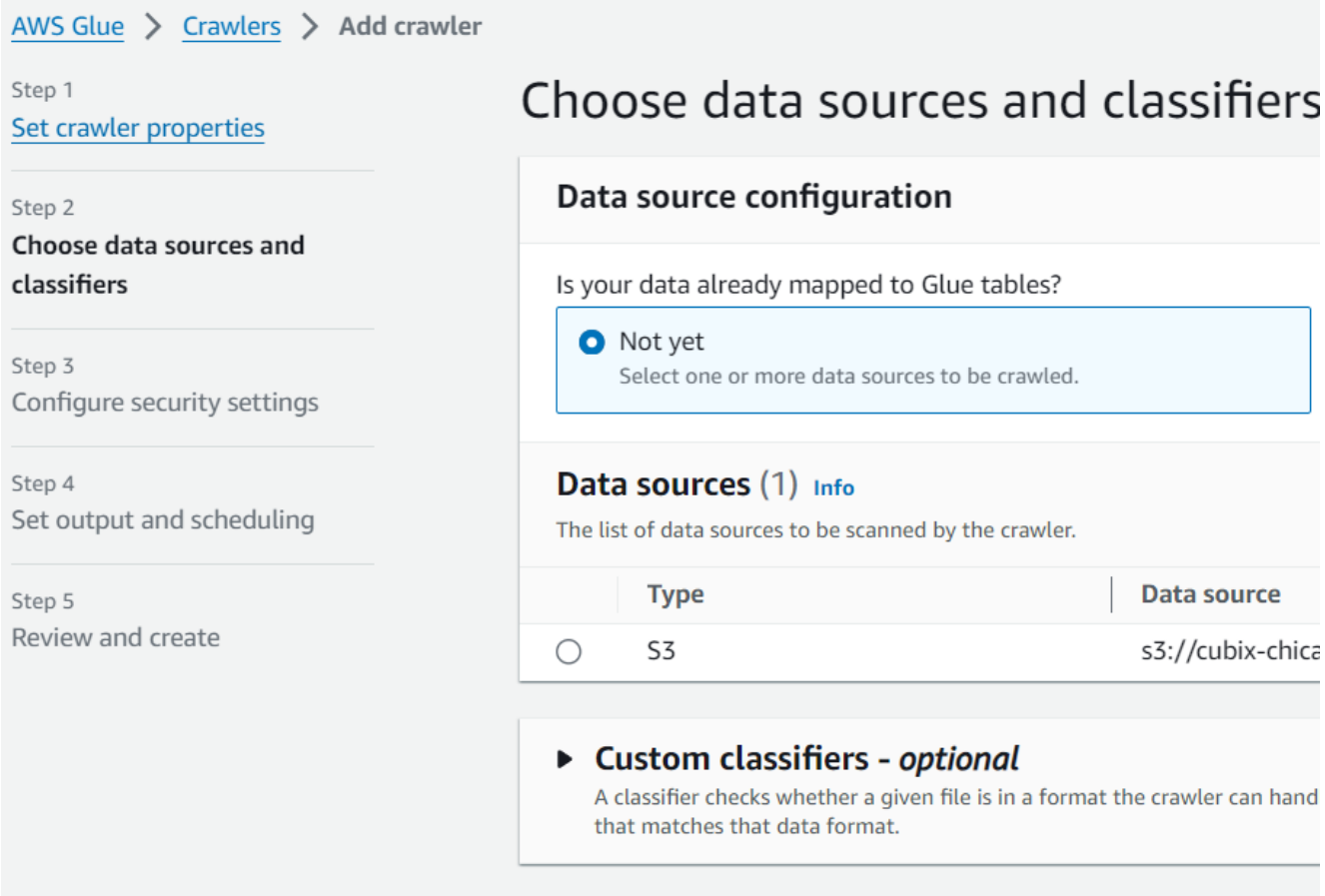
AWS Glue - Crawler

- Create a new crawler, and choose a name for it.



AWS Glue - Crawler

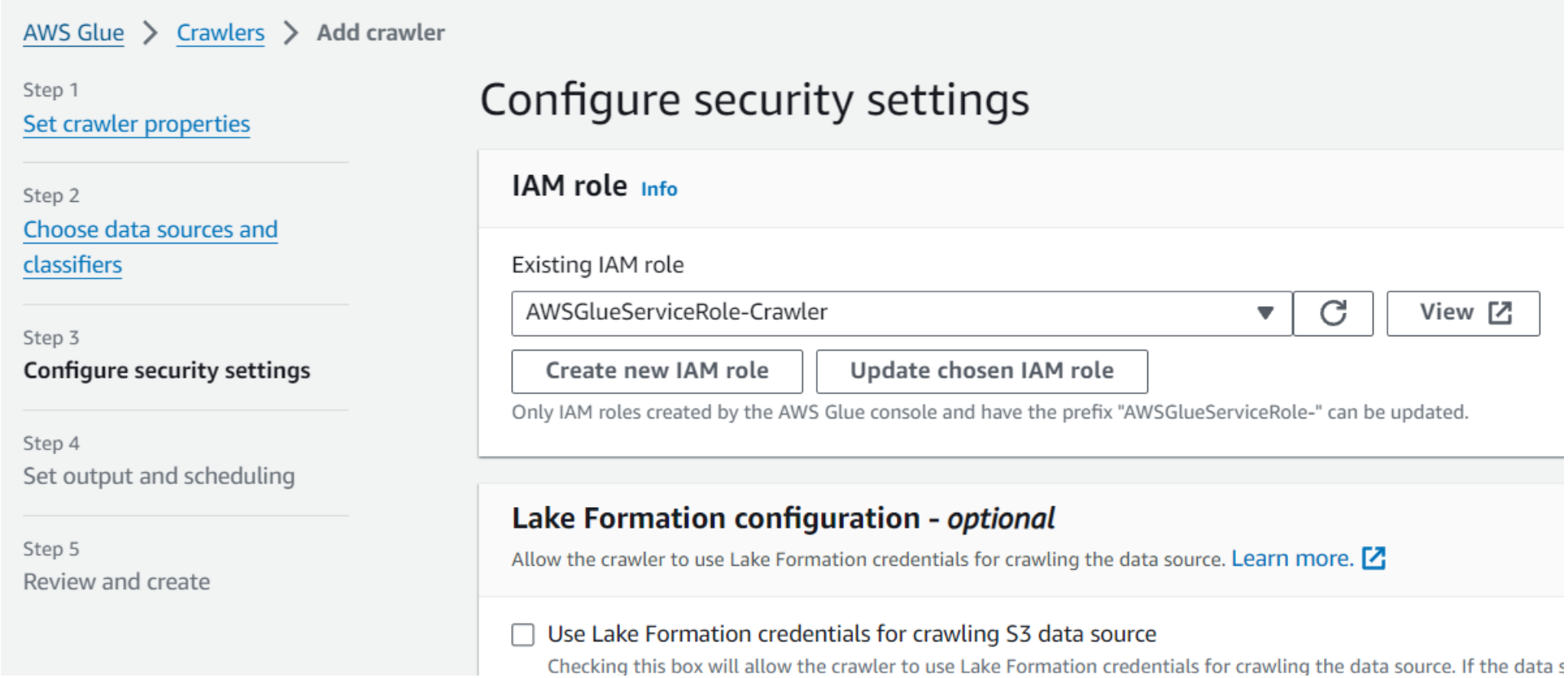
- On the next page, add your data source.



CUBIX
INSTITUTE OF TECHNOLOGY

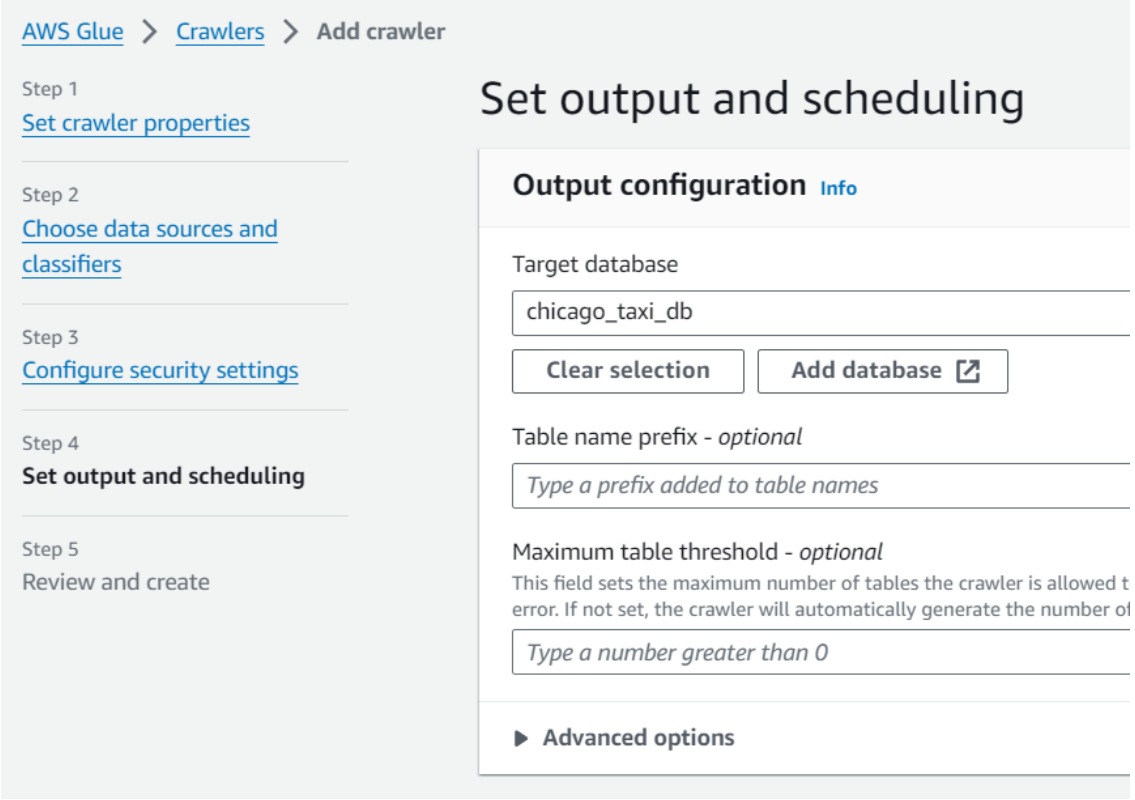
AWS Glue - Crawler

- Create an IAM Role which has S3FullAccess, or extend the default one.



AWS Glue - Crawler

- Choose the target database, if you don’t have one, create it here.
- Choose “On demand” for scheduling, you need this to run only once.



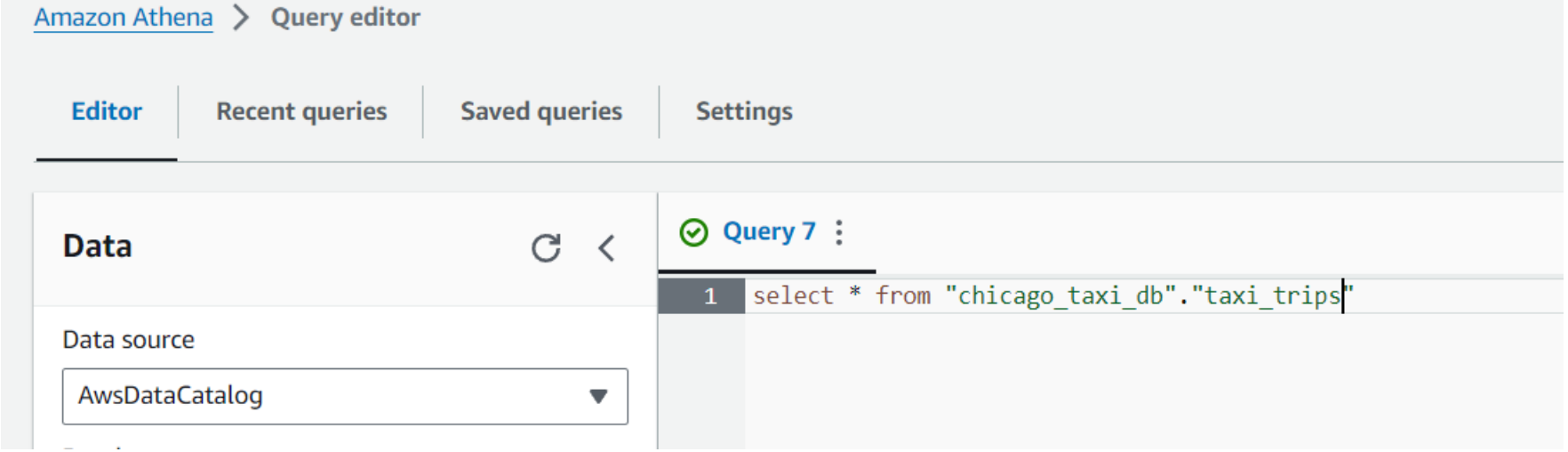
CUBIX
INSTITUTE OF TECHNOLOGY

AWS Glue - Crawler

- When it’s created, click on “Run crawler” to run it.
- It creates the new table under your database.

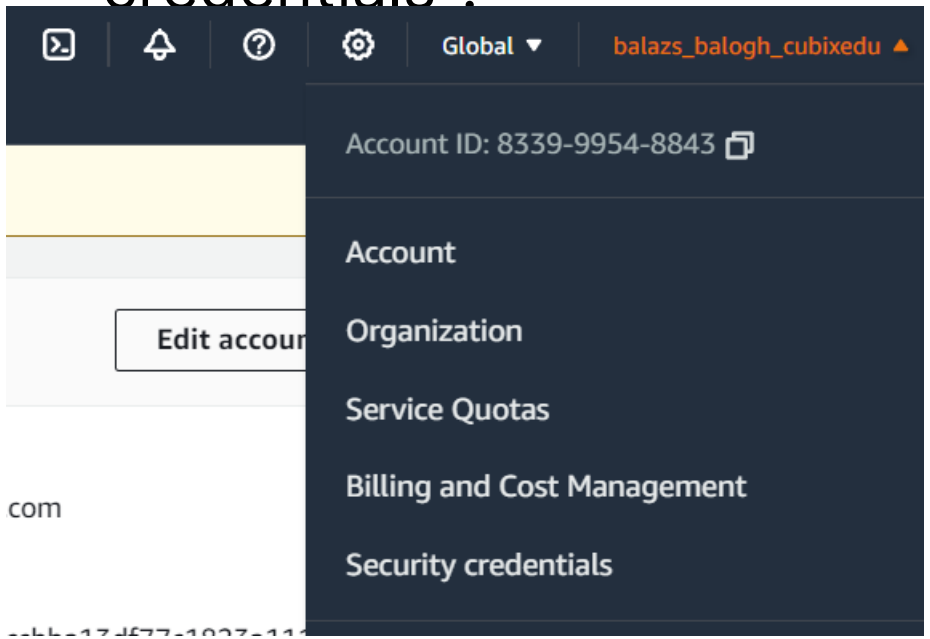
AWS Athena

- Amazon Athena is an interactive query service that makes it simple to analyze data directly in Amazon S3 using standard SQL.
- Athena is serverless, so there is no infrastructure to setup or manage, and you can choose to pay based on the queries you run or compute needed by your queries.



Local visualisations

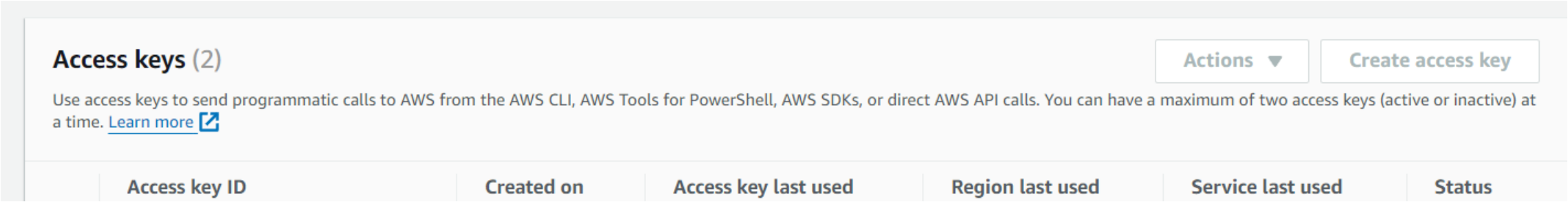
- For local visualisations, you need an ACCESS_KEY and SECRET_KEY from AWS, to be able to download data from your S3 bucket.
- In the top right corner, click on your account, choose “Security credentials”.



CUBIX
INSTITUTE OF TECHNOLOGY

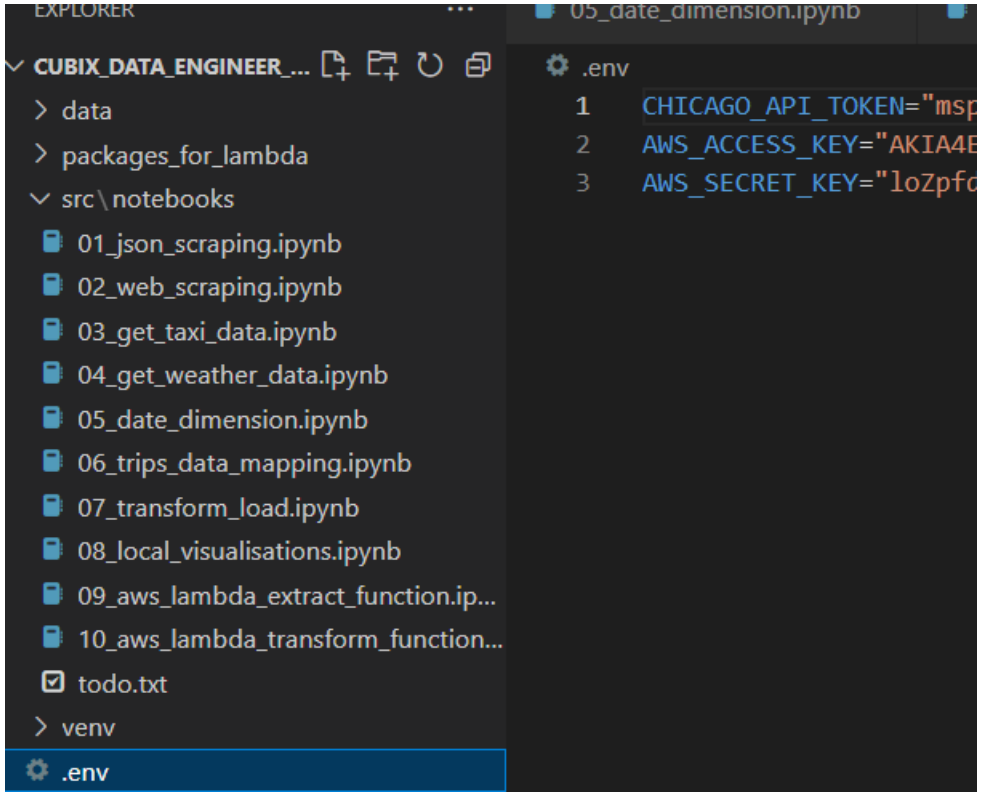
Local visualisations

- Under Access keys, click on “Create access key” and save your Access key ID, and Secret.



Local visualisations

- In your root folder create the .env file, if you don't have it already, and add the "AWS_ACCESS_KEY" and "AWS_SECRET_KEY" variables with the keys you just got. You can choose different names for them.



Local visualisations

- Now you can import the libraries, and use the two new environment variables.

```
[1]
from io import StringIO
import os

import boto3
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

pd.set_option("display.max_columns", 50)
```

```
[2]
aws_acces_key_id = os.getenv("AWS_ACCESS_KEY")
aws_secret_key = os.getenv("AWS_SECRET_KEY")
```

Local visualisations

- After reading the data, and joining all the tables together we have everything to make visualisations.
- This is not the usual task for a Data Engineer, it’s here just to show you, what happens with the data after you oraganized it.

```python
# 1 - Histogram of Trip Durations

sns.histplot(trips_full['trip_seconds'], bins=30, kde=False)
plt.title('Histogram of Trip Durations')
plt.xlabel('Trip Duration (seconds)')
plt.ylabel('Count')
plt.xlim(0, 5000)  # Adjust the x-axis limits for better visibility, clear the outliers
plt.show()
```