Name: Sai Krishna Pinugoti

AWS-GLUE

Job:

CSV------------Transform--------------Parquet

Terminologies:

Data Catalog:

Persistent metadata store in Glue

Crawler:

Program that connects to the data store which will determine the schema of the data using the classifier and writes the metadata in the AWS glue datacatalog

Datastore: Repo where the data is store permanently

Data source: A data store that is used as input to any process or transform

Data Target: A data store where the transformed data is written.

Steps:

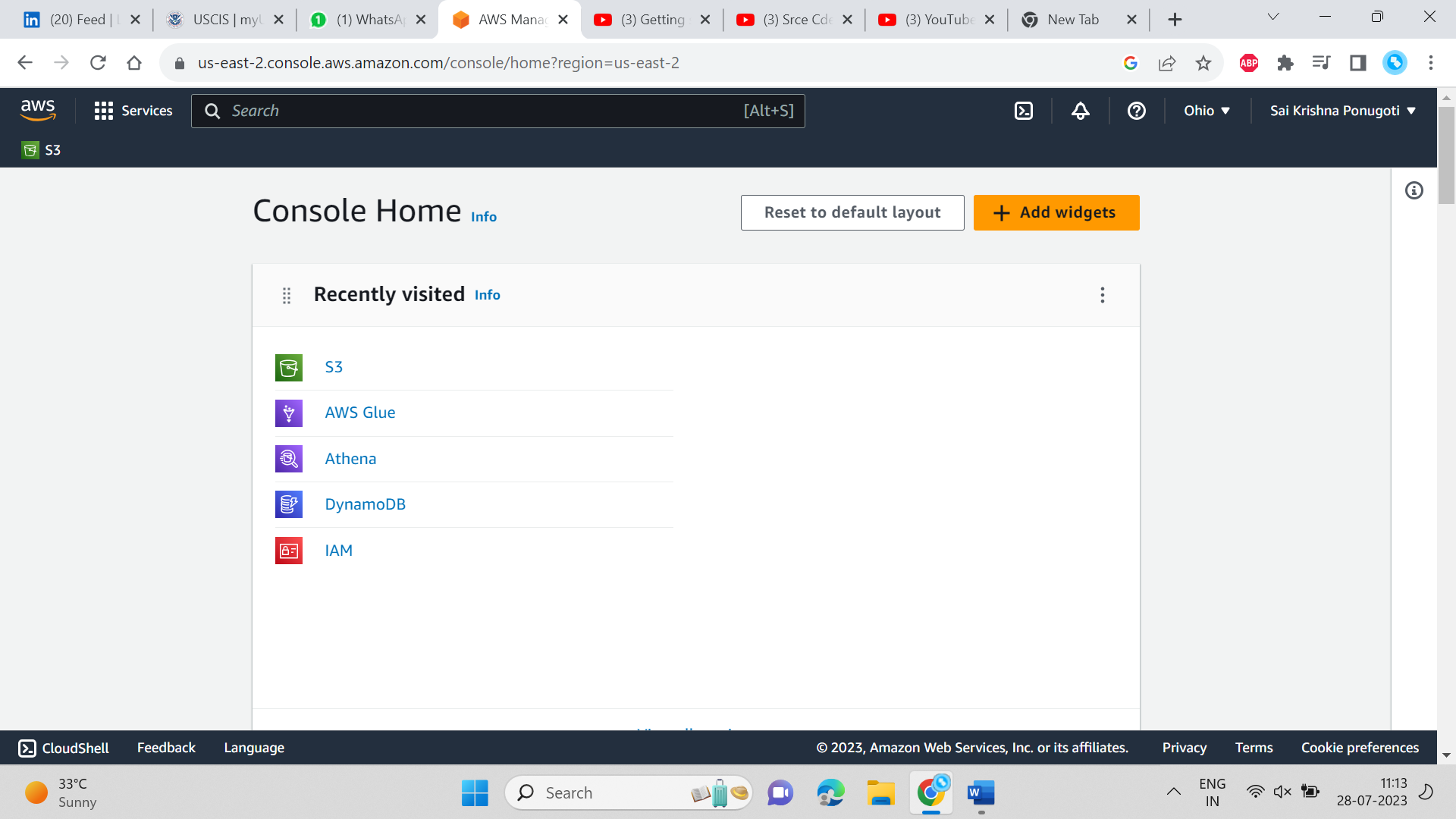
Create S3 Bucket (data source)

Create IAM role with appropriate permission for Aws Glue

Create and run crawler to create table in the database

Create ETL jib to transform csv to parquet.

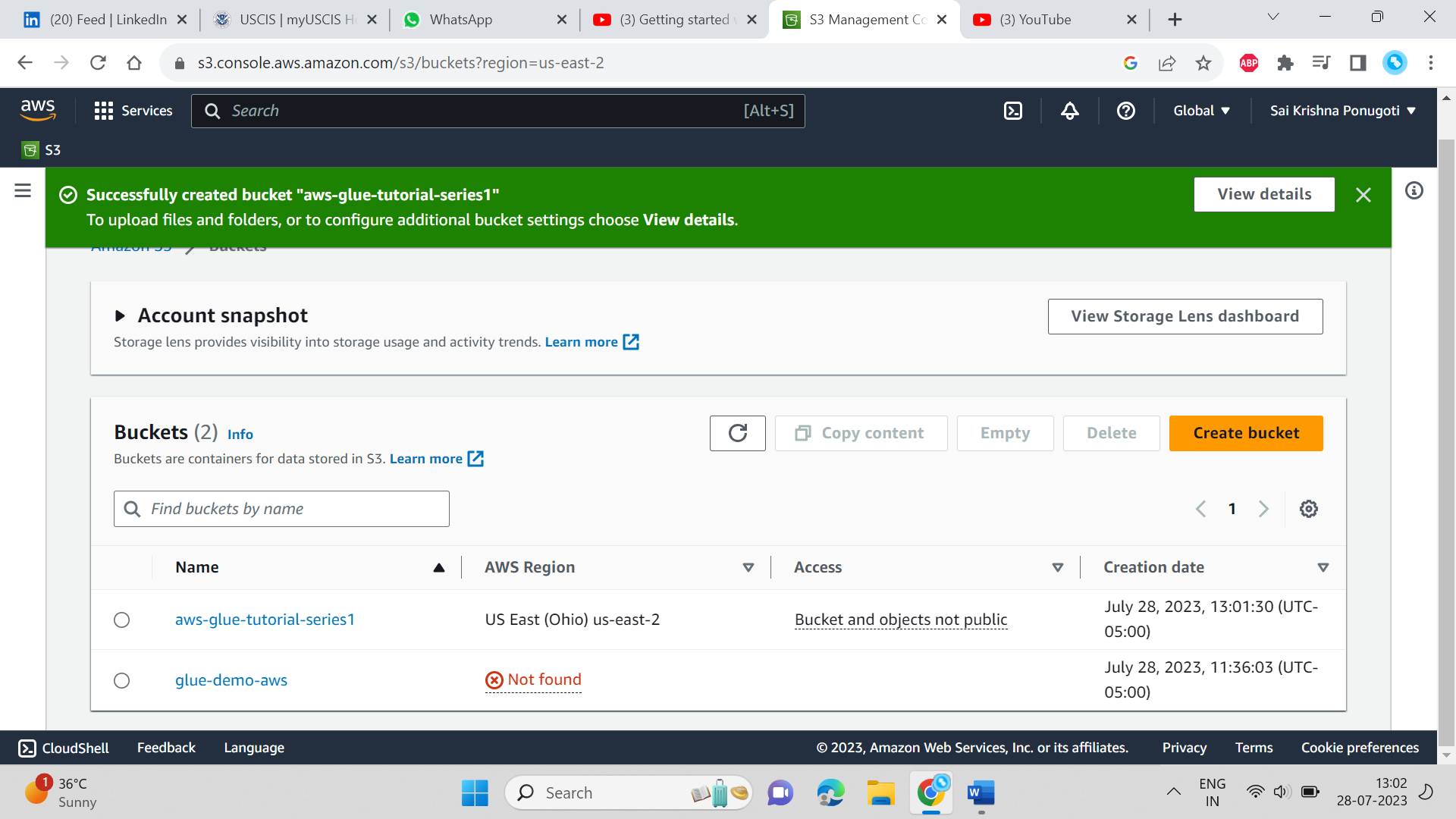
Login to AWS Console:



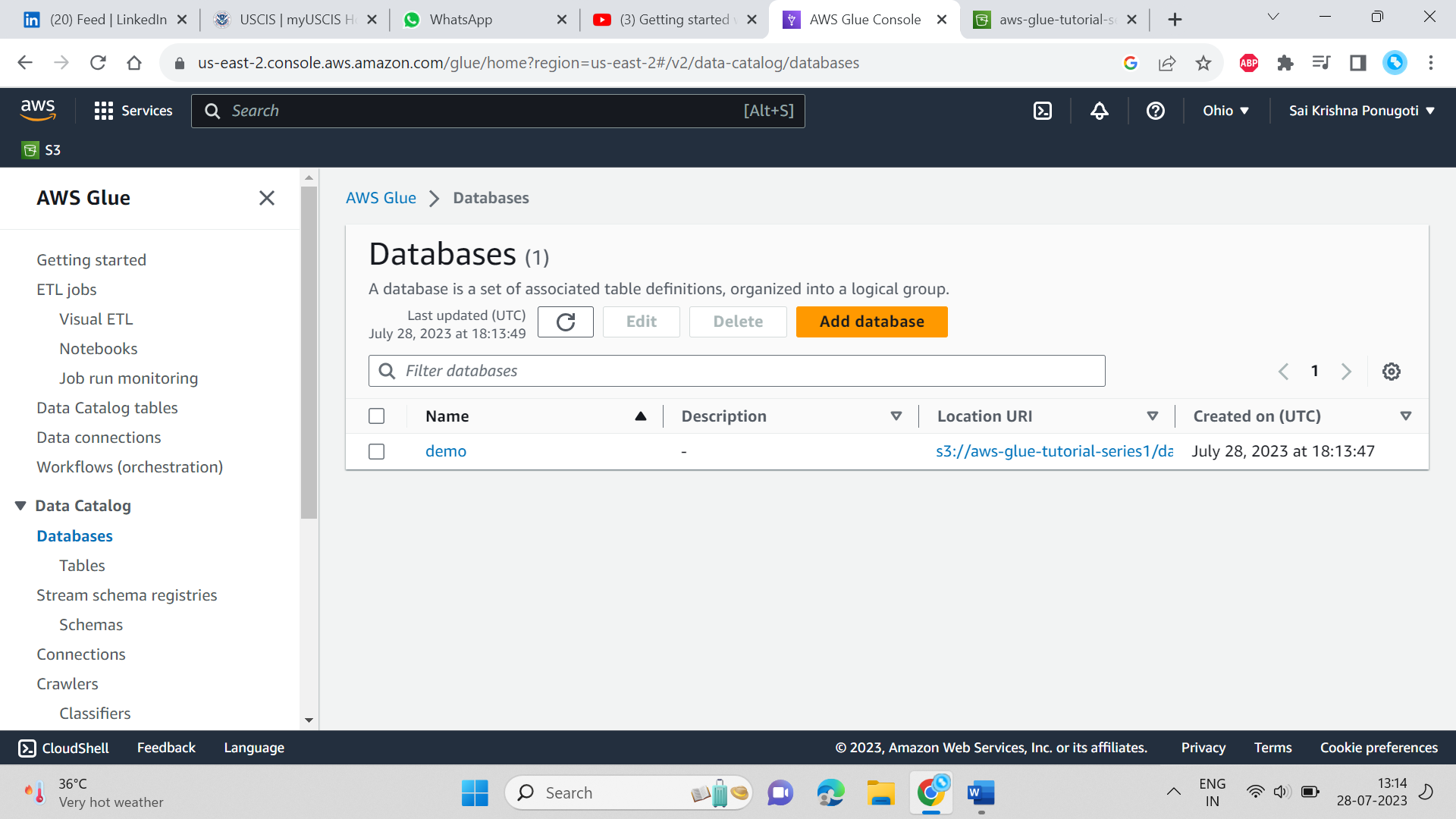
Creating an IAM role for the glue service to have access to this account

Role name: aws-glue-services-demo

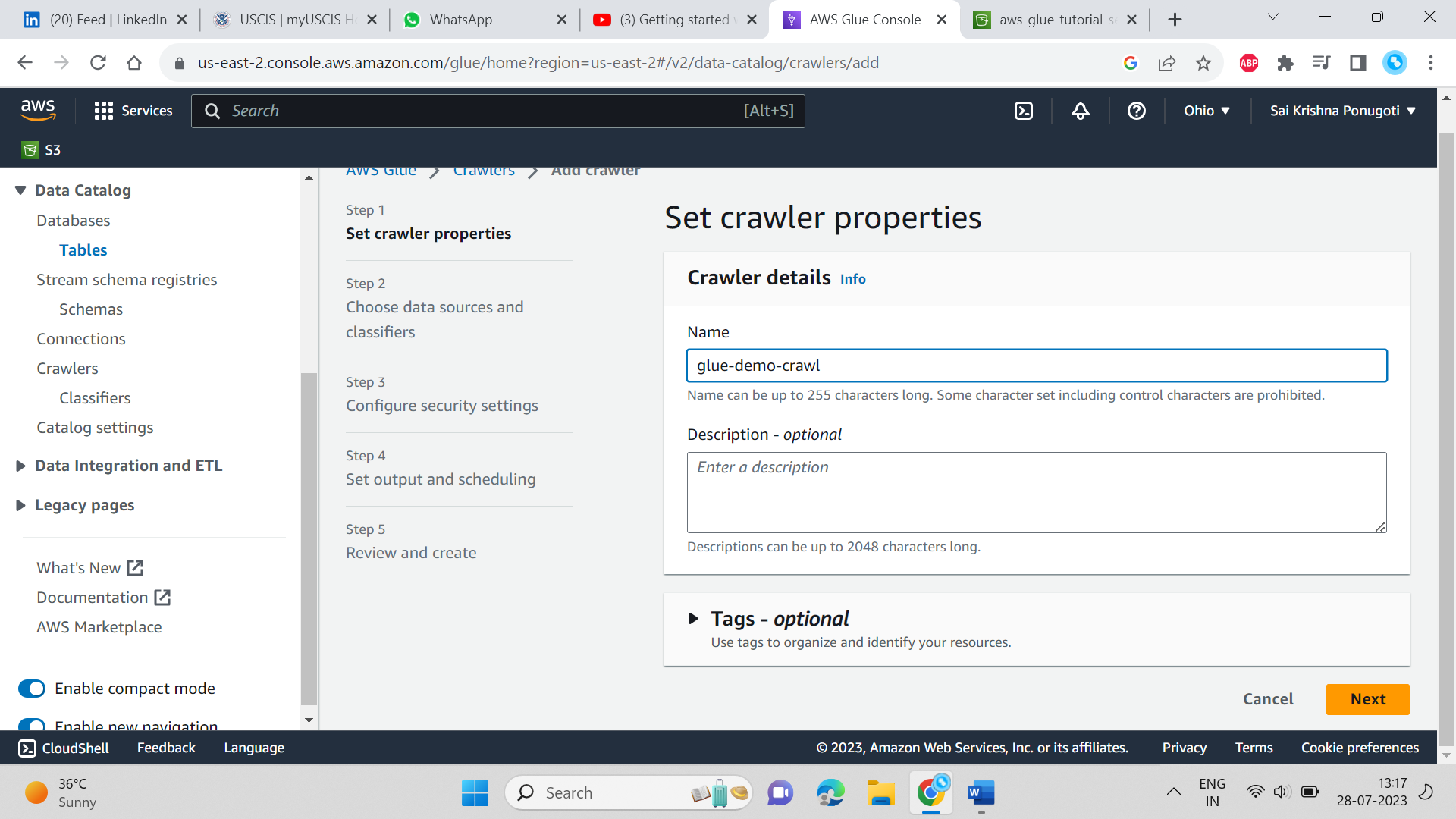
Creating a bucket with name aws-glue-tutorial-series1



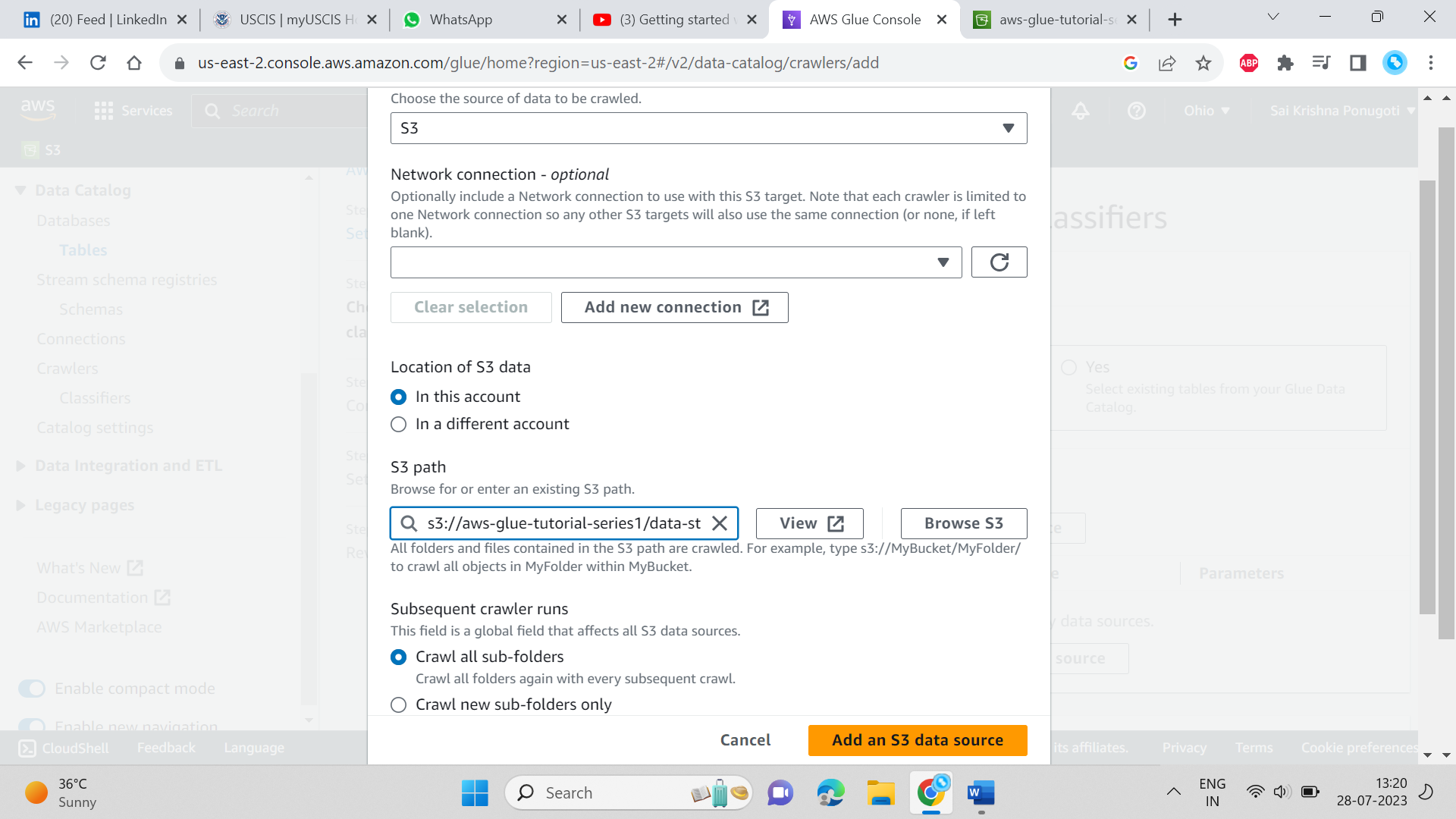
Creating Database with name and giving the location of file where it is located in s3



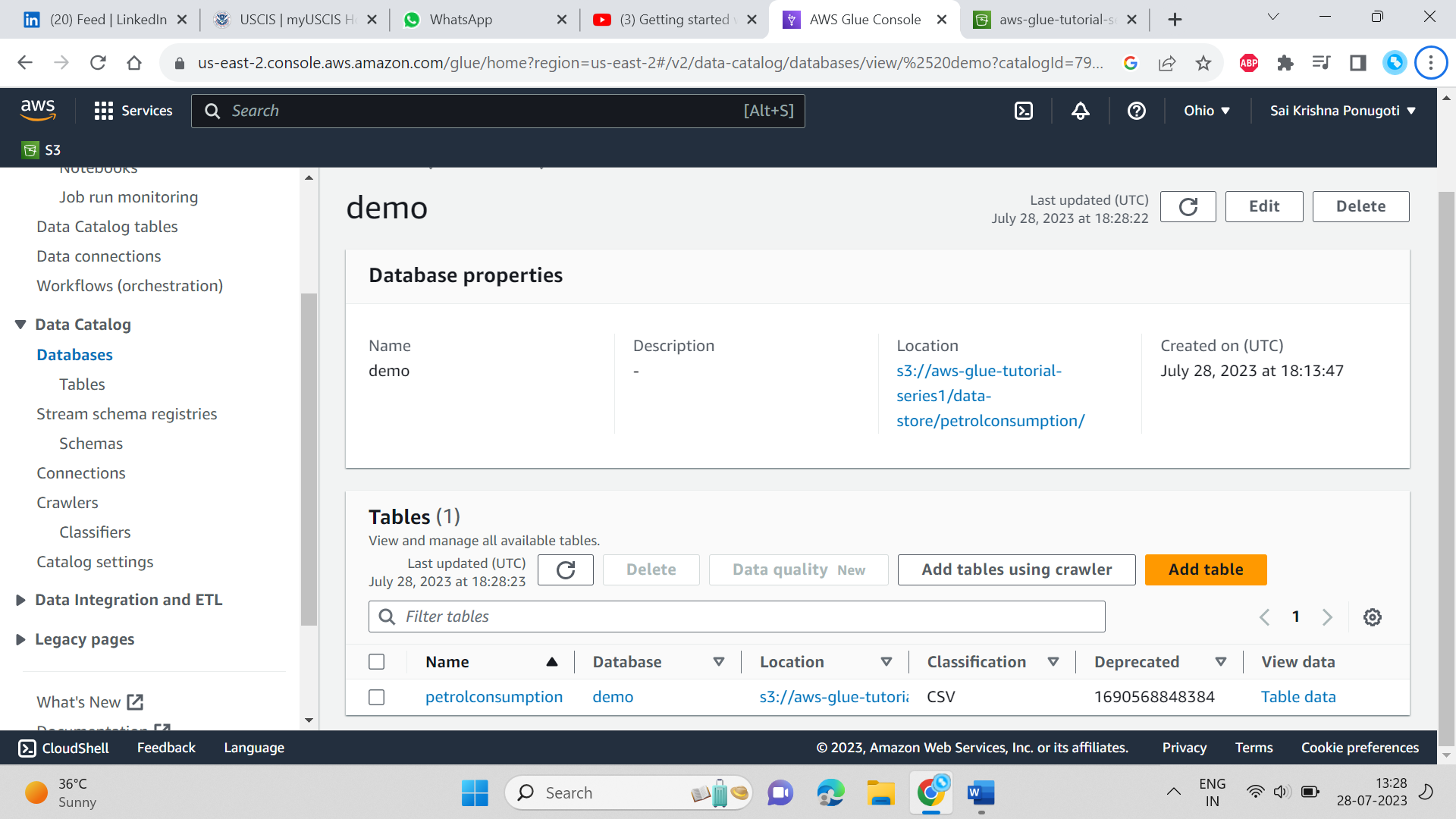
The next step is to create the tables using crawler



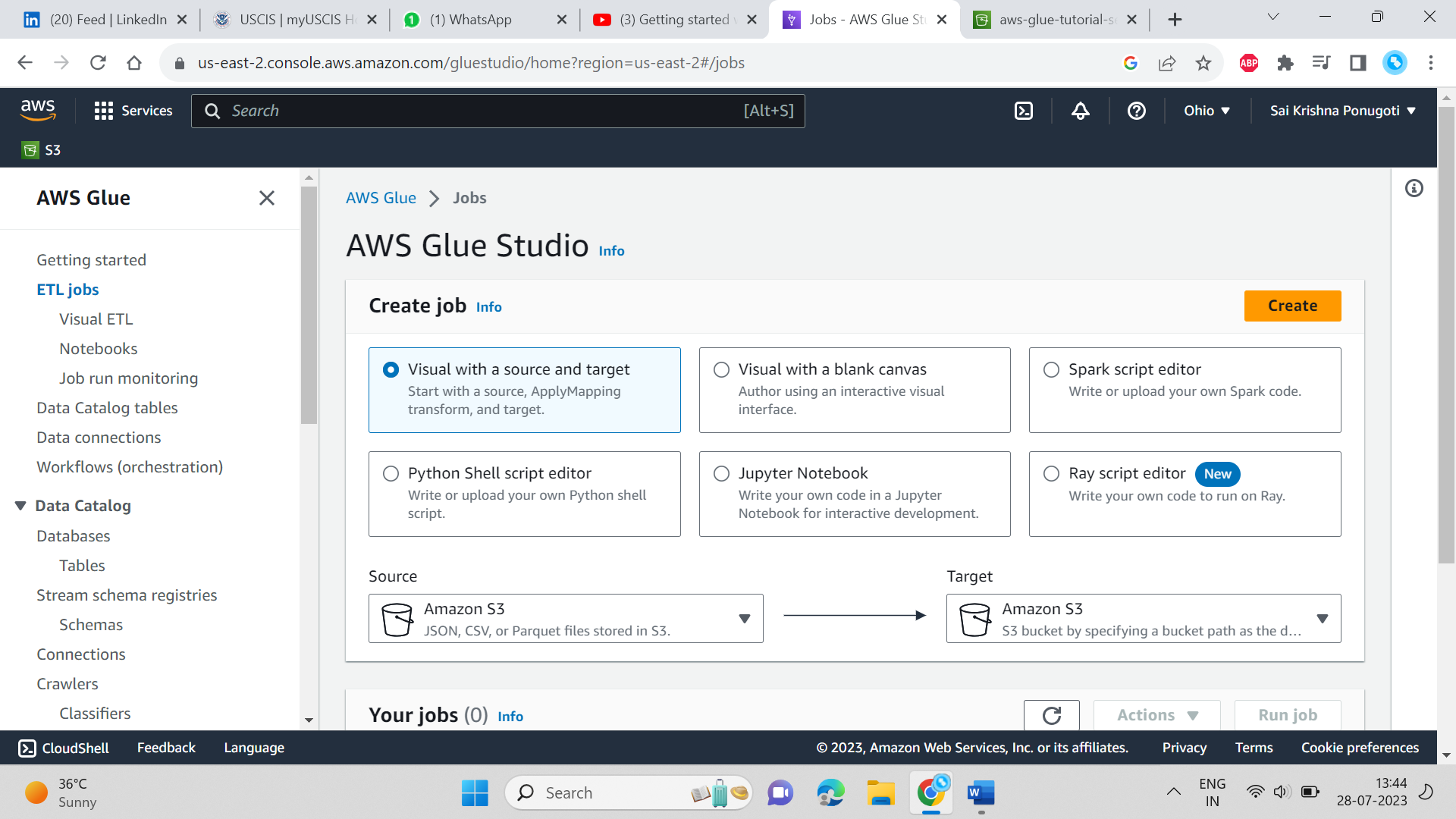
Adding the data source to the table



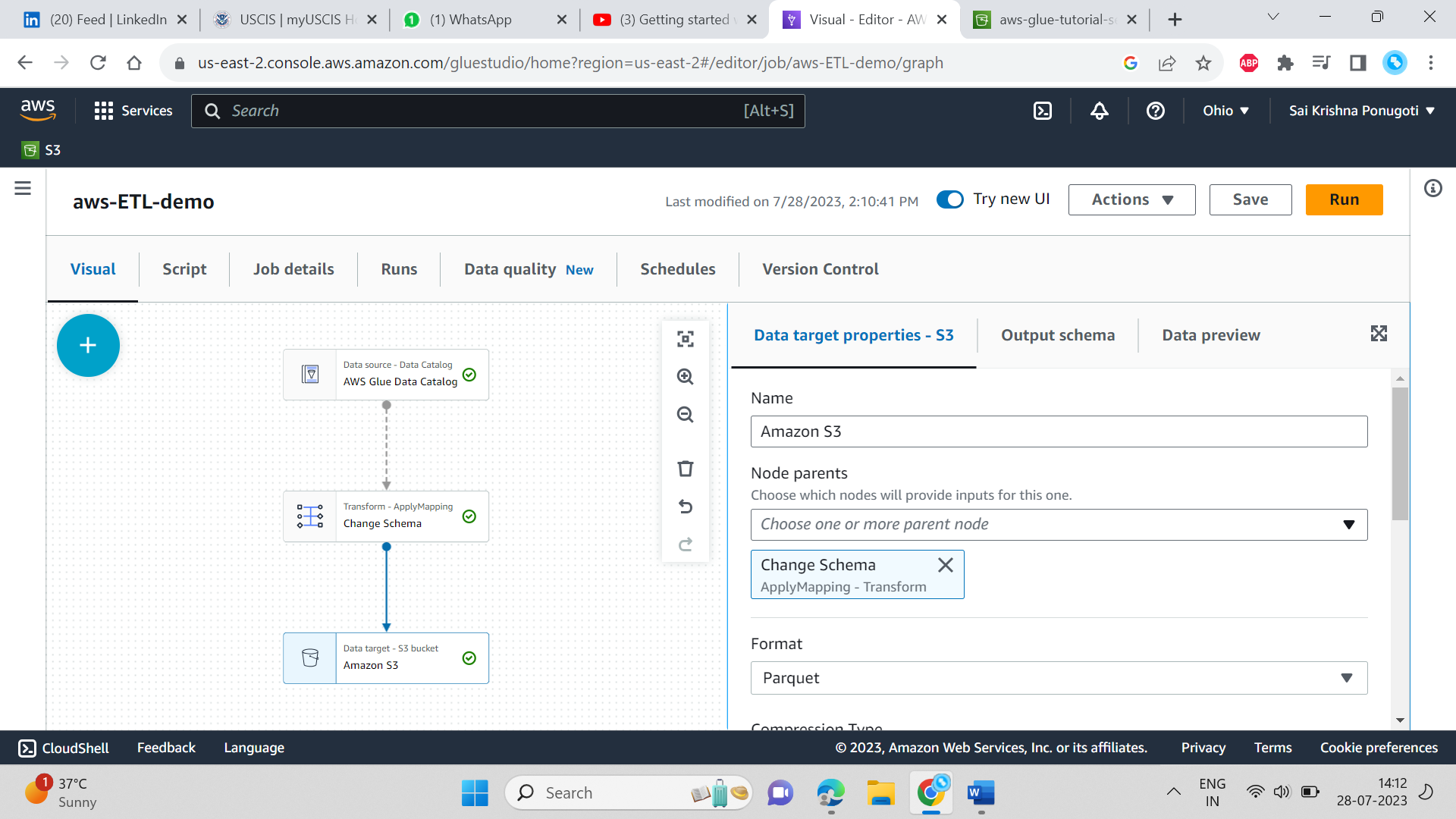
Checking whether the table has created or not in the database after the crawler job is done



Now creating the ETL job



Here I am using visual with a blank canvas to create an ETL job.



After applying transformations based on your requirements save and run the script. And then check the check the target location in you s3 bucket.

