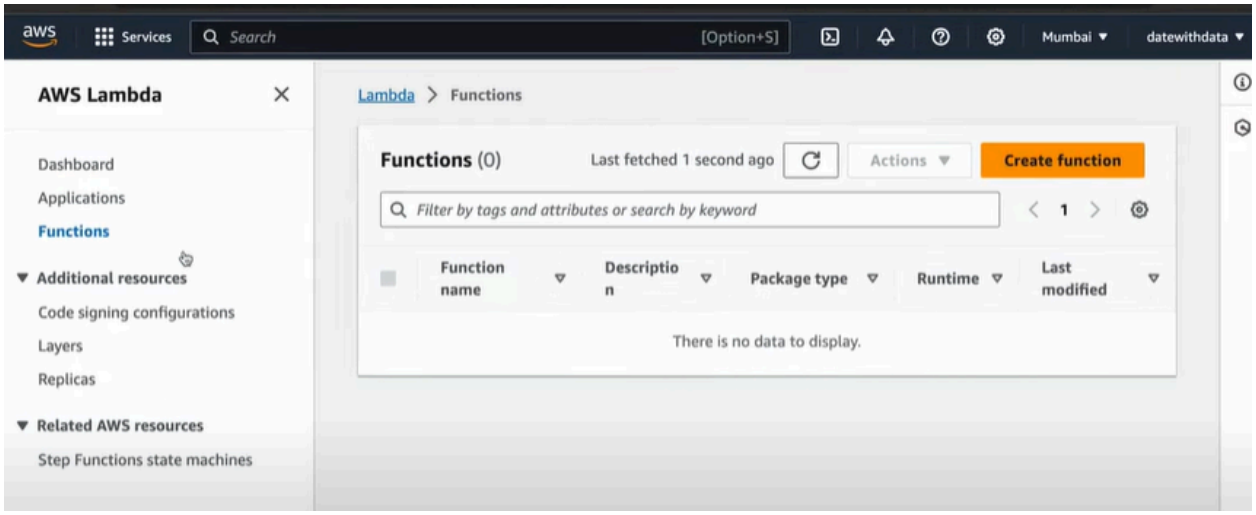
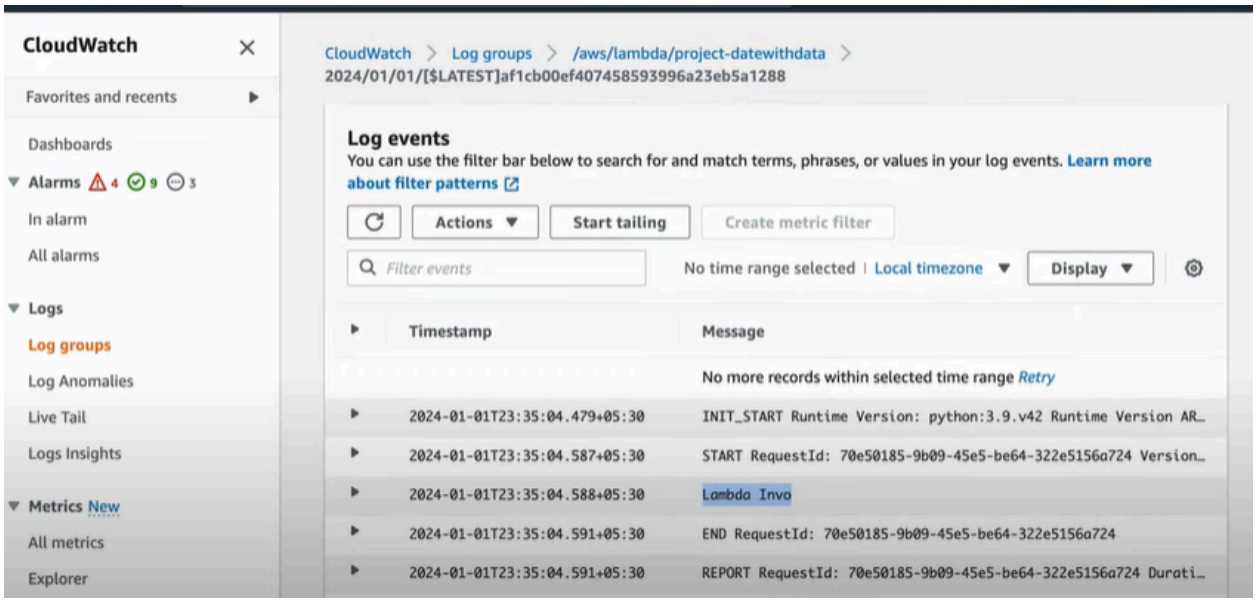
Lambda Function

1. Lets create a lambda function..that takes data from dynamoDB and stores it in S3 Bucket
2. Lets go to lambda and click on create function



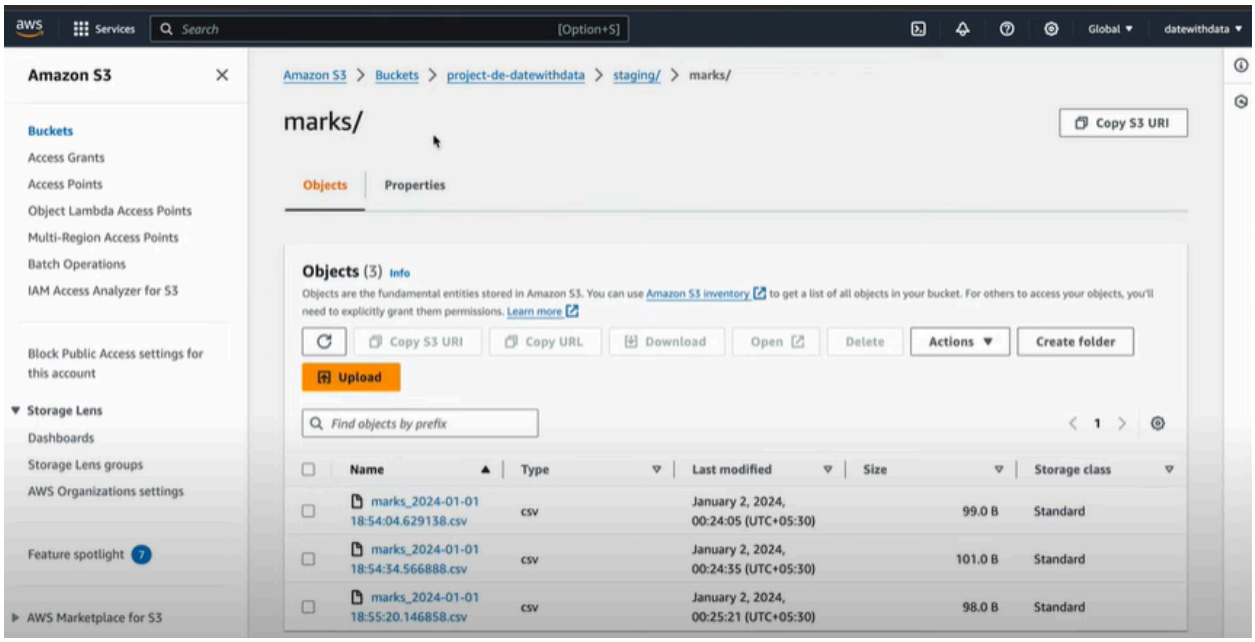
and we create a function with python as runtime

3. Lets setup a dynamoDB trigger
4. Now we need to give access to dynamoDB and S3 for our lambda function…using policies
5. We need to integrate cloud watch with our lambda ..to check the logs of the trigger
6. Next we'll add sample data in our dynamoDB and it triggers the lamda function…which we can see thru the cloudwatch logs



7. Now the data will be reflected in our S3 buckets

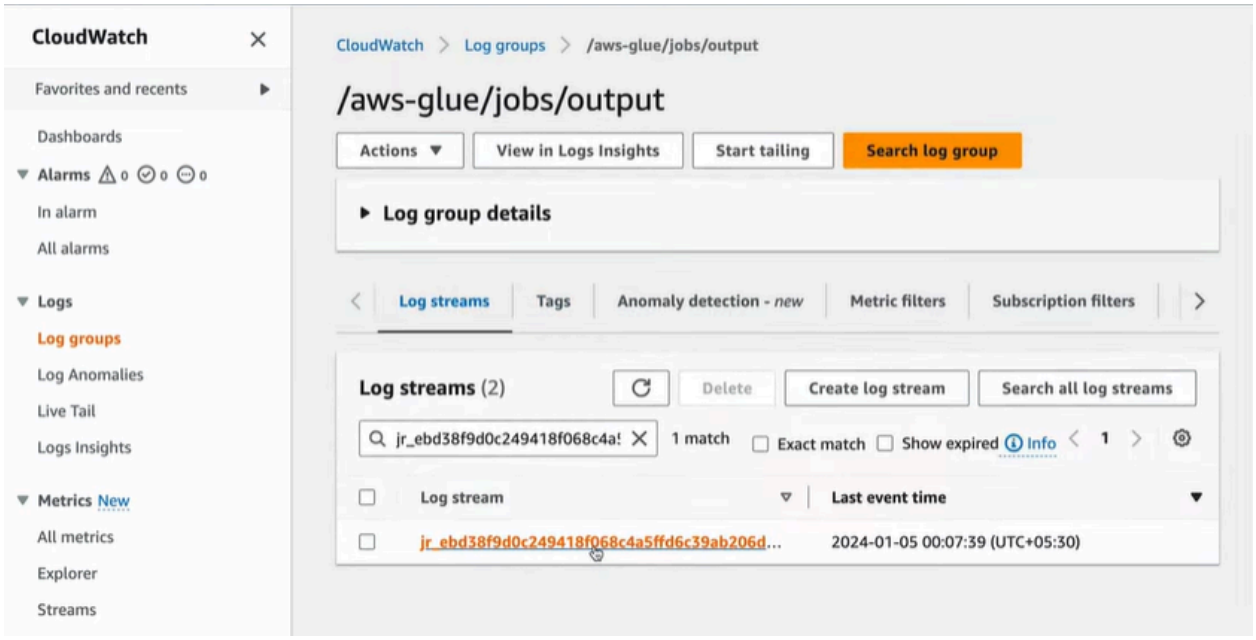
8. And we can see here..the data which we created in our marks table in dynamoDB is reflected in the marksS3 bucket

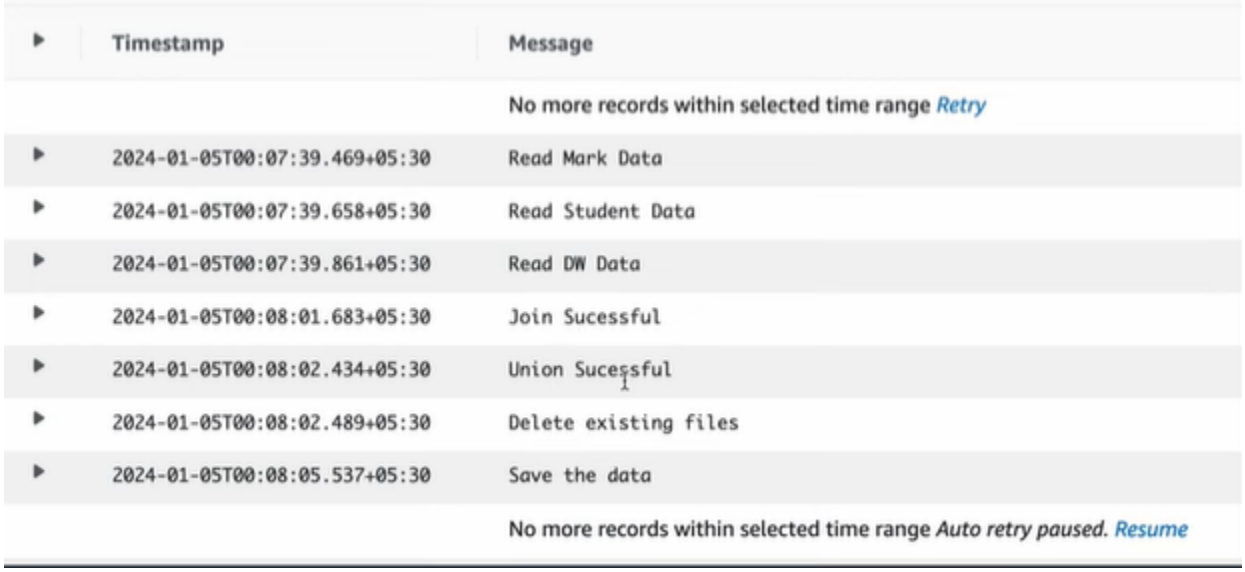


9. Now we have successfully created a lambda function that takes data from dynamoDB streams and load them in S3 buckets

AWS Glue Job

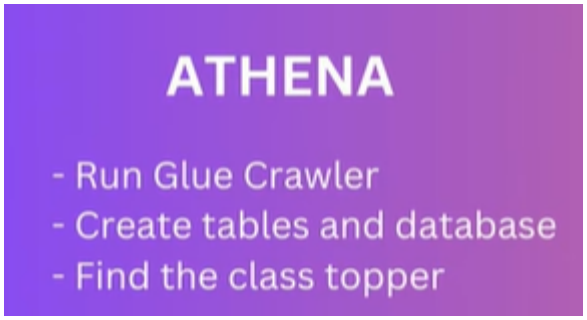
1. Now we'll create a Gluejob which takes data from s3buckets joins and save in data warehouse
2. Please watch the ETL code on youtube
3. Next we'll pste our code in Glue Script and provide all the required Job Details
4. After successfully running the GLue job…we can also see the output logs

5. 

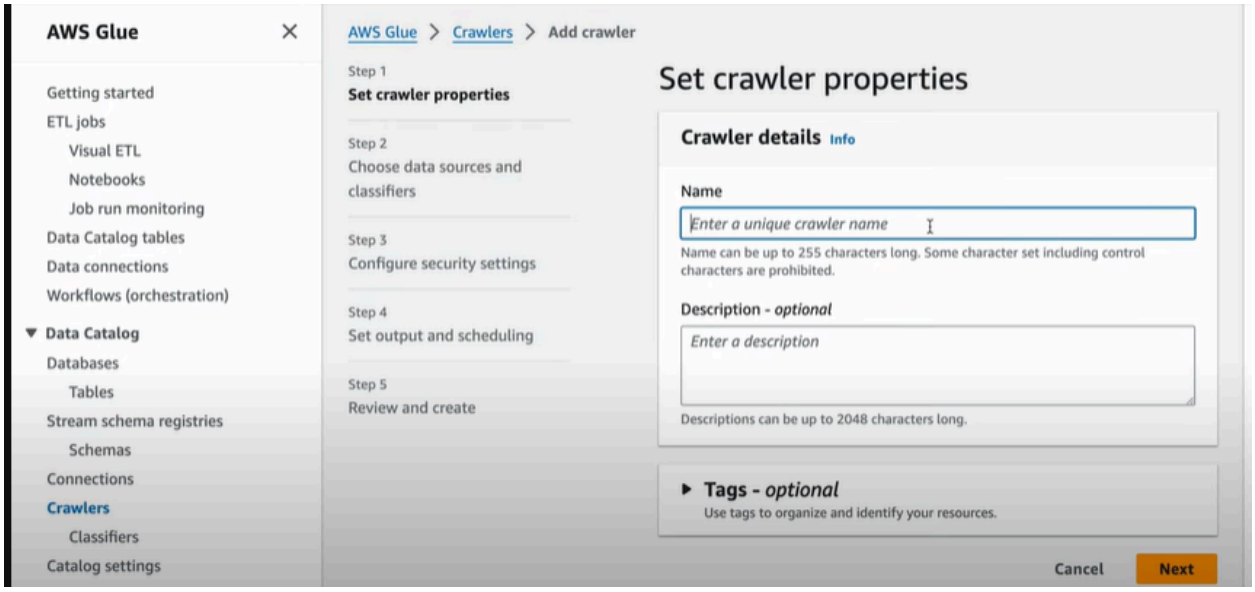


Timestamp	Message
	No more records within selected time range *Retry*
2024-01-05T00:07:39.469+05:30	Read Mark Data
2024-01-05T00:07:39.658+05:30	Read Student Data
2024-01-05T00:07:39.861+05:30	Read DW Data
2024-01-05T00:08:01.683+05:30	Join Sucessful
2024-01-05T00:08:02.434+05:30	Union Sucessful
2024-01-05T00:08:02.489+05:30	Delete existing files
2024-01-05T00:08:05.537+05:30	Save the data
	No more records within selected time range *Auto retry paused. Resume*

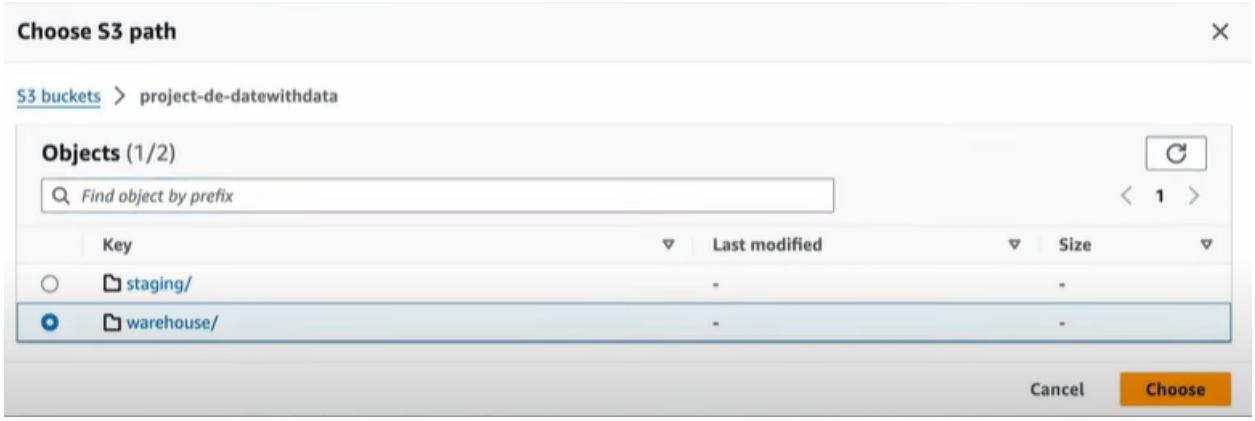
6. Next we schedule the job run..to run the ETL job on regular intervals
7. We also can use version control in AWS glue
8. Now we have successfully created a ETL job..that takes data from s3 buckets and saves in data warehouse

AWS Athena 

1. Here we'll run Glue crawler…which helps us to created tables and databases
2. Then we can use AWS athena…to query the data created by Glue crawler
3. Adding a Glue Crawler

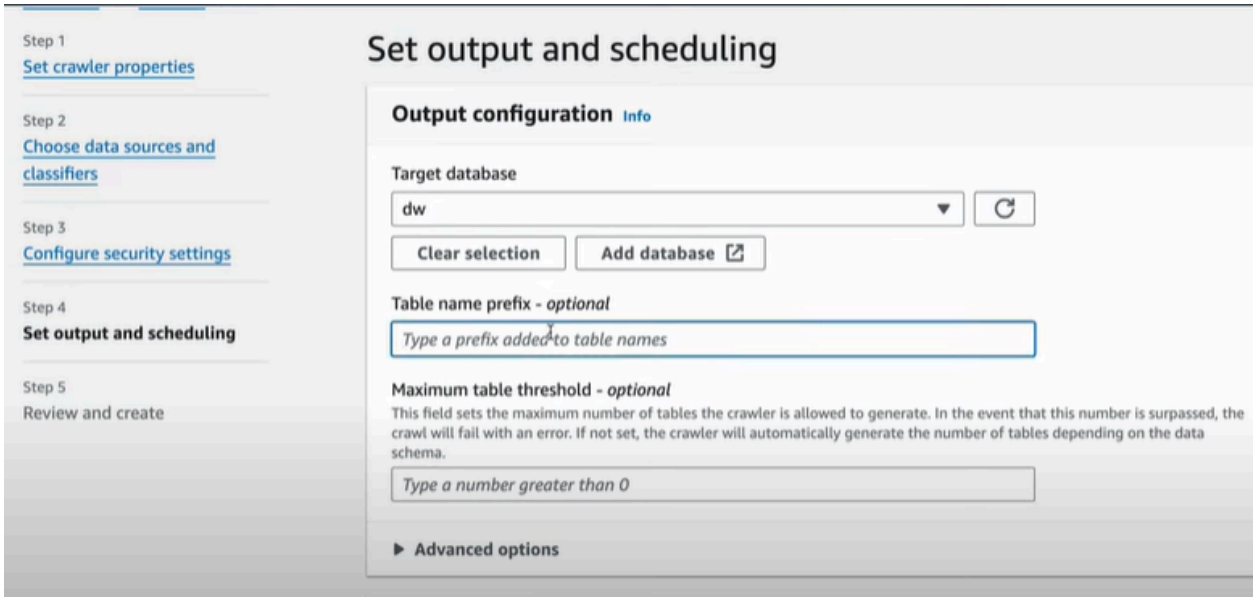


4. Next we'll add the data source…which is our warehouse



5. Next we have to provide the IAM role

6. Next we have create a db and assign a db



7. An AWS Glue crawler is an automated tool that discovers and analyzes data stored in various data sources. It acts like a librarian in a giant data warehouse, automatically sorting and cataloging your information. Here's how it works with an example:

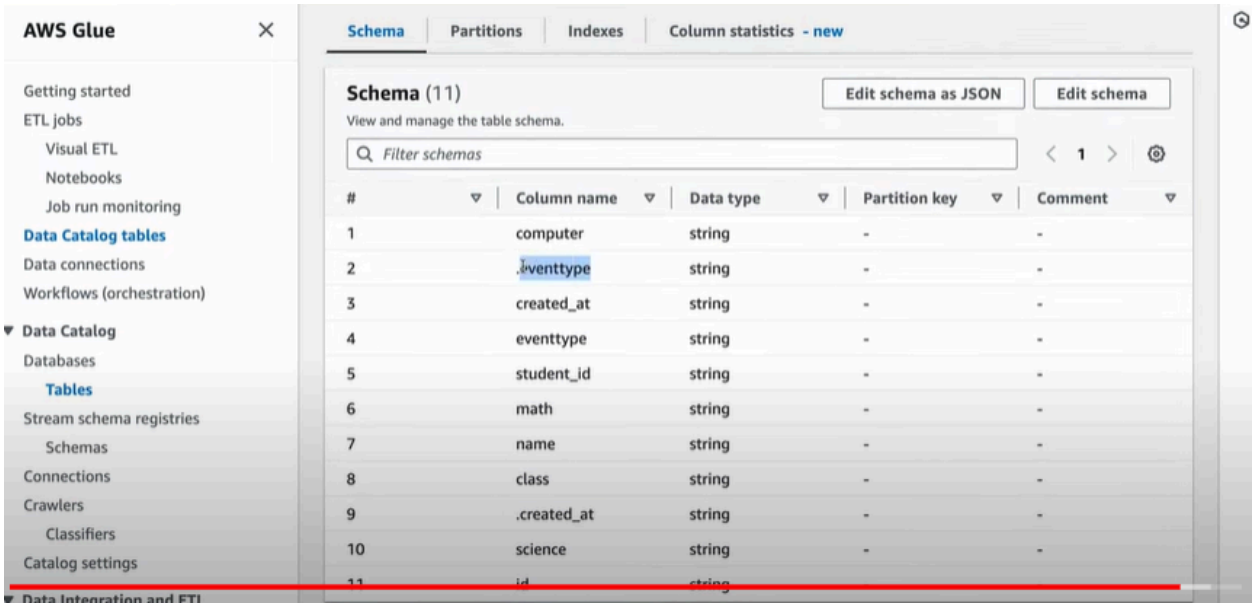
   Imagine you have a bunch of CSV files containing sales data in an Amazon S3 bucket (think of it as a cloud storage space). The data is scattered across different folders, and there might be new files added every day.

   An AWS Glue crawler can be configured to scan this S3 bucket. Here's what it typically does:

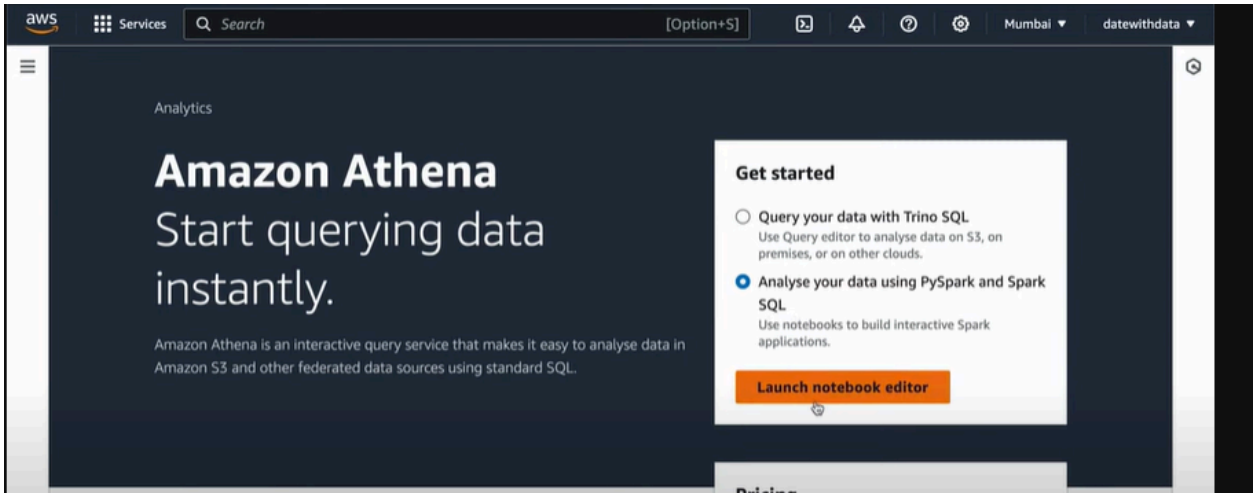
   1. **Connects and Discovers:** The crawler connects to your S3 bucket and starts discovering the files.
   2. **Classifies and Schemas:** It uses built-in classifiers (like a predefined format identifier) to understand the data format (CSV in this case). It then analyzes a sample of the data to figure out the structure, like which column represents "customer name" and which one represents "sales amount."
   3. **Catalogs the Data:** Based on its findings, the crawler creates a schema (blueprint) describing your data and creates an entry in the AWS Glue Data Catalog. This catalog acts as an index, storing information about your data's location, format, and structure.
   4. **Updates Regularly (Optional):** You can schedule the crawler to run periodically. This way, whenever new data files are added to the S3 bucket, the crawler automatically crawls them, updates the schema if needed, and keeps your Data Catalog synchronized.

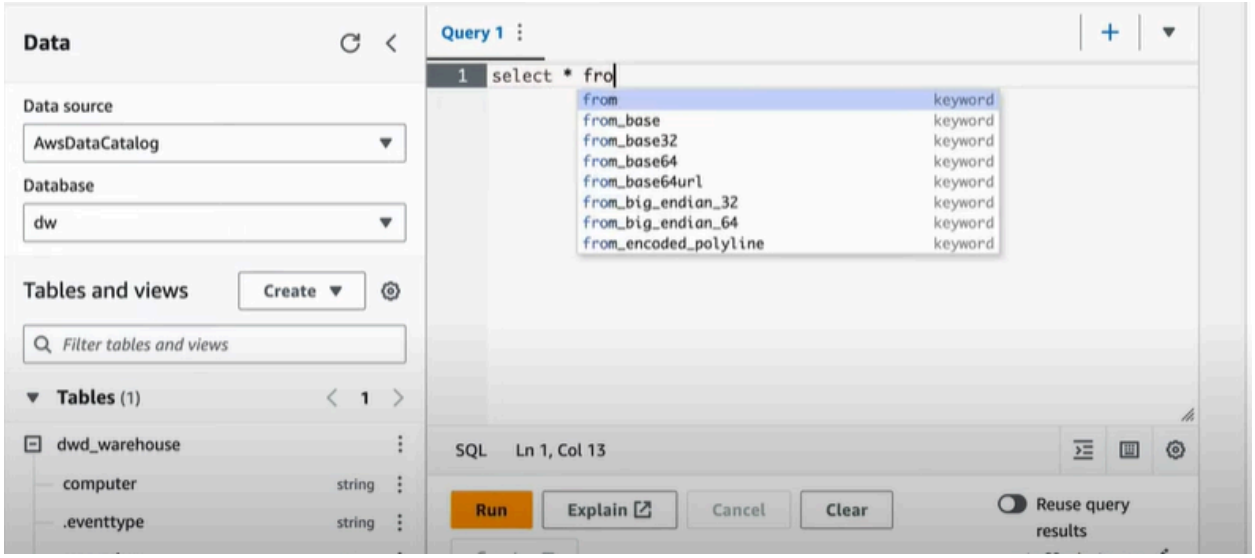
8. Now we'll create the crawler

9. Once our table are created..we can go to athena and query our table



10. Athena



11. 

here by default we got AWS data catalog as source ..and now we can query(using SQL) our data from dw database