

**ETL FLOW USING AWS GLUE**

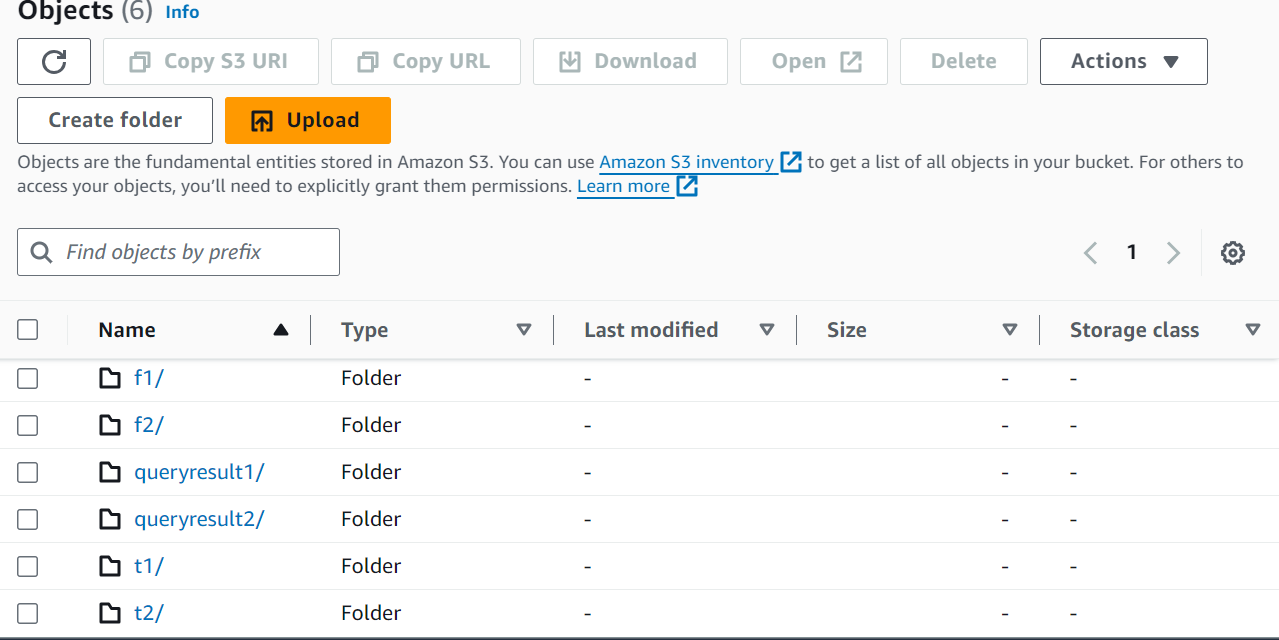
1. **Create S3 Bucket and Folders:**

- Go to the AWS Management Console and navigate to the S3 service.

- Create a new bucket named "statefinance1".

- Inside the bucket, create two folders named "f1" and "f2".

- Upload the source data files into the "f1" and "f2" folders respectively.

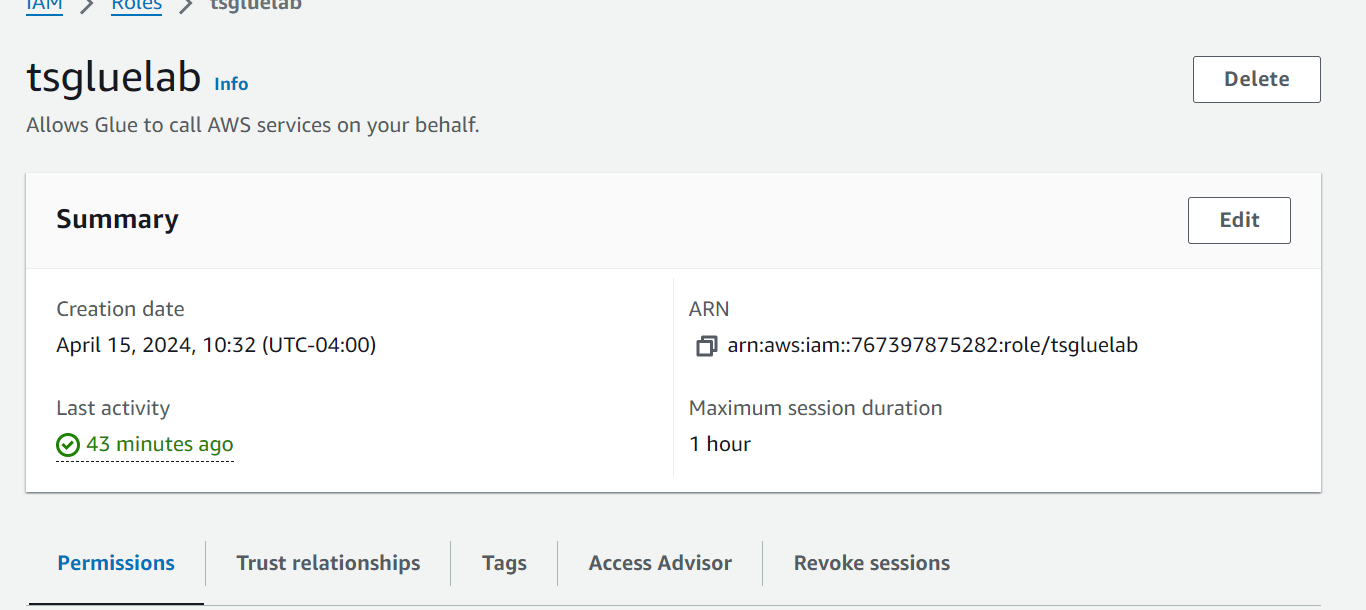
****

2. **Create IAM Role and Policy:**

- Go to the IAM service in the AWS Management Console.

- Create a new IAM role named "tsglue".

- Attach a policy (e.g., "tsgluepolicy") to the role with full access to S3.



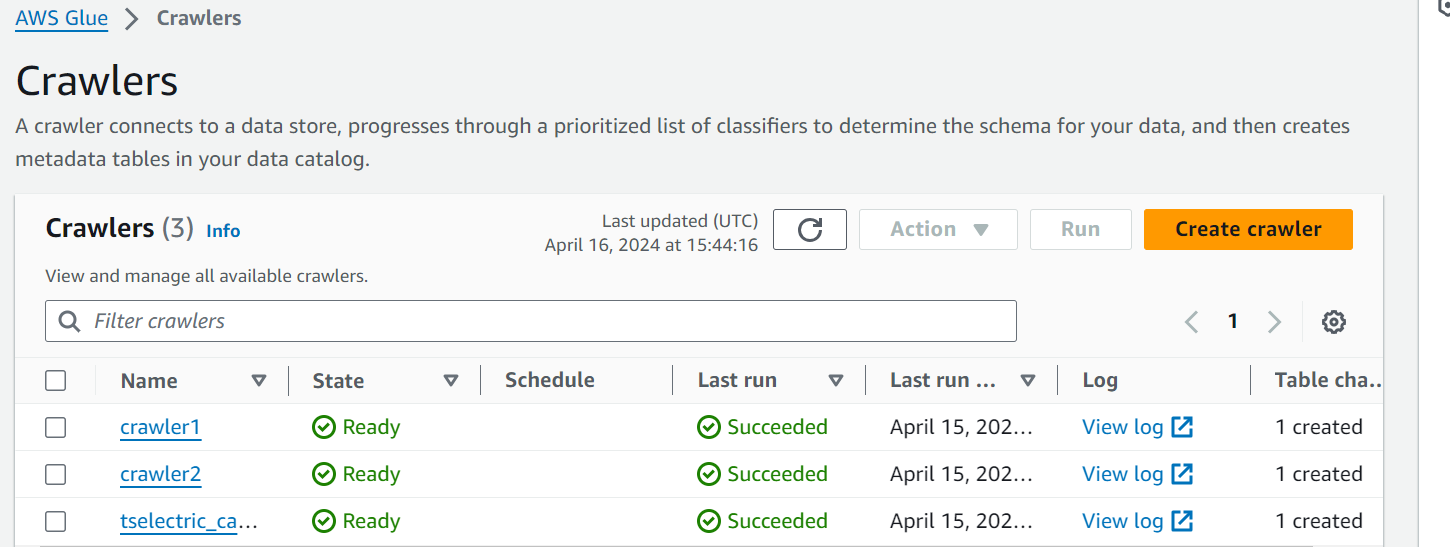
3. **Create AWS Glue Crawlers:**

- Go to the AWS Glue service in the AWS Management Console.

- Create two crawlers: "crawler1" and "crawler2".

- Configure each crawler to add the respective data source (f1 for crawler1 and f2 for crawler2).

- Create and add a database in AWS Glue.



4. **Create ETL Job using AWS Glue Visual ETL:**

- Navigate to AWS Glue and access the Visual ETL tool.

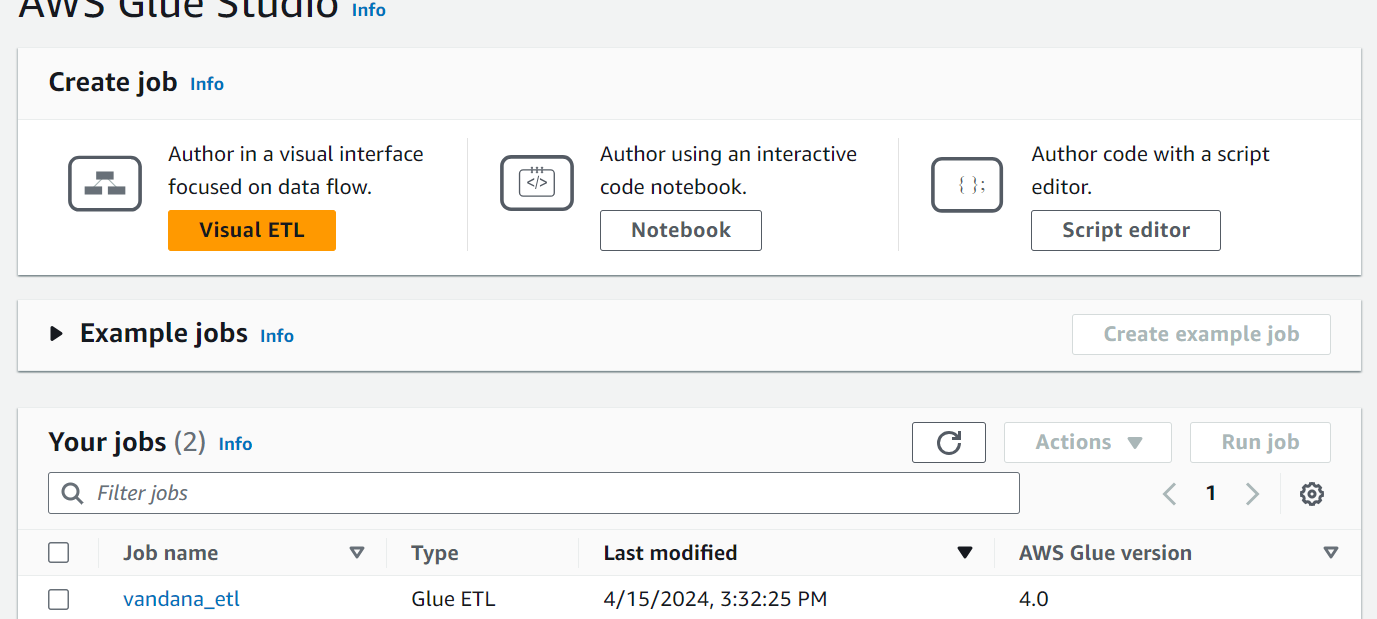
- Create nodes as required for your ETL job and specify each node.

- Add two data sources from the S3 bucket and join them.

- Implement a conditional router to route the data based on the condition (<= 2000 and >= 2000).

- Thus, we should provide the router with one condition, and the other will be saved in the default output.

- Store the results in the S3 bucket by creating folders for each query result (e.g., "query1" and "query2").



5. **Create Tables in AWS Glue:**

- Navigate to the Tables section in AWS Glue.

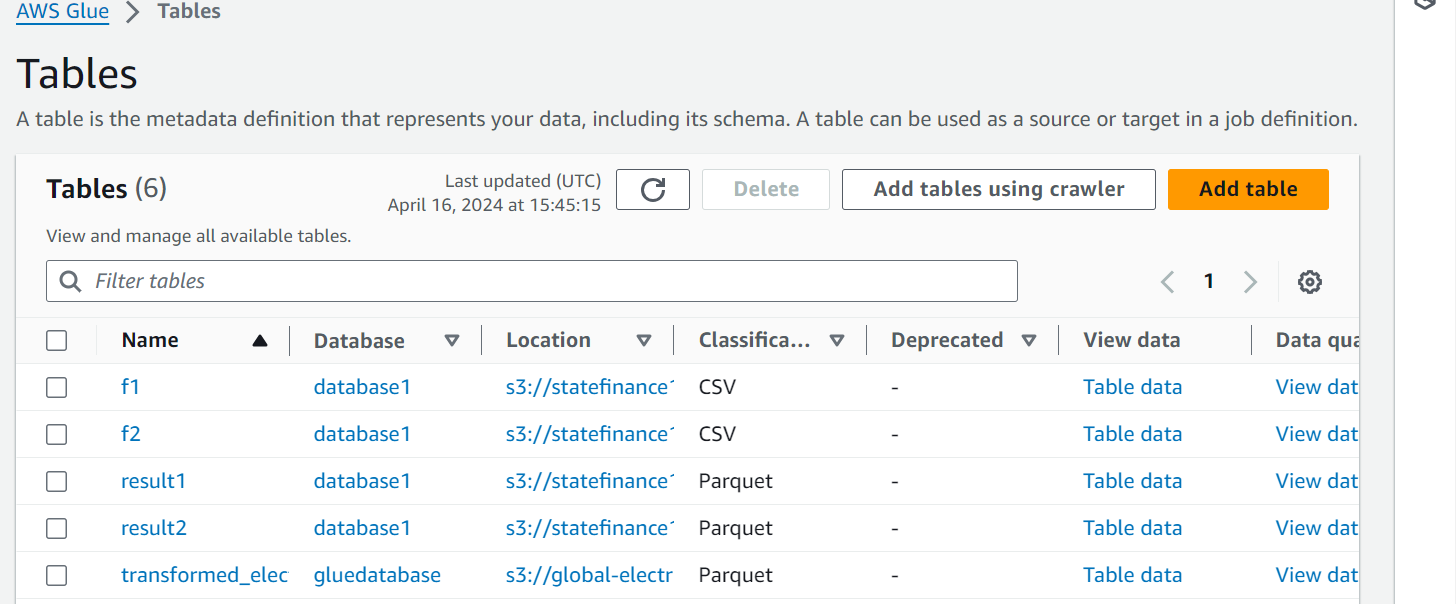
- Click on "Add tables using a crawler" and select the databases created earlier.

- AWS Glue will automatically create tables based on the data crawled by the crawlers (result1 and result2).

6. **View Table Data in AWS Glue:**

- Navigate to the Tables section in AWS Glue.

- Click on the table names (result1 and result2) to view the table data.

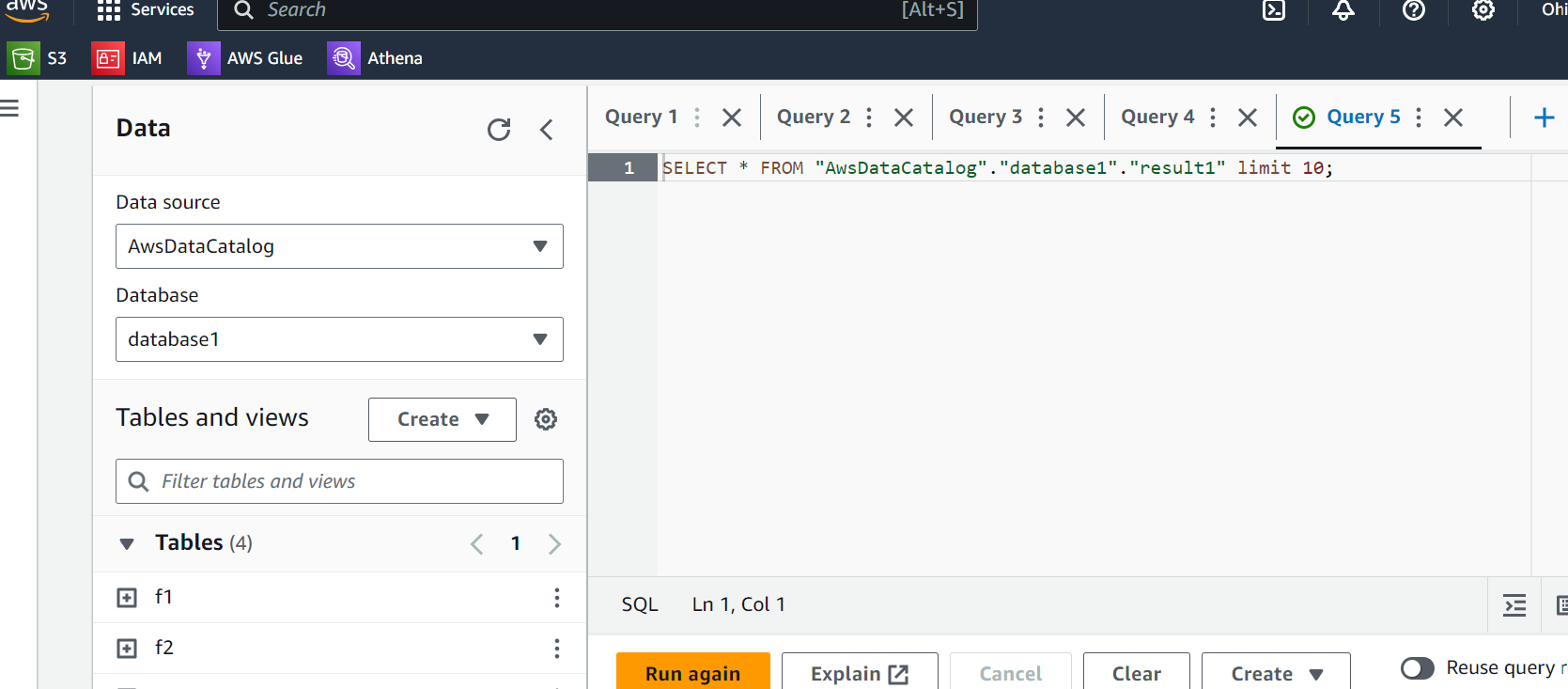


7. **Query Data Using Athena:**

- Navigate to the Athena service in the AWS Management Console.

- Write SQL queries to query the data from result1 and result2.

- Store the query results in the S3 bucket by creating folders for each query result (e.g., "query1" and "query2").



8. **Download CSV Files:**

- Once the query results are stored in the S3 bucket, you can download the CSV files from the respective folders (query1 and query2) to view them.

