Data Transformation Lab:

1. Create a new ETL flow utilizing what you learned in the last class by following the procedures outlined.

2. Make a copy of the data files from this location and upload them to different S3 folders within the same S3 bucket.

3. Utilizing an Extract Node, read the data from an S3 bucket (two nodes are required to read files from two different folders).

4. Utilizing the inner join condition, merge the files from both nodes using the join transform.

5. To route the files into two locations, use the conditional router transform.

6. output nodes, one with data that is older than 2000 and another with data that is older than or equal to 2000.

7. To generate two S3 folders and list these two data files in catalog database tables, add two S3 output nodes at the end.

Steps:

1. Enroll\_states\_sumarry.csv and States\_finances\_all.csv are the two files. Make a bucket and put a different folder in it for every file. Following the files' upload to S3,

2. Get glue crawler jobs ready. To determine the schema of two files, two crawler jobs are made.

3. Data catalog tables are obtained after the crawler jobs are completed.

4. Now that you have created an ETL Glue Job, choose the two source files from the bucket and begin performing transformations (with in mind that we are using data catalog tables as sources in this case).

5. In the first transformation, the join condition is applied based on the primary key. (using an inner join)

**Select \* from file1**

**inner join file2**

**on file1.primary\_key = file2.primary\_key**

6. We are implementing a router transformation depending on year after the join condition has been applied. Every record with a year less than 2000 is kept apart from every record with a year more than 2000.

7. Every group from the router condition is now maintained in a different data target. This data target will be an S3 folder that is distinct from the others.

8. The target files are visible in the S3 bucket's output directories.

9. With Athena, we may query the data catalog table.

Additional Transformations tried:

1. Attempted to delete the file's null values. The transformations "Remove Null Rows" and "Drop Null Fields" were applied. However, while using Athena for queries, it failed to eliminate the null values. It has been determined that there are several ways to accomplish this, including the use of "SQL Query" and "Evaluate Data Quality" transformations.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated