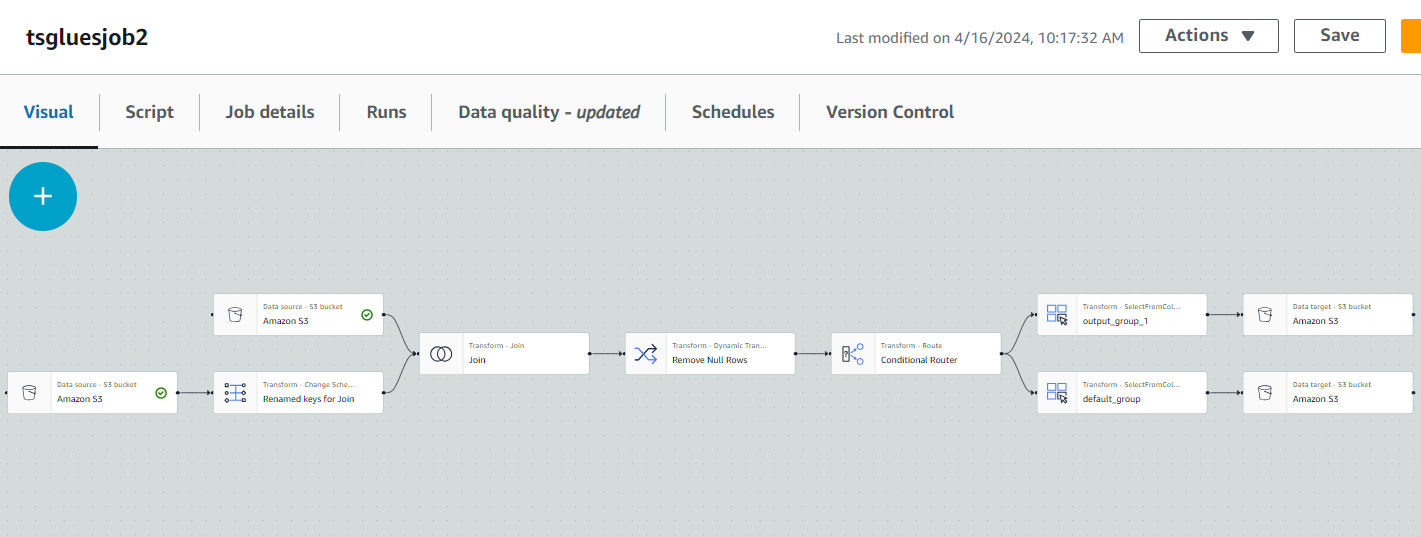
**AWS GLUE ETL PIPELENE :**

* Two files named enroll\_states\_summary and states\_finances\_all were uploaded to Amazon S3 buckets respectively.
* An IAM (Identity and Access Management) role was created using AWS Glue service to provide necessary permissions and access for the ETL process.
* A custom policy was created with all required permissions and access for the ETL process, ensuring the IAM role has appropriate permissions to interact with AWS resources.
* A database was created in AWS Glue to store metadata and schemas for the datasets.
* A Glue crawler was created to crawl the data from the two files stored in the S3 buckets, extracting schema information, and creating a reference table in the AWS Glue Data Catalog.



GLUE JOB:

* A Glue job was created to define the ETL pipeline process.
* Source data was taken from the two Amazon S3 buckets through the data catalog where the reference table was created by crawling the data.
* Name clashes were resolved during data ingestion.
* Data from both sources were joined using a join node on primary key and right primary key.
* A remove null rows node was added to eliminate null values from the dataset.
* A conditional router was included to apply a condition for the join, specifically filtering data where the year is less than 2000.
* Default group and output group were automatically generated.
* Default group and output group generated from the ETL job were assigned to two separate Amazon S3 buckets, which serve as the targets to store the output results.
* A reference table was specified to be created in the Data Catalog for the output results.

AWS ATHENA:

* A query was provided in Athena to analyze and validate the output results stored in the Amazon S3 buckets.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated