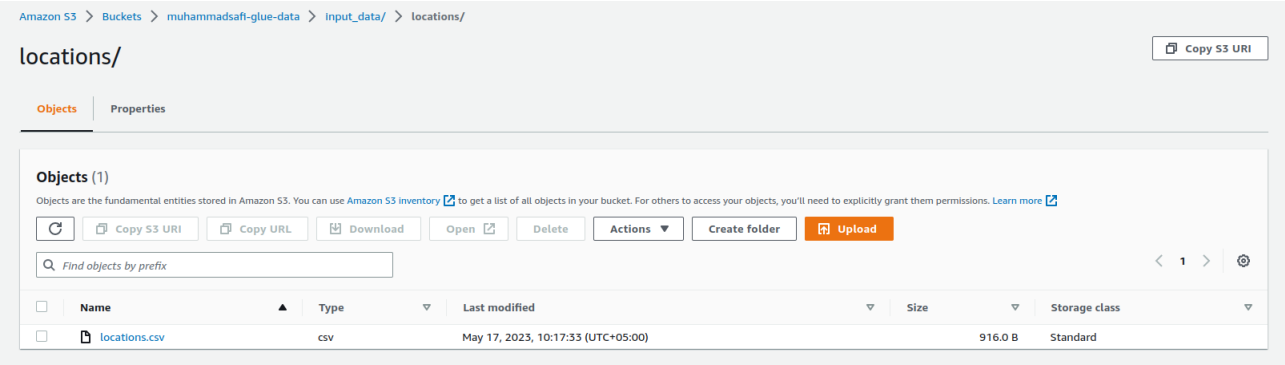
Name : Muhammad Safi (2303.khi.deg.023)

Assignment partners:

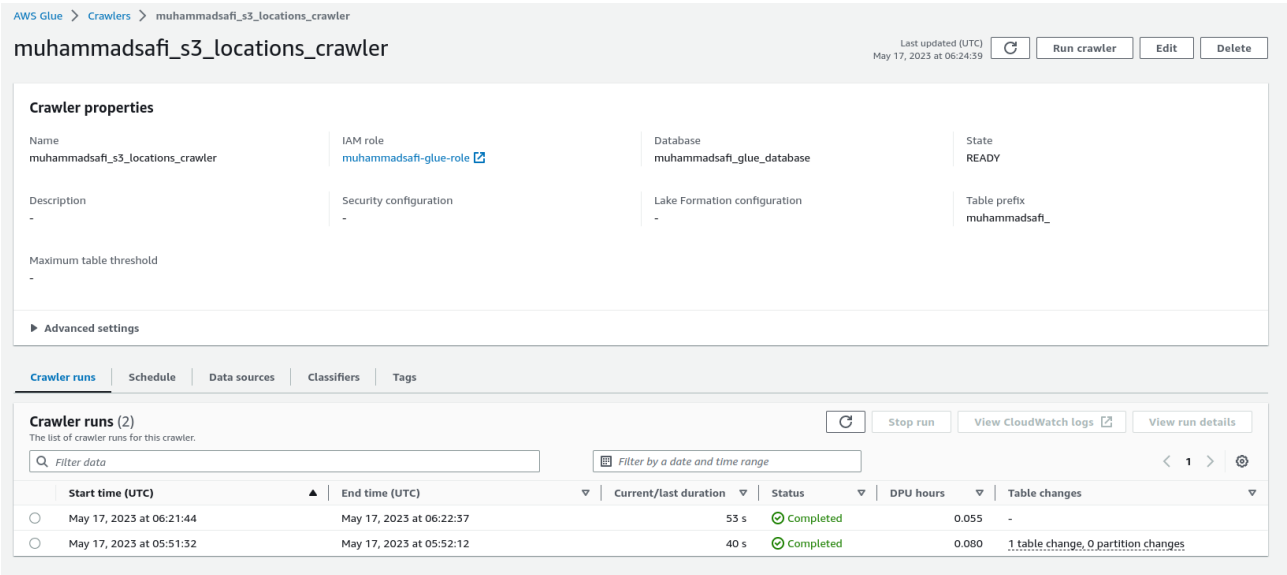
Huzaifa Ali (2303.khi.deg.016)

Shiekh Muhammad Sabih(2303.KHI.DEG.010)

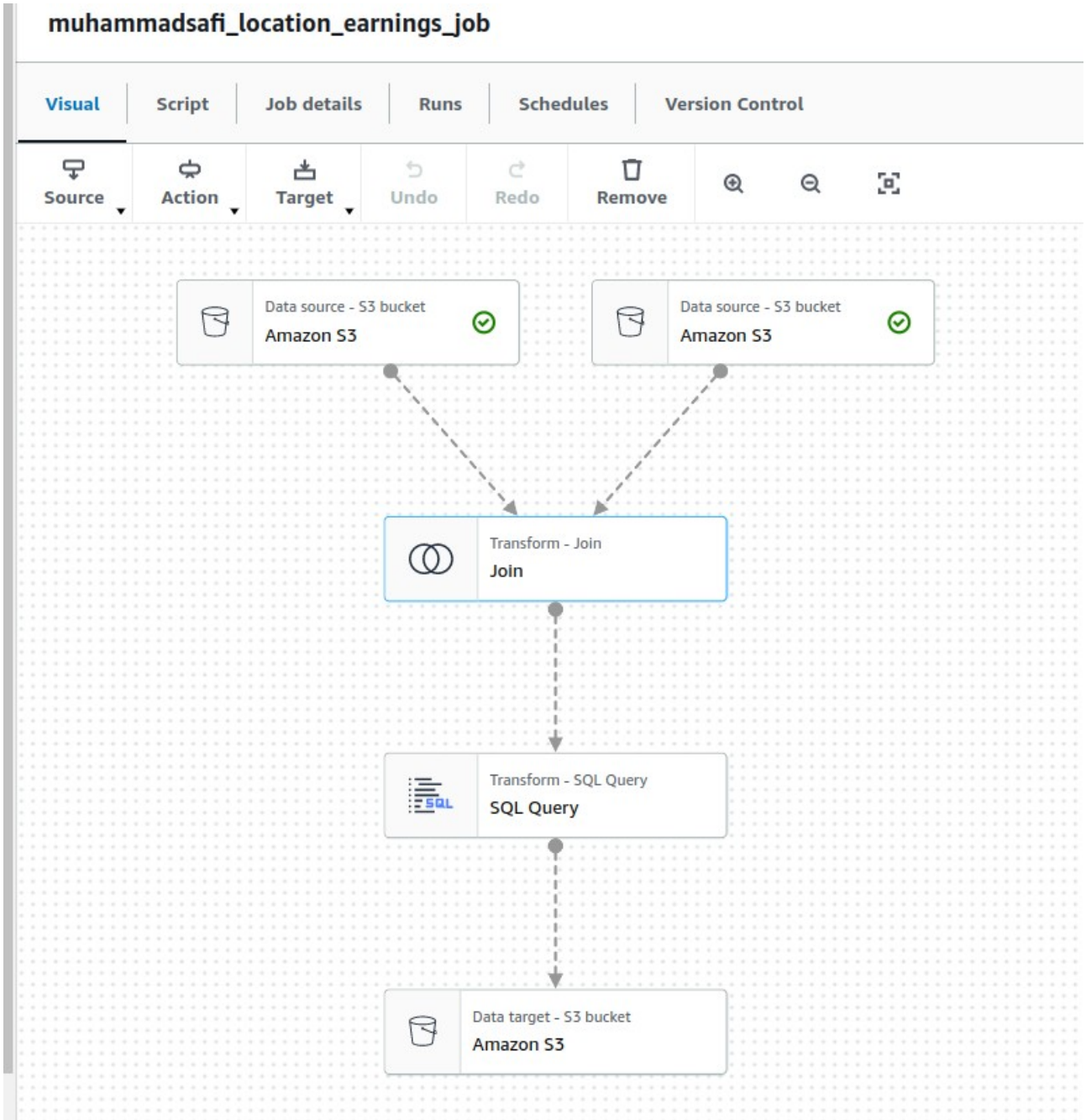
1. first we add locations.csv file uploaded in the s3 bucket.



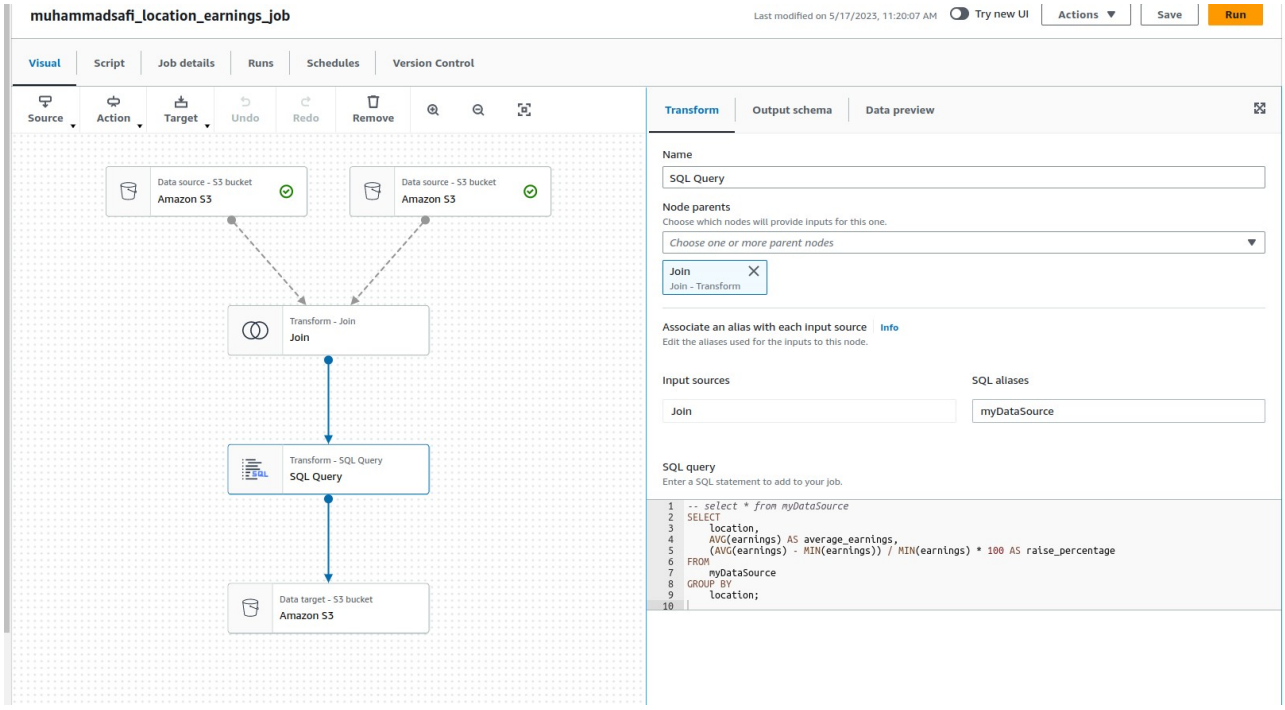
2. Now, we will setup a crawler in order to extract the meta data and generate a catalog.



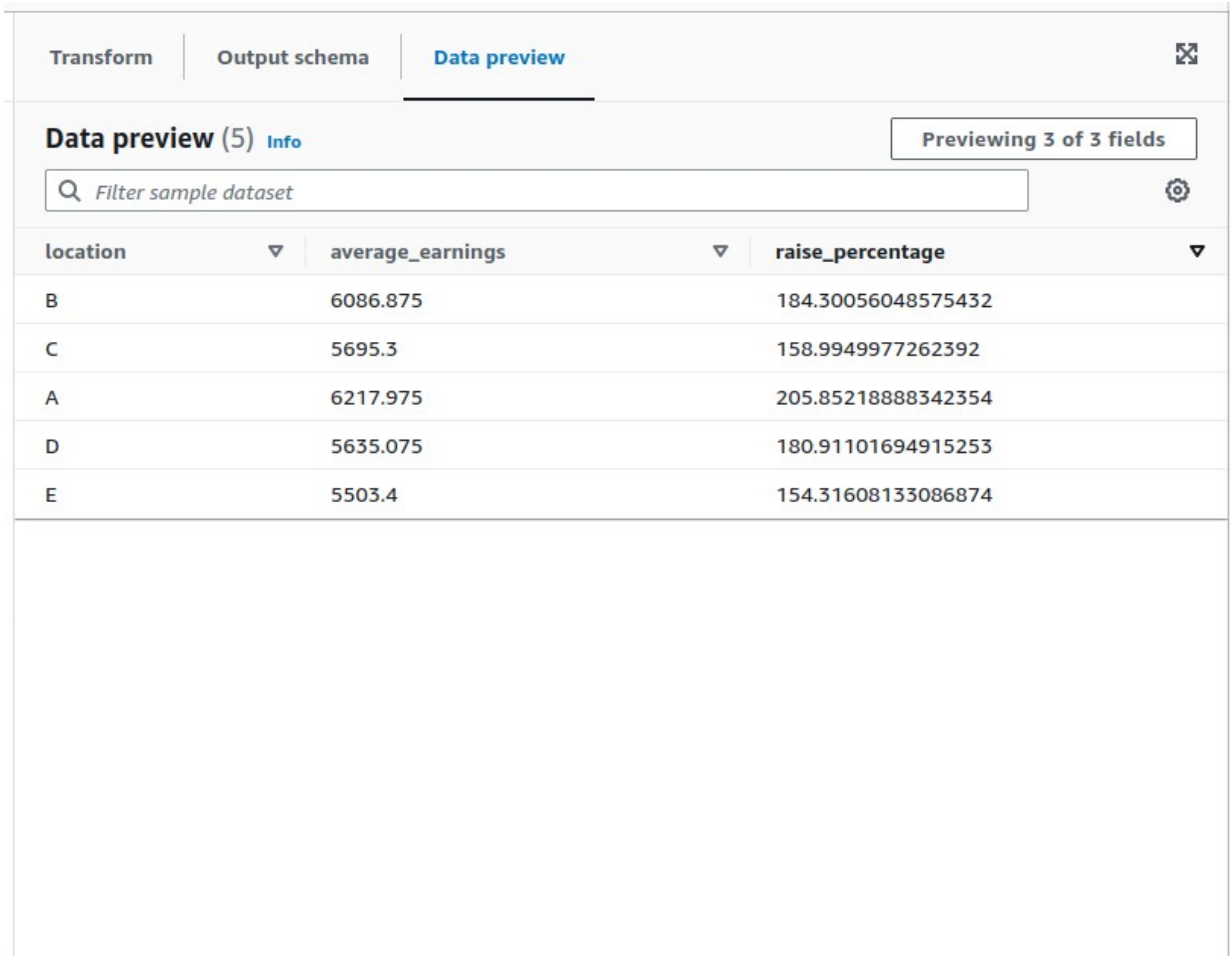
3. We’ll run the crawler and after its done, we’ll start creating the job.



4. The job created here, takes in two tables employee earnings data and other s3 source takesin locations data, we perform an inner join on both data sources on emp_id and prepare the data for querying.

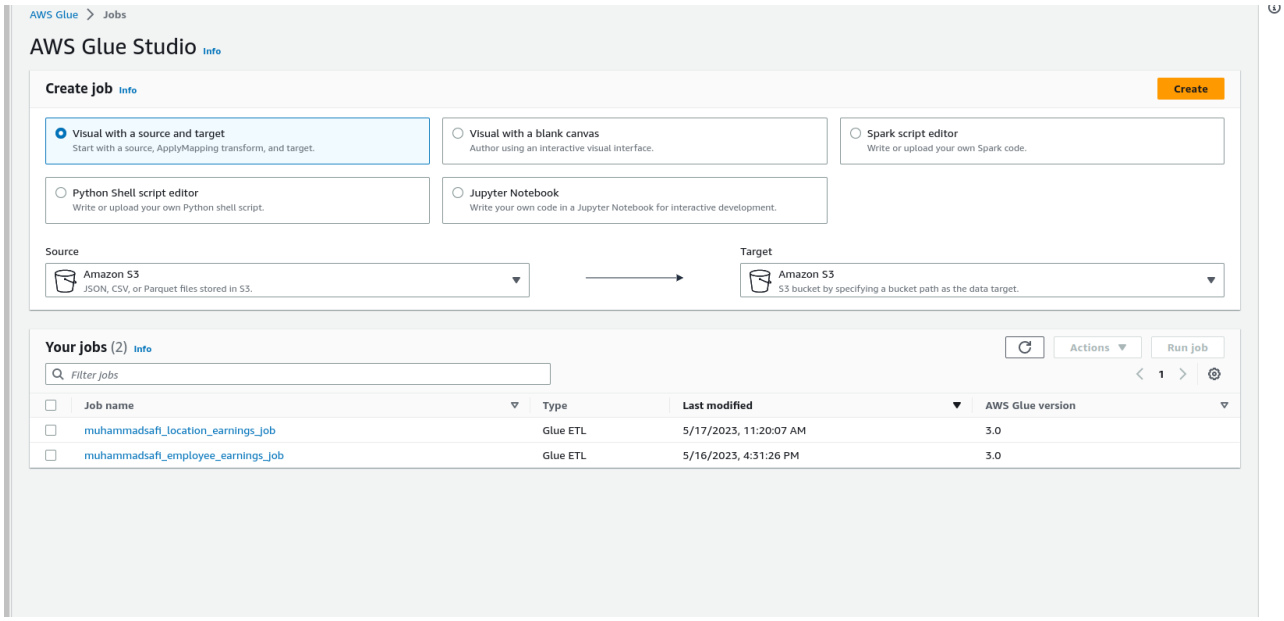


5. The query here aggregates the data based on location and calculates the salaries and percentages of these locations.



location	average_earnings	raise_percentage
B	6086.875	184.30056048575432
C	5695.3	158.9949977262392
A	6217.975	205.85218888342354
D	5635.075	180.91101694915253
E	5503.4	154.31608133086874

6. Now, we'll save this data to our s3 bucket in its output and save this job.



7. After running this job, parquet files compressed with snappy are added to our output location in the s3 bucket.

Here we can see the output schema that is being saved in our output location of s3 bucket.

Data target properties - S3 | Output schema | Data preview

Schema

Key	Data type	Partition
location	string	-
average_earnings	double	-
raise_percentage	double	-

Here we can see the saved files of the outputs generated.