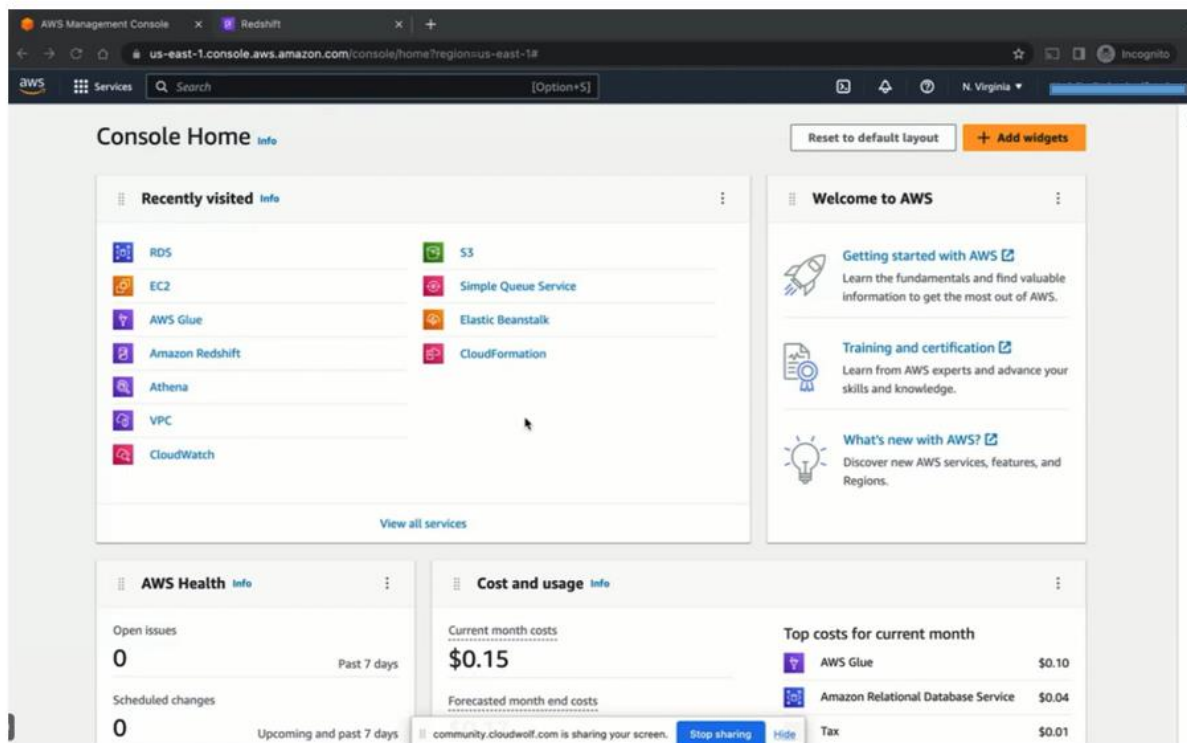
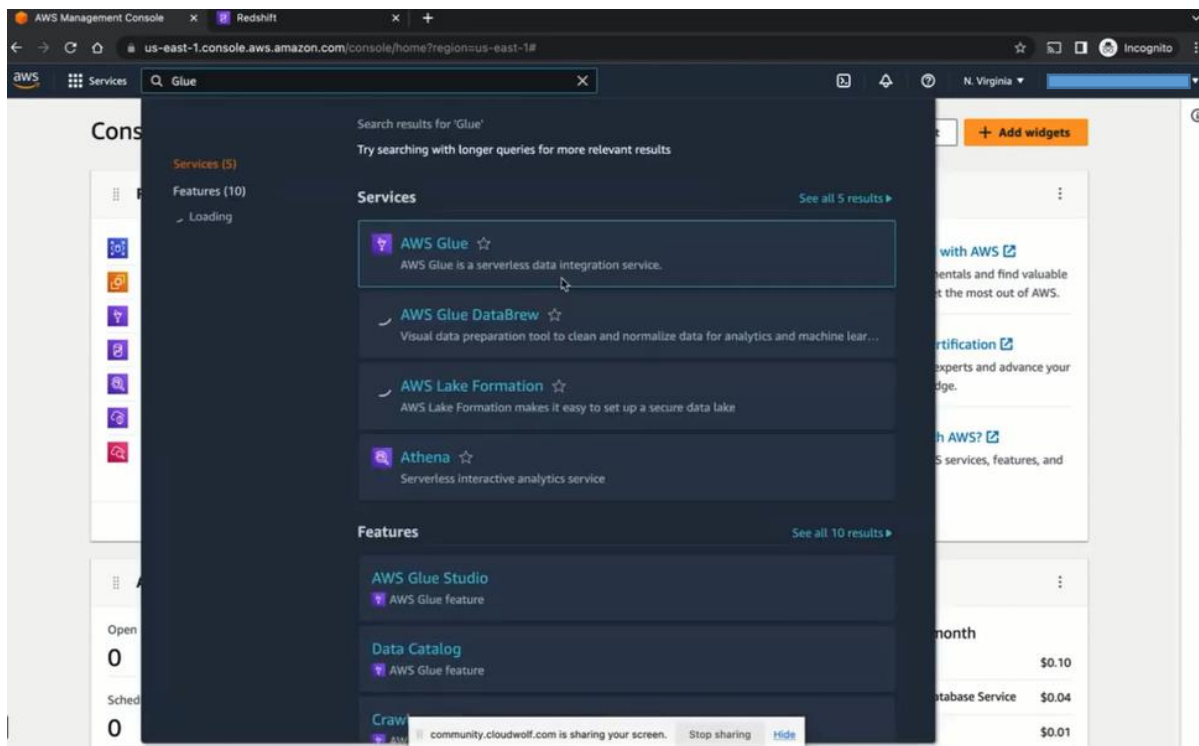


AWS GLUE FOR ETL (EXTRACT, TRANSFORM, AND LOAD) PROCESS

1. GO TO AWS MANAGEMENT CONSOLE. CLICK AWS GLUE.



2. CLICK AWS GLUE.



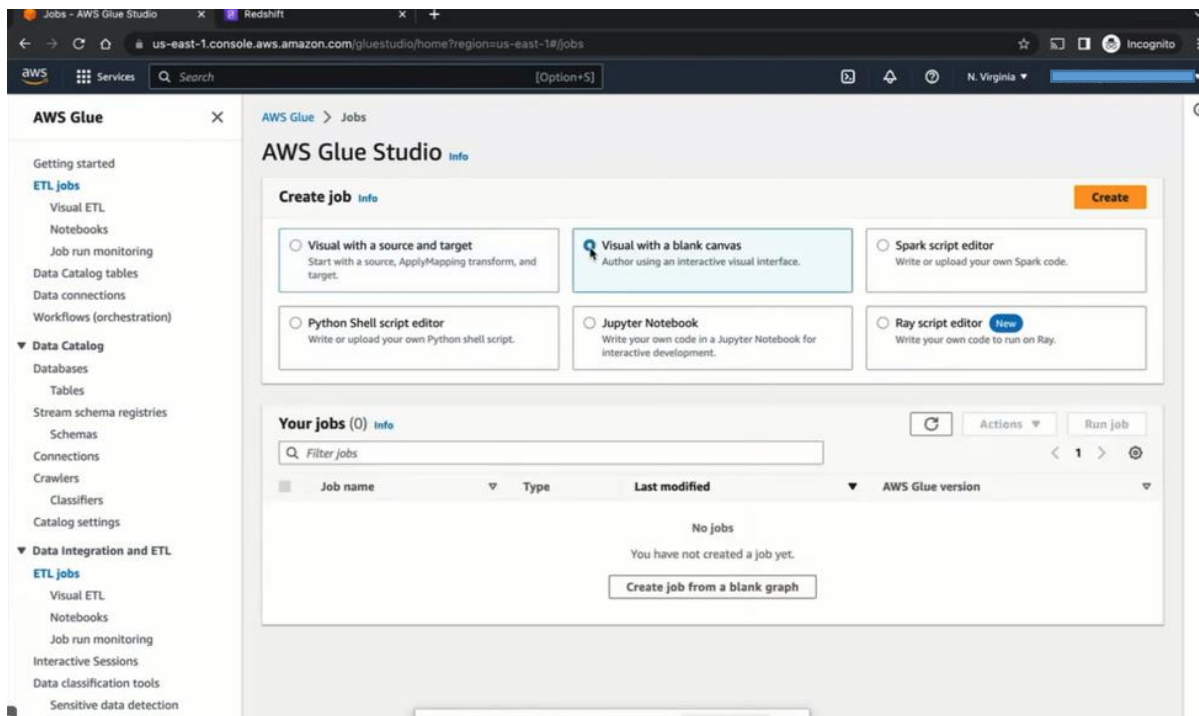
3. CLICK ETL JOBS.

The screenshot shows the AWS Glue console home page. A blue banner at the top says "Welcome to AWS Glue" and "Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data for analytics." Below the banner are three main sections: "Prepare your account for AWS Glue" with a "Set up roles and users" button, "Catalog and search for datasets" with a "Go to the Data Catalog" button, and "Move and transform data" with an "Author and edit ETL jobs" button. On the left is a navigation menu with "ETL jobs" selected. Below the main sections are "Resources and tutorials" and "Data integration and management" sections with various links and buttons like "Go to job run monitoring", "Go to connections", and "Go to workflows".

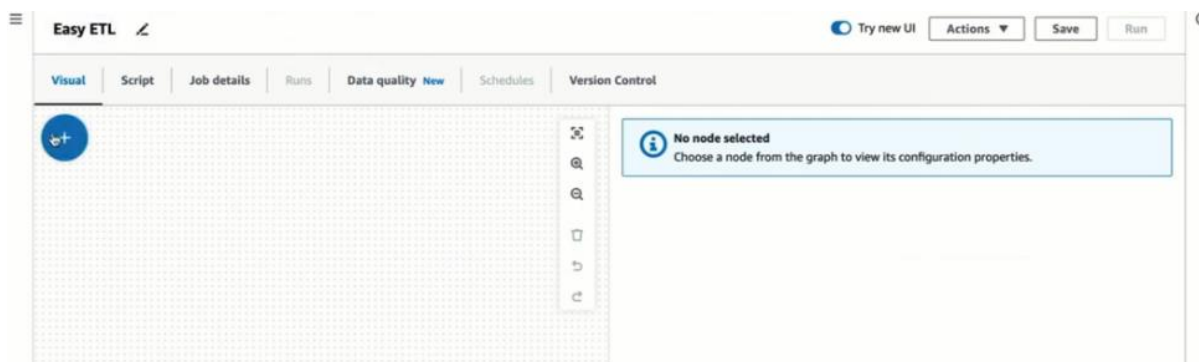
4. ON THE CREATE JOB, CLICK VISUAL WITH A BLANK CANVAS, THEN, CREATE.

The screenshot shows the AWS Glue Studio "Create job" page. The "Visual with a blank canvas" option is selected. Below the options, the "Source" and "Target" are both set to "Amazon S3". At the bottom, there is a section titled "Your jobs (0)" which is currently empty, with a "Create job from a blank graph" button.

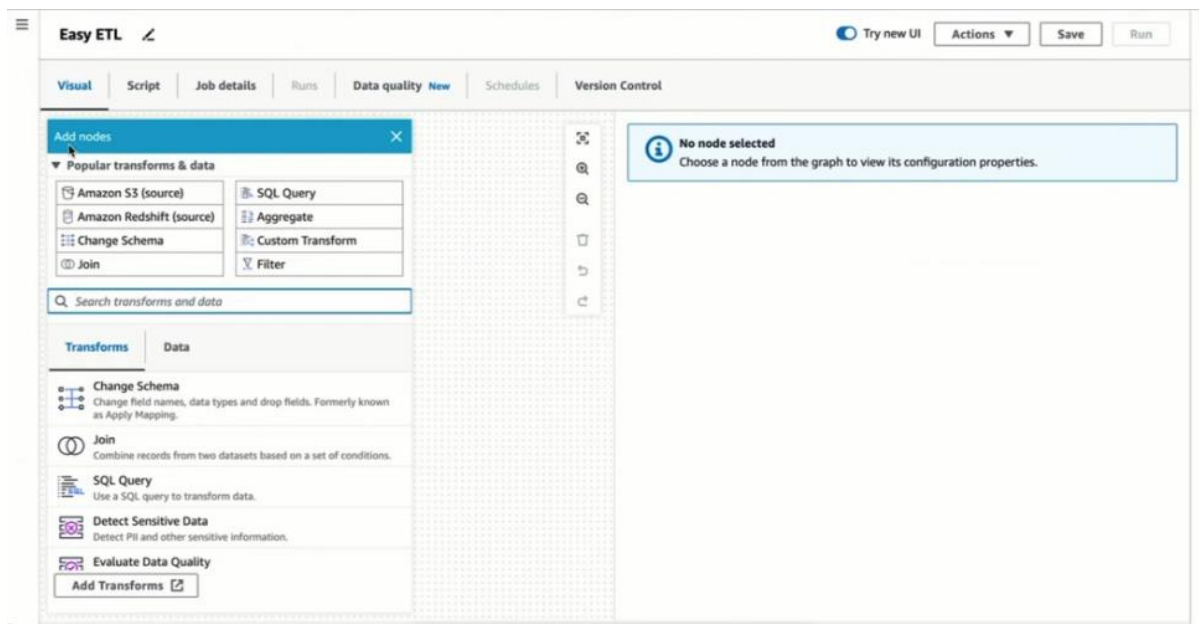
5. CLICK VISUAL WITH A BLANK CANVAS, THEN, CREATE.

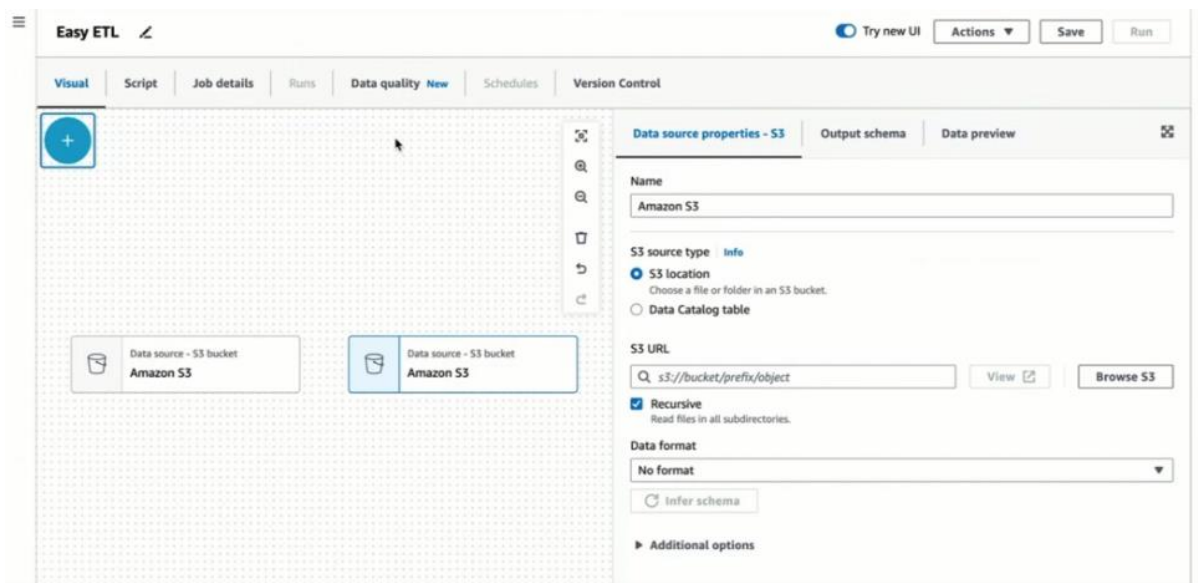


6. NAME IT AS EASY ETL, THEN, CLICK THE PLUS (+) BUTTON TO ADD NOTES.

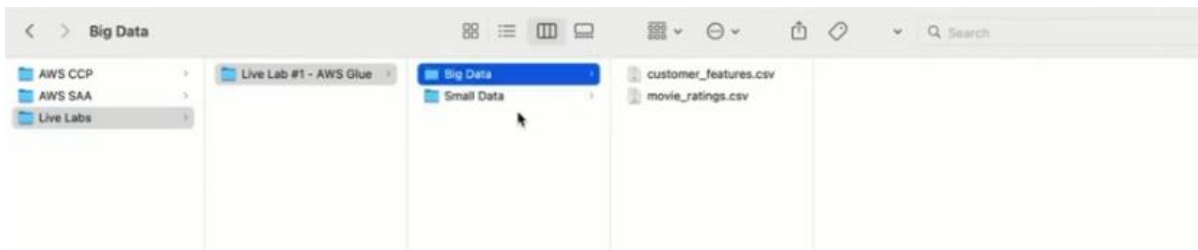


7. CLICK AMAZON S3 – AS OUR DATA SOURCE. DO IT TWICE.

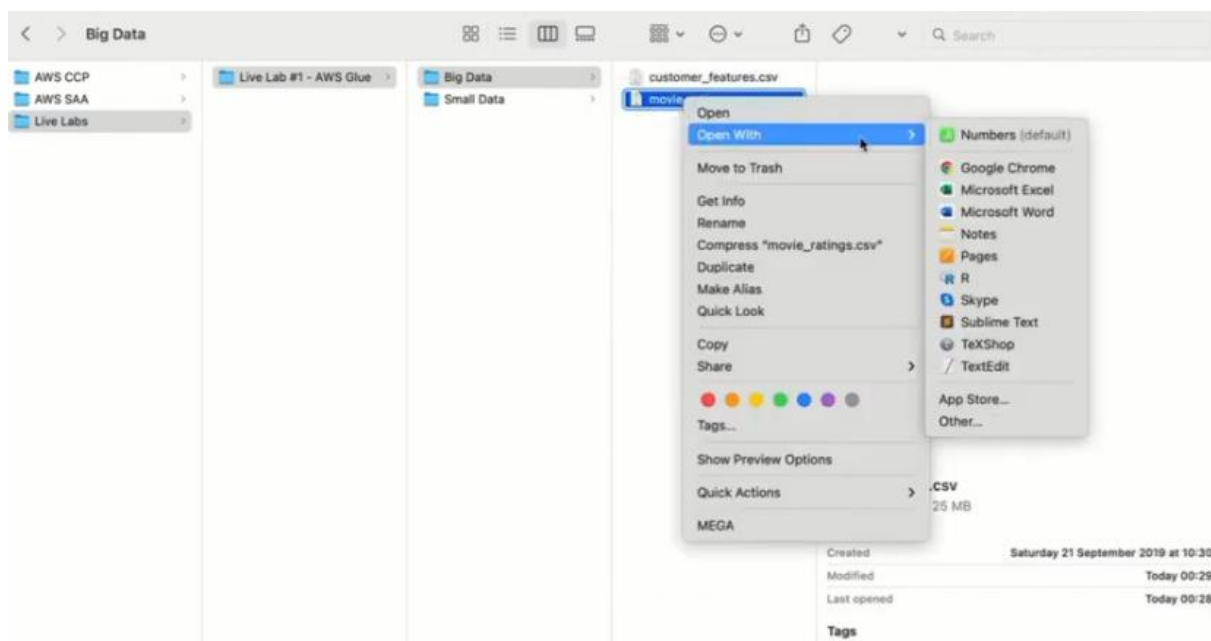




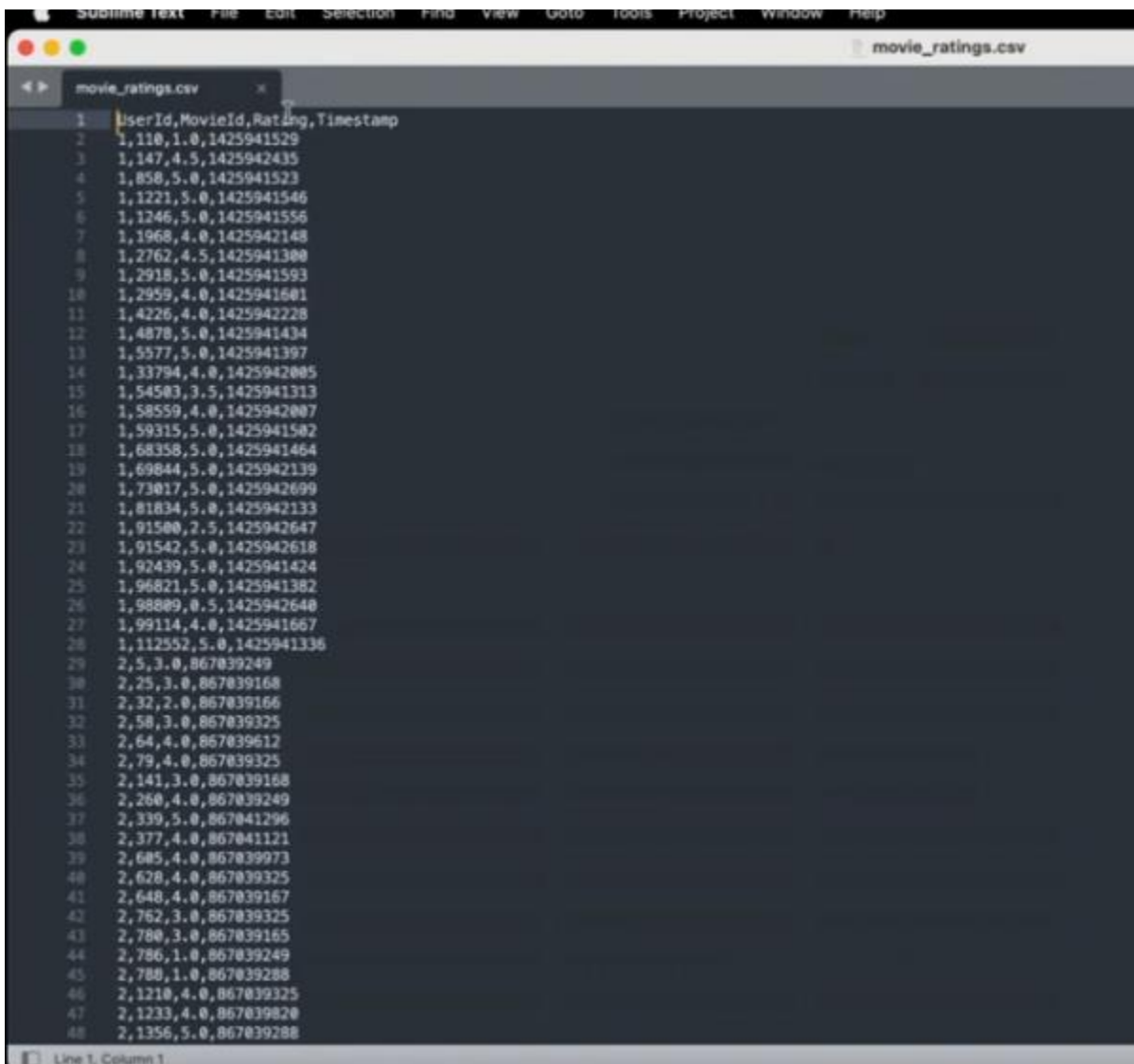
8. GO TO THE SOURCE FOLDER. HERE, WE HAVE THE CUSTOMER AND MOVIE RATINGS SOURCES.



9. OPEN THE MOVIE RATINGS WITH TEXT EDITOR. YOU CAN ALSO OPEN IT WITH EXCEL OR WHATEVER YOU WANT.

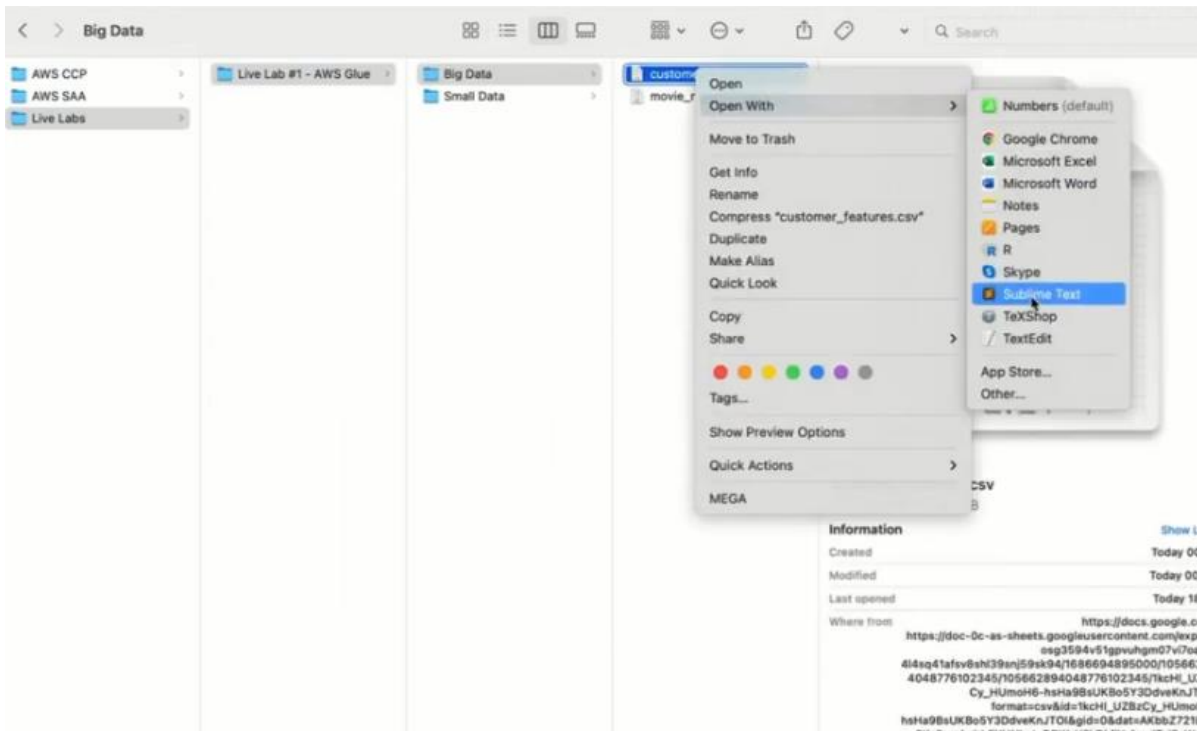


10. HERE, YOU CAN SEE THE CSV FILE CONTAINING FOUR COLUMNS. THE FIRST COLUMN IS USERID, SECOND COLUMN IS MOVIEID, THIRD COLUMN IS THE RATING BY THE USER OF THE MOVIE GOING FROM 1 TO 5, AND THE FOURTH COLUMN IS THE TIMESTAMP. THIS IS OUR FIRST DATABASE.

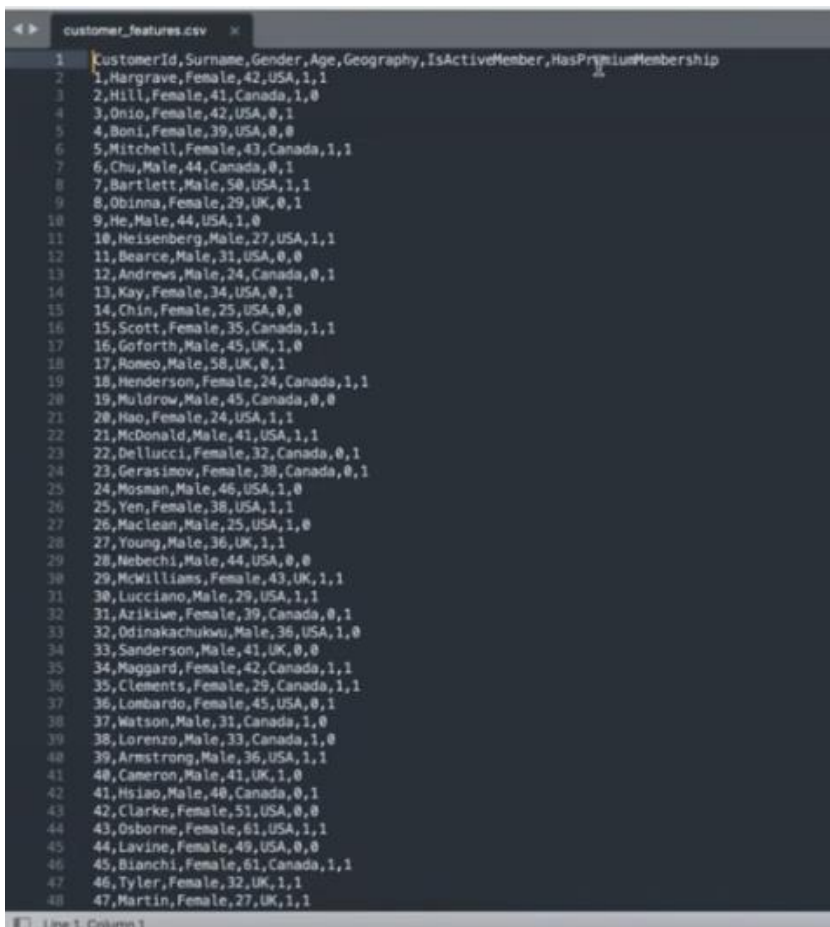


```
Sublime Text  File  Edit  Selection  Find  View  Goto  Tools  Project  Window  Help
movie_ratings.csv
1  UserId,MovieId,Rating,Timestamp
2  1,118,1.0,1425941529
3  1,147,4.5,1425942435
4  1,858,5.0,1425941523
5  1,1221,5.0,1425941546
6  1,1246,5.0,1425941556
7  1,1968,4.0,1425942148
8  1,2762,4.5,1425941300
9  1,2918,5.0,1425941593
10 1,2959,4.0,1425941601
11 1,4226,4.0,1425942228
12 1,4878,5.0,1425941434
13 1,5577,5.0,1425941397
14 1,33794,4.0,1425942005
15 1,54503,3.5,1425941313
16 1,58559,4.0,1425942007
17 1,59315,5.0,1425941502
18 1,68358,5.0,1425941464
19 1,69844,5.0,1425942139
20 1,73017,5.0,1425942699
21 1,81834,5.0,1425942133
22 1,91500,2.5,1425942647
23 1,91542,5.0,1425942618
24 1,92439,5.0,1425941424
25 1,96821,5.0,1425941382
26 1,98809,0.5,1425942640
27 1,99114,4.0,1425941667
28 1,112552,5.0,1425941336
29 2,5,3.0,867039249
30 2,25,3.0,867039168
31 2,32,2.0,867039166
32 2,58,3.0,867039325
33 2,64,4.0,867039612
34 2,79,4.0,867039325
35 2,141,3.0,867039168
36 2,260,4.0,867039249
37 2,339,5.0,867041296
38 2,377,4.0,867041121
39 2,605,4.0,867039973
40 2,620,4.0,867039325
41 2,640,4.0,867039167
42 2,762,3.0,867039325
43 2,780,3.0,867039165
44 2,786,1.0,867039249
45 2,780,1.0,867039288
46 2,1210,4.0,867039325
47 2,1233,4.0,867039820
48 2,1356,5.0,867039288
Line 1, Column 1
```


11. FOR THE SECOND DATABASE, OPEN THE CUSTOMER_FEATURE FILE WITH SUBLIME TEXT.

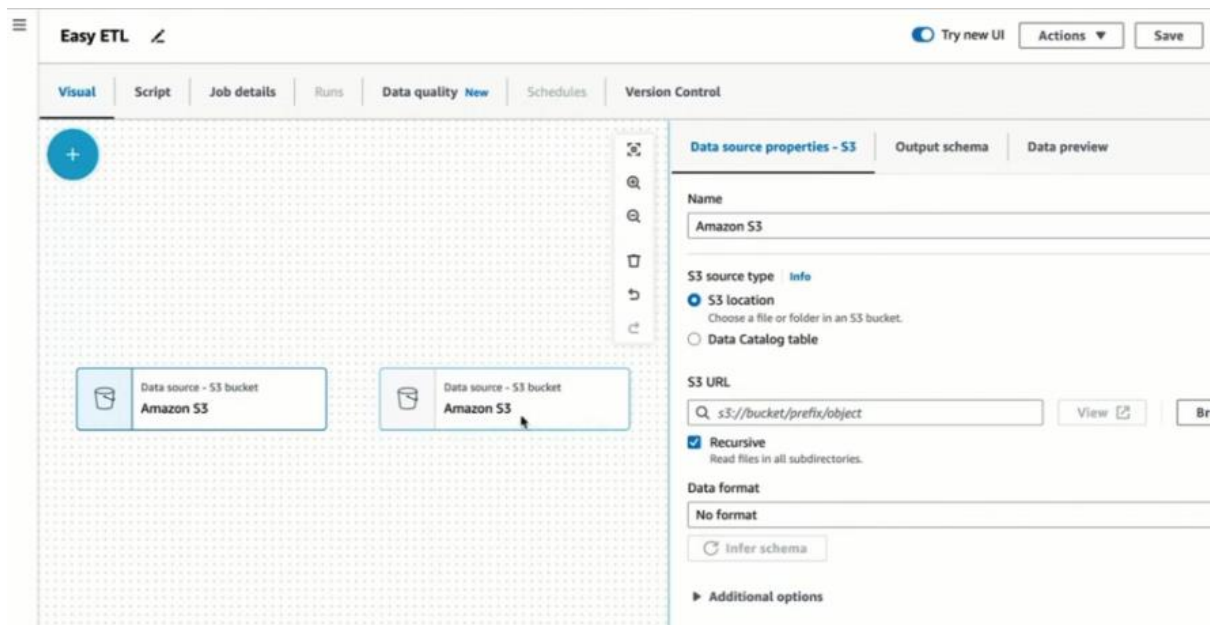


12. HERE, YOU CAN SEE THE CUSTOMER FEATURES SUCH AS THEIR NAME, GENDER, AGE, THE COUNTRY THEY LIVE IN, WHETHER THEY ARE ACTIVE IN THE VIDEO STREAM PLATFORM, AND DO THEY HAVE THE PREMIUM MEMBERSHIP.

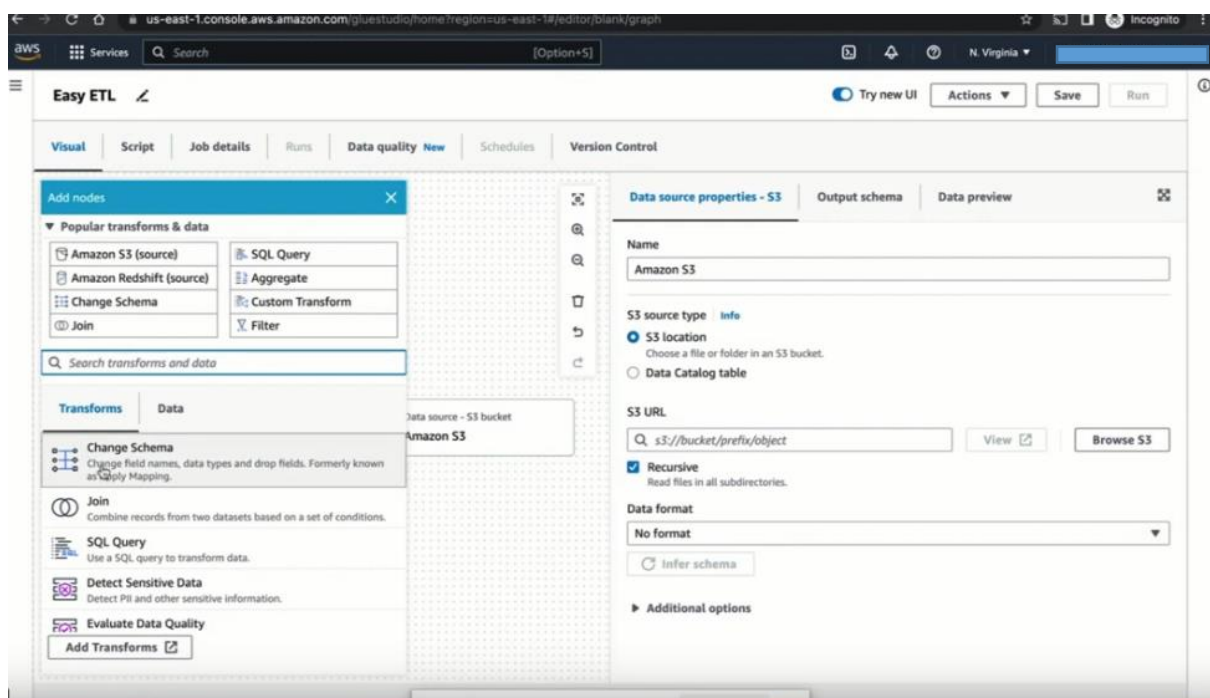


THE CUSTOMERID THAT WE HAVE HERE IS THE SAME AS THE USERID THAT WE HAVE IN THE MOVIE_RATINGS FILE. BUT, AS YOU SEE WE HAVE DIFFERENT DATA SOURCES AND THEY ARE CONNECTED TO THE COMMON DENMINATOR WHICH IS THE USERID IN THE FIRST SOURCE AND THE CUSTOMERID IN THE SECOND SOURCE.

13. GOING BACK TO THE AWS GLUE, WE WILL MAKE THE ETL PROCESS WHEREIN THE EASY ETL WE HAVE TWO DIFFERENT DATA SOURCES AS DATA WERE COLLECTED FROM DIFFERENT DEPARTMENTS COMING BOTH FROM AMAZON S3.



14. CLICK THE PLUS (+) BUTTON. IN EVERY ETL PROCESS, WE HAVE TO DO SOME CLEANING OF THE DATA. IN THE TRANSFORMS FIELD, THERE IS A CHANGE SCHEMA WHICH CAN CHANGE FIELDNAMES, DATA TYPES, AND DROP FIELDS. ALSO COMMONLY KNOWN AS APPLY MAPPING. CLCIK CHANGE SCHEMA.



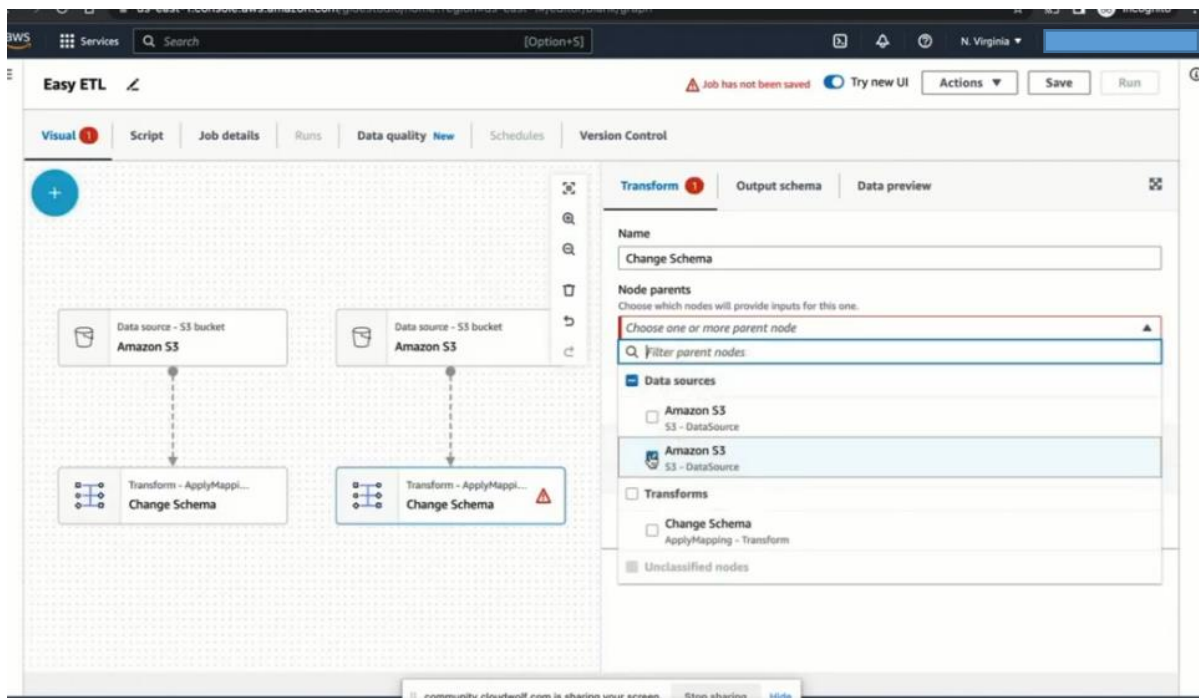
15. IN THE NODE PARENTS, CHOOSE AMAZON S3.

The screenshot shows the Easy ETL interface. The top navigation bar includes 'Visual', 'Script', 'Job details', 'Runs', 'Data quality New', 'Schedules', and 'Version Control'. The main canvas displays a workflow with two 'Data source - S3 bucket Amazon S3' nodes and a 'Transform - Apply Mapping Change Schema' node. The right sidebar is open to the 'Transform' tab, showing the 'Name' field set to 'Change Schema'. Under the 'Node parents' section, the instruction 'Choose one or more parent node' is followed by a search bar containing 'Filter parent nodes'. Below this, the 'Data sources' list shows 'Amazon S3 S3 - DataSource' selected with a checkmark, and 'Amazon S3 S3 - DataSource' listed below it without a checkmark. The 'Transforms' and 'Unclassified nodes' sections are currently empty.

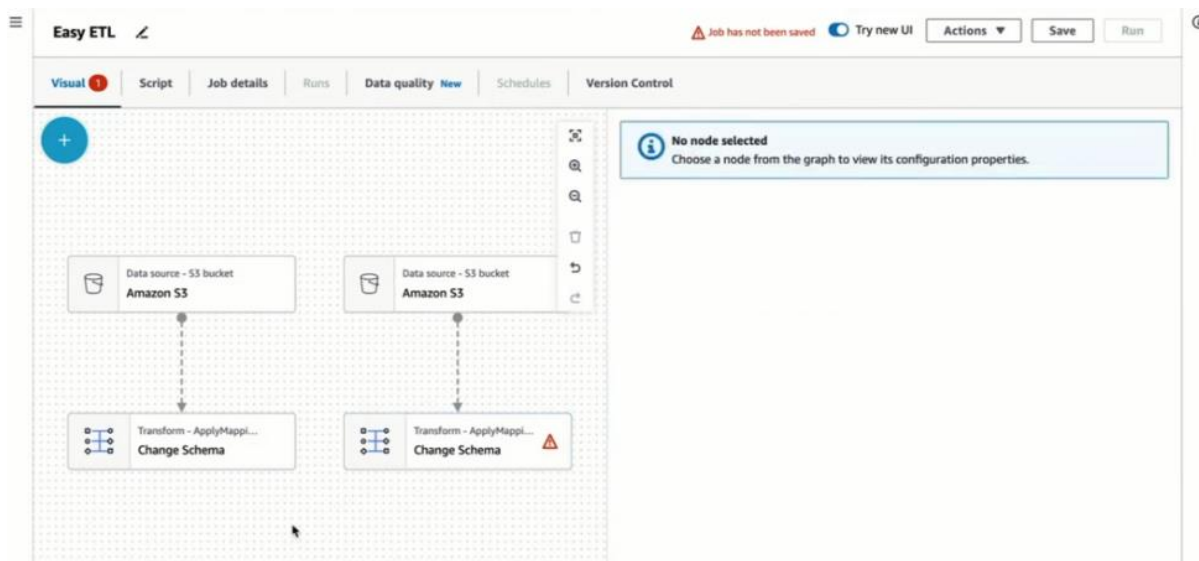
16. DO THE SAME FOR THE SECOND ONE. CLICK CHANGE SCHEMA.

The screenshot shows the Easy ETL interface with the 'Add nodes' panel open on the left. The panel is titled 'Add nodes' and has a search bar. Below the search bar, there are two tabs: 'Transforms' and 'Data'. Under the 'Transforms' tab, several options are listed: 'Change Schema' (with a description: 'Change field names, data types and drop fields. Formerly known as Apply Mapping.'), 'Join' (with a description: 'Combine records from two datasets based on a set of conditions.'), 'SQL Query' (with a description: 'Use a SQL query to transform data.'), 'Detect Sensitive Data' (with a description: 'Detect PII and other sensitive information.'), and 'Evaluate Data Quality'. The 'Change Schema' option is highlighted. The main canvas shows a workflow with two 'Data source - S3 bucket Amazon S3' nodes. The right sidebar is open to the 'Transform' tab, showing a message: 'No node selected. Choose a node from the graph to view its configuration properties.'

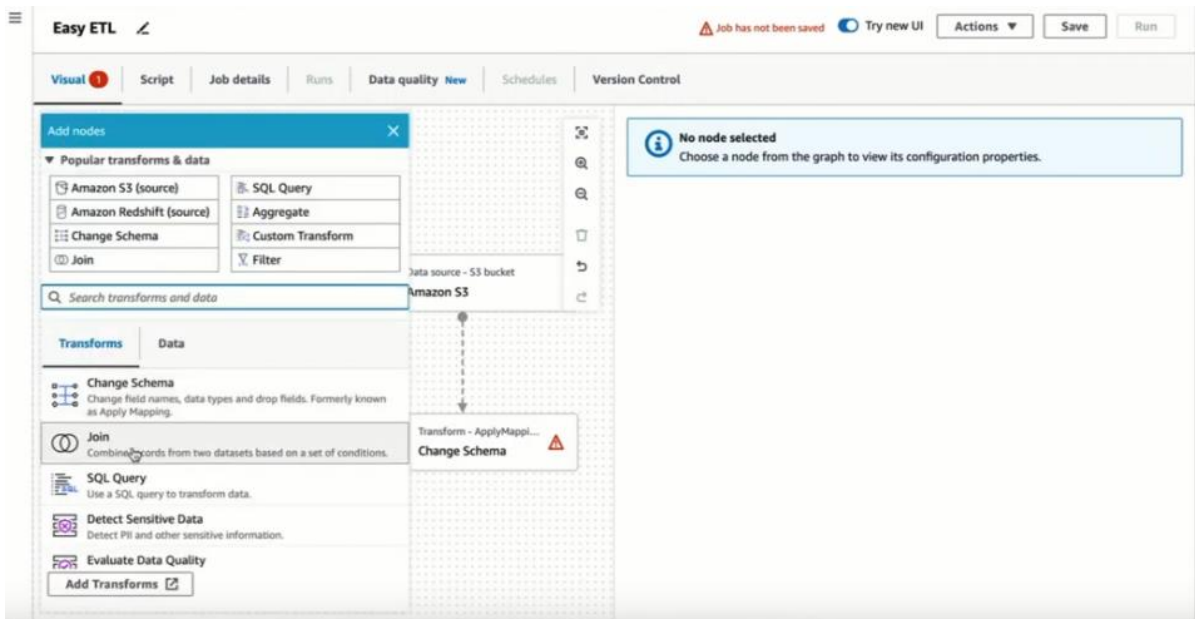
17. IN THE NODE PARENTS. CHOOSE THE SECOND AMAZON S3.



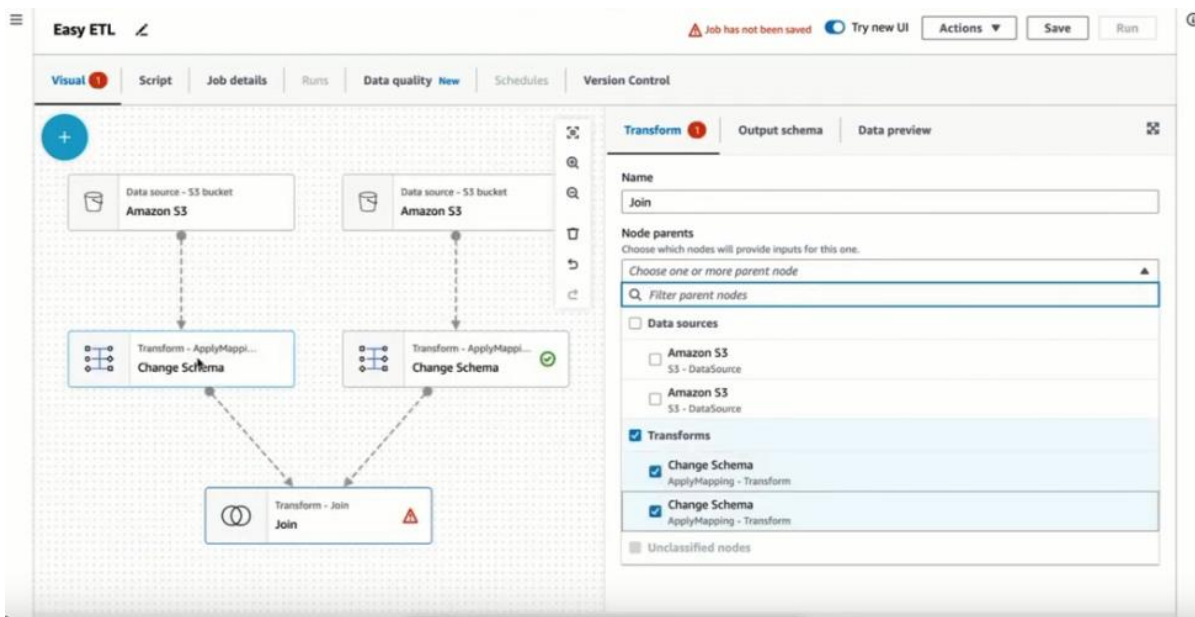
18. THE EXPECTED OUTCOME WILL BE SHOWN LIKE THIS. REMEMBER, WE WANT TO JOIN THE TWO DATA SOURCES. TO DO THIS, CLICK THE PLUS (+) BUTTON.



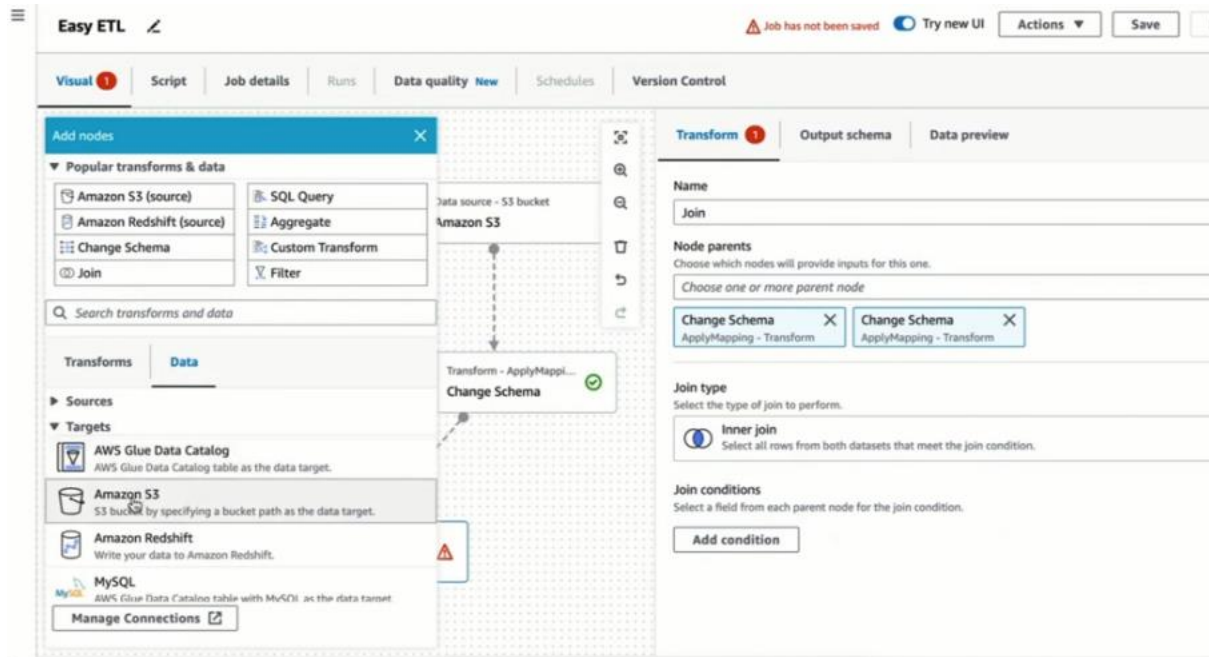
19. CHOOSE JOIN IN THE TRANSFORMS FIELD.



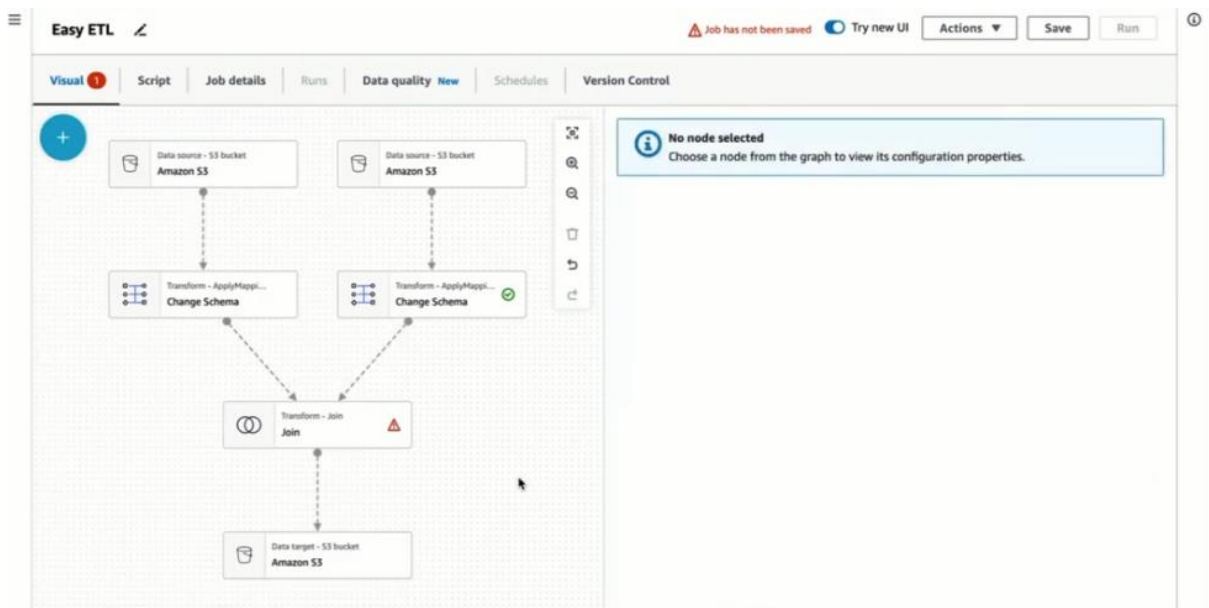
20. IN THE NODE PARENTS, SELECT THE CHANGE SCHEMA.



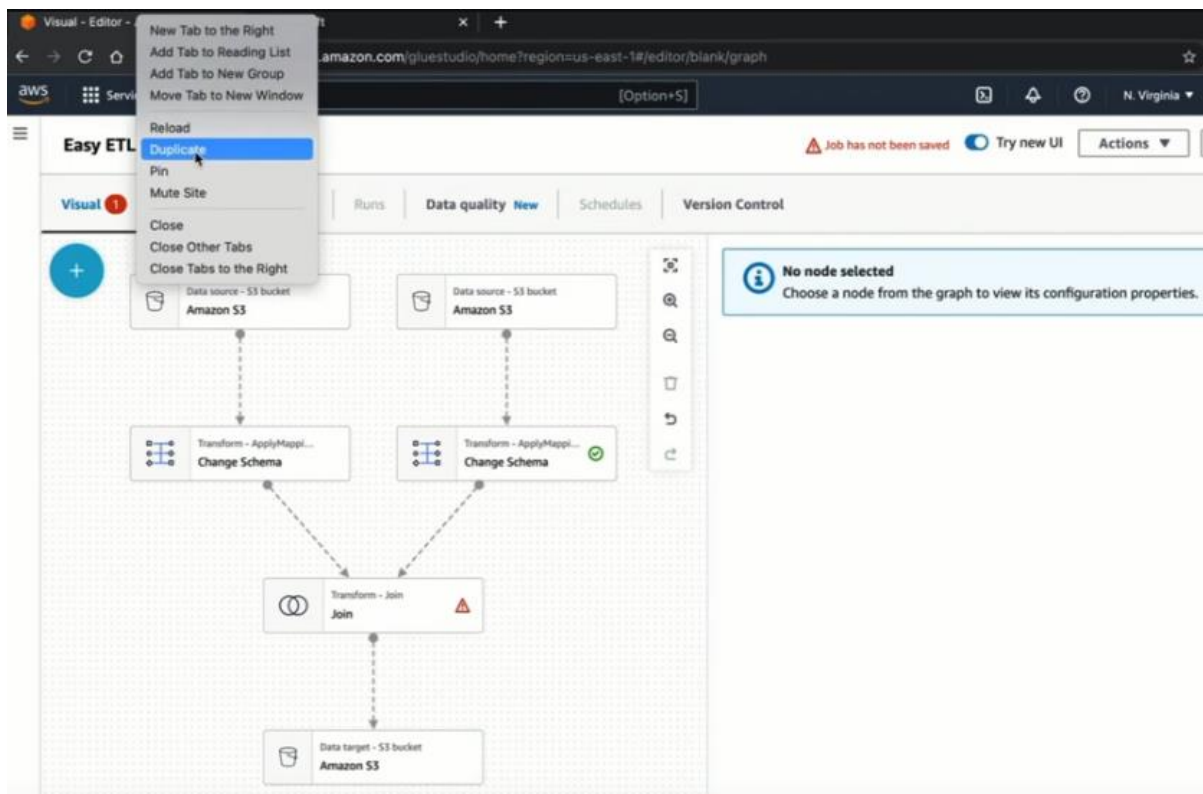
21. FINALLY WE HAVE TO CHOOSE THE DATA TARGET. CLICK DATA, THEN, TARGET, THEN, CHOOSE, AMAZON S3.



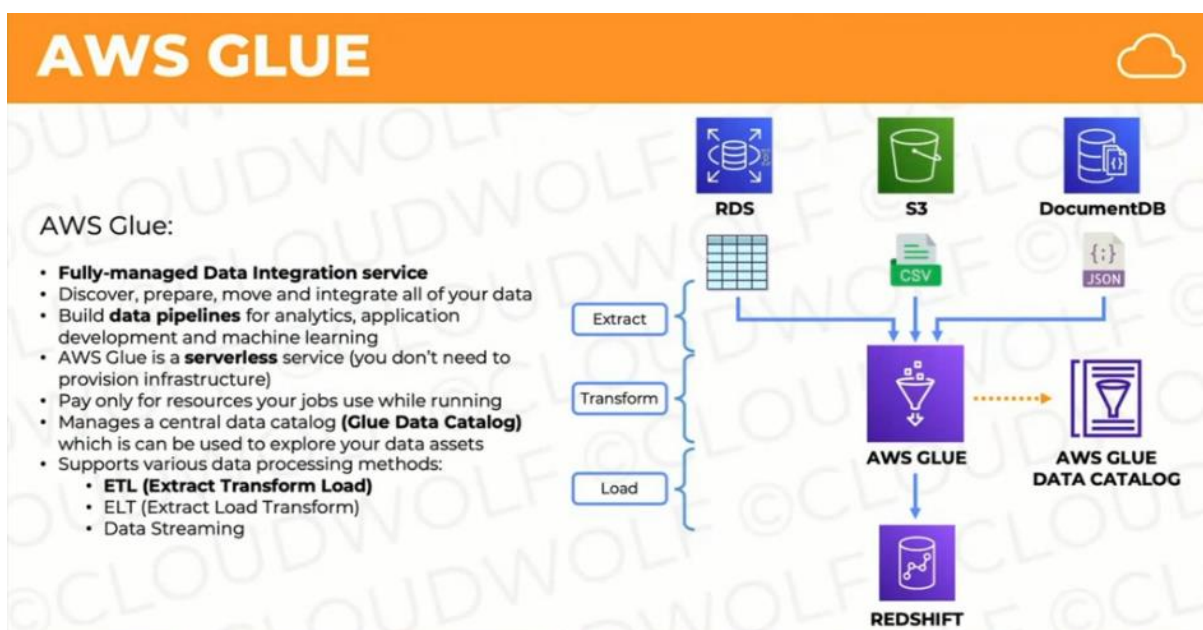
22. AND NOW THIS IS OUR FIRST ETL PROCESS. THE EASY EXTRACT, TRANSFORM, LOAD PROCESS.



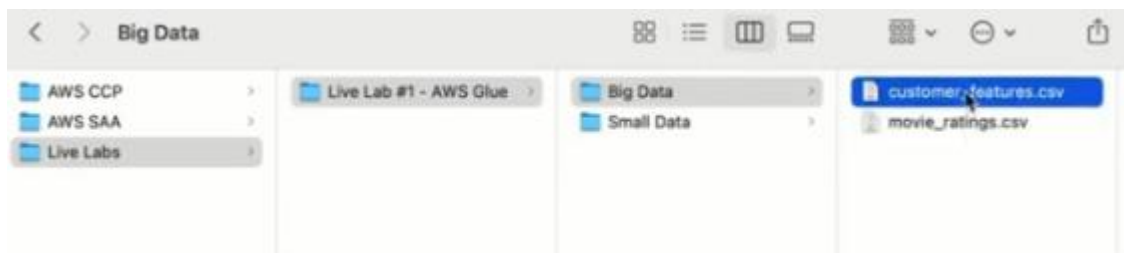
23. NOW, WE WILL DO A HARD ETL. BUT FIRST, LET US DUPLICATE THE TABLE.



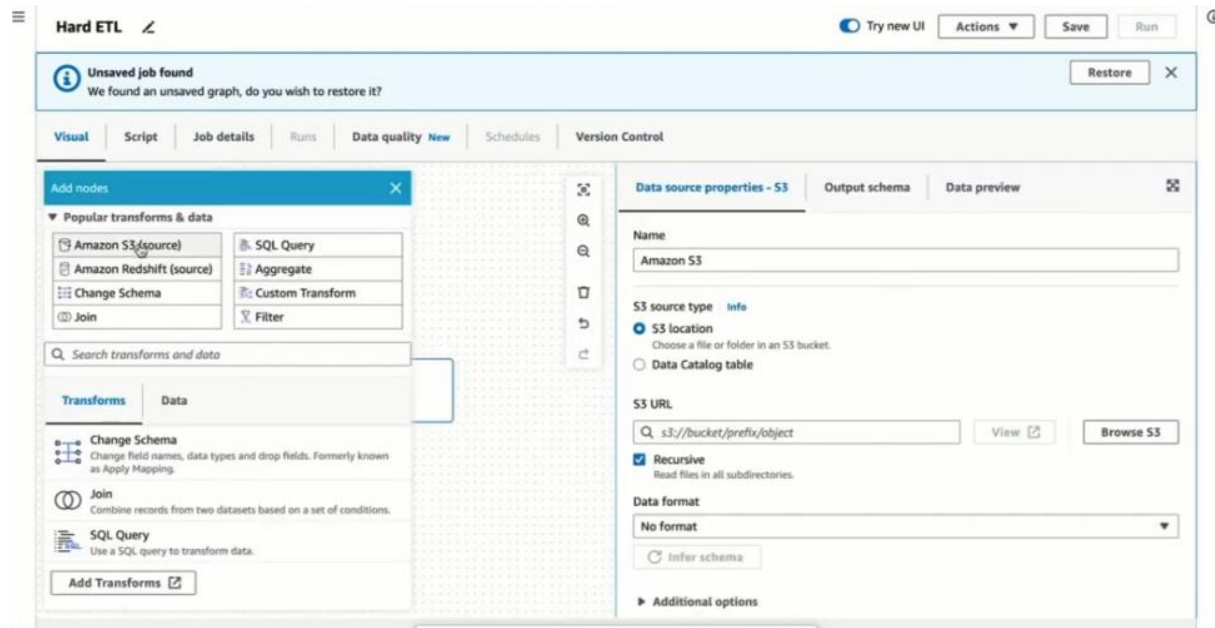
24. WE WILL MAKE THE HARD ETL FOLLOWING THIS DIAGRAM. ONE SOURCE COMING FROM S3, THE OTHER COMING FROM RDS, THEN, THE AWS GLUE ETL PROCESS. FINALLY, THE TARGET OUTPUT IS THE REDSHIFT, NOT THE S3. REDSHIFT IS THE VERY WELL ORGANIZED DATA WAREHOUSE WHERE YOU CAN HAVE EVERYTHING MUCH BETTER ORGANIZED RATHER THAN HAVING DIFFERENT DATA SOURCES WHERE EVERYTHING IS SEPARATE.



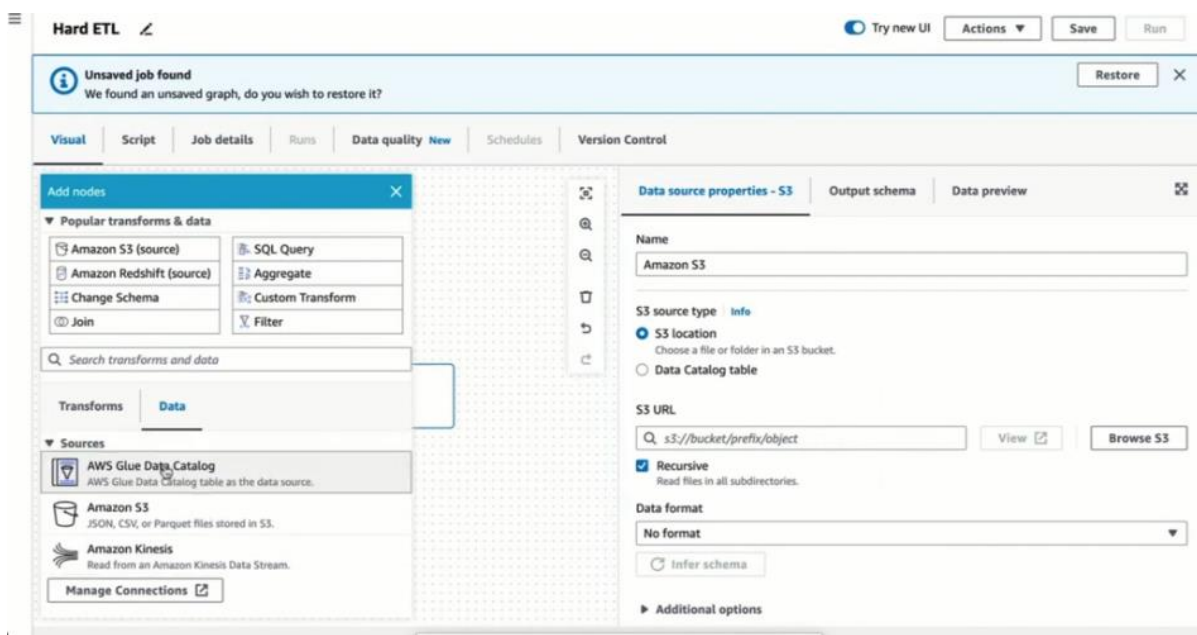
25. LET'S SAY OUR MOVIE RATINGS COMES FROM S3, AND THE CUSTOMER_FEATURES COMES FROM RDS.



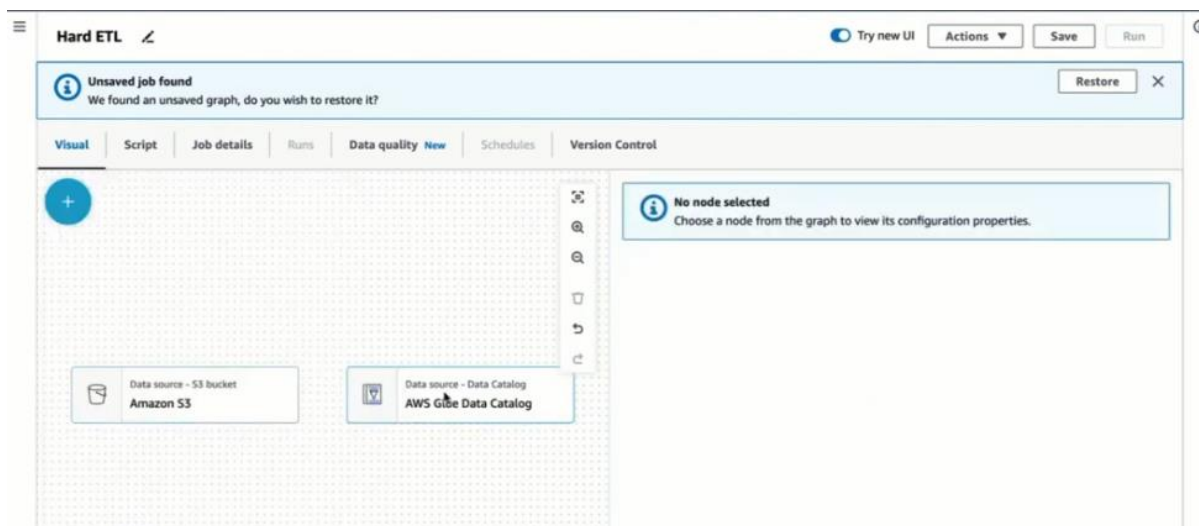
26. MAKE THE FIRST VISUAL. CLICK ADD NODES, CHOOSE S3 FOR THE MOVIE_RATINGS.



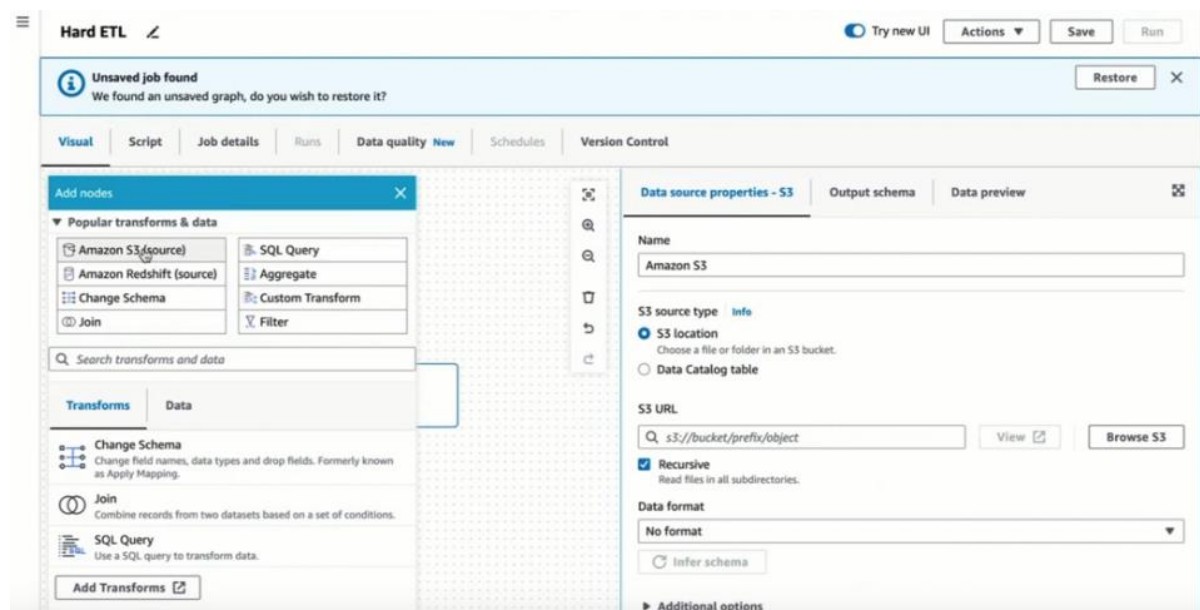
27. FOR THE SECOND ONE, IT IS TO GOING TO BE RDS. IN THE DATA FIELD, CHOOSE AWS GLUE DATA CATALOG.



THIS IS NOW OUR TWO DATA SOURCES, THE S3 AND THE RDS.



28. THEN AGAIN, APPLY THE TRANSFORMATION TO CLEAN THE DATA. MAYBE TO DROP SOME FEATURES THAT MIGHT BE COMPLETELY IRRELEVANT. TO MAKE THE RECOMMENDER SYSTEM. CLICK THE CHANGE SCHEMA IN THE TRANSFORMS FIELD. DO IT TWICE, ONE FOR THE S3, AND FOR THE AWS GLUE DATA CATALOG (RDS).



Hard ETL

Try new UI Actions Save Run

Unsaved job found
We found an unsaved graph, do you wish to restore it? Restore

Visual Script Job details Runs Data quality New Schedules Version Control

Add nodes

Popular transforms & data

- Amazon S3 (source)
- Amazon Redshift (source)
- Change Schema
- Join
- SQL Query
- Aggregate
- Custom Transform
- Filter

Search transforms and data

Transforms Data

Change Schema
Change field names, data types and drop fields. Formerly known as Apply Mapping.

Join
Combine records from two datasets based on a set of conditions.

SQL Query
Use a SQL query to transform data.

Add Transforms

Data source - Data Catalog
AWS Glue Data Catalog

Transform - ApplyMapping - Change Schema

Transform

Name
Change Schema

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node

AWS Glue Data Catalog
Catalog - DataSource

Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
Select your source and target database/table sources.			

29. THEN, DO THE JOIN PROCEDURE. CLICK JOIN.

Unsaved job found
We found an unsaved graph, do you wish to restore it? Restore

Visual Script Job details Runs Data quality New Schedules Version Control

Add nodes

Popular transforms & data

- Amazon S3 (source)
- Amazon Redshift (source)
- Change Schema
- Join
- SQL Query
- Aggregate
- Custom Transform
- Filter

Search transforms and data

Transforms Data

Change Schema
Change field names, data types and drop fields. Formerly known as Apply Mapping.

Join
Combine records from two datasets based on a set of conditions.

SQL Query
Use a SQL query to transform data.

Add Transforms

Data source - Data Catalog
AWS Glue Data Catalog

Transform - ApplyMapping - Change Schema

Transform - Join - Join

Transform

Name
Join

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node

Change Schema
ApplyMapping - Transform

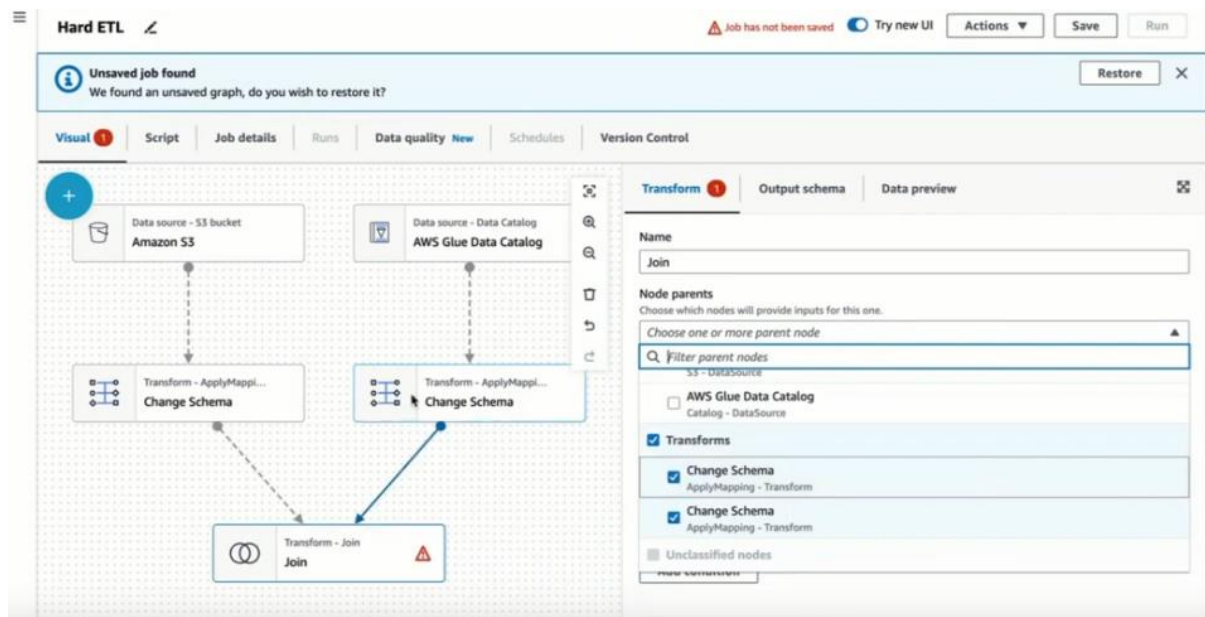
Join type
Select the type of join to perform.

Inner join
Select all rows from both datasets that meet the join condition.

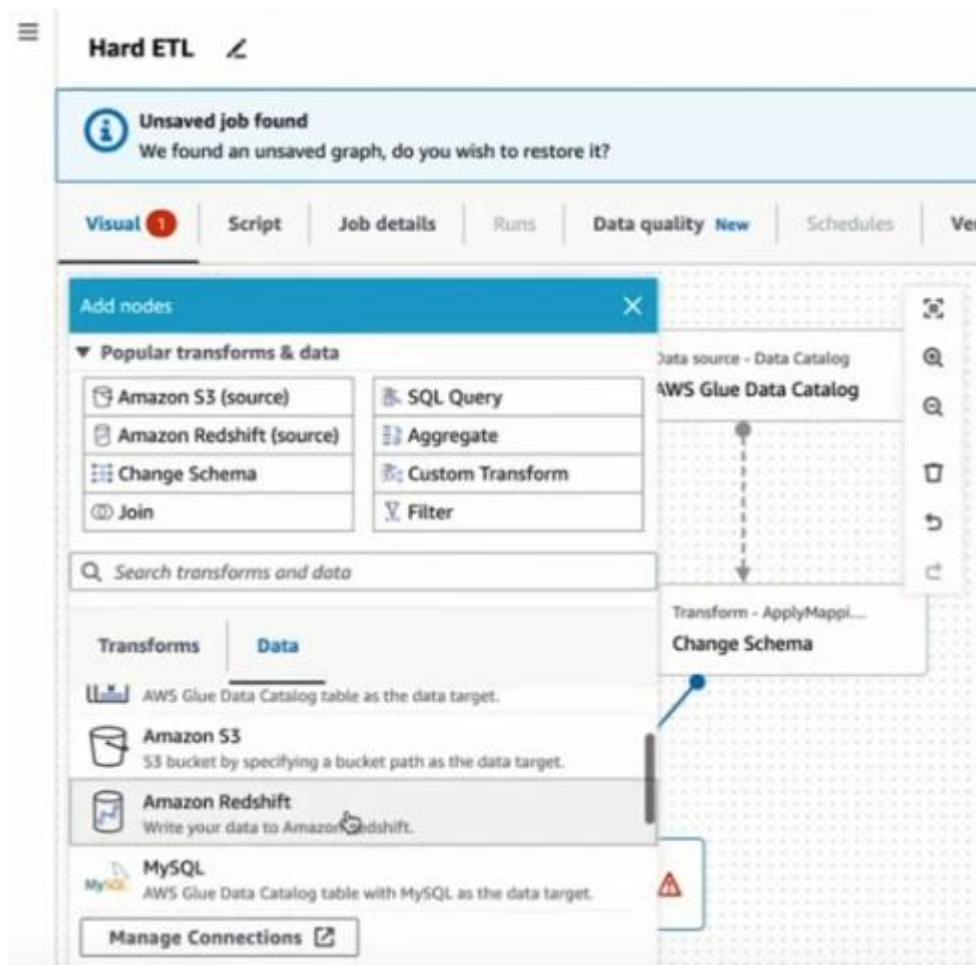
Join conditions
Select a field from each parent node for the join condition.

Insufficient source nodes
The Join transform requires two parent source nodes with selected tables.

30. IN THE NODE PARENTS, SELECT TRANSFORMS – CHANGE SCHEMA.



31. FINALLY, AS THE TARGETS, WE WILL CHOOSE THE REDSHIFT. CLICK PLUS (+) BUTTON. CLICK DATA, CHOOSE AMAZON REDSHIFT.



32. THIS IS NOW OUR HARD ETL PROCESS.

