Usecase 1: Dataset Data Integration using ML

A use case to match records across different datasets as well as identify and remove duplicate records, with little to no human intervention.

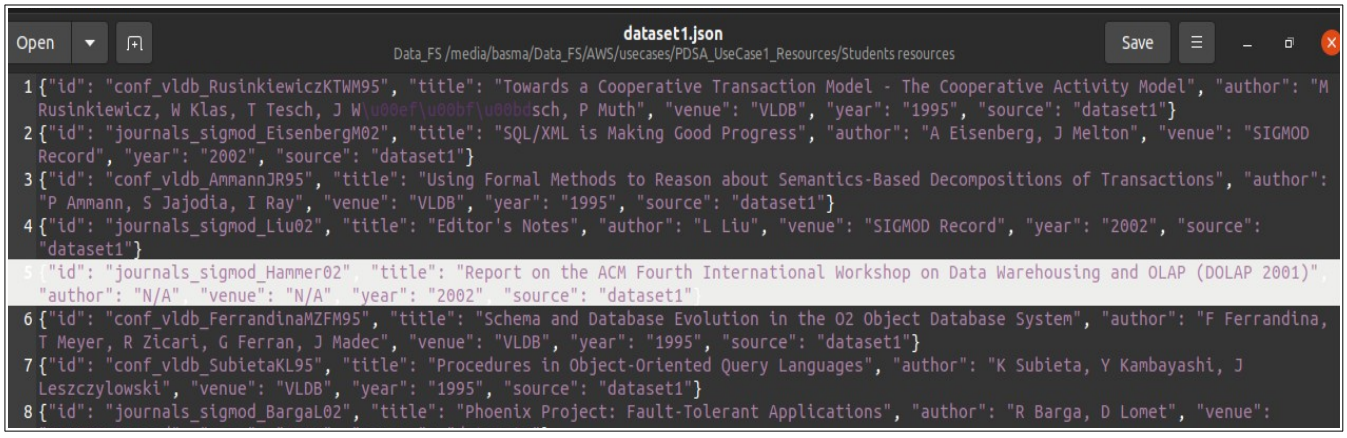
Input

- Neat Dataset1 with some records (~2600).
- Dataset2 with some records (~64,000). Messy,duplicated.
- Matching records (labels) (~130)

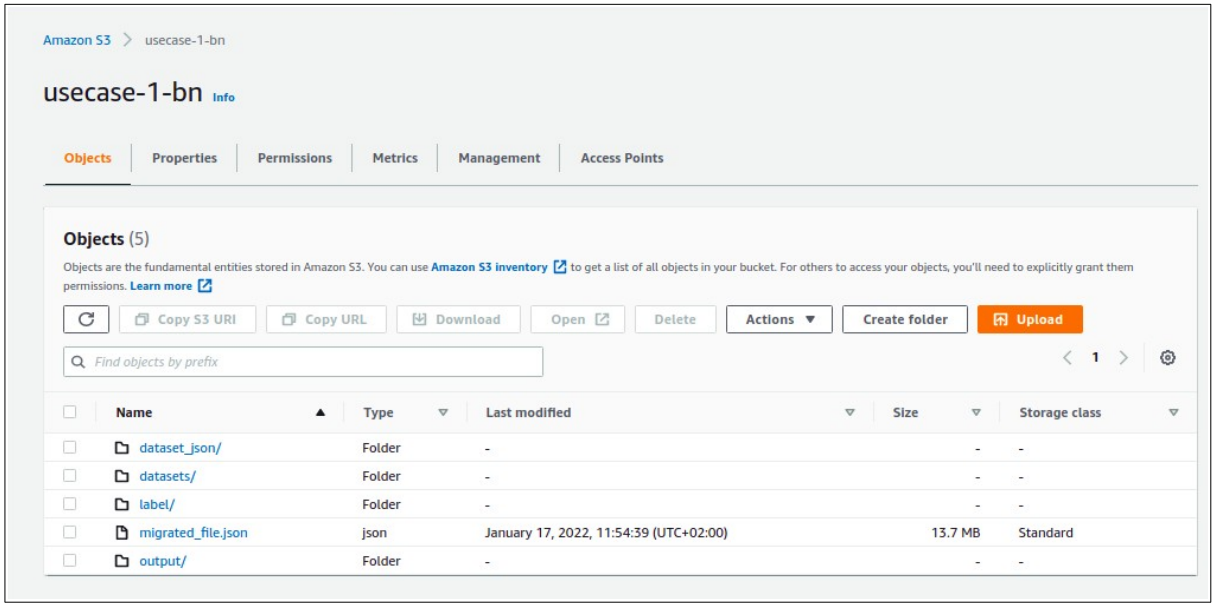
Steps

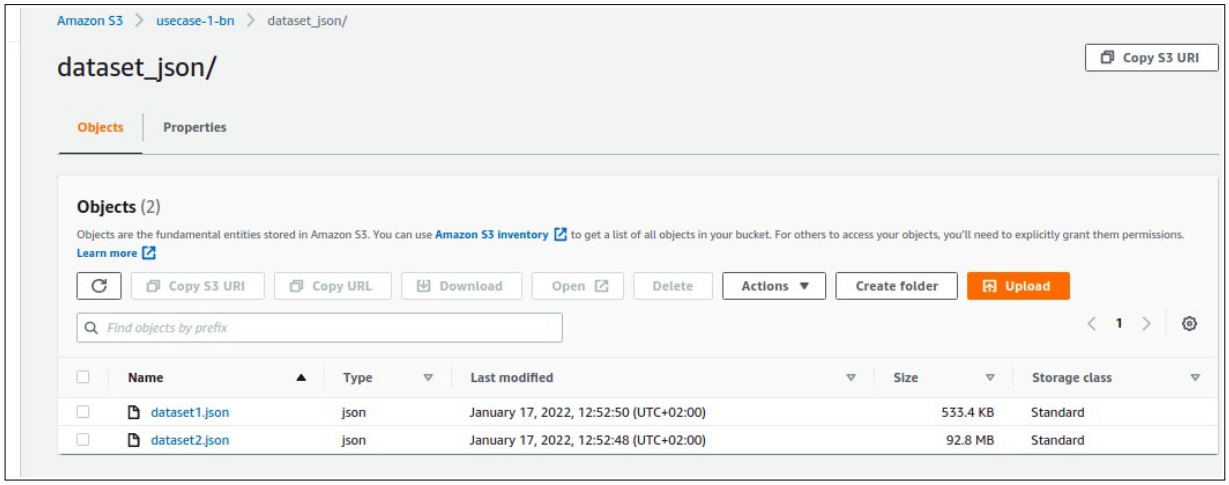
Reformat files → Create IAM role → upload to S3 bucket → Athena datasource → data glue ML → data glue job → get output file from S3

1. Change dataset format to JSON file with encoding **‘utf-8’** and added one column called source to each dataset file using python script.

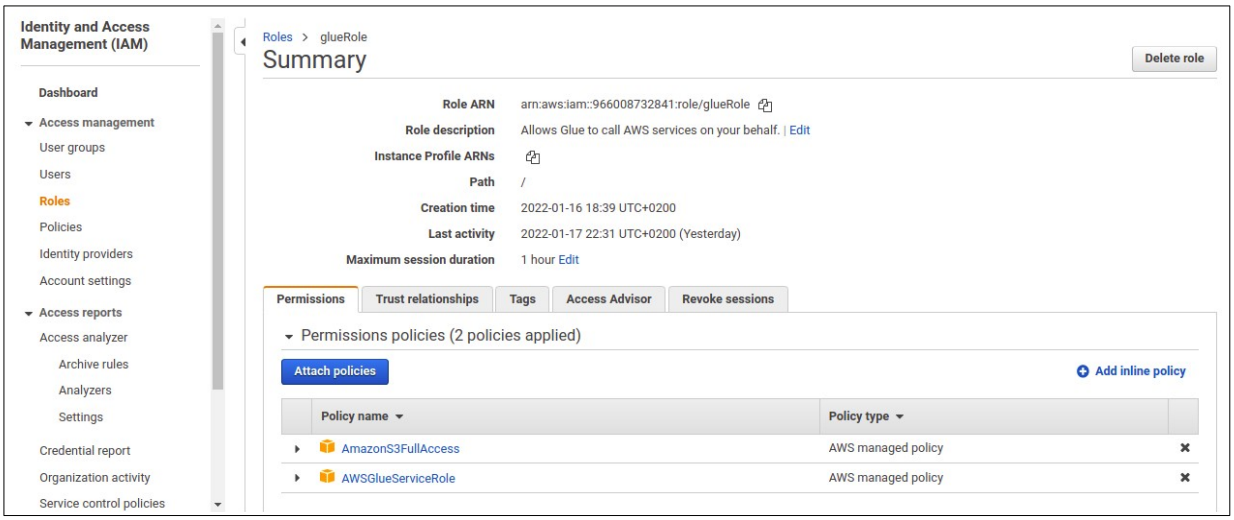


2. Created S3 bucket and uploaded 2 dataset files and label file in CSV format.

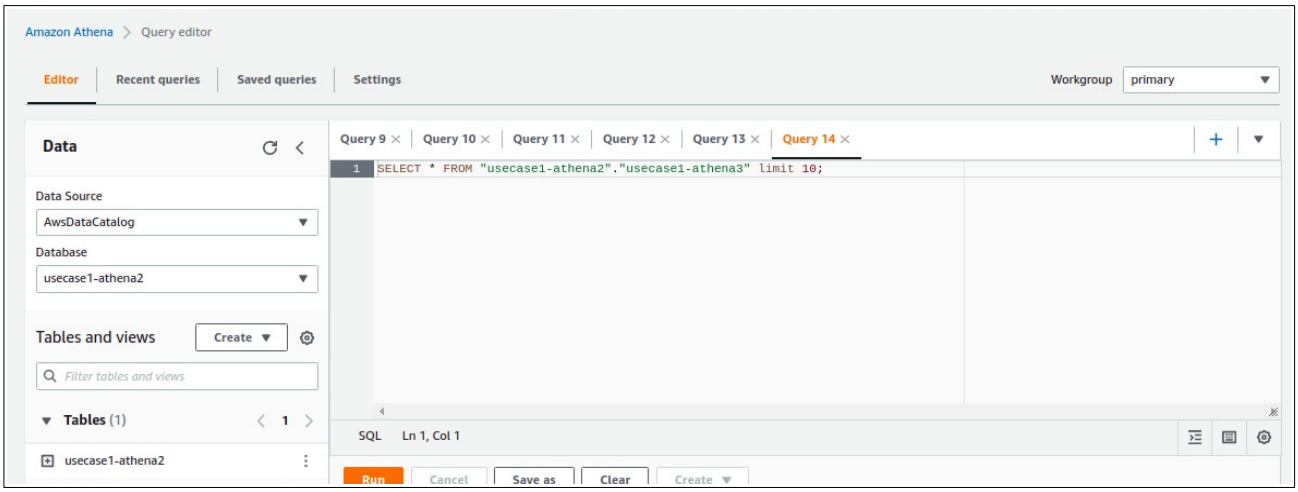




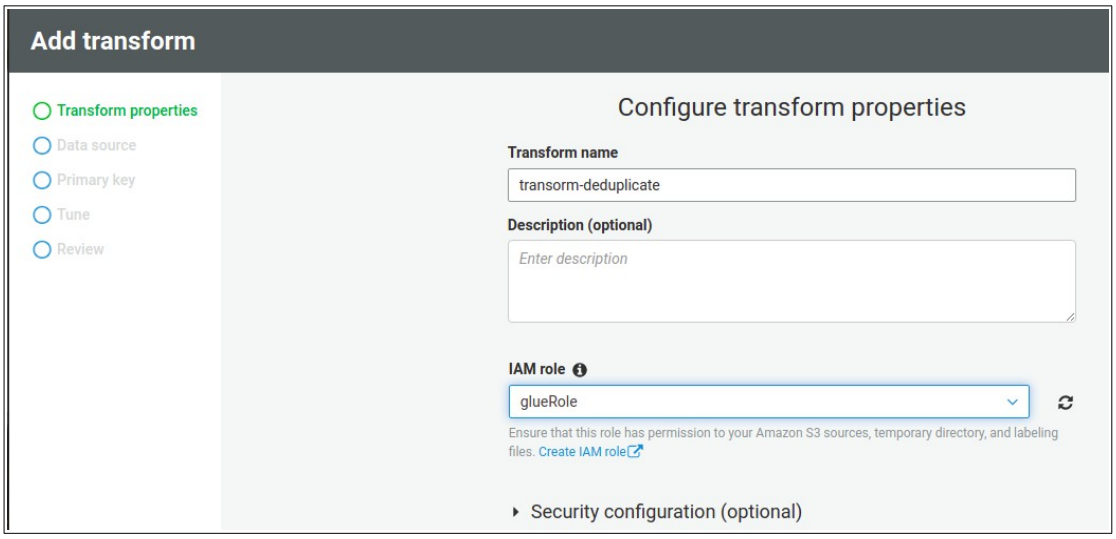
3. created new IAM role for S3 and data glue

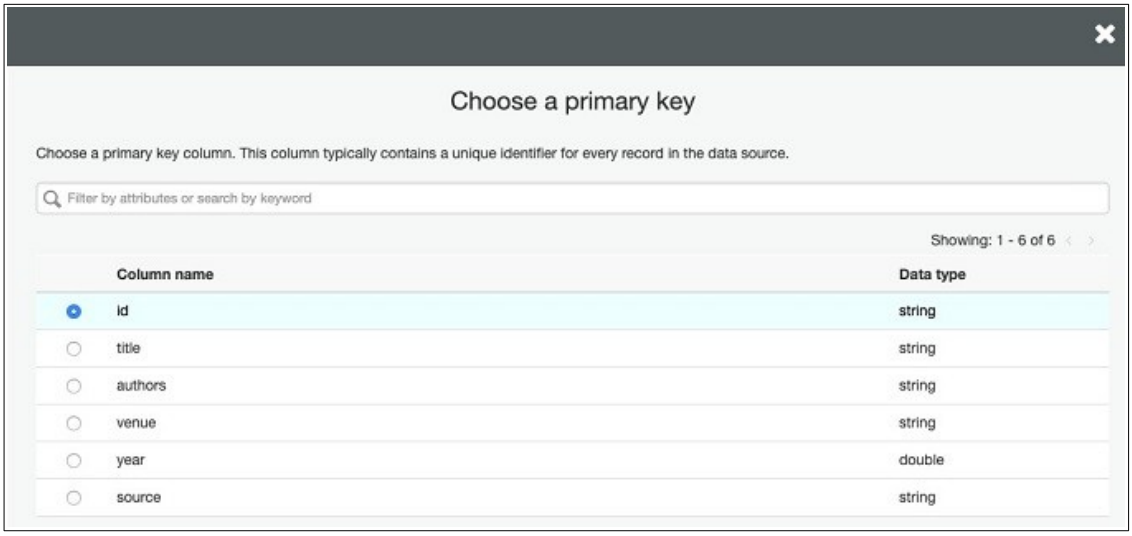


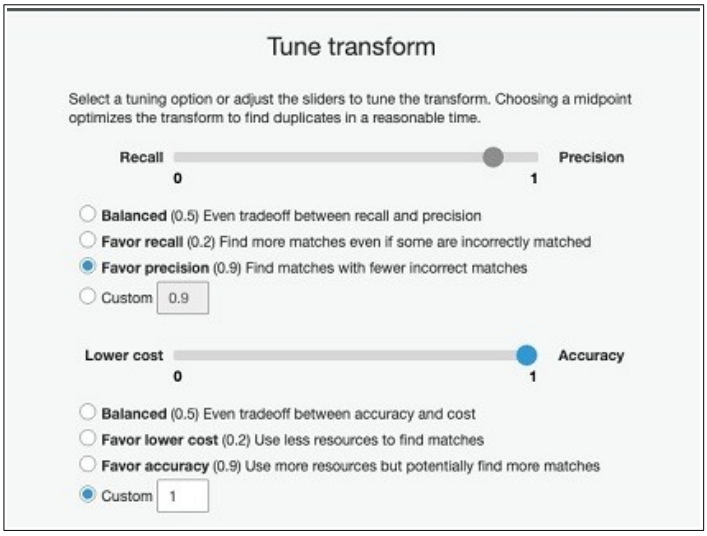
4. Created new database using Athena to merge 2 datasets in one table.

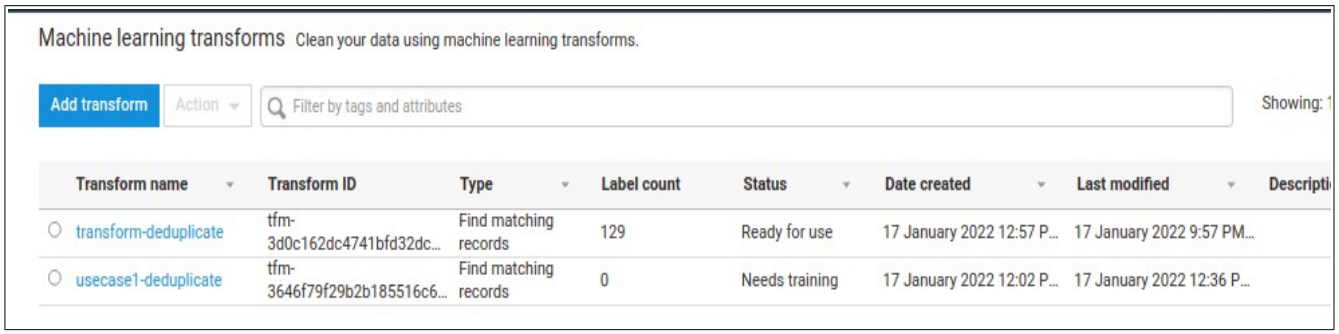


5. From data Glue, created a new FindMatches ML transform with 100% accuracy and 0.9 precision.









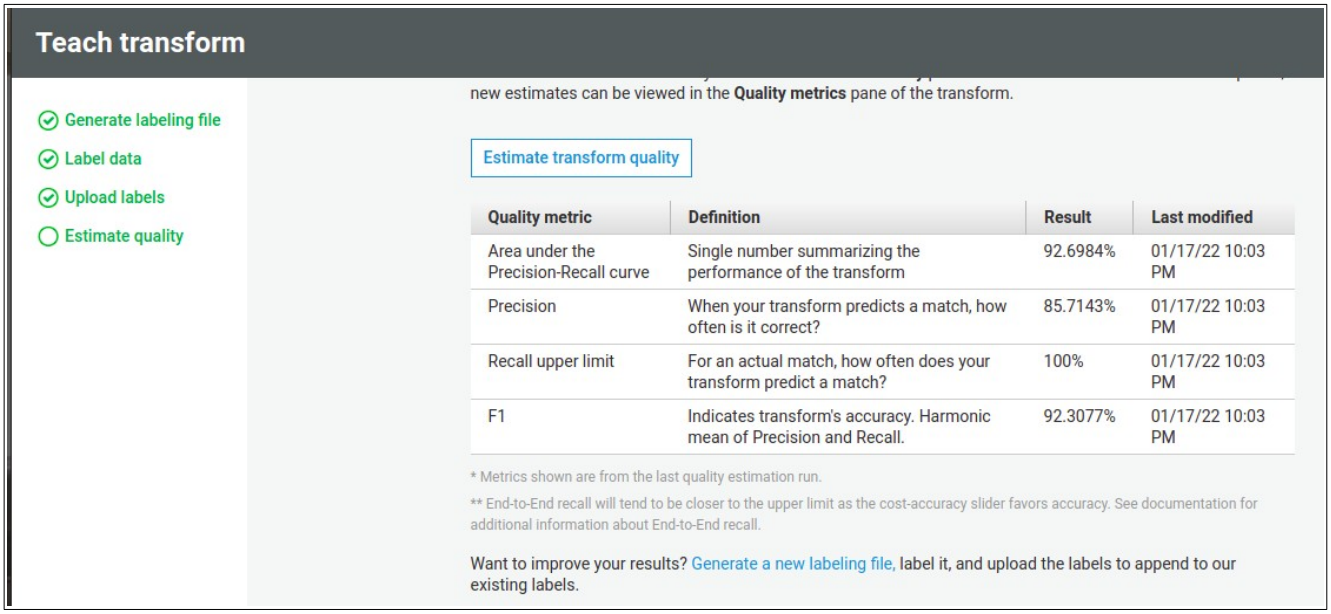
6. Generated sample label file from glue by selecting “I don’t have labels” option.

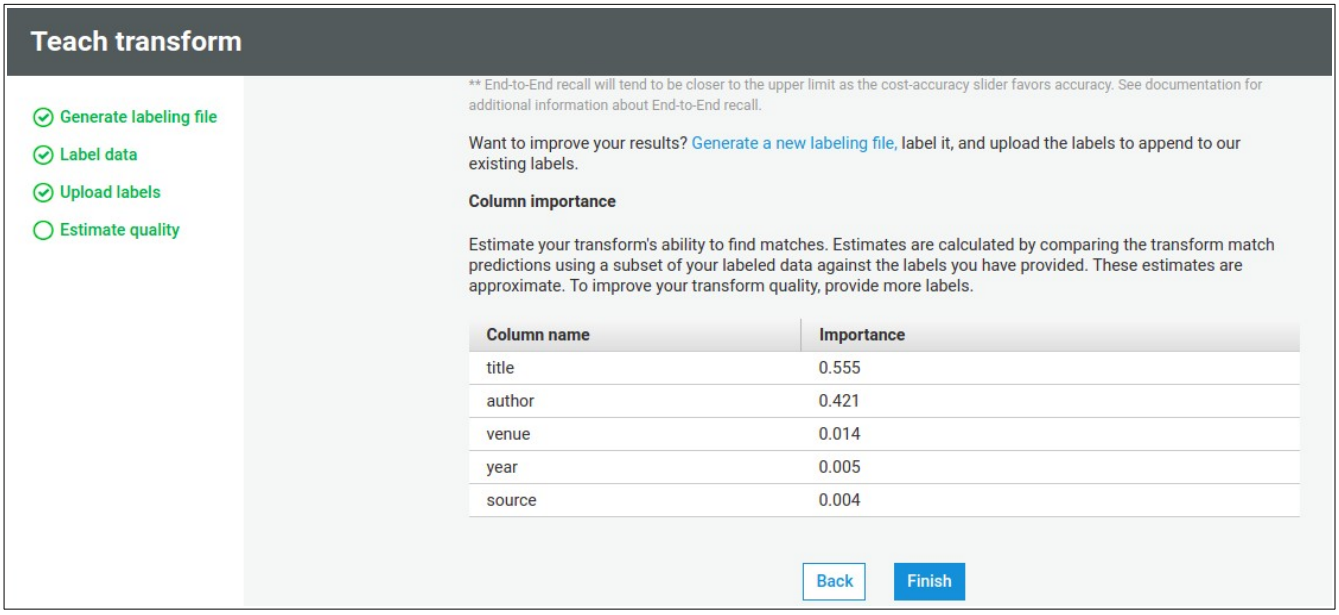
   Then added values to label column and added some matching labels from the shared label file. Then uploaded the new label file to S3.

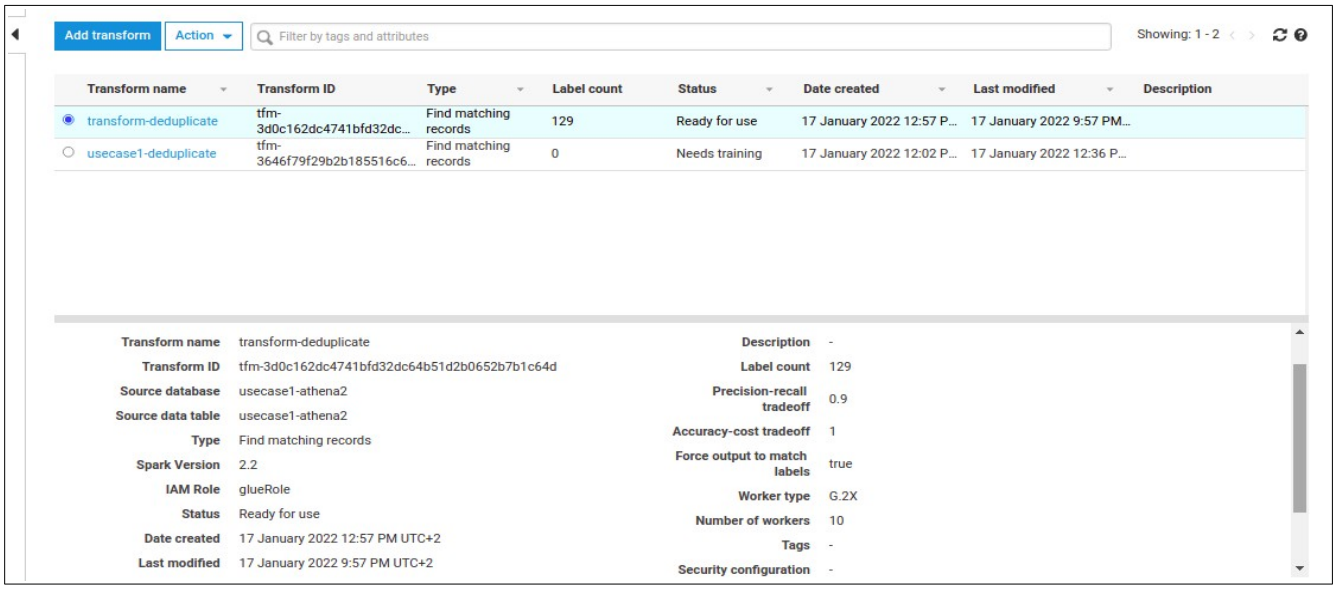
7. Teach FindMatches by providing labeling records from S3.

A	B	C	D	E	F	G	H
labeling_set_id	label	id	title	author	venue	year	source
f570eef9-5ba6-35bd-a393-436e80d1beb6	0	eBnT7IhV2LwJ	Aurora: A Data Stream	D Abadi, D Carney, U Cetii	Proceedings of the 2003 ACM SIGMOD Conference o		dataset2
f570eef9-5ba6-35bd-a393-436e80d1beb6	0	conf_sigmod_Aba	Aurora: A Data Stream	D Abadi, D Carney, U Ã¢t	SIGMOD Conference	2003	dataset1
f570eef9-5ba6-35bd-a393-436e80d1beb6	0	gBVNSFeS4P8J	Aurora: a new model an	DJ Abadi, D Carney, U Ã?	The VLDB Journal The International Journal	2003	dataset2
f570eef9-5ba6-35bd-a393-436e80d1beb6	0	VuY9Y49GqXgJ	Aurora: A Data Stream	DJ Abadi, D Carney, U Cetintemel, M Cherniack, C			dataset2
f570eef9-5ba6-35bd-a393-436e80d1beb6	1	14XNgQag2nQJ	Improving Memory in O	B Levy	JOURNAL OF PERSONALITY AND SOCIA	1996	dataset2
f570eef9-5ba6-35bd-a393-436e80d1beb6	2	0_Lktn3djgAJ	Exploring Windows NT	R Duncan	PC		dataset2
f570eef9-5ba6-35bd-a393-436e80d1beb6	3	d9u9G1duIXgJ	Word for Word	D Haskins	PC		dataset2
f570eef9-5ba6-35bd-a393-436e80d1beb6	4	yEIHjCL5UykJ	Evolution and Change in	ACMS Anthology	SIGMOD Record,	2000	dataset2
f570eef9-5ba6-35bd-a393-436e80d1beb6	5	owCJmT0Mv2IJ	Impulse Detectors for N	R Lukac	RADIOENGINEERING-PRAGUE-,	2001	dataset2
f570eef9-5ba6-35bd-a393-436e80d1beb6	6	qFmPKSqVCEUJ	The common capability	B Levy	BT Technology Journal,	2005	dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	0	AxpQwgyRyLgJ	Active XML Documents	S Abiteboul, A Bonifati, G	ACM SIGMOD,		dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	0	Rjb06zlxbLIJ	Dynamic XML documer	S Abiteboul, A Bonifati, G	SIGMOD Conference,	2003	dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	0	conf_sigmod_Abit	Dynamic XML documer	S Abiteboul, A Bonifati, G	SIGMOD Conference	2003	dataset1
e4cc1af1-1345-3d5d-8551-32997f7582db	2	url:http://www.csa	Modal analysis of aeroe	M Gennaretti, A Corbelli, P	European Rotorcraft Forum, 25 th, Rome, It	1999	dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	3	url:http://portal.ac	Content-Based Routing	P Bizarro, S Babu, D DeW	Proceedings of the 31st international confer	2005	dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	4	url:http://portal.ac	Caching with Ã¢??Good	H Guo, PÃ? Larson, R Ra	Proceedings of the 31st international confer	2005	dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	5	url:http://www.csa	A parametric study of th	S Shaw, N Qin	European Rotorcraft Forum, 25 th, Rome, It	1999	dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	1	WWaxLMIptTMJ	Self-tuning Histograms:	A Aboulnaga, S Chaudhuri	SIGMOD Conference,	1999	dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	1	xnDzelm2t1QJ	Self-tuning Histograms:	AA AC, S Chaudhuri	Proceedings of the ACM SIGMOD International Confe		dataset2
e4cc1af1-1345-3d5d-8551-32997f7582db	1	conf_sigmod_Abo	Self-tuning Histograms:	A Aboulnaga, S Chaudhuri	SIGMOD Conference	1999	dataset1
e4cc1af1-1345-3d5d-8551-32997f758789	0	SFmnhrhoQzgJ	The LyriC Language: Qu	A Brodsky, Y Kornatzky			dataset2
e4cc1af1-1345-3d5d-8551-32997f758789	0	6t_N5axEm7cJ	The LyriC Language: Co	A Brodsky, Y Kornatzky	… SIGMOD International Conference on Manage		dataset2
e4cc1af1-1345-3d5d-8551-32997f758789	0	conf_sigmod_Bro	The LyriC Language: Qu	A Brodsky, Y Kornatzky	SIGMOD Conference	1995	dataset1
e4cc1af1-1345-3d5d-8551-32997f758789	1	7mNCkZ7D-R8J	Copy Detection Mechar	S Brin, J Davis, H Garcia-	SIGMOD Conference,	1995	dataset2
e4cc1af1-1345-3d5d-8551-32997f758789	1	JsKYXyM6IaEJ	SCAM: A Copy Detectio	N Shivakumar, H Garcia-Molina			dataset2

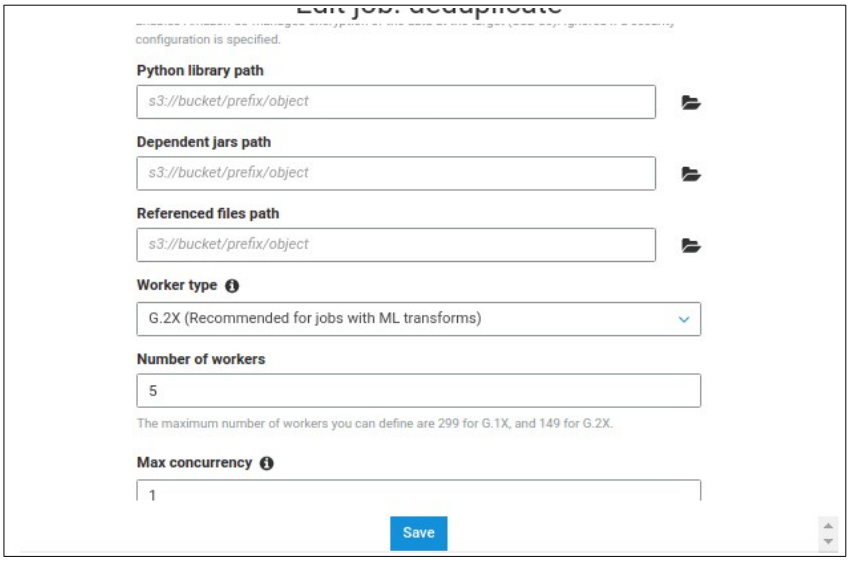

Teach transform
Generate labeling file
Label data
Upload labels
Estimate quality
Upload labels
I do not have labels
I have labels
Upload labels from S3
The completed labeling file must be in the correct format and in Amazon S3.
Upload labeling file from S3
Back
Next

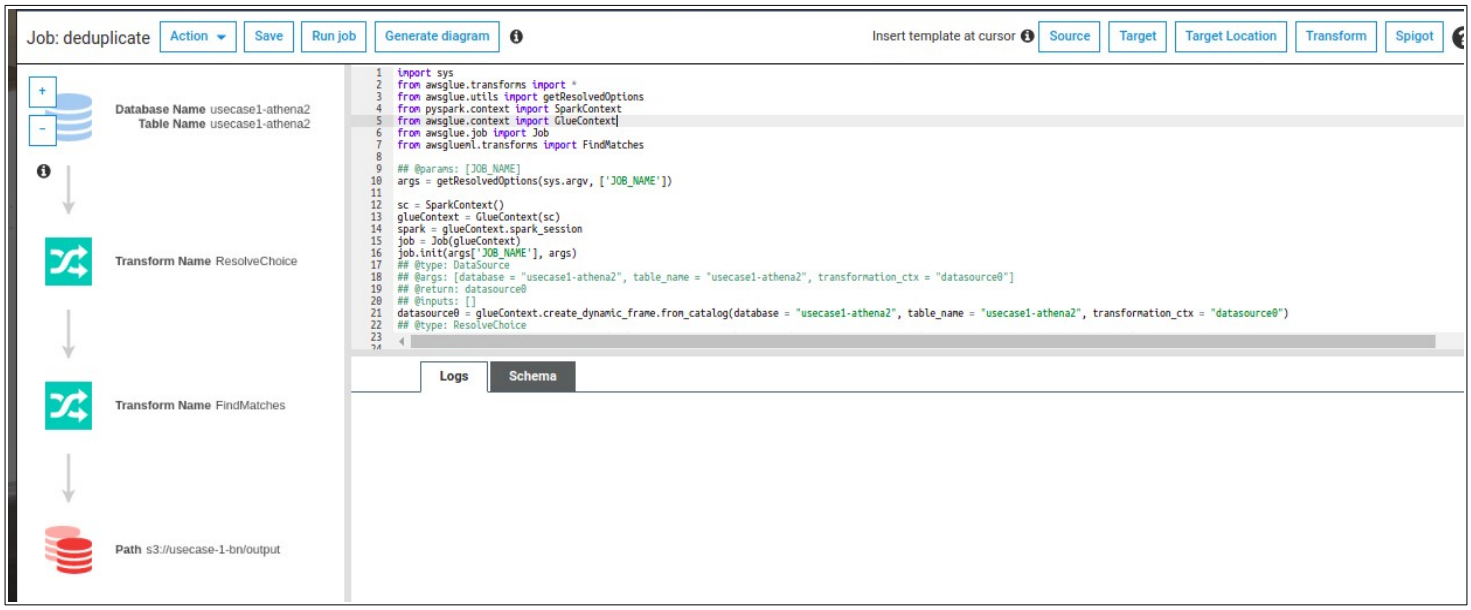

Teach transform
Generate labeling file
Label data
Upload labels
Estimate quality
new estimates can be viewed in the Quality metrics pane of the transform.
Estimate transform quality
Quality metric
Definition
Result
Last modified
Area under the Precision-Recall curve
Single number summarizing the performance of the transform
92.6984%
01/17/22 10:03 PM
Precision
When your transform predicts a match, how often is it correct?
85.7143%
01/17/22 10:03 PM
Recall upper limit
For an actual match, how often does your transform predict a match?
100%
01/17/22 10:03 PM
F1
Indicates transform's accuracy. Harmonic mean of Precision and Recall.
92.3077%
01/17/22 10:03 PM
* Metrics shown are from the last quality estimation run.
** End-to-End recall will tend to be closer to the upper limit as the cost-accuracy slider favors accuracy. See documentation for additional information about End-to-End recall.
Want to improve your results? Generate a new labeling file, label it, and upload the labels to append to our existing labels.


Teach transform
Generate labeling file
Label data
Upload labels
Estimate quality
** End-to-End recall will tend to be closer to the upper limit as the cost-accuracy slider favors accuracy. See documentation for additional information about End-to-End recall.
Want to improve your results? Generate a new labeling file, label it, and upload the labels to append to our existing labels.
Column importance
Estimate your transform's ability to find matches. Estimates are calculated by comparing the transform match predictions using a subset of your labeled data against the labels you have provided. These estimates are approximate. To improve your transform quality, provide more labels.
Column name
Importance
title
0.555
author
0.421
venue
0.014
year
0.005
source
0.004
Back
Finish


Add transform
Action
Filter by tags and attributes
Showing: 1 - 2
Transform name
Transform ID
Type
Label count
Status
Date created
Last modified
Description
transform-deduplicate
tfm-3d0c162dc4741bfd32dc...
Find matching records
129
Ready for use
17 January 2022 12:57 P...
17 January 2022 9:57 PM...
usecase1-deduplicate
tfm-3646f79f29b2b185516c6...
Find matching records
0
Needs training
17 January 2022 12:02 P...
17 January 2022 12:36 P...
Transform name transform-deduplicate
Transform ID tfm-3d0c162dc4741bfd32dc64b51d2b0652b7b1c64d
Source database usecase1-athena2
Source data table usecase1-athena2
Type Find matching records
Spark Version 2.2
IAM Role glueRole
Status Ready for use
Date created 17 January 2022 12:57 PM UTC+2
Last modified 17 January 2022 9:57 PM UTC+2
Description -
Label count 129
Precision-recall tradeoff 0.9
Accuracy-cost tradeoff 1
Force output to match labels true
Worker type G.2X
Number of workers 10
Tags -
Security configuration -

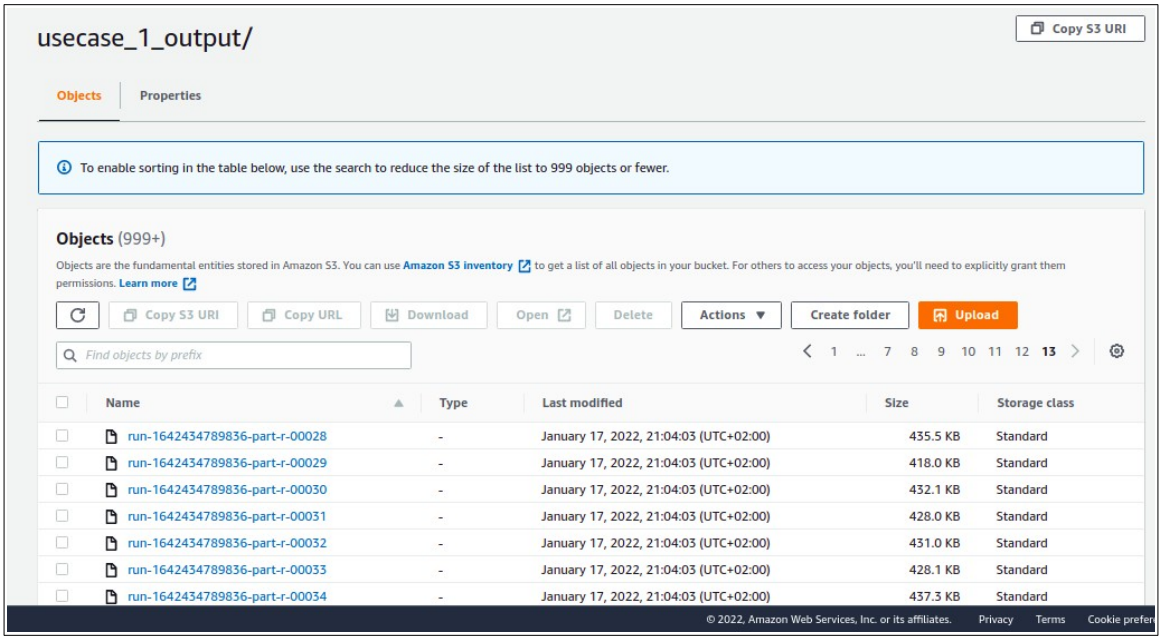
8. Created an AWS Glue ETL job that uses FindMatches ML transform.





9. Downloaded all output files from S3 and merged to one CSV file.

```
aws s3 cp –recursive s3://usecase-1-bn/usecase_1_output/ .
awk '(NR == 1) || (FNR > 1)' * > output.csv
```



id	title	author	venue	year	source	match_
conf_vldb_RusinkiewiczKTWM95	Towards a Cooperative Transaction I	M Rusinkiewicz, W Klas, T Tes	VLDB	1995	dataset1	0
wgK6p4mDSIMJ	SQL/XML is Making Good Progress	A Eisenberg, J Melton	SIGMOD Record,	2002	dataset2	1
journals_sigmod_EisenbergM02	SQL/XML is Making Good Progress	A Eisenberg, J Melton	SIGMOD Record	2002	dataset1	1
wgK6p4mDSIMJ	SQL/XML is Making Good Progress	A Eisenberg, J Melton	SIGMOD Record,	2002	dataset2	1
conf_vldb_AmmannJR95	Using Formal Methods to Reason ab	P Ammann, S Jajodia, I Ray	VLDB	1995	dataset1	2
2e-NAgqt-joJ	Applying Formal Methods to Seman	P Ammann, S Jajodia, I Ray	ACM Transactions on I	1997	dataset2	2
journals_tods_AmmannJR97	Applying Formal Methods to Seman	P Ammann, S Jajodia, I Ray	ACM Trans. Database	1997	dataset1	2
journals_sigmod_Liu02	Editor's Notes	L Liu	SIGMOD Record	2002	dataset1	3
journals_sigmod_Hammer02	Report on the ACM Fourth Internati	N/A	N/A	2002	dataset1	4
conf_vldb_FerrandinaMZFM95	Schema and Database Evolution in t	F Ferrandina, T Meyer, R Zicar	VLDB	1995	dataset1	5
9rofzgQ6HtcJ	Schema and Database Evolution in t	F Ferrandina, T Meyer, R Zicar	Proc. of the 21st IntÃƒÂ¢??l Conf. on Very		dataset2	5
conf_vldb_SubietaKL95	Procedures in Object-Oriented Quer	K Subieta, Y Kambayashi, J Le	VLDB	1995	dataset1	6
conf_vldb_SubietaKL95	Procedures in Object-Oriented Quer	K Subieta, Y Kambayashi, J Le	VLDB	1995	dataset1	6
gEFY87Ma0XUJ	Procedures in Object-Oriented Quer	K Subieta, Y Kambayashi, J Le	PROCEEDINGS OF THE	1995	dataset2	6
hf84fEpX5agJ	Phoenix: Making Applications Robus	R Barga, D Lomet	Philadelphia, PA (June,		dataset2	7
journals_sigmod_BargaL02	Phoenix Project: Fault-Tolerant App	R Barga, D Lomet	SIGMOD Record	2002	dataset1	7
fY3kkkzBLw8J	Phoenix Project: Fault Tolerant Appl	D Lomet, R Barga	SIGMOD Record,		dataset2	7
conf_sigmod_BargaL99	Phoenix: Making Applications Robus	R Barga, D Lomet	SIGMOD Conference	1999	dataset1	7
QGxyK7bJQOMJ	Phoenix: Making Applications Robus	R Barga, DB Lomet			dataset2	7
QTzV3iNq2O8J	Phoenix Project: Fault-Tolerant App	R Barga, D Lomet	SIGMOD Record,	2002	dataset2	7