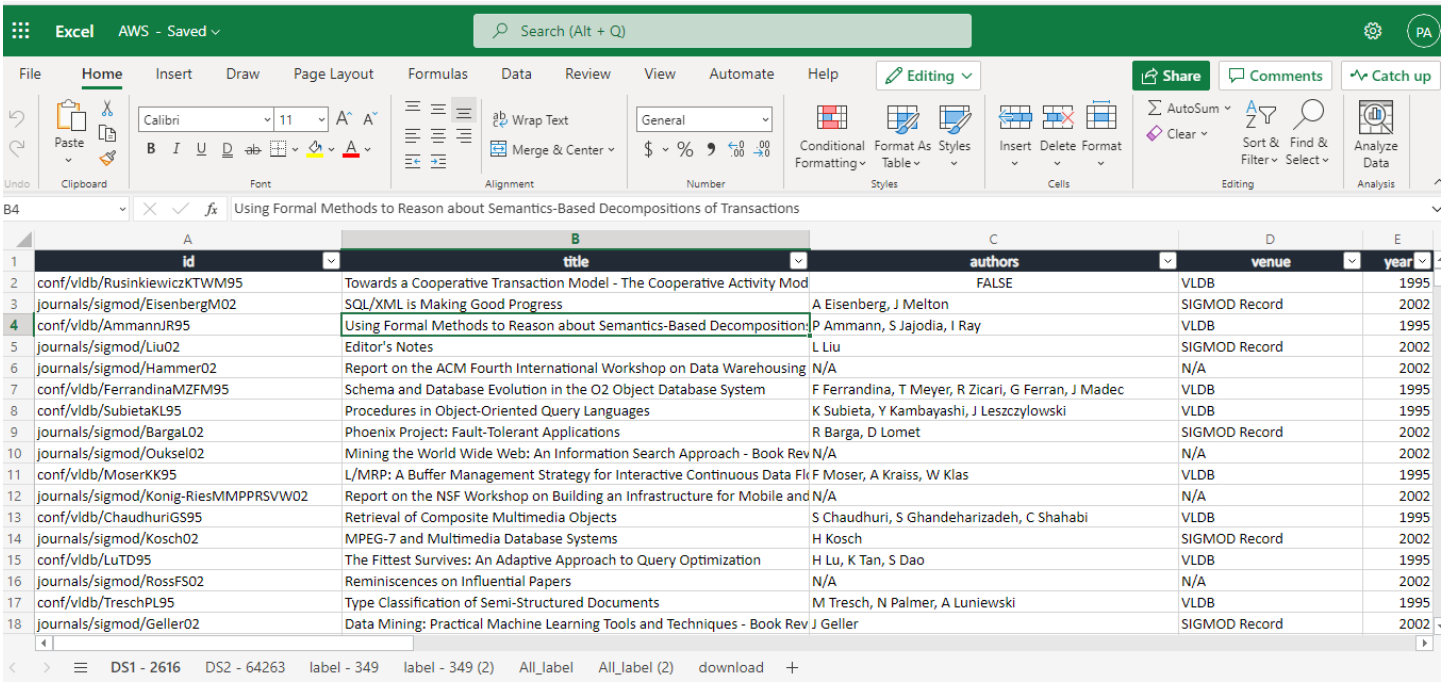
MCIT AWS PDSA Intake 2 - Use Case A

Name: Ahmed Samir Alnaqa
Email: ahmed.samir.mohamed@gmail.com
Group: 6

Steps taken to solve this use case:

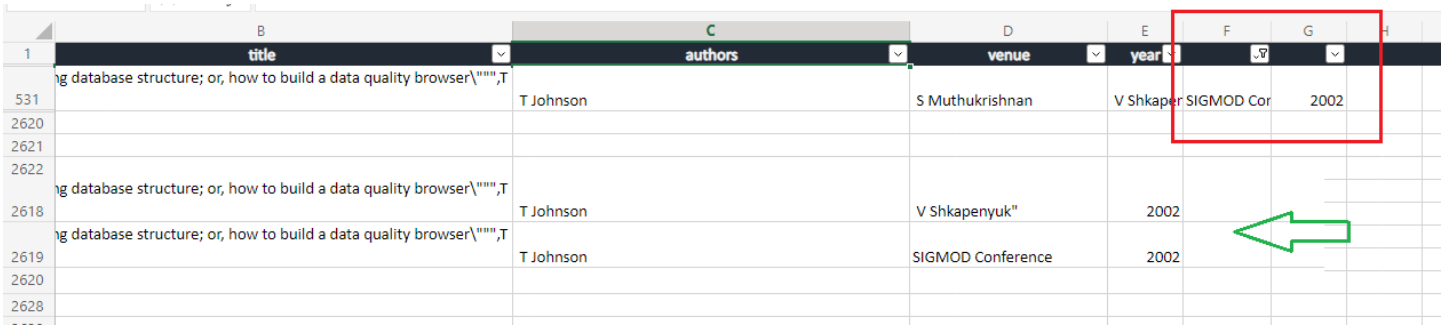
1- Evaluate the Datasets and check it's consistent:

Tool: Excel:



	A	B	C	D	E
1	id	title	authors	venue	year
2	conf/vldb/RusinkiewiczKTWM95	Towards a Cooperative Transaction Model - The Cooperative Activity Mod	FALSE	VLDB	1995
3	journals/sigmod/EisenbergM02	SQL/XML is Making Good Progress	A Eisenberg, J Melton	SIGMOD Record	2002
4	conf/vldb/AmmannJR95	Using Formal Methods to Reason about Semantics-Based Decompositions	P Ammann, S Jajodia, I Ray	VLDB	1995
5	journals/sigmod/Liu02	Editor's Notes	L Liu	SIGMOD Record	2002
6	journals/sigmod/Hammer02	Report on the ACM Fourth International Workshop on Data Warehousing	N/A	N/A	2002
7	conf/vldb/FerrandinaMZFM95	Schema and Database Evolution in the O2 Object Database System	F Ferrandina, T Meyer, R Zicari, G Ferran, J Madec	VLDB	1995
8	conf/vldb/SubietaKL95	Procedures in Object-Oriented Query Languages	K Subieta, Y Kambayashi, J Leszczylowski	VLDB	1995
9	journals/sigmod/BargaL02	Phoenix Project: Fault-Tolerant Applications	R Barga, D Lomet	SIGMOD Record	2002
10	journals/sigmod/Ouksel02	Mining the World Wide Web: An Information Search Approach - Book Rev	N/A	N/A	2002
11	conf/vldb/MoserKK95	L/MRP: A Buffer Management Strategy for Interactive Continuous Data Fl	F Moser, A Kraiss, W Klas	VLDB	1995
12	journals/sigmod/Konig-RiesMMPPRSVW02	Report on the NSF Workshop on Building an Infrastructure for Mobile and	N/A	N/A	2002
13	conf/vldb/ChaudhuriGS95	Retrieval of Composite Multimedia Objects	S Chaudhuri, S Ghandeharizadeh, C Shahabi	VLDB	1995
14	journals/sigmod/Kosch02	MPEG-7 and Multimedia Database Systems	H Kosch	SIGMOD Record	2002
15	conf/vldb/LuTD95	The Fittest Survives: An Adaptive Approach to Query Optimization	H Lu, K Tan, S Dao	VLDB	1995
16	journals/sigmod/RossFS02	Reminiscences on Influential Papers	N/A	N/A	2002
17	conf/vldb/TreschPL95	Type Classification of Semi-Structured Documents	M Tresch, N Palmer, A Luniewski	VLDB	1995
18	journals/sigmod/Geller02	Data Mining: Practical Machine Learning Tools and Techniques - Book Rev	J Geller	SIGMOD Record	2002

I found one problem in columns count in Dataset 1 and changed to have the same column count.

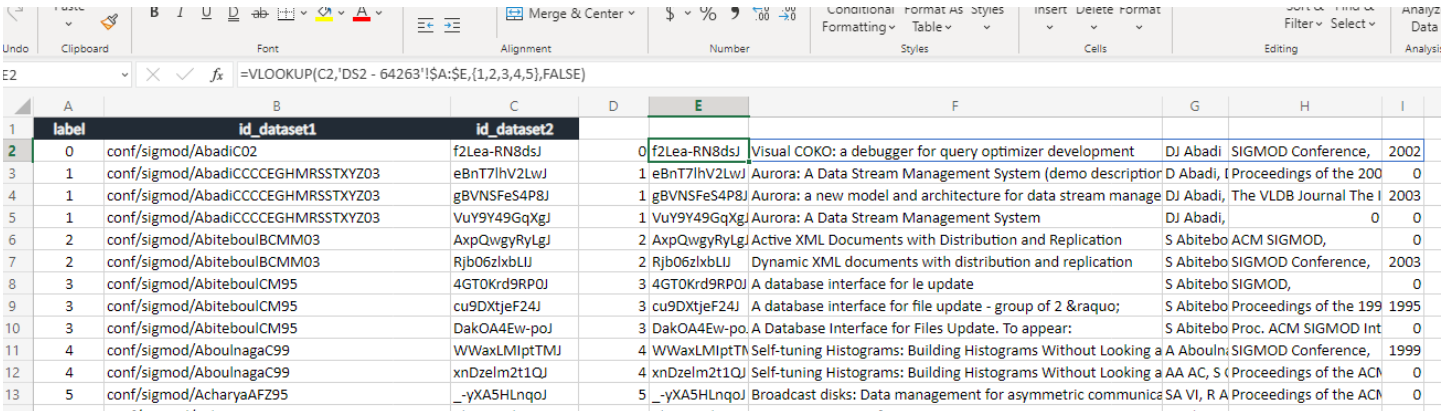


	B	C	D	E	F	G
1	title	authors	venue	year		
531	ng database structure; or, how to build a data quality browser\"""",T	T Johnson	S Muthukrishnan	V Shkaper	SIGMOD Cor	2002
2620						
2621						
2622	ng database structure; or, how to build a data quality browser\"""",T					
2618		T Johnson	V Shkapenyuk"	2002		
2619	ng database structure; or, how to build a data quality browser\"""",T	T Johnson	SIGMOD Conference	2002		
2620						
2628						

2- Check the label file and convert it to match the datasets column count without duplicate:

Tool: Excel

Formula: Vlookup



=VLOOKUP(C2,'DS2 - 64263'!$A:$E,{1,2,3,4,5},FALSE)

	A	B	C	D	E	F	G	H	I
1	label	id_dataset1	id_dataset2						
2	0	conf/sigmod/AbadiC02	f2Lea-RN8dsJ	0	f2Lea-RN8dsJ	Visual COKO: a debugger for query optimizer development	DJ Abadi	SIGMOD Conference,	2002
3	1	conf/sigmod/AbadiCCCCEGHMRSSTXYZ03	eBnT7lhV2LwJ	1	eBnT7lhV2LwJ	Aurora: A Data Stream Management System (demo description	D Abadi, [	Proceedings of the 200	0
4	1	conf/sigmod/AbadiCCCCEGHMRSSTXYZ03	gBVNSFeS4P8J	1	gBVNSFeS4P8J	Aurora: a new model and architecture for data stream manage	DJ Abadi,	The VLDB Journal The I	2003
5	1	conf/sigmod/AbadiCCCCEGHMRSSTXYZ03	VuY9Y49GqXgJ	1	VuY9Y49GqXgJ	Aurora: A Data Stream Management System	DJ Abadi,	0	0
6	2	conf/sigmod/AbiteboulBCMM03	AxpQwgyRyLgJ	2	AxpQwgyRyLgJ	Active XML Documents with Distribution and Replication	S Abitebo	ACM SIGMOD,	0
7	2	conf/sigmod/AbiteboulBCMM03	Rjb06zlxbLIJ	2	Rjb06zlxbLIJ	Dynamic XML documents with distribution and replication	S Abitebo	SIGMOD Conference,	2003
8	3	conf/sigmod/AbiteboulCM95	4GT0Krd9RP0J	3	4GT0Krd9RP0J	A database interface for le update	S Abitebo	SIGMOD,	0
9	3	conf/sigmod/AbiteboulCM95	cu9DXtjeF24J	3	cu9DXtjeF24J	A database interface for file update - group of 2 »	S Abitebo	Proceedings of the 199	1995
10	3	conf/sigmod/AbiteboulCM95	DakOA4Ew-poJ	3	DakOA4Ew-poJ	A Database Interface for Files Update. To appear:	S Abitebo	Proc. ACM SIGMOD Int	0
11	4	conf/sigmod/AboulnagaC99	WWaxLMIptTMJ	4	WWaxLMIptTM	Self-tuning Histograms: Building Histograms Without Looking a	A Aboulna	SIGMOD Conference,	1999
12	4	conf/sigmod/AboulnagaC99	xnDzelm2t1QJ	4	xnDzelm2t1QJ	Self-tuning Histograms: Building Histograms Without Looking a	AA AC, S (	Proceedings of the ACM	0
13	5	conf/sigmod/AcharyaAFZ95	_-yXA5HLnqoJ	5	_-yXA5HLnqoJ	Broadcast disks: Data management for asymmetric communica	SA VI, R A	Proceedings of the ACM	0

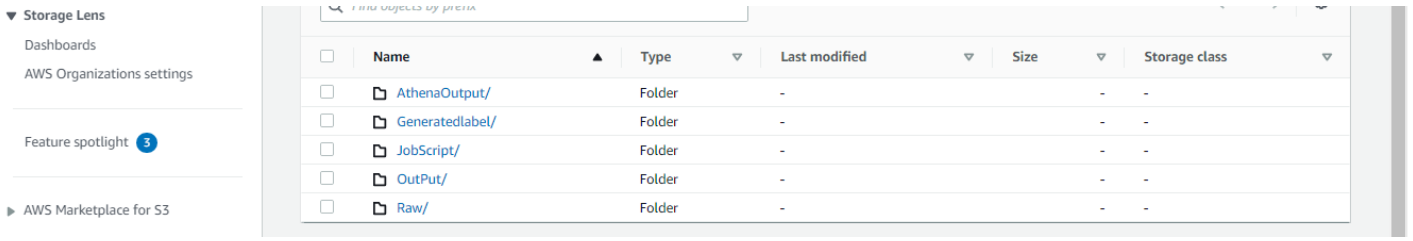
3- Concat the two datasets in one Dataset and include the source column to identify the records, inaddition change the “/” char to “_” then convert it to Json file.

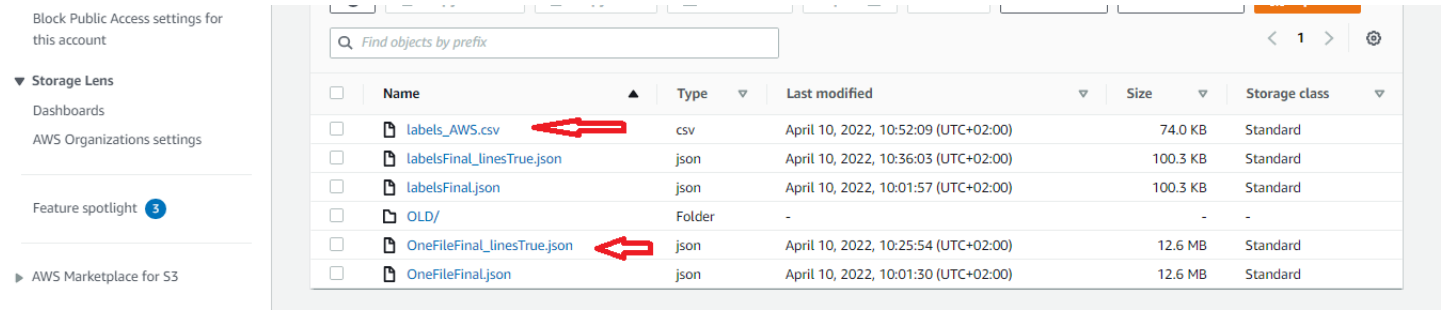
Tool: google colab notebook
Code: Pandas

4- replace the “/” char to “_” in the label file and add the header to it:

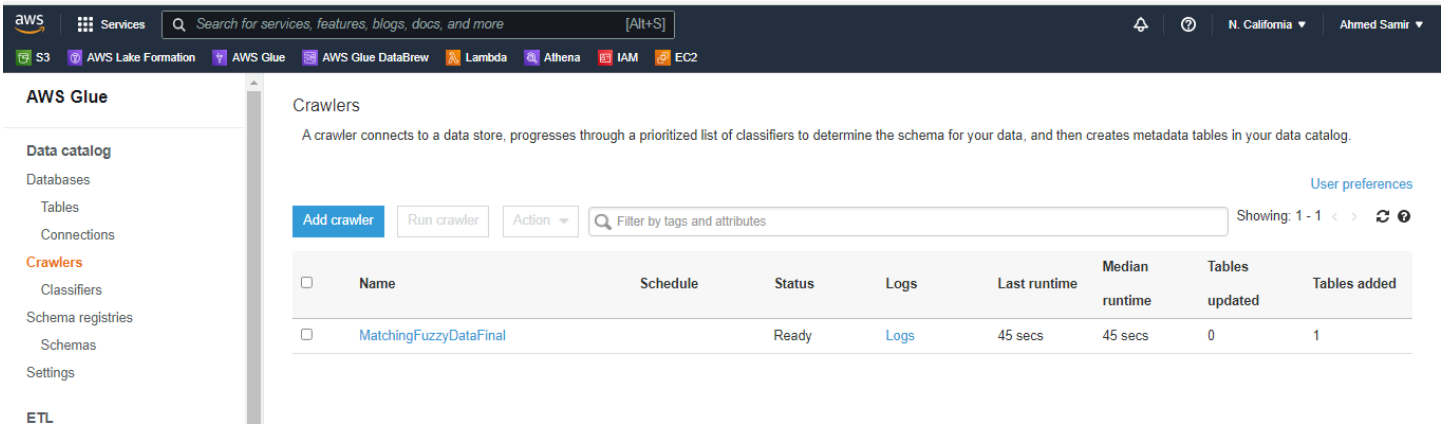
Tool: google colab notebook
Code: Pandas

5- upload the files to S3:

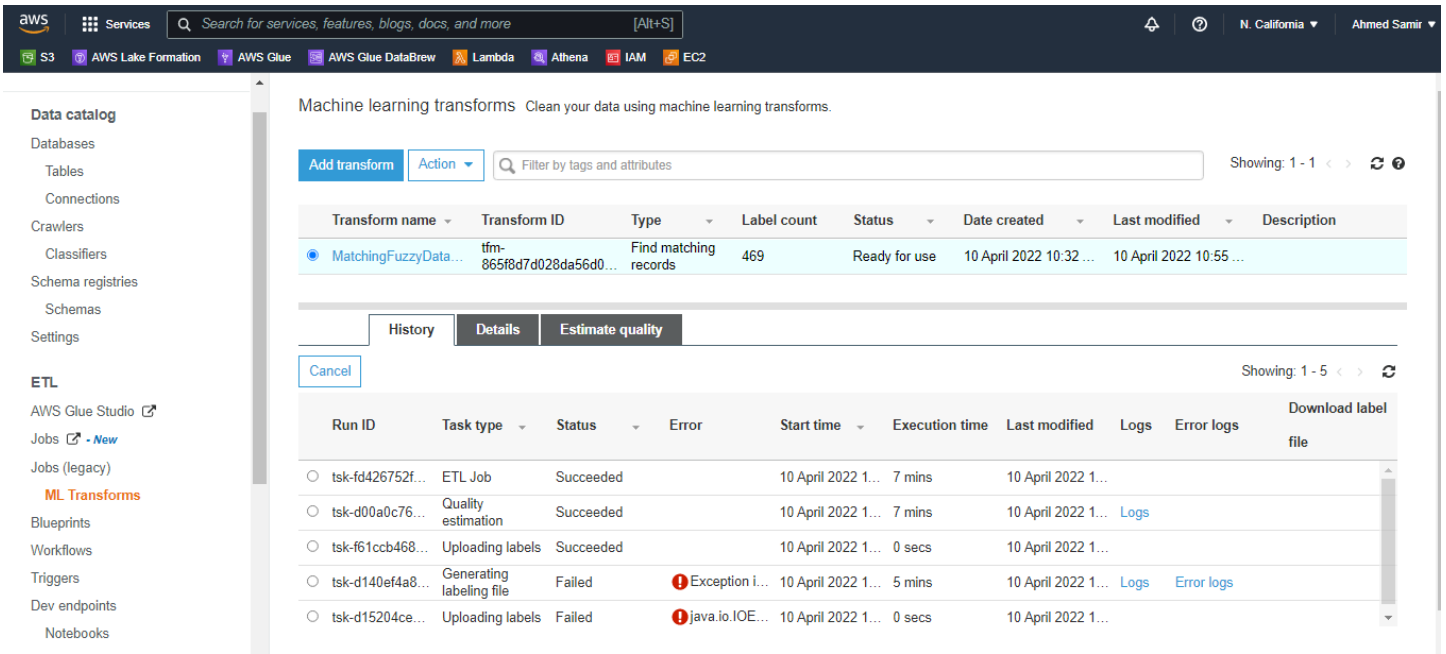


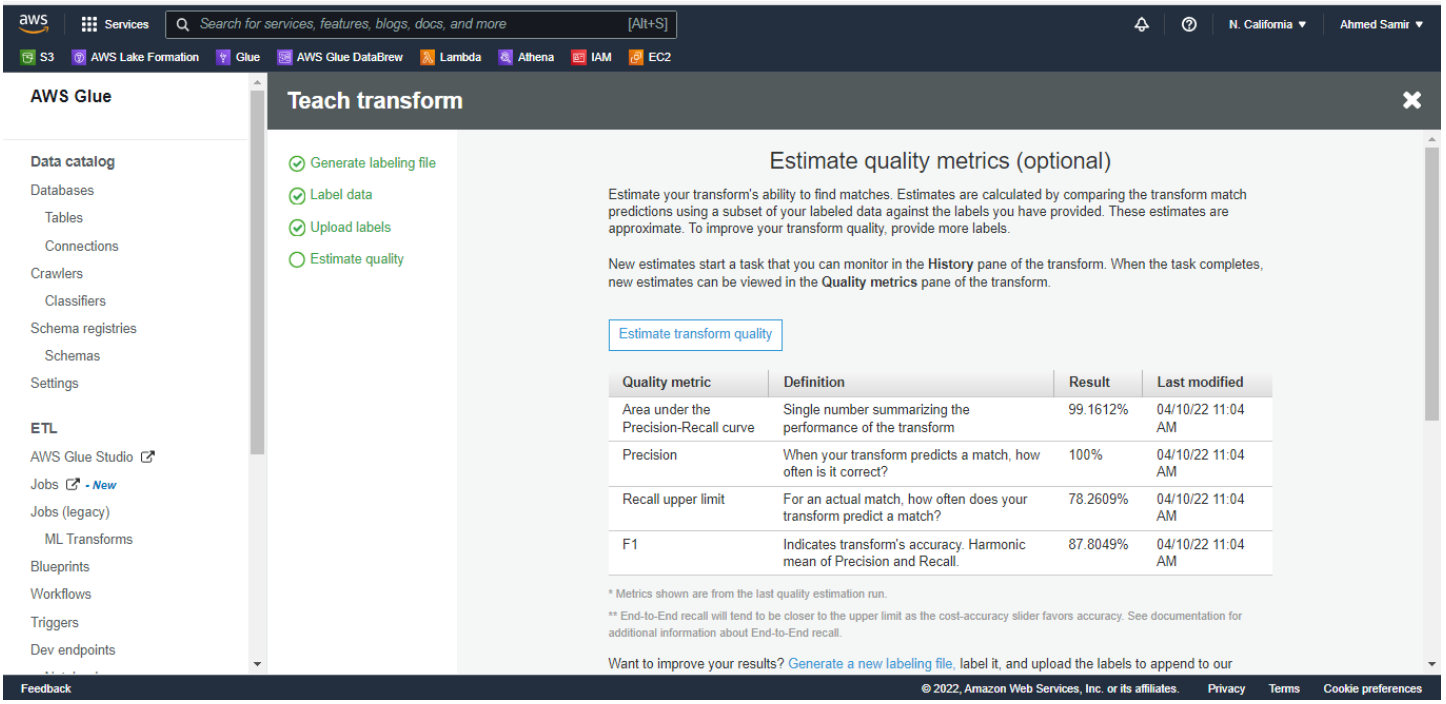


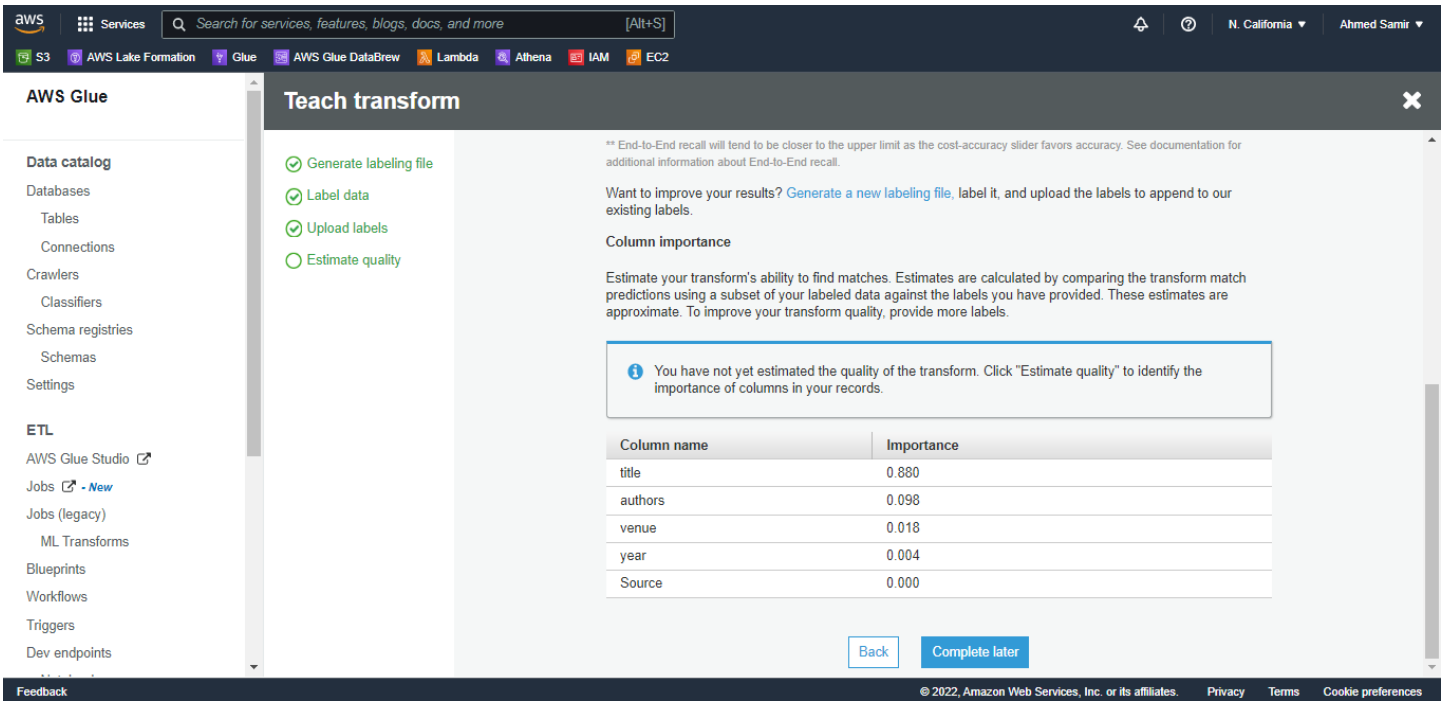
6- start crawler step:



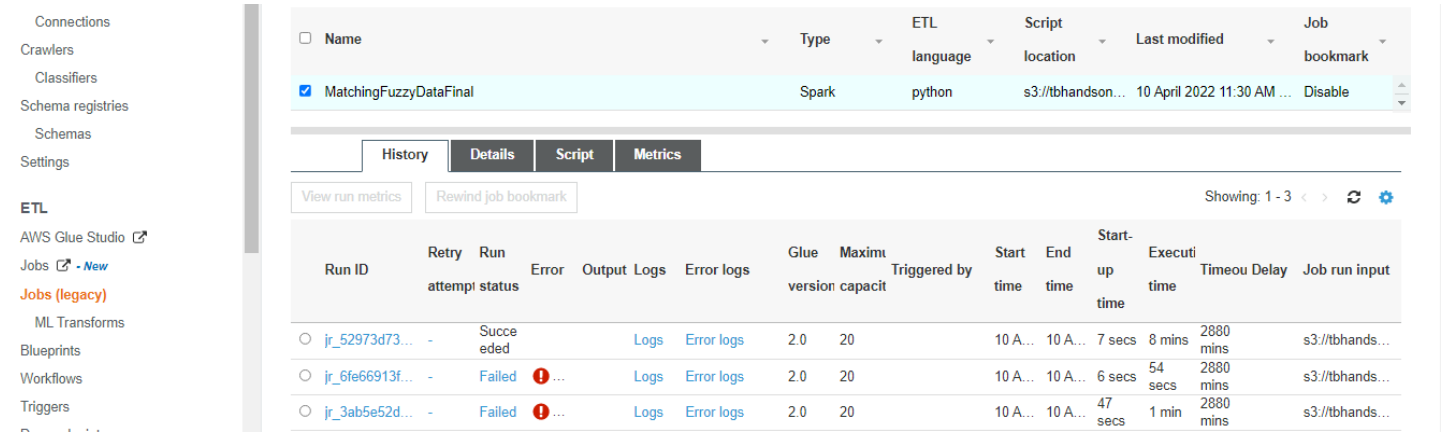
7- ML transform step:

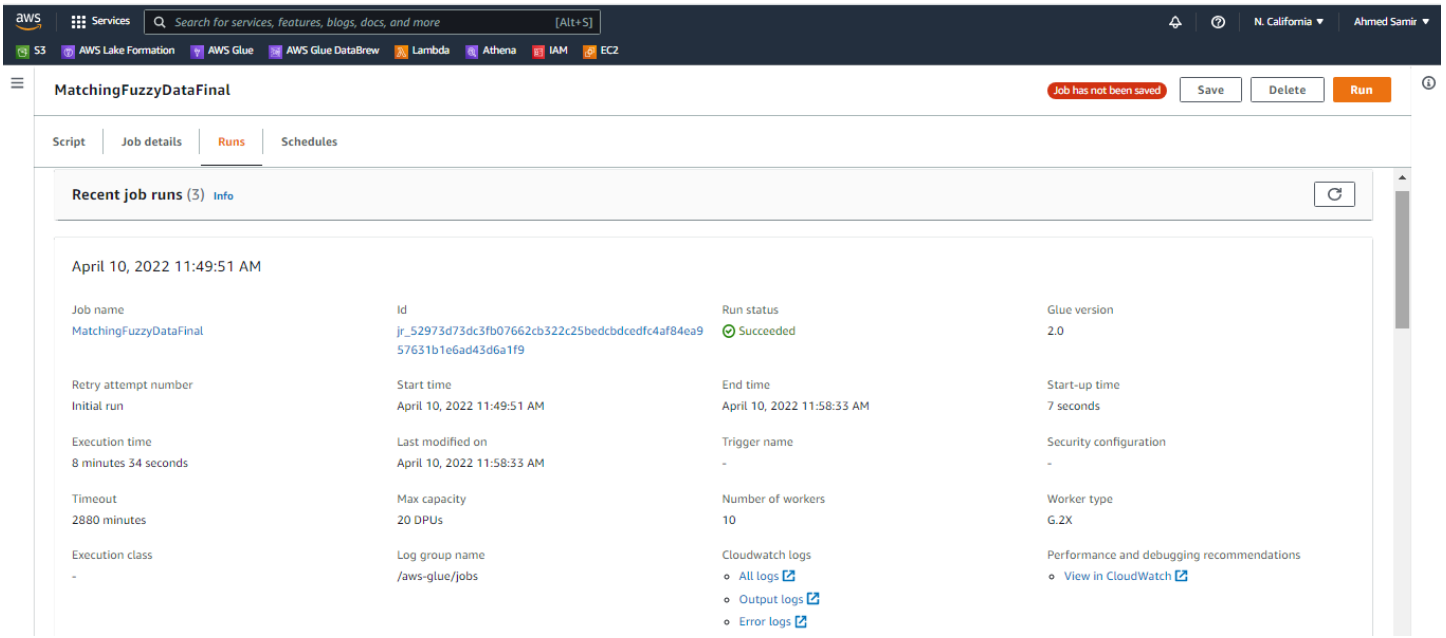




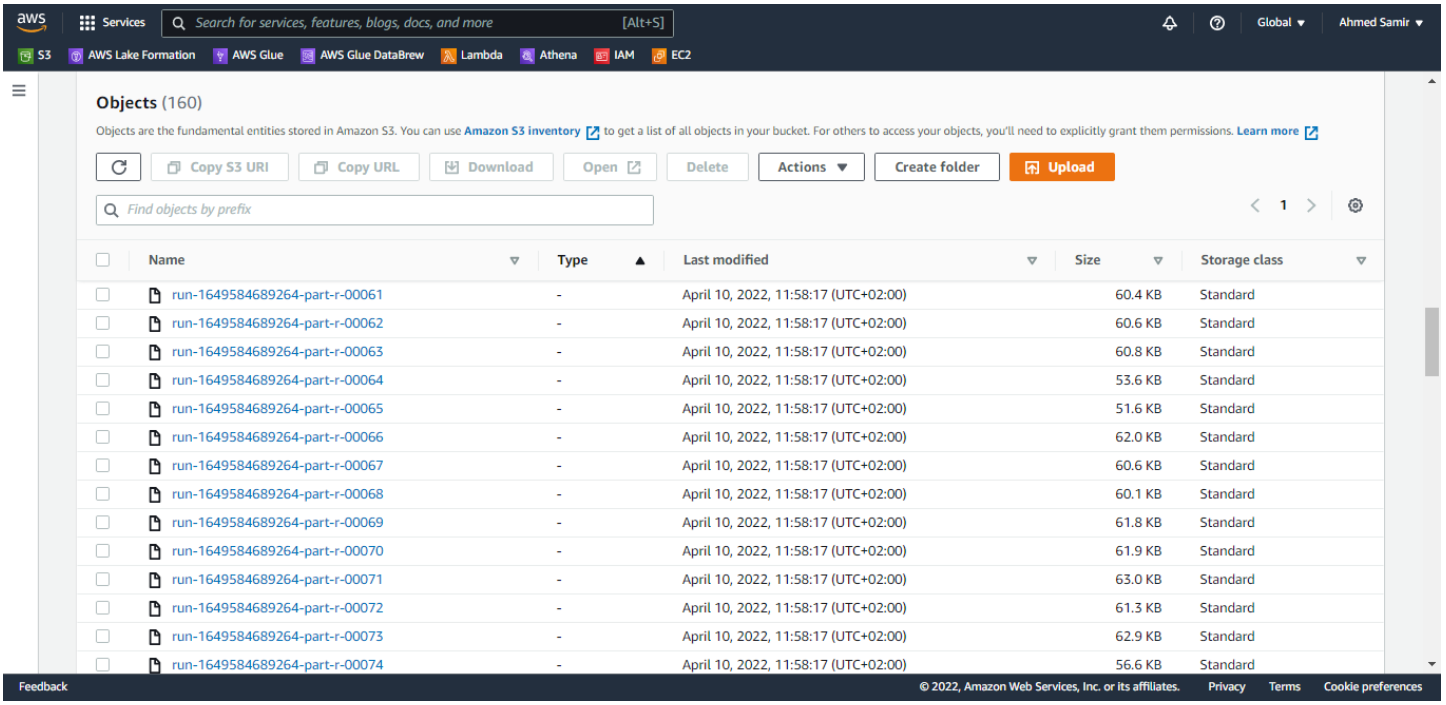


8- Create and run the glue job:





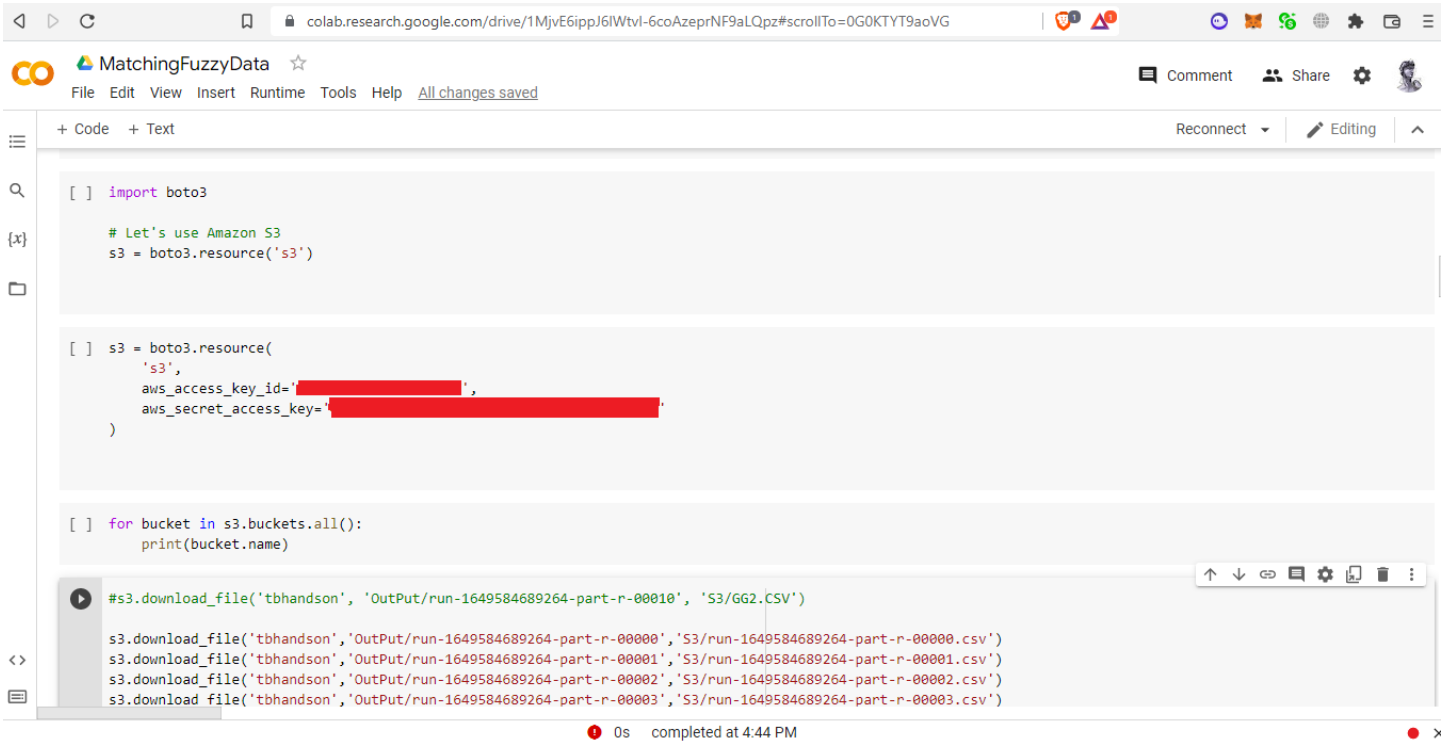
9- check the output files:



10- Download the 160 part:

Tool: google colab notebook

Code: Python (Boto)



11- merage the parts:

Tool: google colab notebook

Code: Python (Pandas)