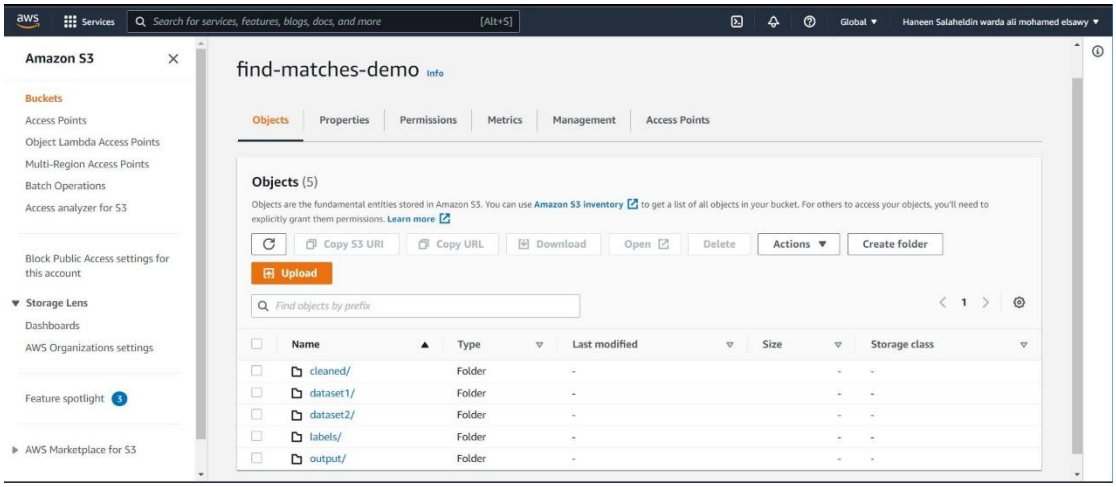
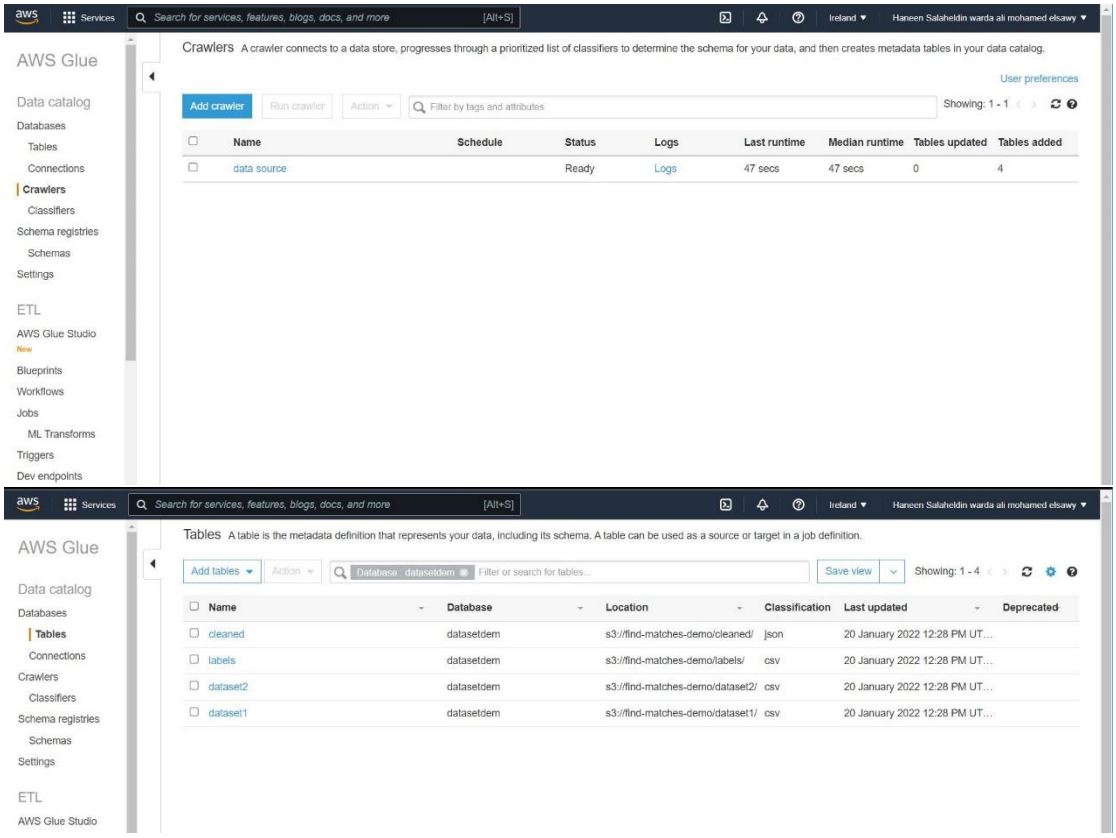
# HANEEN SALAH

AWS-Find-matches

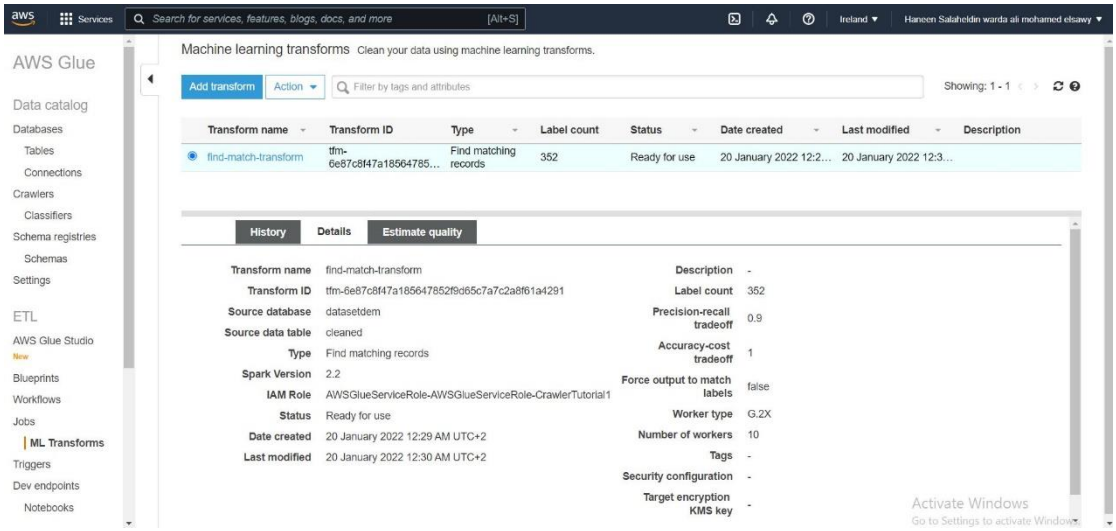
1) Create s3 bucket with folders as directories for dataset1, dataset2, output, labels, cleaned for the merged d1, d2 JSON file



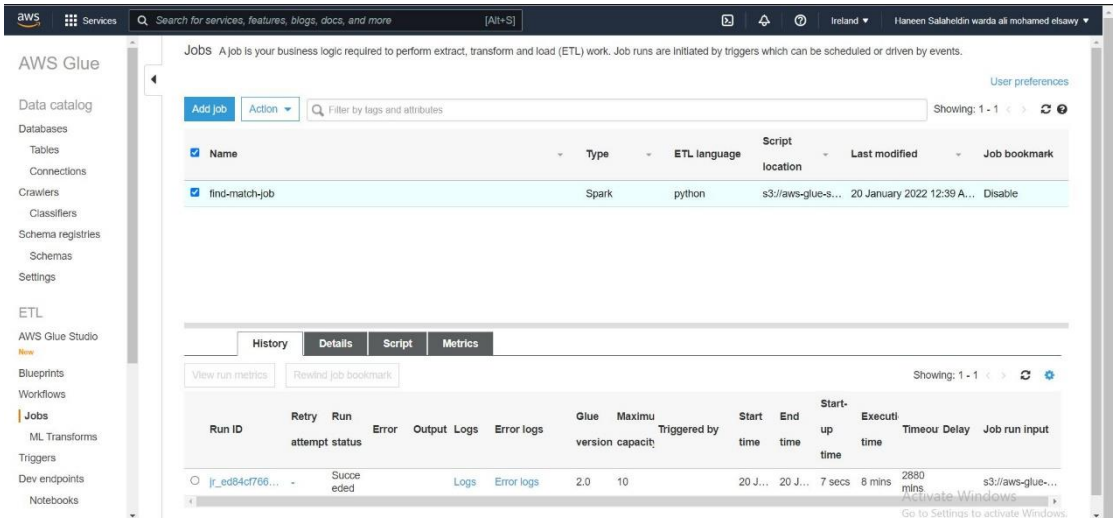

2) Create crawler



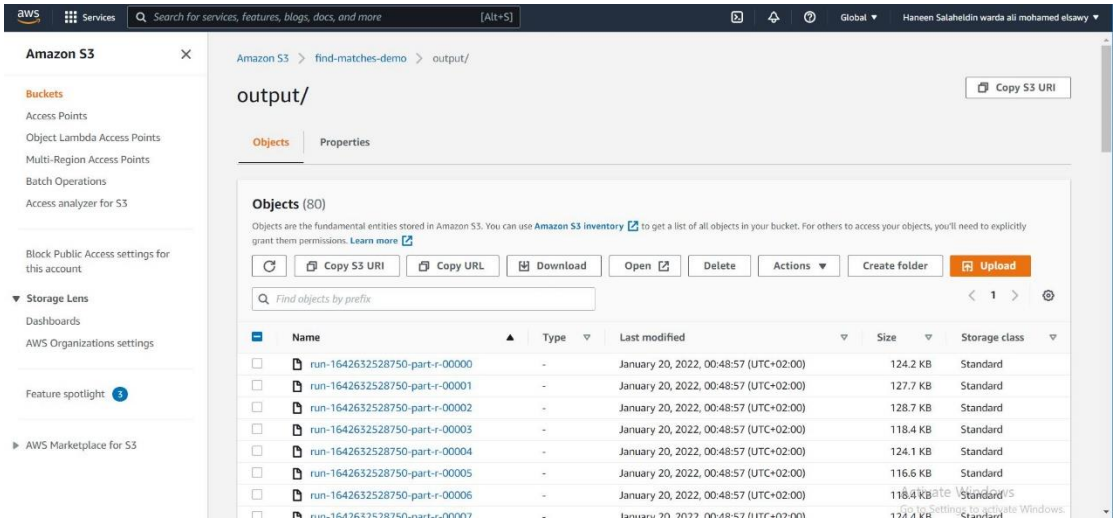
## 3) Create ml-transform and upload labels




## 4) Create and run job




## 5) Output

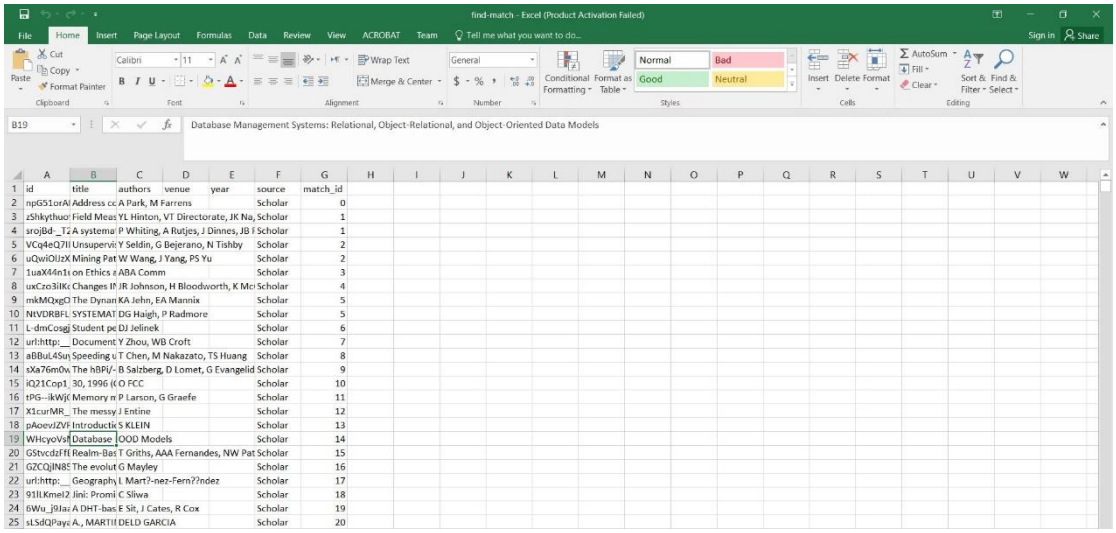


## 6) Merge multiple csv files into a single file

merge-mult-csv-out.py

```
#!/usr/bin/env python
# coding: utf-8

# In[6]:


import os
import glob
import pandas as pd


# In[7]:


os.chdir("D:\AWS\out")


# In[8]:


extension = 'csv'
all_filenames = [i for i in glob.glob('*.{}'.format(extension))]


# In[9]:


#combine all files in the list
combined_csv = pd.concat([pd.read_csv(f) for f in all_filenames ])
#export to csv
combined_csv.to_csv( "find-match.csv", index=False, encoding='utf-8-sig')


# In[ ]:
```

## 7) Csv final output sorted by match_id



find-match - Excel (Product Activation Failed)

B19: Database Management Systems: Relational, Object-Relational, and Object-Oriented Data Models

| | id | title | authors | venue | year | source | match_id |
|---|---|---|---|---|---|---|---|
| 2 | npG51orAl | Address cc | A Park, M Farrens | | | Scholar | 0 |
| 3 | zShkythuo | Field Meas | YL Hinton, VT Directorate, JK Na, | | | Scholar | 1 |
| 4 | srojBd-_T2 | A systema | P Whiting, A Rutjes, J Dinnes, JB I | | | Scholar | 1 |
| 5 | VCq4eQ7II | Unsupervi | Y Seldin, G Bejerano, N Tishby | | | Scholar | 2 |
| 6 | uQwiOIJzX | Mining Pat | W Wang, J Yang, PS Yu | | | Scholar | 2 |
| 7 | 1uaX44n1x | on Ethics a | ABA Comm | | | Scholar | 3 |
| 8 | uxCzo3iIKc | Changes IP | JR Johnson, H Bloodworth, K Mc | | | Scholar | 4 |
| 9 | mkMQxgO | The Dynan | KA Jehn, EA Mannix | | | Scholar | 5 |
| 10 | NtVDRBFL | SYSTEMAT | DG Haigh, P Radmore | | | Scholar | 5 |
| 11 | L-dmCosgj | Student pe | DJ Jelinek | | | Scholar | 6 |
| 12 | url:http:__ | Document | Y Zhou, WB Croft | | | Scholar | 7 |
| 13 | aBBuL4Sus | Speeding u | T Chen, M Nakazato, TS Huang | | | Scholar | 8 |
| 14 | sXa76m0v | The hBPi/- | B Salzberg, D Lomet, G Evangelid | | | Scholar | 9 |
| 15 | iQ21Cop1_ | 30, 1996 (C | O FCC | | | Scholar | 10 |
| 16 | tPG--ikWjC | Memory m | P Larson, G Graefe | | | Scholar | 11 |
| 17 | X1curMR_ | The messy | J Entine | | | Scholar | 12 |
| 18 | pAoevJZVF | Introductic | S KLEIN | | | Scholar | 13 |
| 19 | WHcyoVsf | Database | OOD Models | | | Scholar | 14 |
| 20 | GStvcdzFfL | Realm-Bas | T Griths, AAA Fernandes, NW Pat | | | Scholar | 15 |
| 21 | GZCQjIN8S | The evolut | G Mayley | | | Scholar | 16 |
| 22 | url:http:__ | Geography | L Mart?-nez-Fern??ndez | | | Scholar | 17 |
| 23 | 91ILKmeI2 | Jini: Promi | C Sliwa | | | Scholar | 18 |
| 24 | 6Wu_j9Jaz | A DHT-bas | E Sit, J Cates, R Cox | | | Scholar | 19 |
| 25 | sLSdQPay | A., MARTIN | DELD GARCIA | | | Scholar | 20 |