Take the home test for the Data Engineering position

We were facing a fraud detection problem, and to model that we decided to re-arrange our data into a property graph.

You should do the preprocessing for that purpose. To this purpose, go through the data to understand, clean, and transform it into a graph. I'll provide you with some examples in the following (Fig 2, Fig 3):

First of all, you can find Datasets from the link following (Dataset: banking_transaction)

Dataset description

The data is broken into two files ***identity*** and ***transaction***, which have TransactionID in common.

Note that, not all transactions have corresponding identity information.

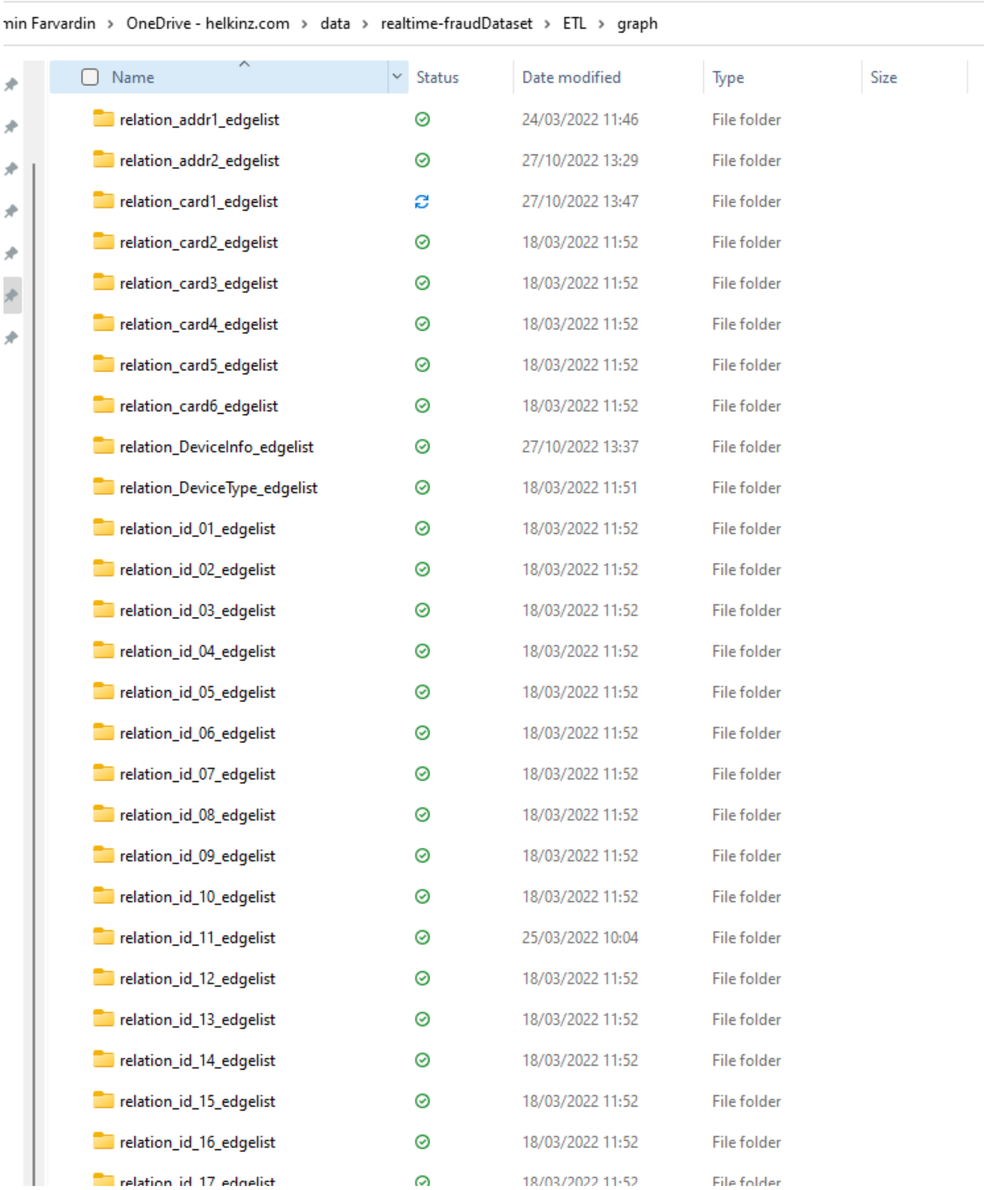
I expected the processed data to be stored in directories such as Fig 1, while every directory contains multiple files (Fig 2 & 3 provide extra information on files order in each folder). You can use either Apache Spark or PySpark.

This challenge is designed to evaluate a few things:

- clarity in documenting and justifying your approach,
- your ability to craft clear, readable code,
- your choice of library, method, or algorithms.
- feel free to come up with your methods or solutions (I like to see any creativity)

Submission format:

Please send me a zip file containing your implemented files (**excluding the datasets**).

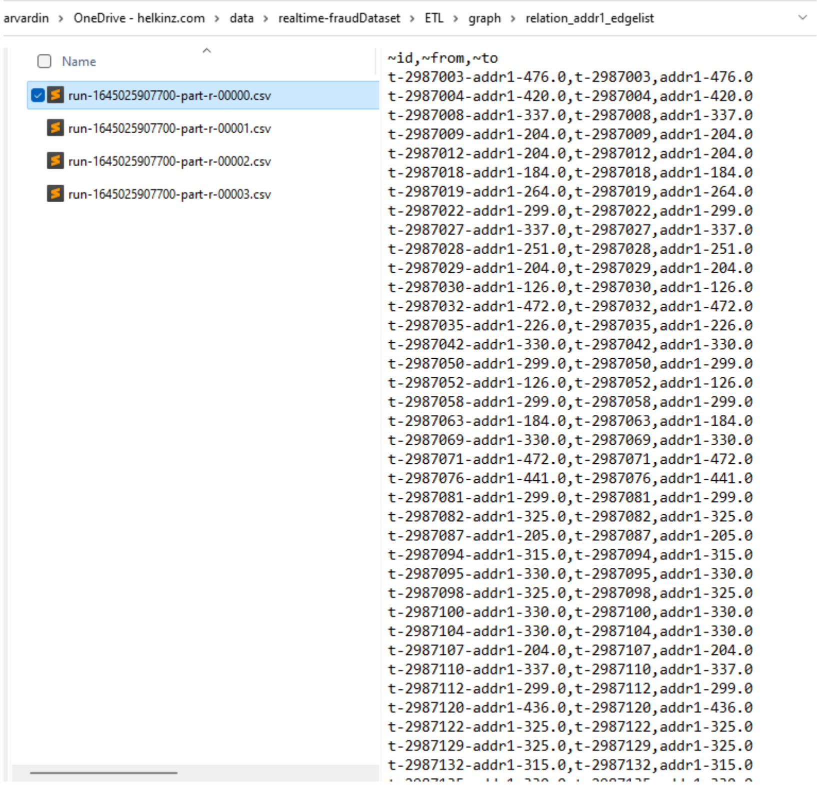


nin Farvardin › OneDrive - helkinz.com › data › realtime-fraudDataset › ETL › graph

Name	Status	Date modified	Type	Size
relation_addr1_edgelist	✓	24/03/2022 11:46	File folder	
relation_addr2_edgelist	✓	27/10/2022 13:29	File folder	
relation_card1_edgelist	⟳	27/10/2022 13:47	File folder	
relation_card2_edgelist	✓	18/03/2022 11:52	File folder	
relation_card3_edgelist	✓	18/03/2022 11:52	File folder	
relation_card4_edgelist	✓	18/03/2022 11:52	File folder	
relation_card5_edgelist	✓	18/03/2022 11:52	File folder	
relation_card6_edgelist	✓	18/03/2022 11:52	File folder	
relation_DeviceInfo_edgelist	✓	27/10/2022 13:37	File folder	
relation_DeviceType_edgelist	✓	18/03/2022 11:51	File folder	
relation_id_01_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_02_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_03_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_04_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_05_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_06_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_07_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_08_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_09_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_10_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_11_edgelist	✓	25/03/2022 10:04	File folder	
relation_id_12_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_13_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_14_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_15_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_16_edgelist	✓	18/03/2022 11:52	File folder	
relation_id_17_edgelist	✓	18/03/2022 11:52	File folder	

Fig 1

For examples:



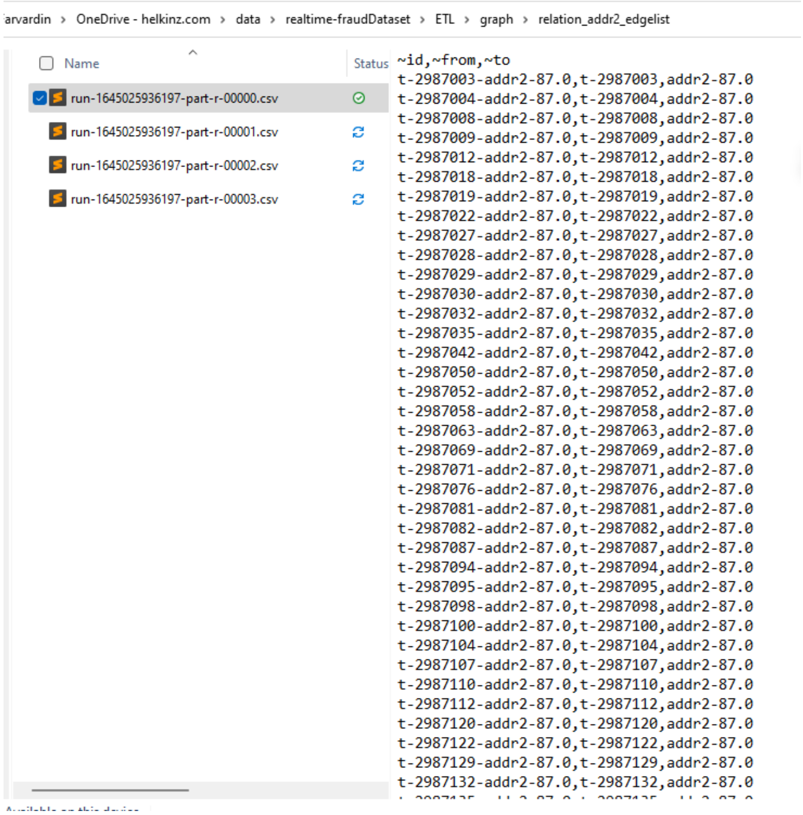
arvardin › OneDrive - helkinz.com › data › realtime-fraudDataset › ETL › graph › relation_addr1_edgelist

□ Name

☑ run-1645025907700-part-r-00000.csv
run-1645025907700-part-r-00001.csv
run-1645025907700-part-r-00002.csv
run-1645025907700-part-r-00003.csv

```
~id,~from,~to
t-2987003-addr1-476.0,t-2987003,addr1-476.0
t-2987004-addr1-420.0,t-2987004,addr1-420.0
t-2987008-addr1-337.0,t-2987008,addr1-337.0
t-2987009-addr1-204.0,t-2987009,addr1-204.0
t-2987012-addr1-204.0,t-2987012,addr1-204.0
t-2987018-addr1-184.0,t-2987018,addr1-184.0
t-2987019-addr1-264.0,t-2987019,addr1-264.0
t-2987022-addr1-299.0,t-2987022,addr1-299.0
t-2987027-addr1-337.0,t-2987027,addr1-337.0
t-2987028-addr1-251.0,t-2987028,addr1-251.0
t-2987029-addr1-204.0,t-2987029,addr1-204.0
t-2987030-addr1-126.0,t-2987030,addr1-126.0
t-2987032-addr1-472.0,t-2987032,addr1-472.0
t-2987035-addr1-226.0,t-2987035,addr1-226.0
t-2987042-addr1-330.0,t-2987042,addr1-330.0
t-2987050-addr1-299.0,t-2987050,addr1-299.0
t-2987052-addr1-126.0,t-2987052,addr1-126.0
t-2987058-addr1-299.0,t-2987058,addr1-299.0
t-2987063-addr1-184.0,t-2987063,addr1-184.0
t-2987069-addr1-330.0,t-2987069,addr1-330.0
t-2987071-addr1-472.0,t-2987071,addr1-472.0
t-2987076-addr1-441.0,t-2987076,addr1-441.0
t-2987081-addr1-299.0,t-2987081,addr1-299.0
t-2987082-addr1-325.0,t-2987082,addr1-325.0
t-2987087-addr1-205.0,t-2987087,addr1-205.0
t-2987094-addr1-315.0,t-2987094,addr1-315.0
t-2987095-addr1-330.0,t-2987095,addr1-330.0
t-2987098-addr1-325.0,t-2987098,addr1-325.0
t-2987100-addr1-330.0,t-2987100,addr1-330.0
t-2987104-addr1-330.0,t-2987104,addr1-330.0
t-2987107-addr1-204.0,t-2987107,addr1-204.0
t-2987110-addr1-337.0,t-2987110,addr1-337.0
t-2987112-addr1-299.0,t-2987112,addr1-299.0
t-2987120-addr1-436.0,t-2987120,addr1-436.0
t-2987122-addr1-325.0,t-2987122,addr1-325.0
t-2987129-addr1-325.0,t-2987129,addr1-325.0
t-2987132-addr1-315.0,t-2987132,addr1-315.0
```

Fig 2.



arvardin › OneDrive - helkinz.com › data › realtime-fraudDataset › ETL › graph › relation_addr2_edgelist

□ Name | Status

☑ run-1645025936197-part-r-00000.csv
run-1645025936197-part-r-00001.csv
run-1645025936197-part-r-00002.csv
run-1645025936197-part-r-00003.csv

```
~id,~from,~to
t-2987003-addr2-87.0,t-2987003,addr2-87.0
t-2987004-addr2-87.0,t-2987004,addr2-87.0
t-2987008-addr2-87.0,t-2987008,addr2-87.0
t-2987009-addr2-87.0,t-2987009,addr2-87.0
t-2987012-addr2-87.0,t-2987012,addr2-87.0
t-2987018-addr2-87.0,t-2987018,addr2-87.0
t-2987019-addr2-87.0,t-2987019,addr2-87.0
t-2987022-addr2-87.0,t-2987022,addr2-87.0
t-2987027-addr2-87.0,t-2987027,addr2-87.0
t-2987028-addr2-87.0,t-2987028,addr2-87.0
t-2987029-addr2-87.0,t-2987029,addr2-87.0
t-2987030-addr2-87.0,t-2987030,addr2-87.0
t-2987032-addr2-87.0,t-2987032,addr2-87.0
t-2987035-addr2-87.0,t-2987035,addr2-87.0
t-2987042-addr2-87.0,t-2987042,addr2-87.0
t-2987050-addr2-87.0,t-2987050,addr2-87.0
t-2987052-addr2-87.0,t-2987052,addr2-87.0
t-2987058-addr2-87.0,t-2987058,addr2-87.0
t-2987063-addr2-87.0,t-2987063,addr2-87.0
t-2987069-addr2-87.0,t-2987069,addr2-87.0
t-2987071-addr2-87.0,t-2987071,addr2-87.0
t-2987076-addr2-87.0,t-2987076,addr2-87.0
t-2987081-addr2-87.0,t-2987081,addr2-87.0
t-2987082-addr2-87.0,t-2987082,addr2-87.0
t-2987087-addr2-87.0,t-2987087,addr2-87.0
t-2987094-addr2-87.0,t-2987094,addr2-87.0
t-2987095-addr2-87.0,t-2987095,addr2-87.0
t-2987098-addr2-87.0,t-2987098,addr2-87.0
t-2987100-addr2-87.0,t-2987100,addr2-87.0
t-2987104-addr2-87.0,t-2987104,addr2-87.0
t-2987107-addr2-87.0,t-2987107,addr2-87.0
t-2987110-addr2-87.0,t-2987110,addr2-87.0
t-2987112-addr2-87.0,t-2987112,addr2-87.0
t-2987120-addr2-87.0,t-2987120,addr2-87.0
t-2987122-addr2-87.0,t-2987122,addr2-87.0
t-2987129-addr2-87.0,t-2987129,addr2-87.0
t-2987132-addr2-87.0,t-2987132,addr2-87.0
```

Fig 3.