Part 2: Data Staging

CSI 4142 - Fundamentals of Data Science
Winter 2023

uOttawa

Faculté de génie
Faculty of Engineering

School of Electrical Engineering and Computer Science
University of Ottawa

Yazan Otoum, Ph.D

Group #30 : Deliverable #2

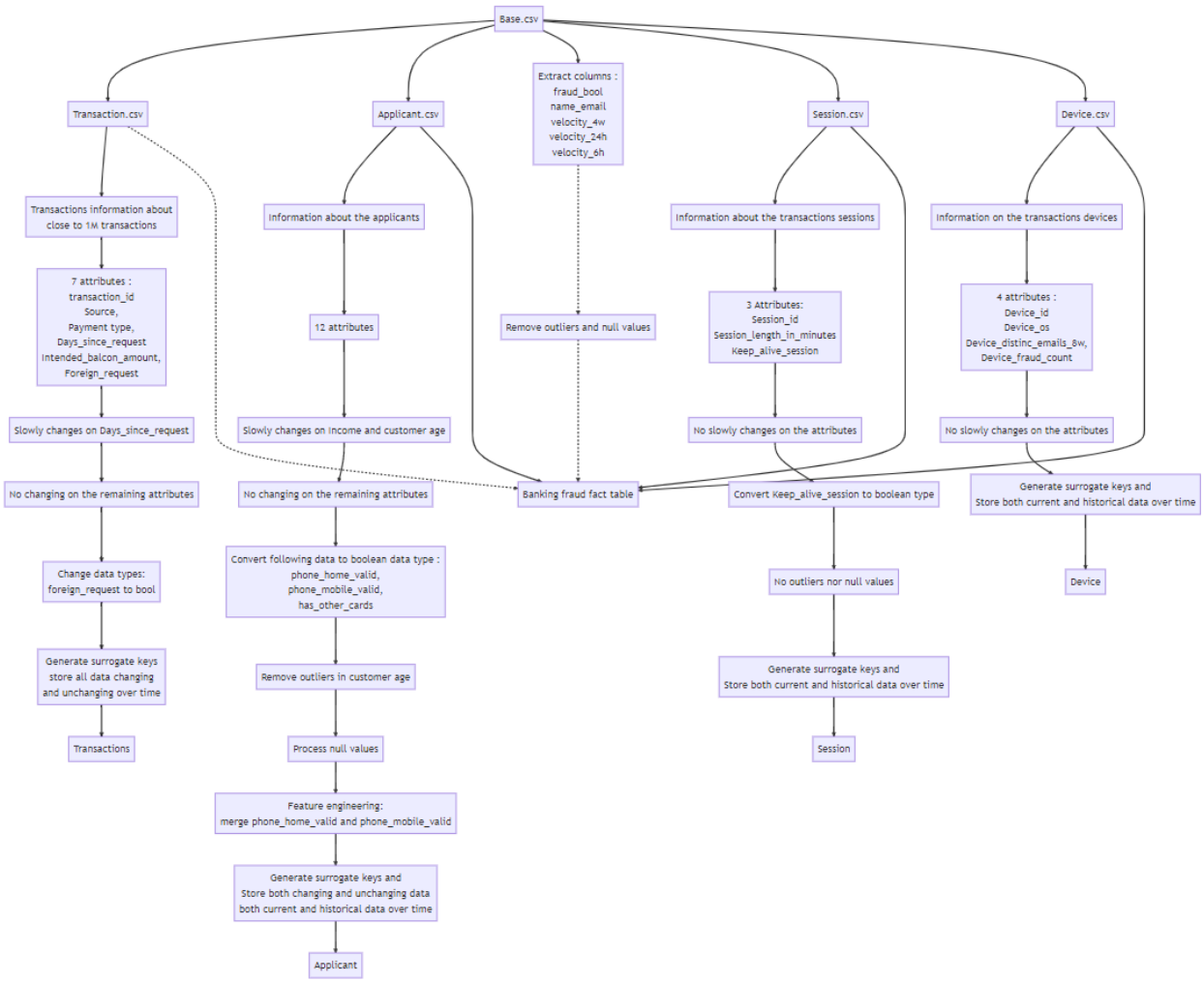
Ishanveer Gobin 300135454

Andie Samadoulougou 300209487

Kate Sin Yan Chun 300144923

Table of Content

A: High-level data staging



A clearer version of the high level data staging plan can be found in the deliverableC_group30 folder.

B : Other details

Downloading and Installing PostgreSQL

The DBMS was created using postgresql.

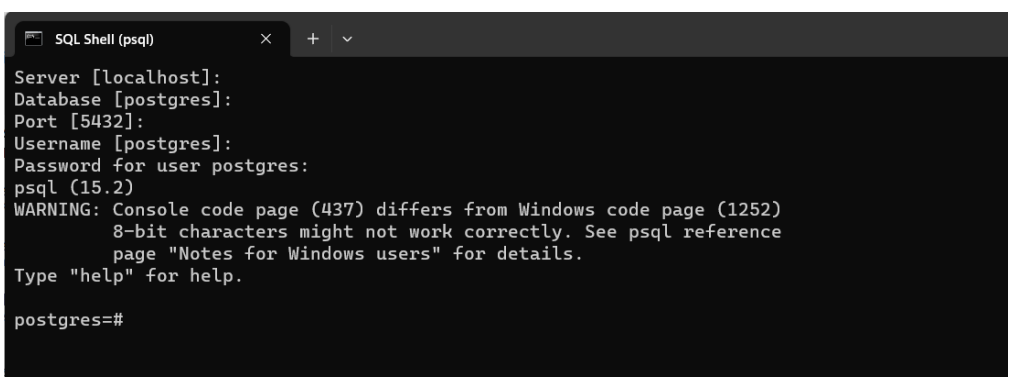
Click here to download the latest version of postgresql and follow these steps for installation. Make sure to install PgAdmin correctly.

Navigate into the deliverableC_group30 folder using `cd <folder path>` and use this command to install all the required libraries on your machine: `pip install -r requirements.txt`

Creating Database and Running Jupyter Notebook

Before running the jupyter notebook (DataStaging.ipynb) provided in the deliverableC_group30 folder, open the psql command Shell to connect to pgAdmin. Press enter for the default values for server, database, port and username. Then enter the password you saved while first installing postgresql.

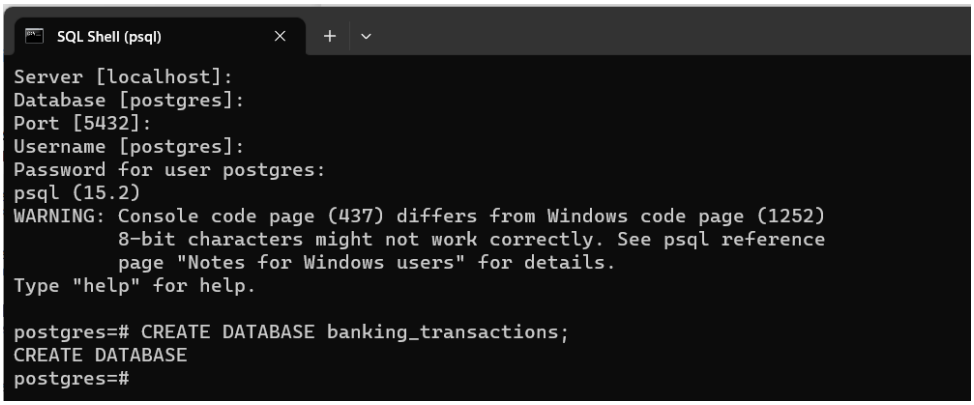
After successfully entered your password, you should be able to see this screen:



Now, create a database called “ banking_transactions” using this command:

```
CREATE DATABASE banking transactions;
```

This is what it should look like after successfully creating the database:



Now, Open the DataStaging.ipynb file and change all the <USERNAME> and <PASSWORD> to your username and password.
Note: the default username is postgres.

You should now be able to execute the DataStaging file without any errors. Note that the file takes ~ 20 mins maximum to run since there are around 1 millions rows to be loaded into the database.

Screenshot of database

```
banking_transactions=# \dt
                List of relations
 Schema |         Name          | Type  |  Owner
--------+-----------------------+-------+----------
 public | applicant_dimension   | table | postgres
 public | device_dimension      | table | postgres
 public | fact_table            | table | postgres
 public | session_dimension     | table | postgres
 public | transaction_dimension | table | postgres
(5 rows)
```

List of tables

```
banking_transactions=# select * from transaction_dimension order by transaction_id asc;
 surrogate_keys |  source  | payment_type |   days_since_request   | intended_balcon_amount | foreign_request | transaction_id
----------------+----------+--------------+------------------------+------------------------+-----------------+----------------
              1 | INTERNET | AA           |     0.0209251728365947 |     -1.3313449634902534 | f               |              1
              2 | INTERNET | AB           |     0.0054175383255355 |     -0.8162237547762208 | f               |              2
              3 | INTERNET | AC           |     3.1085487925698936 |     -0.7557277006560229 | f               |              3
              4 | INTERNET | AB           |     0.0190794348274206 |     -1.2051241582867218 | f               |              4
              5 | INTERNET | AB           |     0.0044405216421238 |     -0.7732757002884915 | f               |              5
              6 | INTERNET | AD           |     0.0282305408875112 |     -0.7482819034992085 | f               |              6
              7 | INTERNET | AB           |     0.0306797774422929 |     -0.2789936040147724 | f               |              7
              8 | INTERNET | AB           |     0.0345566227128476 |     -1.2657210121008342 | t               |              8
              9 | INTERNET | AB           |     0.0206907071020872 |     -1.4420821736875382 | f               |              9
             10 | INTERNET | AB           |     0.0168093817593339 |      -1.070271419761631 | f               |             10
             11 | INTERNET | AA           |      6.735898071202028 |       40.49766199331813 | f               |             11
             12 | INTERNET | AD           |     0.0099104470688059 |     -1.0134485689621835 | f               |             12
             13 | INTERNET | AC           |     0.0355594497190684 |     -0.0509913528587939 | t               |             13
             14 | INTERNET | AA           |         7.391355195863 |       35.18835166664694 | f               |             14
             15 | INTERNET | AB           |     0.0014826980676734 |     -0.3713518978809396 | f               |             15
             16 | INTERNET | AB           |      23.81917314185727 |     -1.5629995366296117 | f               |             16
             17 | INTERNET | AB           |     0.0421030519029779 |     -1.2160216193591868 | f               |             17
             18 | INTERNET | AB           |     0.0125625748427899 |     -1.1824886597059392 | f               |             18
             19 | INTERNET | AD           |      0.000970042231772 |     -0.9311672482319386 | f               |             19
             20 | INTERNET | AC           |     0.0015611431759258 |     -0.5400282566552649 | f               |             20
             21 | INTERNET | AC           |     0.0071755927852316 |     -0.8039460382841104 | f               |             21
             22 | INTERNET | AC           |     0.0009367142424911 |      -1.313046149017068 | f               |             22
             23 | INTERNET | AB           |     0.0243404283948482 |     -0.4976780412860044 | t               |             23
             24 | INTERNET | AD           |     0.0074835148893795 |     -0.6941221327065037 | f               |             24
             25 | INTERNET | AD           |     0.0154273374775722 |     -0.2364950634927995 | f               |             25
             26 | INTERNET | AC           |     0.0019609415308556 |     -0.7039938462322493 | f               |             26
             27 | INTERNET | AC           |     0.0077648524294124 |     -1.3573855754665622 | f               |             27
             28 | INTERNET | AB           |     0.0086988502643446 |     -0.4804367327066812 | f               |             28
             29 | INTERNET | AD           |     0.0051328957015646 |     -0.3779346838062124 | f               |             29
             30 | INTERNET | AA           |     0.0066218406638321 |      31.221784648037247 | f               |             30
```

Transaction Dimension

```
banking_transactions=# select * from session_dimension order by session_id asc;
 surrogate_keys | session_length_in_minutes | keep_alive_session | session_id
----------------+--------------------------+--------------------+------------
              1 |         3.888114604789093 | f                  |          1
              2 |         31.79881936362456 | f                  |          2
              3 |         4.728704865428253 | f                  |          3
              4 |         2.047904421972764 | f                  |          4
              5 |         3.775224949895108 | t                  |          5
              6 |         4.815073224292104 | f                  |          6
              7 |        1.5589774670276988 | t                  |          7
              8 |         2.637471764405503 | f                  |          8
              9 |          2.17541930838834 | t                  |          9
             10 |         24.04072646710152 | f                  |         10
             11 |        17.023723899591435 | t                  |         11
             12 |         5.099527595473899 | f                  |         12
             13 |         9.608223134325051 | f                  |         13
             14 |        11.233453093193631 | t                  |         14
             15 |        1.2082326969138937 | f                  |         15
             16 |        10.399931490451776 | f                  |         16
             17 |         2.826891925752153 | f                  |         17
             18 |         4.708033436050976 | t                  |         18
             19 |         3.761719379192329 | f                  |         19
             20 |         2.129957136905311 | t                  |         20
             21 |         3.635425225263532 | f                  |         21
             22 |        1.3831280043607412 | f                  |         22
             23 |         5.365338002018468 | t                  |         23
             24 |         3.745341970593069 | t                  |         24
             25 |        11.610120366926893 | t                  |         25
             26 |        2.2616695671393114 | t                  |         26
             27 |           33.686044144782 | f                  |         27
             28 |        3.1170719188707223 | f                  |         28
             29 |         6.108820685131532 | f                  |         29
-- More  --
```

Session Dimension

```
banking_transactions=# select * from device_dimension order by device_id asc;
 surrogate_keys | device_os | device_distinct_emails_8w | device_id
----------------+-----------+---------------------------+-----------
              1 | windows   |                         1 |         1
              2 | windows   |                         1 |         2
              3 | other     |                         1 |         3
              4 | linux     |                         1 |         4
              5 | macintosh |                         1 |         5
              6 | windows   |                         1 |         6
              7 | windows   |                         1 |         7
              8 | linux     |                         1 |         8
              9 | windows   |                         1 |         9
             10 | windows   |                         1 |        10
             11 | windows   |                         1 |        11
             12 | windows   |                         1 |        12
             13 | linux     |                         1 |        13
             14 | other     |                         1 |        14
             15 | windows   |                         1 |        15
             16 | windows   |                         1 |        16
             17 | windows   |                         1 |        17
             18 | windows   |                         1 |        18
             19 | windows   |                         1 |        19
             20 | other     |                         1 |        20
             21 | linux     |                         1 |        21
             22 | other     |                         1 |        22
             23 | macintosh |                         1 |        23
             24 | windows   |                         1 |        24
             25 | windows   |                         1 |        25
             26 | windows   |                         2 |        26
             27 | windows   |                         1 |        27
             28 | windows   |                         1 |        28
-- More  --
```

Device Dimension

```
banking_transactions=# select * from applicant_dimension order by applicant_id asc;
```

surrogate_keys	income	current_address_months_count	customer_age	employment_status	credit_risk_score	housing_status	phone_valid	has_other_cards	proposed_credit_limit	applicant_id
1	0.9	88	50	CA	185	BA	t	f	500	1
2	0.9	144	50	CA	259	BA	f	f	1500	2
3	0.9	132	40	CB	177	BA	t	f	200	3
4	0.9	22	50	CA	110	BA	t	t	200	4
5	0.9	218	50	CA	295	BA	t	f	1500	5
6	0.3	30	30	CA	199	BB	t	f	200	6
7	0.7000000000000001	152	30	CA	272	BA	t	f	1500	7
8	0.9	18	50	CB	83	BB	t	f	200	8
9	0.7000000000000001	64	40	CC	222	BA	t	f	1500	9
10	0.9	60	40	CB	118	BC	t	f	200	10
11	0.9	131	30	CB	229	BA	t	f	200	11
12	0.9	109	40	CA	95	BC	t	f	200	12
13	0.9	107	30	CA	296	BA	t	f	2000	13
14	0.3	123	20	CB	172	BA	t	f	200	14
15	0.9	37	40	CA	229	BD	t	f	1500	15
16	0.6000000000000001	55	40	CA	119	BC	t	f	200	16
17	0.9	173	40	CA	12	BA	t	f	200	17
18	0.7000000000000001	50	30	CA	302	BA	t	f	2000	18
19	0.9	94	30	CA	199	BA	t	f	500	19
20	0.4	153	30	CA	181	BA	t	f	1500	20
21	0.9	82	50	CA	234	BA	t	f	1900	21
22	0.7000000000000001	9	30	CA	115	BC	t	f	200	22
23	0.7000000000000001	19	30	CA	308	BC	t	f	500	23
24	0.9	184	40	CA	201	BA	t	f	1000	24
25	0.9	56	30	CD	201	BC	t	f	200	25
26	0.6000000000000001	5	50	CC	125	BB	t	f	200	26
27	0.2	90	40	CA	299	BB	t	f	1500	27

```
-- More --
```

Applicant Dimension

```
banking_transactions=# select * from fact_table order by surrogate_keys asc;
```

surrogate_keys	fraud_bool	name_email_similarity	velocity_4w	velocity_24h	velocity_6h	transaction_id	session_id	applicant_id	device_id
1	t	0.166827734424336	3863.647739528353	3134.319630490106	10650.765522770173	1	1	1	1
2	t	0.296286005233516	3124.298165591961	2670.918291734359	534.0473189424272	2	2	2	2
3	t	0.0449854869684285	3159.590678849287	2893.6214979593	4048.534263154862	3	3	3	3
4	t	0.1595111751427926	3022.261811936421	4054.908411692511	3457.064063279491	4	4	4	4
5	t	0.5964137247529342	3087.6709516945257	2728.2371590193657	5020.341679022825	5	5	5	5
6	t	0.1439210346980546	5078.692467258773	3804.803402020777	3223.248405534034	6	6	6	6
7	t	0.3215543951837989	3089.788378522448 5	2653.430035483565	5515.771547984858	7	7	7	7
8	t	0.0648171008513512	3826.129170291501	6733.703889966787	4736.214496039524	8	8	8	8
9	t	0.0659378725949685	3089.271130320971	3849.76142563304	6101.250655067535	9	9	9	9
10	t	0.7000961673506892	3061.2458893457847	3793.831786058914	4504.470395606259	10	10	10	10
11	t	0.7030827386076245	3163.35224038338	3828.922429330861	4980.0932930021645	11	11	11	11
12	t	0.8569579993939715	3117.306377050442	2770.175851100297	4167.44526266585	12	12	12	12
13	t	0.4518019823578503	3653.963027761709	2295.067639242088 7	3017.526730719104	13	13	13	13
14	t	0.3096802961867089	3134.914840963275 7	3798.301047510335	3809.557167188104	14	14	14	14
15	t	0.3913997154223886	3050.927479888197	4426.267514045966	3757.1740082439496	15	15	15	15
16	t	0.302342109459103	3155.497547915961	3709.06126794654	5442.985207294885	16	16	16	16
17	t	0.0666139199012617	3078.978861507933	2385.245786638145	4422.169075009073	17	17	17	17
18	t	0.130321705784085	3082.0498521472973	4084.073289912525	1628.561552729206	18	18	18	18
19	t	0.160811463557145	3148.48489517542	3033.649388971293	875.4912957745943	19	19	19	19
20	t	0.292941279293387 3	3148.747547918706 5	2889.219783857625	3572.932904410458 67	20	20	20	20
21	t	0.0666950068310928	3094.241540502333	3134.771051755900 3	4350.7386839740775	21	21	21	21
22	t	0.2416509836820354	3131.254603980454 7	2434.593014111932 5	1521.957453727831	22	22	22	22
23	t	0.1067045988478192	3103.816648476350 6	3862.791692429976	4520.152626671901	23	23	23	23
24	t	0.1046187555660303	3134.100875689164	2703.462605598934	917.2410886296524	24	24	24	24
25	t	0.4911980715003746	3088.516083082824	2410.635195311260 3	4365.424835886931	25	25	25	25
26	t	0.4906319314396381	3163.076447326457	5970.211899445721	486.3320356111658 5	26	26	26	26
27	t	0.269909074732069	3048.9793599054547	3697.051520767014	5337.83767279693	27	27	27	27
28	t	0.7813964307437908	3083.44618733104	2819.958016657288	3937.468847385452	28	28	28	28
29	t	0.2954980873655803	3101.860075719691	3247.9966316225987	2912.3320515538453	29	29	29	29
30	t	0.1324448252064571	3110.975607997639	2439.519488866076	4258.2978691919225	30	30	30	30
31	t	0.6115531528304505	3120.622077808958 7	2766.342590239584 3	1369.459700076725	31	31	31	31
32	t	0.619040299363078	3062.834533696928	3125.331498586220 7	3481.820194315718 3	32	32	32	32
33	t	0.8327996039283624	3159.591136426860 4	2854.426100713988	5936.17438298968	33	33	33	33
34	t	0.3766713545174072	3072.159144094069	2890.576109671339	4540.208790983512	34	34	34	34
35	t	0.0887248567947381	3093.913405948688	2303.101067098046	1713.244552460546	35	35	35	35
36	t	0.0682209180415192	3097.9990335066864	4517.082313240875	4784.848851905773	36	36	36	36
37	t	0.2019241273123628	3177.3515602545417	2891.699206907177	2004.1554653281632	37	37	37	37
38	t	0.3166236519152105	3112.540696498754	3310.546120521229	2973.519286589496	38	38	38	38
39	t	0.8102625246156132	3073.967976752266	3170.985393644718	4762.717777744596	39	39	39	39

Fact Table

Note: Photos of dimensions and the fact table can be seen in the Results image folder.
Note: All resulting csv can be found in Results csv files.

C : Data Quality Issues

Since our dataset is pretty complete, we didn't face a lot of data quality issues.

One of the issues we had was that the data type for some columns where the values were either 1 or 0 was int64 instead of being boolean. For example: fraud_bool column, has_other_cards column, 'keep_alive_session' column, etc.. We handled this issue by changing the data type to boolean for all those columns.

There was some data missing in the 'prev_address_months_count' (value = -1). We decided to drop this column since those rows make up more than 71% of the dataset, therefore there is more unimportant information than important ones.

We noticed that 'phone_mobile_valid' and 'phone_home_valid' columns are nearly the same thing. So, we decided to combine those two columns into a single column named 'phone_valid' where the value is either 1 or 0 depending on the value for 'phone_mobile_valid' or 'phone_home_valid'. Therefore, 'phone_valid' is 1 if 'phone_mobile_valid' or 'phone_home_valid' is 1 and 0 if 'phone_mobile_valid' and 'phone_home_valid' are both 0.

We didn't have the data duplicate issue since the number of duplicated rows is 0.

We found out that there were some outliers on 'customer_age' and 'velocity_6h' columns. So, we decided to only keep the customers' whose age are less than 70 and for the velocity, we kept the values which were under 13,000.

D : Team Planning

Deliverable Checklist	Responsible Team Member(s)
Create database instance	Ishanveer
Create Transaction dimension	Kate
Create Applicant dimension	Andie
Create Session dimension	Kate
Create device dimension	Andie
Create fact table dimension	Ishanveer
Staging of dimension transaction	Ishanveer
Staging of dimension applicant	Andie
Staging of dimension Session	Kate
Staging of dimension Device	Andie
Surrogate key pipeline	Kate
Staging of fact table-including FKS and measures	Andie
Data quality handling and reporting	Ishanveer
Learning how to use postgresql	Andie, Kate, Ishanveer

A more informative version of the team planning can be found in the Phase2-Team Planning_W23_.xlsx file.

Meeting with TA

Meeting Date	Meeting Duration	Meeting Attendees
23/03/2023	15 minutes	Kate, Andie, Ishanveer

References

High Level data staging plan using mermaid
https://towardsdatascience.com/sqlalchemy-python-tutorial-79a577141a91
https://www.geeksforgeeks.org/introduction-to-psycopg2-module-in-python/
CSI4142 Practical Session 1- Data Staging
https://chat.openai.com/chat
https://www.postgresql.org/