Data Audit

November 3, 2023

In [1]:
```
import pandas as pd
from pathlib import Path
from abc import ABC, abstractmethod
```

In [2]:
```
# create an interface
class DiscrepanciesInterface(ABC):
    @abstractmethod
    def uploading_discrepancies(self, path_raw_csv: Path, path_database_A: Path) -> pd
        pass
    @abstractmethod
    def streaming_discrepancies(self, path_database_A: Path, path_database_B: Path) ->
        pass
```

In [3]:
```
class Discrepancies(DiscrepanciesInterface):
    def uploading_discrepancies(self, path_raw_csv: Path, path_database_A: Path) -> pd
        try:
            df_raw = pd.read_csv(path_raw_csv, index_col=0)
            df_db_a = pd.read_csv(path_database_A)

            """Transform raw csv data from wide to longitude and attach col headers si

            # Fill missing values with 0
            df_raw = df_raw.fillna(0)
            # We use Stack to transform the DataFrame and reset the index
            melted_df = df_raw.stack().reset_index()

            # Rename the columns
            melted_df.columns = ['deliveryDate', 'farmerId', 'quantity']
            # Reorder the columns to match db A
            melted_df = melted_df[['farmerId', 'deliveryDate', 'quantity']]

            """Compare Transformed raw csv and db A involved in the Upload process"""

            #merge two dfs and return those that are not on both
            result = melted_df.merge(df_db_a, indicator=True, how='outer').loc[lambda
            return result

        except Exception as e:
```

```
                return f"-Error "+ f"{type(e).__name__} {str(e)}"

            def streaming_discrepancies(self, path_database_A: Path, path_database_B: Path) ->
                try:
                    #compare db A and db B involved in the streaming process
                    df_db_a = pd.read_csv(path_database_A)
                    df_db_b = pd.read_csv(path_database_B)

                    result = df_db_a.merge(df_db_b, indicator=True, how='outer').loc[lambda v:
                    return result
                except Exception as e:
                    return f"-Error "+ f"{type(e).__name__} {str(e)}"
```

In [4]:
```
disc = Discrepancies() #instantiate class
```

In [5]:
```
disc.uploading_discrepancies('Raw_Data.csv', 'Database_A.csv') #check discrepancies in
```

Out[5]:

	farmerId	deliveryDate	quantity	_merge
2	19fb0be7-b2b2-4038-80ba-42baabc21c0a	2019-01-01	149.0	left_only
7	584c4094-a231-4c6d-9c49-63b7192bf8a1	2019-01-01	29.0	left_only
15	82b77709-85cc-4a20-9893-cec85ff6d7bd	2019-01-01	0.0	left_only
20	9eb16e61-5d1b-43c7-91bd-7849280e260f	2019-01-01	0.0	left_only
22	573f09bd-c649-448a-b6c9-68f55cf8aa25	2019-01-01	0.0	left_only
24	e1b85e0e-bbe8-42fc-a564-8d296e35af50	2019-01-01	59.0	left_only
25	ff367fff-2f38-4146-bc0e-d5ca7230caf7	2019-01-01	84.0	left_only
27	5633d279-21ce-46e9-b8f9-80db6b071904	2019-01-01	504.0	left_only
41	64eca1f3-b712-467b-976b-107be26d8b18	2019-01-01	0.0	left_only
44	3584558d-d6f2-4b7d-9769-1f8a751782ab	2019-01-01	141.0	left_only
48	b5eb285b-fae0-4bde-bd32-afc9619c07a3	2019-01-01	747.0	left_only
54	4b498bd4-5658-4800-9ae8-65180acc8f2c	2019-01-01	300.0	left_only
55	86f2a01a-fd9b-413a-8458-738ec57b2970	2019-01-01	0.0	left_only
57	13bcd844-c49b-4c9f-b924-882c58e5ec51	2019-01-01	0.0	left_only
60	39e141a0-607b-4b80-a23d-540271f8bcfd	2019-01-01	0.0	left_only
64	5833a225-7752-471a-8e1f-283ba6176c51	2019-01-01	0.0	left_only
65	3390816a-63c2-4250-8d16-e081cb57f6a3	2019-01-01	139.0	left_only
66	4ceed27f-1b18-4473-892c-8723442c3507	2019-01-01	0.0	left_only
67	caa45b50-4d9d-4de3-bc3b-8e4384750251	2019-01-01	0.0	left_only
68	45e02f44-d9a4-4dff-b11c-0b8be0d8b584	2019-01-01	0.0	left_only
70	749b1fee-a4c5-4c49-af3c-952701e4b7b9	2019-01-01	0.0	left_only
71	c9ae1a8b-52dd-4667-9b3c-fcf975a626e7	2019-01-01	0.0	left_only
75	d619a04f-76a2-4d70-8c5c-f9a1f4a6a545	2019-01-01	0.0	left_only
76	47e9ef48-799c-424e-99de-a0a68ecff0d1	2019-01-01	0.0	left_only
77	c22306fa-b771-4c54-afab-0778ea5667b7	2019-01-01	0.0	left_only
78	d6e5dc47-91c9-44e7-bbfe-fb42eeddb29a	2019-01-01	0.0	left_only
79	3b2b8e35-21f9-4f74-a873-efb0ef3faf1d	2019-01-01	0.0	left_only
82	41c2bcd6-2bfa-4be7-9bef-1832f2756768	2019-01-01	0.0	left_only
84	b8c20b8c-da35-441b-af55-26ec32c684d0	2019-01-01	0.0	left_only

```
86      630df0ec-de5a-45ba-ac5b-17758171a592   2019-01-01       0.0  left_only
...                                      ...          ...       ...        ...
174380  f2d1d469-fed4-45ca-ac2c-9ba07a7ecb01   2022-01-15       0.0  left_only
174381  88723cb4-1d3e-4ee1-95ff-19421a87c41e   2022-01-15       0.0  left_only
174382  15ac7909-50b0-4fdb-beb5-9d773786408e   2022-01-15       0.0  left_only
174383  d65be89a-6e06-4b1e-8231-07088ff9070e   2022-01-15       0.0  left_only
174384  77ba07cc-1333-48af-aeb0-9d633c2d0b52   2022-01-15       0.0  left_only
174385  ec815cf3-8d98-4c9b-ac79-62b1d32945b0   2022-01-15       0.0  left_only
174387  afea71bb-d733-418c-847a-60d5db57ff55   2022-01-15       0.0  left_only
174388  3fa303d3-84d2-4fc9-9d11-1385be5b7b58   2022-01-15       0.0  left_only
174390  fbc94d7f-3b65-4e66-8fd9-6f68e61e7044   2022-01-15       0.0  left_only
174392  2f9d9a45-9b01-49ba-b25a-948aa611c472   2022-01-15       0.0  left_only
174393  e3a2b390-6df7-415d-a2af-ba063333f683   2022-01-15       0.0  left_only
174395  5f27cdf3-6c8b-4022-8f2f-82740bbed517   2022-01-15      68.0  left_only
174397  1167a1e0-421e-4d44-8f25-9291ba95eb08   2022-01-15       0.0  left_only
174398  62eb8c86-117c-436f-8d69-95e68579a0b4   2022-01-15       0.0  left_only
174402  acb75fff-9168-40cb-9f58-23b35d9eaf3a   2022-01-15       0.0  left_only
174403  afd6e467-263b-43c2-922d-2dfcbea28c66   2022-01-15       0.0  left_only
174405  d88e3c2d-d981-41bd-9c2f-ce4a3e4cb672   2022-01-15     518.0  left_only
174407  35b37e89-eb11-43ec-b622-6dd8f51db641   2022-01-15       0.0  left_only
174409  cb4d211a-3cd9-410e-ab20-e7b096912790   2022-01-15       0.0  left_only
174411  4b304df5-7ab3-4bff-ac5e-3a05ae847547   2022-01-15       0.0  left_only
174414  40638063-a7bc-4727-a62b-3a99f5f80415   2022-01-15       0.0  left_only
174415  01aadc39-f551-49a5-87b8-43c84dff0d76   2022-01-15       0.0  left_only
174416  69581b79-269f-4e7a-8740-82dae97e833e   2022-01-15       0.0  left_only
174417  e1a5e626-df06-4fa4-bb0d-8a6c6c749269   2022-01-15       0.0  left_only
174419  68a8d97c-cfed-43b2-bb86-4db25e13a6c3   2022-01-15       0.0  left_only
174421  1b000644-36f4-4e12-93d3-e6750179b51e   2022-01-15       0.0  left_only
174422  418acdde-7afb-483e-8b08-f7f92fb330f8   2022-01-15       0.0  left_only
174423  781c4cba-fd27-4fe4-a278-e19de81be2bf   2022-01-15       0.0  left_only
174424  88ef55fa-219e-45db-b4cc-8f0049696d98   2022-01-15       0.0  left_only
174425  45fc6235-a972-4057-a5b8-eef323bda0e5   2022-01-15       0.0  left_only

[76858 rows x 4 columns]
```

In [6]:
```
"""
    Uploading Discrepancies:
    The discrepancies in uploading data shows that raw csv file has extra rows and val
    left_only merge. The left only merge col shows rows available in raw csv data that
    merges for this discrepancy.

    Action Points for correction:
    1. Implement data integrity checks, such as checksums or hashing to ensure data is
    2. Implement a retry mechanism in your data transfer process to handle incomplete
"""
```

Out[6]: '\n   Uploading Discrepancies:\n   The discrepancies in uploading data shows that raw

In [7]:
```
disc.streaming_discrepancies('Database_A.csv', 'Database_B.csv') #check discrepancies
```

Out[7]:

	farmerId	deliveryDate	quantity	\
1	dc7a4468-9eba-4f0d-aedd-c57189467f18	2021-08-04	91	
10	f05d1876-8002-4c38-a7a8-bbf94bcc0159	2021-07-15	8	
14	dc7a4468-9eba-4f0d-aedd-c57189467f18	2021-10-19	808	
25	58f674c2-662e-4d6e-84de-ed0c486a2a63	2020-05-15	54	
29	ad81ce4b-5686-4a6e-9d7a-d9a112834706	2021-02-15	31	
37	dc7a4468-9eba-4f0d-aedd-c57189467f18	2021-07-15	822	
44	34da3784-d47d-4459-97b5-4f6fc800a503	2019-06-21	107	
47	20a2cdfc-e506-4d2a-80f4-957299729790	2020-09-18	716	
52	9f78b129-d618-4d13-8e68-3adc922e8176	2020-06-15	178	
55	621a6f43-431c-4a08-aca2-2ec4553d82d3	2020-10-07	388	
56	60e00f14-6b52-466d-ab2b-023067b78472	2020-11-10	237	
79	acba7469-ced3-4dbb-818a-09c6cda733a5	2020-02-04	262	
89	4bfd270b-2ab1-46a2-80f0-b1fc40fce073	2021-08-20	61	
90	9f78b129-d618-4d13-8e68-3adc922e8176	2020-08-21	27	
91	24d5bbc9-9979-42d2-bc3b-89cb34a17eb1	2020-05-30	401	
131	112c0602-3515-472c-8e8a-7ef8fcef89ae	2021-02-13	591	
150	acba7469-ced3-4dbb-818a-09c6cda733a5	2021-02-02	375	
154	dc7a4468-9eba-4f0d-aedd-c57189467f18	2020-10-09	808	
160	acba7469-ced3-4dbb-818a-09c6cda733a5	2021-03-08	34	
164	afd6e467-263b-43c2-922d-2dfcbea28c66	2020-07-16	25	
168	58f674c2-662e-4d6e-84de-ed0c486a2a63	2021-07-31	258	
171	acba7469-ced3-4dbb-818a-09c6cda733a5	2021-06-02	267	
175	ad81ce4b-5686-4a6e-9d7a-d9a112834706	2020-09-12	22	
182	20a2cdfc-e506-4d2a-80f4-957299729790	2019-07-02	94	
183	58f674c2-662e-4d6e-84de-ed0c486a2a63	2019-06-09	296	
197	621a6f43-431c-4a08-aca2-2ec4553d82d3	2021-08-08	363	
201	9f78b129-d618-4d13-8e68-3adc922e8176	2021-03-07	176	
212	ad81ce4b-5686-4a6e-9d7a-d9a112834706	2021-07-06	33	
214	afd6e467-263b-43c2-922d-2dfcbea28c66	2021-04-25	33	
216	58f674c2-662e-4d6e-84de-ed0c486a2a63	2019-04-04	31	
...	...	...	...	
110242	3584558d-d6f2-4b7d-9769-1f8a751782ab	2020-07-22	26	
110243	c9ae1a8b-52dd-4667-9b3c-fcf975a626e7	2019-07-13	28	
110244	a235b83b-2308-4bb2-8f9c-e63e3d9df511	2020-11-11	833	
110245	3390816a-63c2-4250-8d16-e081cb57f6a3	2019-10-30	165	
110246	5f27cdf3-6c8b-4022-8f2f-82740bbed517	2019-09-23	67	
110247	584c4094-a231-4c6d-9c49-63b7192bf8a1	2019-06-05	234	
110248	584c4094-a231-4c6d-9c49-63b7192bf8a1	2019-06-08	47	
110249	584c4094-a231-4c6d-9c49-63b7192bf8a1	2019-02-26	25	
110250	3390816a-63c2-4250-8d16-e081cb57f6a3	2019-09-24	809	
110251	584c4094-a231-4c6d-9c49-63b7192bf8a1	2021-05-07	216	
110252	a235b83b-2308-4bb2-8f9c-e63e3d9df511	2021-04-15	843	
110253	b5eb285b-fae0-4bde-bd32-afc9619c07a3	2019-08-18	143	
110254	5633d279-21ce-46e9-b8f9-80db6b071904	2020-07-16	83	
110255	a235b83b-2308-4bb2-8f9c-e63e3d9df511	2020-03-11	858	
110256	d88e3c2d-d981-41bd-9c2f-ce4a3e4cb672	2020-12-07	56	
110257	a235b83b-2308-4bb2-8f9c-e63e3d9df511	2021-09-23	823	

```
110258  a235b83b-2308-4bb2-8f9c-e63e3d9df511   2020-06-09       116
110259  f7cf47b8-886d-4094-8f2c-4a06914a0065   2020-06-27       215
110260  4b498bd4-5658-4800-9ae8-65180acc8f2c   2019-05-21       274
110261  c9ae1a8b-52dd-4667-9b3c-fcf975a626e7   2019-02-04        28
110262  45e02f44-d9a4-4dff-b11c-0b8be0d8b584   2021-04-15       118
110263  e1b85e0e-bbe8-42fc-a564-8d296e35af50   2020-07-06       420
110264  1b000644-36f4-4e12-93d3-e6750179b51e   2019-10-28        73
110265  b5eb285b-fae0-4bde-bd32-afc9619c07a3   2021-04-23       717
110266  a235b83b-2308-4bb2-8f9c-e63e3d9df511   2021-04-16        90
110267  a235b83b-2308-4bb2-8f9c-e63e3d9df511   2019-08-09       163
110268  f7cf47b8-886d-4094-8f2c-4a06914a0065   2019-06-20       214
110269  ec815cf3-8d98-4c9b-ac79-62b1d32945b0   2019-02-17       476
110270  584c4094-a231-4c6d-9c49-63b7192bf8a1   2020-09-21        41
110271  b5eb285b-fae0-4bde-bd32-afc9619c07a3   2019-08-29       747

            _merge
1        left_only
10       left_only
14       left_only
25       left_only
29       left_only
37       left_only
44       left_only
47       left_only
52       left_only
55       left_only
56       left_only
79       left_only
89       left_only
90       left_only
91       left_only
131      left_only
150      left_only
154      left_only
160      left_only
164      left_only
168      left_only
171      left_only
175      left_only
182      left_only
183      left_only
197      left_only
201      left_only
212      left_only
214      left_only
216      left_only
...            ...
110242  right_only
```

```
110243  right_only
110244  right_only
110245  right_only
110246  right_only
110247  right_only
110248  right_only
110249  right_only
110250  right_only
110251  right_only
110252  right_only
110253  right_only
110254  right_only
110255  right_only
110256  right_only
110257  right_only
110258  right_only
110259  right_only
110260  right_only
110261  right_only
110262  right_only
110263  right_only
110264  right_only
110265  right_only
110266  right_only
110267  right_only
110268  right_only
110269  right_only
110270  right_only
110271  right_only

[29258 rows x 4 columns]
```

In [8]:
```
"""
    Streaming Discrepancies:
    The discrepancies in streaming data from db A to db B indicates that there are val
    db B indicated by left_only. There are also values transfered to db B from db A th
    by right_only.
    Some ways to handle streaming discrepancies
    1. Thoroughly test data pipelines and processing logic to catch potential issues.
    2. Data Quality Monitoring: Continuously monitor data quality and set up alerts fo
"""
```

Out[8]:
```
'\n    Streaming Discrepancies:\n'
```