

testcase2_parquet_parquet_match

content comparison

1. Run Summary

Application Name	:sample
Protocol Name	:DATFDemo
Protocol File Path	:/Workspace/Repos/sahil.gupta@tigeranalytics.com/QE_ATF/datf_core/test/testprotocol/testprotocol.xlsx
Testcase Name	:testcase2_parquet_parquet_match
Testcase Type	:content
Test Environment	:dev
Start Time	:30-Aug-2024 12:11:40 UTC
End Time	:30-Aug-2024 12:13:52 UTC
Run Time	:0:02:11
Test Result	:Failed
Reason	:Content mismatched

2. Configuration Details

Compare Type	:s2tcompare
testquerygenerationmode	:Auto
Testcase Type	:content
Source Connection Type	:aws-s3
Source Format	:parquet
Source Path	:/Workspace/Repos/sahil.gupta@tigeranalytics.com/QE_ATF/datf_core/test/data/source/patients_source_parquet
Target Connection Type	:aws-s3
Target Format	:delta
Target Path	:hive_metastore.default.patients_target_modified_new
S2T Path	:test/s2t/source_parquet_target_delta_auto.xlsx
Primary Keys	:id_new

3. Content Summary

Test Result	:Failed
No. of matched columns	:25
No. of columns compared	:24
No. of cols in Source but not in Target	:0
No. of cols in Target but not in Source	:0
No. of rows in Source	:1,171
No. of distinct rows in Source	:1,171
No. of duplicate rows in Source	:0
No. of rows in Target	:1,169
No. of distinct rows in Target	:1,169
No. of duplicate rows in Target	:0
No. of matched rows	:0
No. of mismatched rows	:1,169
No. of rows in Source but not in Target	:2
No. of rows in Target but not in Source	:0

4. SQL Queries

4.1 Source Query

```
readdatadf=spark.read.format('parquet').load('file:/Workspace/Repos/sahil.gupta@tigeranalytics.com/QE_ATF/datf_core/test/data/source/patients_source_parquet')
```

```
readdatadf.createOrReplaceTempView('dataview')
```

```
spark.sql("SELECT src.id as id_new, src.BIRTHDATE as BIRTHDATE, src.DEATHDATE as DEATHDATE, src.SSN as SSN, src.DRIVERS as DRIVERS, src.PASSPORT as PASSPORT, src.PREFIX as PREFIX, src.FIRST as FIRST, src.LAST as LAST, src.SUFFIX as SUFFIX, src.MAIDEN as MAIDEN, src.MARITAL as MARITAL, src.RACE as RACE, src.ETHNICITY as ETHNICITY, src.GENDER as GENDER, src.BIRTHPLACE as BIRTHPLACE, src.ADDRESS as ADDRESS, src.CITY as CITY, src.STATE as STATE, src.COUNTY as COUNTY, src.ZIP as ZIP, src.LAT as LAT, src.LON as LON,
```

src.HEALTHCARE_EXPENSES as HEALTHCARE_EXPENSES, src.HEALTHCARE_COVERAGE as HEALTHCARE_COVERAGE FROM dataview src ")

4.2 Target Query

readdatadf=spark.table('hive_metastore.default.patients_target_modified_new')

readdatadf.createOrReplaceTempView('dataview')

spark.sql("SELECT id_new, BIRTHDATE, DEATHDATE, SSN, DRIVERS, PASSPORT, PREFIX, FIRST, LAST, SUFFIX, MAIDEN, MARITAL, RACE, ETHNICITY, GENDER, BIRTHPLACE, ADDRESS, CITY, STATE, COUNTY, ZIP, LAT, LON, HEALTHCARE_EXPENSES, HEALTHCARE_COVERAGE FROM dataview tgt ")

5. Sample Mismatches 5 rows

5.1 Keys in source but not in target

S.No	Key Columns
1	id_new=88ea8573-863c-47e3-b144-b810c63156a0
2	id_new=0982ef39-7ff9-4c24-8239-e9fc0667e8ca

5.2 Keys in target but not in source

None

5.3 Keys having one or more unequal column values

S.No	Key Columns
1	id_new=00185faa-2760-4218-9bf5-db301acf8274
2	id_new=0042862c-9889-4a2e-b782-fac1e540ecb4
3	id_new=0047123f-12e7-486c-82df-53b3a450e365
4	id_new=010d4a3a-2316-45ed-ae15-16f01c611674
5	id_new=01207ecd-9dff-4754-8887-4652eda231e2

6. Columnwise Mismatch Summary

No. of rows in Source :1171
No. of rows in Target :1169

S.No	Source Column Name	Target Column Name	Rows in Source	Rows in Target	Rows with Common Keys	Rows Matched	Rows Mismatch
1	BIRTHDATE	BIRTHDATE	1171	1169	0	0	1169
2	DEATHDATE	DEATHDATE	1171	1169	0	998	171
3	MARITAL	MARITAL	1171	1169	0	1167	2
4	RACE	RACE	1171	1169	0	1164	5
5	ETHNICITY	ETHNICITY	1171	1169	0	1163	6
6	ZIP	ZIP	1171	1169	0	541	628
7	LAT	LAT	1171	1169	0	0	1169
8	LON	LON	1171	1169	0	0	1169
9	HEALTHCARE_EXPENSES	HEALTHCARE_EXPENSES	1171	1169	0	1084	85
10	HEALTHCARE_COVERAGE	HEALTHCARE_COVERAGE	1171	1169	0	550	619

7. Columnwise Mismatch Details

7.1 BIRTHDATE

S.No	Key Columns	Source value	Target value
1	id_new=7b51cca8-284b-42a5-8117-cc21f48454f5	1979-12-25	1979-12-25 00:00:00
2	id_new=d95287b4-213f-4b3b-b68e-b23b9f3bac90	1937-09-07	1937-09-07 00:00:00
3	id_new=58100357-85b6-4334-bc2a-8998c24a4812	1971-04-15	1971-04-15 00:00:00
4	id_new=cf12424d-63da-4fbb-b93f-22308793cafa	1964-02-18	1964-02-18 00:00:00
5	id_new=eedb8836-d94c-4a70-8491-bea52efc4ce8	1972-03-15	1972-03-15 00:00:00

7.2 DEATHDATE

S.No	Key Columns	Source value	Target value
1	id_new=d95287b4-213f-4b3b-b68e-b23b9f3bac90	1999-09-07	1999-09-07 00:00:00
2	id_new=9e5a9e59-cfe3-42de-8c2e-76ef7e2efa5f	2009-12-01	2009-12-01 00:00:00
3	id_new=59cf17d9-6c13-4333-a1cb-cc5fdf63366d	1995-10-14	1995-10-14 00:00:00
4	id_new=44c8b4c8-4457-4ba0-94b6-285ef327a249	1989-01-17	1989-01-17 00:00:00
5	id_new=b48a38ae-3bb7-4e52-add7-bce581c87262	1999-03-24	1999-03-24 00:00:00

7.3 MARITAL

S.No	Key Columns	Source value	Target value
1	id_new=034e9e3b-2def-4559-bb2a-7850888ae060	M	None
2	id_new=1d604da9-9a81-4ba9-80c2-de3375d59b40	M	None

7.4 RACE

S.No	Key Columns	Source value	Target value
1	id_new=1fc8dafd-90e5-44d6-9179-81156c0ea083	white	None
2	id_new=fc817953-cc8b-45db-9c85-7c0ced8fa90d	white	None
3	id_new=6c2b1cfd-887a-43f8-bee0-000ce2fc6ef0	white	None
4	id_new=6d048a56-edb8-4f29-891d-7a84d75a8e78	white	None
5	id_new=fca3178e-fb68-41c3-8598-702d3ca68b96	white	None

7.5 ETHNICITY

S.No	Key Columns	Source value	Target value
1	id_new=73fc00f8-a341-4f62-be27-07c35c3460bc	nonhispanic	None
2	id_new=8a9a9b41-2e52-4805-92ed-3be1c2972c41	nonhispanic	None
3	id_new=1bacaab1-708b-42fe-8f12-fbe97a725440	nonhispanic	None
4	id_new=35e1afdd-e938-441d-b54e-db15db0bbcfd	nonhispanic	None
5	id_new=c1a6f986-01c4-4581-ba74-bf22d28f8ace	nonhispanic	None

7.6 ZIP

S.No	Key Columns	Source value	Target value
1	id_new=cf12424d-63da-4fbb-b93f-22308793cafa	02492	2492
2	id_new=4f9dcdaa-fb47-412b-8288-03f8cbbc321e	01462	1462
3	id_new=431971f5-a39f-44f8-9ef9-10502db973c9	01905	1905
4	id_new=73fc00f8-a341-4f62-be27-07c35c3460bc	01027	1027
5	id_new=5f08cc74-e542-4b93-a565-bf802a03d529	01510	1510

7.7 LAT

S.No	Key Columns	Source value	Target value
1	id_new=d95287b4-213f-4b3b-b68e-b23b9f3bac90	41.60341019541261	41.6034102
2	id_new=7b51cca8-284b-42a5-8117-cc21f48454f5	42.581534819112115	42.58153482
3	id_new=cf12424d-63da-4fbb-b93f-22308793cafa	42.314108546231964	42.31410855
4	id_new=58100357-85b6-4334-bc2a-8998c24a4812	42.56149882305895	42.56149882
5	id_new=eedb8836-d94c-4a70-8491-bea52efc4ce8	42.1606685582146	42.16066856

7.8 LON

S.No	Key Columns	Source value	Target value
1	id_new=cf12424d-63da-4fbb-b93f-22308793cafa	-71.25656899788075	-71.256569
2	id_new=7b51cca8-284b-42a5-8117-cc21f48454f5	-73.11661133251157	-73.11661133
3	id_new=eedb8836-d94c-4a70-8491-bea52efc4ce8	-71.33654032185571	-71.33654032
4	id_new=d95287b4-213f-4b3b-b68e-b23b9f3bac90	-70.87525169645723	-70.8752517
5	id_new=58100357-85b6-4334-bc2a-8998c24a4812	-71.68665525691193	-71.68665526

7.9 HEALTHCARE_EXPENSES

S.No	Key Columns	Source value	Target value
1	id_new=431971f5-a39f-44f8-9ef9-10502db973c9	4682.139999999999	4682.14
2	id_new=2950e244-0ba6-4afe-ac30-e0d5529d1d1b	437988.83999999997	437988.84
3	id_new=bae73dce-20fa-4a6a-8e46-b2ed541f9296	30854.00999999999	30854.01
4	id_new=14289f20-085c-4fc8-bdd8-e2074166d91f	983073.4700000001	983073.47
5	id_new=ab88386a-1c0d-4d1c-89fc-b38f631b3edc	753274.9299999999	753274.93

7.10 HEALTHCARE_COVERAGE

S.No	Key Columns	Source value	Target value
1	id_new=58100357-85b6-4334-bc2a-8998c24a4812	10790.559999999998	10790.56
2	id_new=7b51cca8-284b-42a5-8117-cc21f48454f5	6857.5999999999985	6857.6
3	id_new=9e5a9e59-cfe3-42de-8c2e-76ef7e2efa5f	19327.539999999997	19327.54
4	id_new=4f9dcdaa-fb47-412b-8288-03f8cbbc321e	51530.63999999999	51530.64
5	id_new=73fc00f8-a341-4f62-be27-07c35c3460bc	15092.529999999999	15092.53