testcase1_parquet_parquet_mismatch

content comparison

1. Run Summary

Application Name	:etl_pipeline_goldlayer
Protocol Name	:DATFDemo
Protocol File Path	:/Workspace/Shared/QE_ATF_Enhanced_03/datf_core/test/testprotocol/traversedtestprotocol_schema_loman.xlsx
Testcase Name	:testcase1_parquet_parquet_mismatch
Testcase Type	:content
Test Environment	:SIT
Start Time	:10-Apr-2025 05:36:29 UTC
End Time	:10-Apr-2025 05:40:17 UTC
Run Time	:0:03:48
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	:file:/Workspace/Shared/QE_ATF_Enhanced_03/datf_core/test/data/source/patients_source_parquet
Target Connection Type	:aws-s3
Target Format	:parquet
Target Path	:file:/Workspace/Shared/QE_ATF_Enhanced_03/datf_core/test/data/target/patients_target_parquet_mismatch
S2T Path	:test/s2t/s2t_1_parquet_parquet_mismatch.xlsx
Primary Keys	:id

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	:1,169
No. of mismatched rows	:0
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/Shared/QE_ATF_Enhanced_03/datf_core/test/data/source/patients_source_parquet')

readdatadf.createOrReplaceTempView('dataview')

spark.sql("SELECT src.id as id, 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.read.format('parquet').load('file:/Workspace/Shared/QE_ATF_Enhanced_03/datf_core/test/data/target/patients_target_parquet_mismatch')

readdatadf.createOrReplaceTempView('dataview')

spark.sql("SELECT id, 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=034e9e3b-2def-4559-bb2a-7850888ae060
2	id=1d604da9-9a81-4ba9-80c2-de3375d59b40

5.2 Keys in target but not in source

None

5.3 Keys having one or more unequal column values

None

6. Columnwise Mismatch Summary

None

7. Columnwise Mismatch Details

None

testcase12_parquet_parquet_mismatch_manual

content comparison

1. Run Summary

Application Name	:etl_pipeline_goldlayer
Protocol Name	:DATFDemo
Protocol File Path	:/Workspace/Shared/QE_ATF_Enhanced_03/datf_core/test/testprotocol/traversedtestprotocol_schema_loman.xlsx
Testcase Name	:testcase12_parquet_parquet_mismatch_manual
Testcase Type	:content
Test Environment	:SIT
Start Time	:10-Apr-2025 05:40:26 UTC
End Time	:10-Apr-2025 05:43:22 UTC
Run Time	:0:02:56
Test Result	:Passed
Reason	:Content matched

2. Configuration Details

Compare Type	:likeobjectcompare
testquerygenerationmode	:Auto
Testcase Type	:content
Source Connection Name	:RawS3Bucket
Source Connection Type	:aws-s3
Source Connection Value	:RawS3Bucket
Source Format	:parquet
Source Name	:patients_source_parquet
Source Path	:test/data/source
Target Connection Name	:CuratedS3Bucket
Target Connection Type	:aws-s3
Target Connection Value	:CuratedS3Bucket
Target Format	:parquet
Target Name	:patients_source_parquet
Target Path	:test/data/source
S2T Path	:test/s2t/s2t_1_parquet_parquet_mismatch.xlsx
Primary Keys	:id

3. Content Summary

Test Result	:Passed
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	:1,171
No. of distinct rows in Target	:1,171
No. of duplicate rows in Target	:0
No. of matched rows	:1,171
No. of mismatched rows	:0
No. of rows in Source but not in Target	:0
No. of rows in Target but not in Source	:0

4. SQL Queries

4.1 Source Query

```
SELECT
ADDRESS,BIRTHDATE,BIRTHPLACE,CITY,COUNTY,DEATHDATE,DRIVERS,ETHNICITY,FIRST,GENDER,HEALTHCARE_COVERAGE,HEALTHCARE_EXPENSES,LAST,LAT,LON,MAIDEN,MARITAL,PASSPORT,PREFIX,RACE,SSN,STATE,SUFFIX,ZIP,id FROM rawpatients
```

4.2 Target Query

```
SELECT
ADDRESS,BIRTHDATE,BIRTHPLACE,CITY,COUNTY,DEATHDATE,DRIVERS,ETHNICITY,FIRST,GENDER,HEALTHCARE_COVERAGE,HEALTHCARE_EXPENSES,LAST,LAT,LON,MAIDEN,MARITAL,PASSPORT,PREFIX,RACE,SSN,STATE,SUFFIX,ZIP,id FROM curatedpatients
```

5. Sample Mismatches 8 rows

5.1 Keys in source but not in target

None

5.2 Keys in target but not in source

None

5.3 Keys having one or more unequal column values

None

6. Columnwise Mismatch Summary

None

7. Columnwise Mismatch Details

None