testcase1_parquet_parquet_mismatch

null comparison

1. Run Summary

Application Name	:etl_pipeline_goldlayer
Protocol Name	:DATFDemo
Protocol File Path	:datf_core//test/testprotocol/testselection_template.xlsx
Testcase Name	:testcase1_parquet_parquet_mismatch
Testcase Type	:null
Test Environment	:SIT
Start Time	:11-Mar-2025 13:54:03 UTC
End Time	:11-Mar-2025 13:55:04 UTC
Run Time	:0:01:01
Test Result	:Failed
Reason	:Nulls are not matching for each column

2. Configuration Details

Compare Type	:s2tcompare
testquerygenerationmode	:Auto
Testcase Type	:null
Source Connection Type	:aws-s3
Source Format	:parquet
Source Path	:datf_core/test/data/source/patients_source_parquet
Target Connection Type	:aws-s3
Target Format	:parquet
Target Path	:datf_core/test/data/target/patients_target_parquet_mismatch
S2T Path	:test/s2t/s2t_1_parquet_parquet_mismatch.xlsx
Primary Keys	:id

3. Null Summary

Test Result	:Failed
No of column has null in source	:8
No of column has null in target	:8
No of columns has null count match	:5
No of columns has null count mismatch	:3

4. SQL Queries

4.1 Source Query

["readdatadf=spark.read.format('parquet').load('datf_core/test/data/source/patients_source_parquet')", "readdatadf.createOrReplaceTempView('dataview')", 'spark.sql("SELECT src.ADDRESS as ADDRESS, src.BIRTHDATE as BIRTHDATE, src.BIRTHPLACE as BIRTHPLACE, src.CITY as CITY, src.COUNTY as COUNTY, src.DEATHDATE as DEATHDATE, src.DRIVERS as DRIVERS, src.ETHNICITY as ETHNICITY, src.FIRST as FIRST, src.GENDER as GENDER, src.HEALTHCARE_COVERAGE as HEALTHCARE_COVERAGE, src.HEALTHCARE_EXPENSES as HEALTHCARE_EXPENSES, src.LAST as LAST, src.LAT as LAT, src.LON as LON, src.MAIDEN as MAIDEN, src.MARITAL as MARITAL, src.PASSPORT as PASSPORT, src.PREFIX as PREFIX, src.RACE as RACE, src.SSN as SSN, src.STATE as STATE, src.SUFFIX as SUFFIX, src.ZIP as ZIP, src.id as id FROM dataview src ")', "]

4.2 Target Query

["readdatadf=spark.read.format('parquet').load('datf_core/test/data/target/patients_target_parquet_mismatch')", "readdatadf.createOrReplaceTempView('dataview')", 'spark.sql("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 dataview tgt ")', "]

5. Columns having nulls at source and target

5.1 Columns having nulls in source

S.No	Column	Count
1	DEATHDATE	1000

2	DRIVERS	213
3	MAIDEN	840
4	MARITAL	380
5	PASSPORT	273
6	PREFIX	244
7	SUFFIX	1159
8	ZIP	543

5.2 Columns having nulls in target

S.No	Column	Count
1	DEATHDATE	998
2	DRIVERS	213
3	MAIDEN	838
4	MARITAL	380
5	PASSPORT	273
6	PREFIX	244
7	SUFFIX	1157
8	ZIP	543

5.3 Columns having null count mismatch between source and target

S.No	Source Column Name	Src Null Count	Target Column Name	Tgt Null Count
1	DEATHDATE	1000	DEATHDATE	998
2	MAIDEN	840	MAIDEN	838
3	SUFFIX	1159	SUFFIX	1157

testcase12_parquet_parquet_mismatch_manual

null comparison

1. Run Summary

Application Name	:etl_pipeline_goldlayer
Protocol Name	:DATFDemo
Protocol File Path	:datf_core//test/testprotocol/testselection_template.xlsx
Testcase Name	:testcase12_parquet_parquet_mismatch_manual
Testcase Type	:null
Test Environment	:SIT
Start Time	:11-Mar-2025 13:55:34 UTC
End Time	:11-Mar-2025 13:57:09 UTC
Run Time	:0:01:34
Test Result	:Failed
Reason	:Nulls are not matching for each column

2. Configuration Details

Compare Type	:likeobjectcompare
testquerygenerationmode	:Manual
Testcase Type	:null
Source Connection Name	:RawS3Bucket
Source Connection Type	:aws-s3
Source Connection Value	:RawS3Bucket
Source Format	:parquet
Source Name	:patient_data_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	:patient_data_target_parquet
Target Path	:test/data/target
Primary Keys	:id

3. Null Summary

Test Result	:Failed
No of column has null in source	:5
No of column has null in target	:5
No of columns has null count match	:0
No of columns has null count mismatch	:5

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. Columns having nulls at source and target

5.1 Columns having nulls in source

S.No	Column	Count
1	DEATHDATE	900877

2	DRIVERS	200539
3	MAIDEN	800319
4	PASSPORT	300807
5	SUFFIX	800558

5.2 Columns having nulls in target

S.No	Column	Count
1	DEATHDATE	901273
2	DRIVERS	200008
3	MAIDEN	800625
4	PASSPORT	299989
5	SUFFIX	800405

5.3 Columns having null count mismatch between source and target

S.No	Source Column Name	Src Null Count	Target Column Name	Tgt Null Count
1	DEATHDATE	900877	DEATHDATE	901273
2	DRIVERS	200539	DRIVERS	200008
3	MAIDEN	800319	MAIDEN	800625
4	PASSPORT	300807	PASSPORT	299989
5	SUFFIX	800558	SUFFIX	800405

testcase31_snowflake_parquet_validation

null comparison

1. Run Summary

Application Name	:etl_pipeline_goldlayer
Protocol Name	:DATFDemo
Protocol File Path	:datf_core//test/testprotocol/testselection_template.xlsx
Testcase Name	:testcase31_snowflake_parquet_validation
Testcase Type	:null
Test Environment	:SIT
Start Time	:11-Mar-2025 13:57:40 UTC
End Time	:11-Mar-2025 13:58:14 UTC
Run Time	:0:00:33
Test Result	:Failed
Reason	:Nulls are not matching for each column

2. Configuration Details

Compare Type	:likeobjectcompare
testquerygenerationmode	:Auto
Testcase Type	:null
Source Connection Name	:raw_snowflake_sql_connection
Source Connection Type	:snowflake
Source Connection Value	:raw_snowflake_sql_connection
Source Format	:table
Source Name	:rtpcr_diagnostic_lab_testing
Target Connection Type	:aws-s3
Target Format	:parquet
Target Name	:patients_target_parquet_mismatch
Target Path	:test/data/target
S2T Path	:test/s2t/s2t_31_snowflake_parquet_val.xlsx
Primary Keys	:id,state

3. Null Summary

Test Result	:Failed
No of column has null in source	:0
No of column has null in target	:8
No of columns has null count match	:0
No of columns has null count mismatch	:2

4. SQL Queries

4.1 Source Query

SELECT DATE,FEMA_REGION,NEW_RESULTS_REPORTED,OVERALL_OUTCOME,STATE,STATE_FIPS,STATE_NAME,TOTAL_RESULTS_REPORTED FROM rtpcr_diagnostic_lab_testing

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 rtpcr_target

5. Columns having nulls at source and target

5.1 Columns having nulls in source

S.No	Column	Count

5.2 Columns having nulls in target

S.No	Column	Count
1	DEATHDATE	998
2	DRIVERS	213
3	MAIDEN	838
4	MARITAL	380
5	PASSPORT	273
6	PREFIX	244
7	SUFFIX	1157
8	ZIP	543

5.3 Columns having null count mismatch between source and target

S.No	Source Column Name	Src Null Count	Target Column Name	Tgt Null Count
1	STATE_FIPS	0	DEATHDATE	998
2	STATE_NAME	0	DRIVERS	213