Load S3 files in Dataframe

test\_df=spark.read.text("s3://abacus-clover-data-lake-dev-892395949777-us-east-1/raw/s3-replication/ongoing/cms/RH5141.DTRRD.D240402.T0009073")

Create view from dataframe

df.createOrReplaceTempView("people")

Convert table into data frame

main\_df=spark.table("mgln02\_dev\_catalog.gold.ihp\_rx\_flat\_claims")

ref\_ihp\_tcoc\_members = spark.table("mgln02\_dev\_catalog.bronze.ihp\_tcoc\_members")

Join two dataframes

main\_df = main\_df.join(

ref\_ihp\_tcoc\_members.select(

F.col("member\_id").alias("ref\_health\_toc\_\_member\_id"),

F.col("first\_date\_of\_oncology\_dx").alias("ref\_first\_date\_of\_oncology\_dx"),

F.col("current\_master\_member\_id").alias("ref\_current\_master\_member\_id"),

F.col("Filename").alias("ref\_file\_name"),

).distinct(),

(F.col("patient\_member\_source\_id") == F.col("ref\_health\_toc\_\_member\_id"))

,

"left",

)

import json

import os

from pyspark.sql import SparkSession, functions as F

from pyspark.sql.types import BooleanType, StringType

from pyspark.sql.window import Window

from air.pipeline.sinks import OverwriteSink

With Column

main\_df=main\_df.withColumn("filename\_dt", F.get\_json\_object("source\_view\_source\_map", "$.FILENAME"))