import sys

from awsglue.transforms import \*

from awsglue.utils import getResolvedOptions

from awsglue.context import GlueContext

from awsglue.job import Job

from pyspark.context import SparkContext

from pyspark.sql.types import StructType, StructField, StringType

from pyspark.sql import functions as F

args = getResolvedOptions(sys.argv, ['JOB\_NAME'])

sc = SparkContext()

glueContext = GlueContext(sc)

spark = glueContext.spark\_session

job = Job(glueContext)

job.init(args['JOB\_NAME'], args)

# Define schema explicitly as all string

schema = StructType([

StructField("id", StringType(), True),

StructField("case number", StringType(), True),

StructField("block", StringType(), True),

StructField("nibrs code", StringType(), True),

StructField("primary type", StringType(), True),

StructField("description", StringType(), True),

StructField("location description", StringType(), True),

StructField("arrest", StringType(), True),

StructField("domestic", StringType(), True),

StructField("beat", StringType(), True),

StructField("area", StringType(), True),

StructField("ward", StringType(), True),

StructField("community area", StringType(), True),

StructField("fbi code", StringType(), True),

StructField("x coordinate", StringType(), True),

StructField("y coordinate", StringType(), True),

StructField("year", StringType(), True),

StructField("latitude", StringType(), True),

StructField("longitude", StringType(), True),

StructField("location", StringType(), True),

StructField("weapon description", StringType(), True),

StructField("vict age", StringType(), True),

StructField("vict sex", StringType(), True),

StructField("victim race", StringType(), True),

StructField("date occ", StringType(), True),

StructField("time occ", StringType(), True),

StructField("date rptd", StringType(), True),

StructField("time rptd", StringType(), True),

StructField("date arrested", StringType(), True),

StructField("time arrested", StringType(), True),

StructField("premises desc", StringType(), True),

StructField("district", StringType(), True),

StructField("suspect age", StringType(), True),

StructField("suspect sex", StringType(), True),

StructField("suspect race", StringType(), True),

StructField("case status", StringType(), True),

StructField("crime category", StringType(), True),

StructField("secondary description", StringType(), True),

StructField("census tract", StringType(), True),

StructField("zip code", StringType(), True),

StructField("incident narrative", StringType(), True),

StructField("priority level", StringType(), True),

StructField("repeat offense flag", StringType(), True)

])

# Read CSVs from S3

df = spark.read.format("csv") \

.option("header", True) \

.schema(schema) \

.load("s3://final-project-bucket-group-5/rawData/")

# Optional: Reorder columns to match schema

df = df.select([field.name for field in schema.fields])

# Write as a single Parquet file

df.coalesce(1).write \

.mode("overwrite") \

.parquet("s3://final-project-bucket-group-5/master-data/")

job.commit()