from pyspark.sql import SparkSession

from pyspark.sql.functions import col, avg

def movierating():

# Sample data

data = [

(1, 101, 4, 1622388000000),

(1, 102, 3, 1622388020000),

(2, 101, 5, 1622388040000),

(2, 103, 4, 1622388060000),

(3, 101, 3, 1622388080000),

(3, 102, 4, 1622388100000),

(3, 103, None, 1622388120000),

(4, 101, 2, 1622388140000),

(4, 101, 2, 1622388140000)

]

# Initialize Spark session

columns = ["user\_id", "movie\_id", "rating", "timestamp"]

spark = SparkSession.builder.appName("Movie Rating").getOrCreate()

df = spark.createDataFrame(data, schema=columns)

# print(df.show())

# Find all ratings that are greater than or equal to 4.

#df\_filtered = df.filter(df.rating >= 4)

#print(df\_filtered.show())

#Replace null values in the 'rating' column with the average rating.

#average\_rating = df.selectExpr('avg(rating)').collect()[0][0]

#df\_filled = df.na.fill({'rating': average\_rating})

#print(df\_filled.show())

#Remove duplicate ratings by 'user\_id' and 'movie\_id'.

#df\_no\_duplicates = df.dropDuplicates(['user\_id', 'movie\_id'])

#print(df\_no\_duplicates.show())

#Select 'user\_id' and 'rating' columns.

#df\_selected = df.select('user\_id', 'rating')

#print(df\_selected.show())

#Calculate the average rating per movie.

#df\_grouped = df.groupBy('movie\_id').agg({'rating': 'avg'})

#print(df\_grouped.show())

moviedata=[

(101,"Nepali babu"),

(102,"Ta taa sarai bigris ni badri") ,

(103,"Prem geet") ,

(104,"Bahubali")

]

# Create a movie details DataFrame

columns = ["movie\_id", "movie\_name"]

df\_movies=spark.createDataFrame(moviedata,columns)

# Assuming df\_movies is another DataFrame that contains movie details

#df\_joined = df.join(df\_movies, on='movie\_id', how='inner')

#print(df\_joined.show())

#df\_union = df.union(df)

#print(df\_union.show())

#Create a temp view and find the top - rated movies.

df.createOrReplaceTempView('ratings')

sql\_result = spark.sql(

'SELECT movie\_id, AVG(rating) as avg\_rating FROM ratings GROUP BY movie\_id ORDER BY avg\_rating DESC LIMIT 10')

print(sql\_result.show())

if \_\_name\_\_ == '\_\_main\_\_':

movierating()