CSCI.620.01/03

Homework 8 Report (ss7495)Suraj Sureshkumar

Python Program:

This is the initial python program which contains all the imports and the files are loaded into the dataframes. Then joining the three data frames and not including the adult movies.

```
import time

from pyspark.sql import SparkSession
from pyspark.sql.functions import col, split, array_contains

spark = SparkSession.builder.getOrCreate()

dataframe = spark.read.csv('name.basics (1).tsv.gz', sep='\t', header=True,
inferSchema=True)
dataframe2 = spark.read.csv('title.basics.tsv.gz', sep='\t', header=True,
inferSchema=True)
dataframe3 = spark.read.csv('title.principals (1).tsv.gz', sep='\t', header=True,
inferSchema=True)

df1 = dataframe.join(dataframe3, on="nconst", how="inner")
final_df = df1.join(dataframe2, on="tconst", how="inner")
final_df = final_df.filter(final_df.isAdult == 0)
```

Query1

```
start_time = time.time()
query1 = final_df.filter(
    (col("deathYear") == "\\N") & (col("primaryName").startswith("Phi")) &
(col("titleType") == "movie")
    & (col("startYear") != 2014)).select("primaryName")
query1.show(10)
elapsed_time = time.time() - start_time
print('Time taken to run the query', elapsed_time)
```

Output:

```
+------------------+
|       primaryName|
+------------------+
|Philip J. Spinelli|
|      Phil Schmidt|
|     Philippe Welt|
|      Phil Katzman|
|      Philip Quast|
|    Phil Nibbelink|
|      Philip Glass|
| Philip van Volsem|
|     Philippe Diaz|
|     Philippe Diaz|
+------------------+
only showing top 10 rows

Time taken to run the query 79.60628271102905
```

Time is in seconds.

Query2

```
start_time = time.time()
genre_array = final_df.withColumn("genres", split(final_df.genres, ","))
query2 = genre_array.filter(
   (col("category") == "producer") & (col("primaryName").contains("Gill")) &
(col("startYear") == 2017) & (
       array_contains(genre_array.genres, "Talk-Show"))).groupBy(
   "primaryName").count().sort(col('count').desc())
query2.show(10)
elapsed_time = time.time() - start_time
print('Time taken to run the query', elapsed_time)
```

Output:

```
+----------------+-----+
|     primaryName|count|
+----------------+-----+
|       Ryan Gill|   81|
|Dominic Gillette|   73|
|Corinne Gilliard|   14|
|      Shane Gill|   13|
|   Gilles Bérard|    1|
+----------------+-----+

Time taken to run the query 56.36176896095276
```

Time is in seconds.

Query3:

```
start_time = time.time()
query3 = final_df.filter(
    (col("category") == "producer") & (col("runtimeMinutes") > 120) &
(col("deathYear") == "\\N")).groupBy(
    "primaryName").count().sort(col('count').desc())
query3.show(10)
elapsed_time = time.time() - start_time
print('Time taken to run the query', elapsed_time)
```

Output:

```
+------------------+-----+
|       primaryName|count|
+------------------+-----+
|      Acun Ilicali|  241|
|     Maxwell James|  176|
|   Wade Baverstock|  143|
|     Vince McMahon|  141|
|John Michael Flynn|  114|
|Christopher Lockey|  114|
|        Kyle Shire|  104|
|         Efe Irvül|  102|
|      Nick Rylance|  102|
|      Matt Spencer|   95|
+------------------+-----+
only showing top 10 rows

Time taken to run the query 58.33132314682007
```

Time is in seconds

Query4:

```
start_time = time.time()
query4 = final_df.filter(
    (col("deathYear") == "\\N") & ((col("category") == "actor") |
(col("category") == "actress")) &
    ((col("characters").contains("Jesus")) |
(col("characters").contains("Christ")))).select("primaryName").distinct()
query4.show(10)
elapsed_time = time.time() - start_time
print('Time taken to run the query', elapsed_time)
```

Output:

```
+----------------+
|     primaryName|
+----------------+
|     Lana O'Kell|
|    Shane Willis|
|Emily Rutherfurd|
|     Daisy Louve|
|   Julia Richter|
|    Richie Boyko|
|      Liron Levo|
|     Al Madrigal|
|  Jane Krakowski|
|   Daniel Eckert|
+----------------+
only showing top 10 rows

Time taken to run the query 74.36318397521973
```

The time is in seconds.