In [1]:
```
pip install findspark
```

```
Requirement already satisfied: findspark in c:\users\pc\anaconda3\lib\site-packages
(2.0.1)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
import findspark
findspark.init()
```

In [3]:
```
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql.functions import *

# Configure spark session
spark = SparkSession\
    .builder\
    .master('local[2]')\
    .appName('AMAZON_BOOK')\
    .config('spark.jars.packages', 'org.mongodb.spark:mongo-spark-connector_2.12:2.4.1
    .config("spark.driver.memory", "4g")\
    .getOrCreate()
```

In [4]:
```
# Load the dataset
df_book = spark.read.csv(r"C:\Users\Pc\Downloads\amazon_book\books_data.csv", header=T
df_rate = spark.read.csv(r"C:\Users\Pc\Downloads\amazon_book\Books_rating.csv", header
```

In [5]:
```
print(df_book.count())
print(df_rate.count())
```

```
212404
3000000
```

In [6]:
```
# Preview df_book
df_book.show(5)
```

```
+--------------------+--------------------+--------------------+--------------------+
--------------------+--------------------+-------------+--------------------+-------
------------+------------+
|               Title|         description|             authors|               image|
previewLink|           publisher| publishedDate|            infoLink|          catego
ries|ratingsCount|
+--------------------+--------------------+--------------------+--------------------+
--------------------+--------------------+-------------+--------------------+-------
------------+------------+
|Its Only Art If I...|                NULL|    ['Julie Strain']|http://books.goog...|
http://books.goog...|                NULL|          1996|http://books.goog...|['Comic
s & Graphi...|        NULL|
|Dr. Seuss: Americ...|"Philip Nel takes...| like that of Lew...| has changed lang...|
giving us new wo...| inspiring artist...|['Philip Nel']|http://books.goog...|http://b
ooks.goog...|   A&C Black|
|Wonderful Worship...|This resource inc...|    ['David R. Ray']|http://books.goog...|
http://books.goog...|                NULL|          2000|http://books.goog...|
['Religion']|        NULL|
|Whispers of the W...|Julia Thomas find...| ['Veronica Haddon']|http://books.goog...|
http://books.goog...|           iUniverse|       2005-02|http://books.goog...|
['Fiction']|        NULL|
|Nation Dance: Rel...|                NULL|     ['Edward Long']|                NULL|
http://books.goog...|                NULL|    2003-03-01|http://books.goog...|
NULL|        NULL|
+--------------------+--------------------+--------------------+--------------------+
--------------------+--------------------+-------------+--------------------+-------
------------+------------+
only showing top 5 rows
```

In [7]:
```
# Preview df_book
df_rate
```

Out[7]:
```
DataFrame[Id: string, Title: string, Price: string, User_id: string, profileName: str
ing, review/helpfulness: string, review/score: string, review/time: string, review/su
mmary: string, review/text: string]
```

In [8]:
```
import csv
# Generate a user.csv file from the df_rate
# Select the 'user_id' and 'profileName' columns
user_data = df_rate.select('User_id', 'profileName')

# Collect the data from the DataFrame
data_to_write = user_data.collect()

# Write the data to the CSV file
with open("users.csv", 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header
    csv_writer.writerow(["userId", "profileName"])

    # Write the data
    csv_writer.writerows(data_to_write)
```

In [9]:
```
import pandas as pd
d=pd.read_csv("users.csv")
d.shape
```

Out[9]: (3000000, 2)

In [10]:
```
#drop the users profileName from the df_rate
df_rate = df_rate.drop('profileName')
# Preview df_rate
df_rate.show(10)
```

```
+----------+--------------------+-----+--------------+------------------+------------
+-----------+--------------------+--------------------+
|        Id|               Title|Price|       User_id|review/helpfulness|review/score
|review/time|      review/summary|         review/text|
+----------+--------------------+-----+--------------+------------------+------------
+-----------+--------------------+--------------------+
|1882931173|Its Only Art If I...| NULL| AVCGYZL8FQQTD|               7/7|         4.0
|  940636800|Nice collection o...|This is only for ...|
|0826414346|Dr. Seuss: Americ...| NULL|A30TK6U7DNS82R|             10/10|         5.0
| 1095724800|   Really Enjoyed It|I don't care much...|
|0826414346|Dr. Seuss: Americ...| NULL|A3UH4UZ4RSVO82|             10/11|         5.0
| 1078790400|Essential for eve...|"If people become...|
|0826414346|Dr. Seuss: Americ...| NULL|A2MVUWT453QH61|               7/7|         4.0
| 1090713600|Phlip Nel gives s...|Theodore Seuss Ge...|
|0826414346|Dr. Seuss: Americ...| NULL|A22X4XUPKF66MR|               3/3|         4.0
| 1107993600|Good academic ove...|"Philip Nel - Dr....|
|0826414346|Dr. Seuss: Americ...| NULL|A2F6NONFUDB6UK|               2/2|         4.0
| 1127174400|One of America's ...|"""Dr. Seuss: Ame...|
|0826414346|Dr. Seuss: Americ...| NULL|A14OJS0VWMOSWO|               3/4|         5.0
| 1100131200|A memorably excel...|Theodor Seuss Gie...|
|0826414346|Dr. Seuss: Americ...| NULL|A2RSSXTDZDUSH4|               0/0|         5.0
| 1231200000|Academia At It's ...|"When I recieved ...|
|0826414346|Dr. Seuss: Americ...| NULL|A25MD5I2GUIW6W|               0/0|         5.0
| 1209859200|And to think that...|"Trams (or any pu...|
|0826414346|Dr. Seuss: Americ...| NULL|A3VA4XFS5WNJO3|               3/5|         4.0
| 1076371200|Fascinating accou...|As far as I am aw...|
+----------+--------------------+-----+--------------+------------------+------------
+-----------+--------------------+--------------------+
only showing top 10 rows
```

In [11]:
```
#drop the review/time colomn from the df_rate
df_rate = df_rate.drop('review/time')
#drop the review/helpfulness colomn from the df_rate
df_rate = df_rate.drop('review/helpfulness')
# Overwrite df_rate with the renamed columns
df_rate = df_rate.withColumnRenamed('review/score', 'score') \
                 .withColumnRenamed('review/time', 'time') \
                 .withColumnRenamed('review/summary', 'summary') \
                 .withColumnRenamed('review/text', 'text')

# Preview df_rate
df_rate.show(5)
```

```
+----------+--------------------+-----+-------------+-----+--------------------+----
----------------+
|        Id|               Title|Price|      User_id|score|             summary|
text|
+----------+--------------------+-----+-------------+-----+--------------------+----
----------------+
|1882931173|Its Only Art If I...| NULL| AVCGYZL8FQQTD|  4.0|Nice collection o...|This
is only for ...|
|0826414346|Dr. Seuss: Americ...| NULL|A30TK6U7DNS82R|  5.0|   Really Enjoyed It|I do
n't care much...|
|0826414346|Dr. Seuss: Americ...| NULL|A3UH4UZ4RSVO82|  5.0|Essential for eve...|"If
people become...|
|0826414346|Dr. Seuss: Americ...| NULL|A2MVUWT453QH61|  4.0|Phlip Nel gives s...|Theo
dore Seuss Ge...|
|0826414346|Dr. Seuss: Americ...| NULL|A22X4XUPKF66MR|  4.0|Good academic ove...|"Phi
lip Nel - Dr....|
+----------+--------------------+-----+-------------+-----+--------------------+----
----------------+
only showing top 5 rows
```

In [12]:
```
# Drop the rows where "Title" is not empty
df_rate = df_rate.na.drop(subset=["Title"])
```

In [13]:
```
df_rate = df_rate.na.drop(subset=["score"])
```

In [14]:
```
pip install pymongo
```

```
Requirement already satisfied: pymongo in c:\users\pc\anaconda3\lib\site-packages (4.
6.0)Note: you may need to restart the kernel to use updated packages.

Requirement already satisfied: dnspython<3.0.0,>=1.16.0 in c:\users\pc\anaconda3\lib
\site-packages (from pymongo) (2.4.2)
```

In [15]:
```
# MongoDB connection settings
mongo_uri = "mongodb://localhost:27017/AMAZON_BOOK"

# Write df_rate to MongoDB
df_rate.write.format("mongo").mode("overwrite").option("uri", mongo_uri+".RATE").save(

# Write df_book to MongoDB
df_book.write.format("mongo").mode("overwrite").option("uri", mongo_uri+".BOOK").save(

# Write user_data to MongoDB
user_data.write.format("mongo").mode("overwrite").option("uri", mongo_uri+".USER").sav
```

In [16]:
```
# Check for null or NaN values in the "score" column
df_rate.select([count(when(isnan('score') | col('score').isNull(), 'score'))]).show()
```

```
+-------------------------------------------------------------------+
|count(CASE WHEN (isnan(score) OR (score IS NULL)) THEN score END)|
+-------------------------------------------------------------------+
|                                                                  0|
+-------------------------------------------------------------------+
```

In [ ]: