Amazon_Reviews_ETL.ipynb using Outdoors dataset

```
import os
# Find the latest version of spark 3.0  from http://www.apache.org/dist/spark/ and enter as t
# For example:
# spark_version = 'spark-3.0.3'
spark_version = 'spark-3.1.3'
os.environ['SPARK_VERSION']=spark_version



# Install Spark and Java
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget https://downloads.apache.org/spark/spark-3.1.3/spark-3.1.3-bin-hadoop2.7.tgz
!tar -xvf spark-3.1.3-bin-hadoop2.7.tgz
!pip install -q findspark

import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content//spark-3.1.3-bin-hadoop2.7"

# Start a SparkSession
import findspark
findspark.init()
```

```
# Download the Postgres driver that will allow Spark to interact with Postgres.
!wget https://jdbc.postgresql.org/download/postgresql-42.2.16.jar
```

```
--2022-09-16 10:21:35--  https://jdbc.postgresql.org/download/postgresql-42.2.16.jar
Resolving jdbc.postgresql.org (jdbc.postgresql.org)... 72.32.157.228, 2001:4800:3e1:1::
Connecting to jdbc.postgresql.org (jdbc.postgresql.org)|72.32.157.228|:443... connected
HTTP request sent, awaiting response... 200 OK
Length: 1002883 (979K) [application/java-archive]
Saving to: ‘postgresql-42.2.16.jar.1’

postgresql-42.2.16. 100%[===================>] 979.38K  5.05MB/s    in 0.2s

2022-09-16 10:21:36 (5.05 MB/s) - ‘postgresql-42.2.16.jar.1’ saved [1002883/1002883]
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("BigData-Challenge").config("spark.driver.extraClassPath
```

Load Amazon Data into Spark DataFrame

```
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Outdoors_v1_00.tsv.g
spark.sparkContext.addFile(url)
df = spark.read.option("encoding", "UTF-8").csv(SparkFiles.get("amazon_reviews_us_Outdoors_v1
df.show()
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|p
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   18446823|R35T75OLUGHL5C|B000NV6H94|     110804376|Stearns Youth Boa...|
|         US|   13724367|R2BV735O46BN33|B000IN0W3Y|     624096774|Primal Wear Men's...|
|         US|   51001958|R2NBEUGPQQGXP1|B008RBJXFM|     278970944|Osprey Hydraulics...|
|         US|   32866903|R17LLAOJ8ITK0S|B00FK8WUQY|     312877650|CamelBak eddy .75...|
|         US|   30907790|R39PEQBT5ISEF4|B00EZA3VW0|     305567912|Children Black Re...|
|         US|   20232229|R3GNM3SU9VHJFT|B006JA8WEG|     842306035|Ibera Bicycle Tri...|
|         US|   17698862| R2Y81OP0EK467|B002PWFSEO|     451480122|Therm-a-Rest Comp...|
|         US|   38486114|R2LFGSI6HAYH5F|B002DZGKHW|     124386306|Sawyer Products P...|
|         US|   26319572|R297G6ED1IQO7W|B00ABA08F6|     991442421|Zippo Hand Warmer...|
|         US|   27152337| RE27RFC6101N6|B003Z8WIHC|     886483892|Camp Chef Dutch O...|
|         US|   12516845|R3BPDME6E94W8Z|B007CP6UK0|     150224054|3CERA Portable Wi...|
|         US|    3225242|R2P08O1RILUOX3|B003V3U9JK|     343847969|Texsport King Kot...|
|         US|     961839|R37CVAB03PTDVI|B00Y846HN8|     858088629|Wallygadgets 2 Wh...|
|         US|   47796452| RAWNWOGXPCPMD|B00IYQ84VY|     474493517|RainStoppers 34-I...|
|         US|   32004835| R5DYGP6ASX77M|B002MYCKLY|     920014456|Alpha Deluxe Port...|
|         US|   23972939|R1O0SAOOGF2KG7|B00EZV69JG|     128489321|Speedfil Z4 BTA B...|
|         US|   40889047|R35NJUT0U3MU3V|B00AWOT3T8|     571303876|O'Brien Kids Plat...|
|         US|   11244387|R242C08MF9D1AH|B0000AXTID|     739769424|Kwik-Tek F-5R Pla...|
|         US|   20121211| R3RYG8TJTO4E2|B00IFHFJXI|     984009972|Ivation Portable ...|
|         US|   25657249|R3IKH1DNY0CP9F|B00KFILTWU|     405521681|GreenInsync Repla...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 20 rows
```

Create DataFrames to match tables

```
from pyspark.sql.functions import to_date
# Read in the Review dataset as a DataFrame
df.show(5)
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|p
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   18446823|R35T75OLUGHL5C|B000NV6H94|     110804376|Stearns Youth Boa...|
|         US|   13724367|R2BV735O46BN33|B000IN0W3Y|     624096774|Primal Wear Men's...|
|         US|   51001958|R2NBEUGPQQGXP1|B008RBJXFM|     278970944|Osprey Hydraulics...|
|         US|   32866903|R17LLAOJ8ITK0S|B00FK8WUQY|     312877650|CamelBak eddy .75...|
|         US|   30907790|R39PEQBT5ISEF4|B00EZA3VW0|     305567912|Children Black Re...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
```

```
only showing top 5 rows
```

```
# Create the customers_table DataFrame
customers_df = df.groupby("customer_id").agg({"customer_id":"count"}).withColumnRenamed("coun
customers_df.show(5)
```

```
+-----------+--------------+
|customer_id|customer_count|
+-----------+--------------+
|   43679767|             1|
|   32024654|             1|
|   52913169|             1|
|   24297214|             1|
|   26096454|             1|
+-----------+--------------+
only showing top 5 rows
```

```
# Create the products_table DataFrame and drop duplicates.
products_df = df.select(["product_id", "product_title"]).drop_duplicates(["product_id"])
products_df.show()
products_df.count()
```

```
+----------+--------------------+
|product_id|       product_title|
+----------+--------------------+
|6040161299|Santa Cruz Scream...|
|B0000AXTWB|Kwik Tek SF-1C Su...|
|B0000BYDQX|Bellows Foot Pump...|
|B0000DYNXN|DIN &amp; Yoke Ta...|
|B0000E66XJ|Black Diamond Zio...|
|B00012O8B4|Razor E100 Electr...|
|B00029NCOK|ATV Logic ATVHP-B...|
|B0002BJZ6W|Pro XL-C Corragat...|
|B0002CTKO8|Heat Factory Flee...|
|B0002LKM8W|Weekender Men's T...|
|B0002M9DO0|STABILicers Maxx ...|
|B000650ZMW|Panaracer, 700x32...|
|B00066Z29W|             Fairing|
|B00068NHMO|Marmot Men's Midw...|
|B0006FYKCI|Transpack Edge Is...|
|B0006UOKFA|FireOne Windproof...|
|B0006VSTOW|Flowlab Blue Logo...|
|B0007IS6BA|Columbia Bugaboo ...|
|B0007QCO4M|              Bantam|
|B0007QCOP6|Victorinox Swiss ...|
+----------+--------------------+
only showing top 20 rows

391729
```

```
# Create the products_table DataFrame and drop duplicates.
products_df = df.select(["product_id", "product_title"]).drop_duplicates()
products_df.show()
products_df.count()
```

```
+----------+--------------------+
|product_id|       product_title|
+----------+--------------------+
|B00IFHFJXI|Ivation Portable ...|
|B00WG0J0D0|JanSport Superbre...|
|B00V15AUN0|Nickelodeon Paw P...|
|B00FUWSTI8|Bago Lightweight ...|
|B003FV94NA|Michelin Lithion ...|
|B00WIK04HO|Ultra Bright Camp...|
|B00J2HSCM0|High Sierra Tank ...|
|B009I6NSR4|Black Veil Brides...|
|B001GSHSLE|Stansport 191 App...|
|B00L2IO9M4|Columbia Sportswe...|
|B00KY7IM7W|Nalgene 32 Oz Wid...|
|B00TV5JCTK|Rollerblade ABEC ...|
|B00B9D071Y|BUFF UV Multifunc...|
|B00F9IGIKO|Condor Tactical F...|
|B004X55L9I|Hydro Flask Insul...|
|B00LORROIY|Scuba Choice Divi...|
|B00AATRU8G|Kelty Redwing 44 ...|
|B00HMCYWEO|Dakine Explorer L...|
|B004DK1CM8|Hot Headz 12V Hea...|
|B00T4W6SSS|Fits Sock Light H...|
+----------+--------------------+
only showing top 20 rows

391733
```

```
# Create the products_table DataFrame and drop duplicates.
products_df = df.select(["product_id", "product_title"]).drop_duplicates(["product_id"])
products_df.show()
products_df.count()
```

```
+----------+--------------------+
|product_id|       product_title|
+----------+--------------------+
|6040161299|Santa Cruz Scream...|
|B0000AXTWB|Kwik Tek SF-1C Su...|
|B0000BYDQX|Bellows Foot Pump...|
|B0000DYNXN|DIN &amp; Yoke Ta...|
|B0000E66XJ|Black Diamond Zio...|
|B00012O8B4|Razor E100 Electr...|
|B00029NCOK|ATV Logic ATVHP-B...|
|B0002BJZ6W|Pro XL-C Corragat...|
|B0002CTKO8|Heat Factory Flee...|
|B0002LKM8W|Weekender Men's T...|
|B0002M9DO0|STABILicers Maxx ...|
|B000650ZMW|Panaracer, 700x32...|
|B00066Z29W|             Fairing|
|B00068NHMO|Marmot Men's Midw...|
|B0006FYKCI|Transpack Edge Is...|
|B0006UOKFA|FireOne Windproof...|
|B0006VSTOW|Flowlab Blue Logo...|
|B0007IS6BA|Columbia Bugaboo ...|
|B0007QCO4M|              Bantam|
|B0007QCOP6|Victorinox Swiss ...|
+----------+--------------------+
only showing top 20 rows

391729
```

```
# Create the review_id_table DataFrame.
# Convert the 'review_date' column to a date datatype with to_date("review_date", 'yyyy-MM-dd
review_id_df = df.select(["review_id", "customer_id", "product_id", "product_parent", to_date
review_id_df.show(5)
```

```
+--------------+-----------+----------+--------------+-----------+
|     review_id|customer_id|product_id|product_parent|review_date|
+--------------+-----------+----------+--------------+-----------+
|R35T75OLUGHL5C|   18446823|B000NV6H94|     110804376| 2015-08-31|
|R2BV735O46BN33|   13724367|B000IN0W3Y|     624096774| 2015-08-31|
|R2NBEUGPQQGXP1|   51001958|B008RBJXFM|     278970944| 2015-08-31|
|R17LLAOJ8ITK0S|   32866903|B00FK8WUQY|     312877650| 2015-08-31|
|R39PEQBT5ISEF4|   30907790|B00EZA3VW0|     305567912| 2015-08-31|
+--------------+-----------+----------+--------------+-----------+
only showing top 5 rows
```

```
# Create the vine_table. DataFrame
vine_df = df.select(["review_id", "star_rating", "helpful_votes", "total_votes", "vine", "ver
vine_df.show(5)
```

```
+--------------+-----------+-------------+-----------+----+-----------------+
|     review_id|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+--------------+-----------+-------------+-----------+----+-----------------+
|R35T75OLUGHL5C|          4|            0|          0|   N|                Y|
|R2BV735O46BN33|          5|            0|          0|   N|                Y|
|R2NBEUGPQQGXP1|          4|            0|          0|   N|                Y|
|R17LLAOJ8ITK0S|          3|            1|          1|   N|                Y|
|R39PEQBT5ISEF4|          1|            0|          0|   N|                Y|
+--------------+-----------+-------------+-----------+----+-----------------+
only showing top 5 rows
```

Connect to the AWS RDS and write each DataFrame to tables

```
# Configure settings for RDS
mode = "append"
jdbc_url="jdbc:postgresql://dataviz.cyjzthdhr8cx.us-east-1.rds.amazonaws.com:5432/dataviz"
config = {"user":"postgres",
          "password": "Module16",
          "driver":"org.postgresql.Driver"}
print (jdbc_url)
```

```
     jdbc:postgresql://dataviz.cyjzthdhr8cx.us-east-1.rds.amazonaws.com:5432/dataviz
```

```
# Write review_id_df to table in RDS
import findspark
findspark.add_packages('mysql:mysql-connector-java:8.0.11')
review_id_df.write.jdbc(url=jdbc_url, table='review_id_table', mode=mode, properties=config)
```

```
# Write products_df to table in RDS

products_df.write.jdbc(url=jdbc_url, table='products_table', mode=mode, properties=config)
```

```
# Write customers_df to table in RDS
# 5 min 14 s
customers_df.write.jdbc(url=jdbc_url, table='customers_table', mode=mode, properties=config)
```

```
# Write vine_df to table in RDS
# 11 minutes
vine_df.write.jdbc(url=jdbc_url, table='vine_table', mode=mode, properties=config)
```

✓ 36s completed at 10:30 AM ● ×