```
import os
# Find the latest version of spark 3.0 from http://www.apache.org/dist/spark/ and enter as the spark version
# For example:
# spark_version = 'spark-3.0.3'
spark_version = 'spark-3.0.3'
os.environ['SPARK_VERSION']=spark_version

# Install Spark and Java
!apt-get update
!apt-get install openjdk-11-jdk-headless -qq > /dev/null
!wget -q http://www.apache.org/dist/spark/$SPARK_VERSION/$SPARK_VERSION-bin-hadoop2.7.tgz
!tar xf $SPARK_VERSION-bin-hadoop2.7.tgz
!pip install -q findspark

# Set Environment Variables
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = f"/content/{spark_version}-bin-hadoop2.7"

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

```
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/ InRelease
Ign:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  InRelease
Ign:3 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64  InRelease
Hit:4 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  Release
Hit:5 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64  Release
Get:6 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Hit:7 http://archive.ubuntu.com/ubuntu bionic InRelease
Hit:8 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic InRelease
Get:10 http://archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]
Hit:12 http://ppa.launchpad.net/cran/libgit2/ubuntu bionic InRelease
Hit:13 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic InRelease
Get:14 http://archive.ubuntu.com/ubuntu bionic-backports InRelease [74.6 kB]
Hit:15 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease
Fetched 252 kB in 4s (68.0 kB/s)
Reading package lists... Done
```

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

```
--2022-04-08 15:32:48--  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::228
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.2’

postgresql-42.2.16. 100%[===================>] 979.38K  1.23MB/s    in 0.8s

2022-04-08 15:32:50 (1.23 MB/s) - ‘postgresql-42.2.16.jar.2’ saved [1002883/1002883]
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("M16-Amazon-Challenge").config("spark.driver.extraClassPath","/content/postgresql-42.2.16.jar").getOrCreate()
```

Load Amazon Data into Spark DataFrame

```
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Watches_v1_00.tsv.gz"
spark.sparkContext.addFile(url)
df = spark.read.option("encoding", "UTF-8").csv(SparkFiles.get("amazon_reviews_us_Watches_v1_00.tsv.gz"), sep="\t", header=True, inferSchema=True)
df.show()
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+------------
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+------------
|         US|    3653882|R3O9SGZBVQBV76|B00FALQ1ZC|     937001370|Invicta Women's 1...|         Watches|          5|            0|          0|   N|                Y|          Fi
|         US|   14661224| RKH8BNC3L5DLF|B00D3RGO20|     484010722|Kenneth Cole New ...|         Watches|          5|            0|          0|   N|                Y|I love thisw
|         US|   27324930|R2HLE8WKZSU3NL|B00DKYC7TK|     361166390|Ritche 22mm Black...|         Watches|          2|            1|          1|   N|                Y|           T
|         US|    7211452|R31U3UH5AZ42LL|B000EQS1JW|     958035625|Citizen Men's BM8...|         Watches|          5|            0|          0|   N|                Y|          Fi
|         US|   12733322|R2SV659OUJ945Y|B00A6GFD7S|     765328221|Orient ER27009B M...|         Watches|          4|            0|          0|   N|                Y|Beautiful fa
|         US|    6576411| RA51CP8TR5A2L|B00EYSOSE8|     230493695|Casio Men's GW-94...|         Watches|          5|            0|          0|   N|                Y|       No co
|         US|   11811565| RB2Q7DLDN6TH6|B00WM0QA3M|     549298279|Fossil Women's ES...|         Watches|          5|            1|          1|   N|                Y|          Fi
|         US|   49401598|R2RHFJV0UYBK3Y|B00A4EYBR0|     844009113|INFANTRY Mens Nig...|         Watches|          1|            1|          5|   N|                N|I was about
|         US|   45925069|R2Z6JOQ94LFHEP|B00MAMPGGE|     263720892|G-Shock Men's Gre...|         Watches|          5|            1|          2|   N|                Y|      Perfec
|         US|   44751341| RX27XIIWY5JPB|B004LBPB7Q|     124278407|Heiden Quad Watch...|         Watches|          4|            0|          0|   N|                Y|Great qualit
|         US|    9962330|R15C7QEZT0LGZN|B00KGTVGKS|      28017857|Fossil Women's ES...|         Watches|          4|            2|          2|   N|                Y|           S
|         US|   16097204|R361XSS37V0NCZ|B0039UT5OU|     685450910|Casio General Men...|         Watches|          1|            0|          0|   N|                N|I do not thi
|         US|   51330346| ROTNLALUAJAUB|B00MPF0XJQ|     767769082|2Tone Gold Silver...|         Watches|          3|            0|          0|   N|                Y|         Thr
|         US|    4201739|R2DYX7QU6BGOHR|B003P1OHHS|     648595227|Bulova Men's 98B1...|         Watches|          5|            0|          0|   N|                Y|          Fi
|         US|   26339765| RWASY7FKI7QOT|B00R70YEOE|     457338020|Casio - G-Shock -...|         Watches|          5|            2|          3|   N|                Y|  Worth it -
|         US|    2692576|R2KKYZIN3CCL21|B000FVE3BG|     824370661|Invicta Men's 332...|         Watches|          5|            0|          0|   N|                Y|This is when
|         US|   44713366|R22H4FGVD5O52O|B008X6JB12|     814431355|Seiko Women's SUT...|         Watches|          4|            1|          1|   N|                Y|Thewatch is
|         US|   32778769|R11UACZERCM4ZY|B0040UOFPW|     187700878|Anne Klein Women'...|         Watches|          5|            0|          0|   N|                Y|          Fi
|         US|   27258523|R1AT8NQ38UQOL6|B00UR2R5UY|     594315262|Guess U13630G1 Me...|         Watches|          5|            0|          0|   N|                Y|          Fi
|         US|   42646538|R2NCZRQGIF1Q75|B00HFF57L0|     520810507|Nixon Men's Geo V...|         Watches|          4|            0|          0|   N|                Y|        Very
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+------------
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
#reviews_df = df
#reviews_df.show()

df.show()
```

```
+-----------+----------+--------------+----------+--------------+--------------------+---------------+----------+-------------+-----------+----+-----------------+------------
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_
+-----------+----------+--------------+----------+--------------+--------------------+---------------+----------+-------------+-----------+----+-----------------+------------
|         US|    3653882|R3O9SGZBVQBV76|B00FALQ1ZC|     937001370|Invicta Women's 1...|        Watches|         5|            0|          0|   N|                Y|          Fi
|         US|   14661224| RKH8BNC3L5DLF|B00D3RGO20|     484010722|Kenneth Cole New ...|        Watches|         5|            0|          0|   N|                Y|I love thisw
|         US|   27324930|R2HLE8WKZSU3NL|B00DKYC7TK|     361166390|Ritche 22mm Black...|        Watches|         2|            1|          1|   N|                Y|           T
|         US|    7211452|R31U3UH5AZ42LL|B000EQS1JW|     958035625|Citizen Men's BM8...|        Watches|         5|            0|          0|   N|                Y|          Fi
|         US|   12733322|R2SV659OUJ945Y|B00A6GFD7S|     765328221|Orient ER27009B M...|        Watches|         4|            0|          0|   N|                Y|Beautiful fa
|         US|    6576411| RA51CP8TR5A2L|B00EYSOSE8|     230493695|Casio Men's GW-94...|        Watches|         5|            0|          0|   N|                Y|       No co
|         US|   11811565| RB2Q7DLDN6TH6|B00WM0QA3M|     549298279|Fossil Women's ES...|        Watches|         5|            1|          1|   N|                Y|          Fi
|         US|   49401598|R2RHFJV0UYBK3Y|B00A4EYBR0|     844009113|INFANTRY Mens Nig...|        Watches|         1|            1|          5|   N|                N|I was about
|         US|   45925069|R2Z6JOQ94LFHEP|B00MAMPGGE|     263720892|G-Shock Men's Gre...|        Watches|         5|            1|          2|   N|                Y|      Perfec
|         US|   44751341| RX27XIIWY5JPB|B004LBPB7Q|     124278407|Heiden Quad Watch...|        Watches|         4|            0|          0|   N|                Y|Great qualit
|         US|    9962330|R15C7QEZT0LGZN|B00KGTVGKS|      28017857|Fossil Women's ES...|        Watches|         4|            2|          2|   N|                Y|           S
|         US|   16097204|R361XSS37V0NCZ|B0039UT5OU|     685450910|Casio General Men...|        Watches|         1|            0|          0|   N|                N|I do not thi
|         US|   51330346| ROTNLALUAJAUB|B00MPF0XJQ|     767769082|2Tone Gold Silver...|        Watches|         3|            0|          0|   N|                Y|         Thr
|         US|    4201739|R2DYX7QU6BGOHR|B003P1OHHS|     648595227|Bulova Men's 98B1...|        Watches|         5|            0|          0|   N|                Y|          Fi
|         US|   26339765| RWASY7FKI7QOT|B00R70YEOE|     457338020|Casio - G-Shock -...|        Watches|         5|            2|          3|   N|                Y|  Worth it -
|         US|    2692576|R2KKYZIN3CCL21|B000FVE3BG|     824370661|Invicta Men's 332...|        Watches|         5|            0|          0|   N|                Y|This is when
|         US|   44713366|R22H4FGVD5O52O|B008X6JB12|     814431355|Seiko Women's SUT...|        Watches|         4|            1|          1|   N|                Y|Thewatch is
|         US|   32778769|R11UACZERCM4ZY|B0040UOFPW|     187700878|Anne Klein Women'...|        Watches|         5|            0|          0|   N|                Y|          Fi
|         US|   27258523|R1AT8NQ38UQOL6|B00UR2R5UY|     594315262|Guess U13630G1 Me...|        Watches|         5|            0|          0|   N|                Y|          Fi
|         US|   42646538|R2NCZRQGIF1Q75|B00HFF57L0|     520810507|Nixon Men's Geo V...|        Watches|         4|            0|          0|   N|                Y|        Very
+-----------+----------+--------------+----------+--------------+--------------------+---------------+----------+-------------+-----------+----+-----------------+------------
only showing top 20 rows
```

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

```
+-----------+--------------+
|customer_id|customer_count|
+-----------+--------------+
|    1567510|             1|
|   19502021|             1|
|   12819130|             1|
```

```
|   35329257|            2|
|     108460|            1|
|    5453476|            1|
|   29913055|            1|
|   30717305|            1|
|    1570030|            1|
|   19032020|            1|
|   44178035|            1|
|   26079415|            2|
|   14230926|            1|
|   43478048|            2|
|   43694941|            1|
|   12318815|            3|
|   13731855|            1|
|     740134|            1|
|   41956754|            1|
|   20324070|            3|
+-----------+-------------+
only showing top 20 rows
```

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

```
+----------+--------------------+
|product_id|       product_title|
+----------+--------------------+
|B00EVX7V1I|Game Time Women's...|
|B009S4DODY|XOXO Women's XO55...|
|B00LBKXQRW|Anne Klein Women'...|
|B0009P679Y|Invicta Men's 993...|
|B00DHF30RU|M&c Women's | Cla...|
|B00NIDA43Y|GuTe Classic Skel...|
|B008EQDDPQ|Nautica Men's N13...|
|B004VRBZ66|Timex Men's T2N63...|
|B009BEO81I|        Fossil Riley|
|B008B39MTI|XOXO Women's XO55...|
|B00TGPM8PU|Handmade Wooden W...|
|B00VNXQQQ0|Eterna 2520-41-64...|
|B00B1PV1C4|Nautica Men's N19...|
|B00N1Y8TQ4|Tissot Men's T095...|
|B00G6DBTY6|red line Men's RL...|
|B00HM04AYI|Columbia Men's Fi...|
|B00VI8HB96|GUESS I90176L1 Wo...|
|B00IT25WJU|LanTac DGN556B Dr...|
|B0106S12XE|Skmei S Shock Ana...|
|B00FPSJ63Y|Michael Kors Ladi...|
+----------+--------------------+
```

only showing top 20 rows

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

```
+--------------+-----------+----------+--------------+-----------+
|     review_id|customer_id|product_id|product_parent|review_date|
+--------------+-----------+----------+--------------+-----------+
|R3O9SGZBVQBV76|    3653882|B00FALQ1ZC|     937001370| 2015-08-31|
| RKH8BNC3L5DLF|   14661224|B00D3RGO20|     484010722| 2015-08-31|
|R2HLE8WKZSU3NL|   27324930|B00DKYC7TK|     361166390| 2015-08-31|
|R31U3UH5AZ42LL|    7211452|B000EQS1JW|     958035625| 2015-08-31|
|R2SV659OUJ945Y|   12733322|B00A6GFD7S|     765328221| 2015-08-31|
| RA51CP8TR5A2L|    6576411|B00EYSOSE8|     230493695| 2015-08-31|
| RB2Q7DLDN6TH6|   11811565|B00WM0QA3M|     549298279| 2015-08-31|
|R2RHFJV0UYBK3Y|   49401598|B00A4EYBR0|     844009113| 2015-08-31|
|R2Z6JOQ94LFHEP|   45925069|B00MAMPGGE|     263720892| 2015-08-31|
| RX27XIIWY5JPB|   44751341|B004LBPB7Q|     124278407| 2015-08-31|
|R15C7QEZT0LGZN|    9962330|B00KGTVGKS|      28017857| 2015-08-31|
|R361XSS37V0NCZ|   16097204|B0039UT5OU|     685450910| 2015-08-31|
| ROTNLALUAJAUB|   51330346|B00MPF0XJQ|     767769082| 2015-08-31|
|R2DYX7QU6BGOHR|    4201739|B003P1OHHS|     648595227| 2015-08-31|
| RWASY7FKI7QOT|   26339765|B00R70YEOE|     457338020| 2015-08-31|
|R2KKYZIN3CCL21|    2692576|B000FVE3BG|     824370661| 2015-08-31|
|R22H4FGVD5O52O|   44713366|B008X6JB12|     814431355| 2015-08-31|
|R11UACZERCM4ZY|   32778769|B0040UOFPW|     187700878| 2015-08-31|
|R1AT8NQ38UQOL6|   27258523|B00UR2R5UY|     594315262| 2015-08-31|
|R2NCZRQGIF1Q75|   42646538|B00HFF57L0|     520810507| 2015-08-31|
+--------------+-----------+----------+--------------+-----------+
only showing top 20 rows
```

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

```
+--------------+-----------+-------------+-----------+----+-----------------+
|     review_id|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+--------------+-----------+-------------+-----------+----+-----------------+
|R3O9SGZBVQBV76|          5|            0|          0|   N|                Y|
| RKH8BNC3L5DLF|          5|            0|          0|   N|                Y|
|R2HLE8WKZSU3NL|          2|            1|          1|   N|                Y|
|R31U3UH5AZ42LL|          5|            0|          0|   N|                Y|
|R2SV659OUJ945Y|          4|            0|          0|   N|                Y|
| RA51CP8TR5A2L|          5|            0|          0|   N|                Y|
| RB2Q7DLDN6TH6|          5|            1|          1|   N|                Y|
|R2RHFJV0UYBK3Y|          1|            1|          5|   N|                N|
```

```
|R2Z6JOQ94LFHEP|          5|            1|          2|   N|                Y|
| RX27XIIWY5JPB|          4|            0|          0|   N|                Y|
|R15C7QEZT0LGZN|          4|            2|          2|   N|                Y|
|R361XSS37V0NCZ|          1|            0|          0|   N|                N|
| ROTNLALUAJAUB|          3|            0|          0|   N|                Y|
|R2DYX7QU6BGOHR|          5|            0|          0|   N|                Y|
| RWASY7FKI7QOT|          5|            2|          3|   N|                Y|
|R2KKYZIN3CCL21|          5|            0|          0|   N|                Y|
|R22H4FGVD5O52O|          4|            1|          1|   N|                Y|
|R11UACZERCM4ZY|          5|            0|          0|   N|                Y|
|R1AT8NQ38UQOL6|          5|            0|          0|   N|                Y|
|R2NCZRQGIF1Q75|          4|            0|          0|   N|                Y|
+--------------+-----------+-------------+-----------+----+-----------------+
only showing top 20 rows
```

Connect to the AWS RDS instance and write each DataFrame to its table.

```
from getpass import getpass
password = getpass('Enter database password')
# Configure settings for RDS
mode = "append"
jdbc_url="jdbc:postgresql://datachallenge.cweam0dq6cy0.us-east-1.rds.amazonaws.com:5432/postgres"
config = {"user":"postgres",
          "password": password,
          "driver":"org.postgresql.Driver"}
```

```
Enter database password·········
```

```
# Write review_id_df to table in RDS
review_id_df.write.jdbc(url=jdbc_url, table='review_id_table', mode=mode, properties=config)
```

```
# Write products_df to table in RDS
# about 3 min
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)
```

✓ 12m 26s completed at 12:15 PM