```
# Install Java, Spark, and Findspark
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q http://www-us.apache.org/dist/spark/spark-2.4.6/spark-2.4.6-bin-hadoop2.7.tgz
!tar xf spark-2.4.6-bin-hadoop2.7.tgz
!pip install -q findspark

# Set Environment Variables
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-2.4.6-bin-hadoop2.7"

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

```
!wget https://jdbc.postgresql.org/download/postgresql-42.2.9.jar
```

```
--2020-08-06 17:06:13--  https://jdbc.postgresql.org/download/postgresql-42.2.9.jar
Resolving jdbc.postgresql.org (jdbc.postgresql.org)... 72.32.157.228, 2001:4800:3e1:1::2
Connecting to jdbc.postgresql.org (jdbc.postgresql.org)|72.32.157.228|:443... connected
HTTP request sent, awaiting response... 200 OK
Length: 914037 (893K) [application/java-archive]
Saving to: ‘postgresql-42.2.9.jar’

postgresql-42.2.9.j 100%[===================>] 892.61K  4.77MB/s    in 0.2s

2020-08-06 17:06:13 (4.77 MB/s) - ‘postgresql-42.2.9.jar’ saved [914037/914037]
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Level1").config("spark.driver.extraClassPath","/content
```

```
# Read in data from S3 files
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Home_Improvement_v1_
spark.sparkContext.addFile(url)
home_imp_df = spark.read.csv(SparkFiles.get("amazon_reviews_us_Home_Improvement_v1_00.tsv.gz"
home_imp_df.show()
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|p
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   48881148|R215C9BDXTDQOW|B00FR4YQYK|     381800308|SadoTech Model C ...|H
|         US|   47882936|R1DTPUV1J57YHA|B00439MYYE|     921341748|iSpring T32M 3.2 ...|H
|         US|   44435471| RFAZK5EWKJWOU|B00002N762|      56053291|Schlage F10CS V E...|H
|         US|   28377689|R2XT8X0O0WS1AL|B000QFCP1G|     595928517|Citri-Strip QCG73...|H
|         US|   50134766|R14GRNANKO2Y2J|B00WRCRKOI|     417053744|SleekLighting Bul...|H
|         US|   14066511|R2BLF9VYL24LCQ|B00NIH88EW|     275395071|VDOMUS&reg;Exquis...|H
|         US|   15211046|R1GI9UW5KJ671O|B005B9CI96|     856617815|Frigidaire 316075...|H
|         US|   14862498|R2H5CEJN863M86|B008L0OMWI|     125102494|Anyray® 5-Bulbs 7...|H
|         US|   23617292| R5PPDHFOZ3SMU|B00P9FTC6O|     523110842|Cambridge 100 pcs...|H
|         US|   35820485| RE1L9IENKJJ7Y|B00K6BQEHQ|     797306964|EUBUY Silver Tone...|H
|         US|   47162350|R3CZ0990QC2Z0H|B008BYQCWM|     865874404|Legend 809125 Leg...|H
|         US|   31884789|R3UMMD2IO29QSP|B003BLHTOU|     183592595|Forearm Forklift ...|H
|         US|   43835770| R541LE5J3OJH2|B0065I114K|     185006358|Newer Technology ...|H
|         US|   26212294|R10KD9FIF6Y1AS|B00J9YB6RO|      53422207|Bok Hardware 35 P...|H
```

```
# Count rows
home_imp_df.count()
```

```
2634781
```

```
|         US|   48001709|R1QTQSE2BHU2LJ|B000BQQX88|     498022954|Intermatic T101 2...|H
```

```
# Chage datatypes of individual columns
from pyspark.sql.types import IntegerType
home_imp_df = home_imp_df.withColumn("customer_id",home_imp_df["customer_id"].cast(IntegerTyp
home_imp_df = home_imp_df.withColumn("product_parent",home_imp_df["product_parent"].cast(Inte
home_imp_df = home_imp_df.withColumn("star_rating",home_imp_df["star_rating"].cast(IntegerTyp
home_imp_df = home_imp_df.withColumn("helpful_votes",home_imp_df["helpful_votes"].cast(Intege
home_imp_df = home_imp_df.withColumn("total_votes",home_imp_df["total_votes"].cast(IntegerTyp
```

```
# Clean up DataFrames to match tables
from pyspark.sql.functions import to_date

# Review DataFrame
review_df = home_imp_df.select(["review_id", "customer_id", "product_id", "product_parent", t

review_df.show()
```

```
+--------------+-----------+----------+--------------+-----------+
|     review_id|customer_id|product_id|product_parent|review_date|
+--------------+-----------+----------+--------------+-----------+
|R215C9BDXTDQOW|   48881148|B00FR4YQYK|     381800308| 2015-08-31|
|R1DTPUV1J57YHA|   47882936|B00439MYYE|     921341748| 2015-08-31|
| RFAZK5EWKJWOU|   44435471|B00002N762|      56053291| 2015-08-31|
|R2XT8X0O0WS1AL|   28377689|B000QFCP1G|     595928517| 2015-08-31|
|R14GRNANKO2Y2J|   50134766|B00WRCRKOI|     417053744| 2015-08-31|
|R2BLF9VYL24LCQ|   14066511|B00NIH88EW|     275395071| 2015-08-31|
|R1GI9UW5KJ671O|   15211046|B005B9CI96|     856617815| 2015-08-31|
|R2H5CEJN863M86|   14862498|B008L0OMWI|     125102494| 2015-08-31|
| REPPDHEOZ2SMU|   22617292|B00P9FTC6O|     523110842| 2015-08-31|
```

```
# Products DataFrame
products_df = home_imp_df.select(["product_id", "product_title"]).drop_duplicates()
products_df.show()
```

```
+----------+--------------------+
|product_id|       product_title|
+----------+--------------------+
|B0136K7XN2|Brightown 72ft/22...|
|B013UDQE26|Amdirect Folding ...|
|B005WQIDHY|Woods 50007 Indoo...|
|B000H5QLJC|Dap 18013 Kwik-Se...|
|B002I0GRRA|KOHLER Cimarron C...|
|B00O4CUB7M|Kraus Nola Single...|
|B00857R9PY|Park Madison Ligh...|
|B009IJ2NBI|TOTO Washlet S350...|
|B00IZA2TAQ|XKTTSUEERCRR Wate...|
|B008UDP4PA|35265B Genie 90FT...|
|B004TSYJSI|Moen LR2356DBN 16...|
|B003YUGQWE|General Finishes ...|
|B0006VVN1I|Culligan Certifie...|
|B00CJ5EO2E|Gorilla Super Glu...|
|B00RBYV0JS|Decor Star TPC11 ...|
|B007VP7QCC|LiftMaster 1345 C...|
|B00106L9AS|Moonrays 91251 Co...|
|B011SXJCN8|Dean Sports Alumi...|
|B000R89WZO|Jaw Inserts for 4...|
|B000Z6G008|Westinghouse 6751...|
+----------+--------------------+
only showing top 20 rows
```

```
# Reviews DataFrame
reviews_df = home_imp_df.select(["review_id", "review_headline", "review_body"])
reviews_df.show()
```

```
+--------------+--------------------+--------------------+
|     review_id|     review_headline|         review_body|
+--------------+--------------------+--------------------+
|R215C9BDXTDQOW|          Four Stars|        good product|
|R1DTPUV1J57YHA|Good price, quick...|Good price, quick...|
| RFAZK5EWKJWOU|          Five Stars|       Excellent...!|
|R2XT8X0O0WS1AL|Although *slightl...|Although *slightl...|
|R14GRNANKO2Y2J|      Great Adapters|These adapters ar...|
|R2BLF9VYL24LCQ|                nice|awesome and great...|
|R1GI9UW5KJ671O|          Five Stars|Perfect. Exactly ...|
|R2H5CEJN863M86|So far working gr...|So far working gr...|
| R5PPDHFOZ3SMU|           Ties tie.|Not much to say a...|
| RE1L9IENKJJ7Y|             GARBAGE|       crap quality.|
|R3CZ0990QC2Z0H|          Five Stars|  Good. As expected.|
|R3UMMD2IO29QSP|           It works!|My husband and I ...|
| R541LE5J3OJH2|          Five Stars|     works perfectly|
|R10KD9FIE6Y1AS|Banging cabinet d...|Banging cabinet d...|
|R1YT5YG0QG5DCG|Great product for...|Great product for...|
```

```
# Customers DataFrame
customers_df = home_imp_df.groupby("customer_id").agg({"customer_id": "count"}).withColumnRen
customers_df.show()
```

```
+-----------+--------------+
|customer_id|customer_count|
+-----------+--------------+
|   28377689|             1|
|   28258386|             2|
|    9967574|             1|
|   25153155|             1|
|   10088068|             1|
|   45657423|             1|
|   19021463|             1|
|   41413793|             2|
|   43789873|             2|
|   12406466|             1|
|   41045019|             1|
|   35535911|             2|
|   47108763|             1|
|   48113150|             1|
|   18201417|             9|
|   26079415|             3|
|   36114891|             7|
|    3712628|             1|
|   37499901|             2|
|   47321438|             2|
+-----------+--------------+
only showing top 20 rows
```

```
# Vine DataFrame
vine_df = home_imp_df.select(["review_id", "star_rating", "helpful_votes", "total_votes", "vi
vine_df.show()
```

```
+--------------+-----------+-------------+-----------+----+
|     review_id|star_rating|helpful_votes|total_votes|vine|
+--------------+-----------+-------------+-----------+----+
|R215C9BDXTDQOW|          4|            0|          0|   N|
|R1DTPUV1J57YHA|          5|            0|          0|   N|
| RFAZK5EWKJWOU|          5|            0|          0|   N|
|R2XT8X0O0WS1AL|          5|            0|          0|   N|
|R14GRNANKO2Y2J|          5|            0|          0|   N|
|R2BLF9VYL24LCQ|          5|            1|          1|   N|
|R1GI9UW5KJ671O|          5|            0|          0|   N|
|R2H5CEJN863M86|          5|            0|          1|   N|
| R5PPDHFOZ3SMU|          5|            0|          0|   N|
| RE1L9IENKJJ7Y|          1|            0|          0|   N|
|R3CZ0990QC2Z0H|          5|            0|          0|   N|
|R3UMMD2IO29QSP|          5|            0|          0|   N|
| R541LE5J3OJH2|          5|            0|          0|   N|
|R10KD9FIE6Y1AS|          5|            0|          0|   N|
|R1YT5YG0QG5DCG|          5|            0|          0|   N|
|R207LXJWL40V1S|          4|            0|          0|   N|
| RIDP0ZD7WT9DE|          5|            0|          0|   N|
|R2XJSNZ9219U1Z|          5|            0|          0|   N|
|R2M9F1FVVDOGFL|          5|            0|          0|   N|
|R1QTOSE2BHU2LJ|          3|            1|          2|   N|
+--------------+-----------+-------------+-----------+----+
only showing top 20 rows
```

```
# Postgres Setup
```

```
# Configure settings for RDS
mode = "append"
jdbc_url="jdbc:postgresql://mypostgresdb.ckmpkdemgqsk.us-east-2.rds.amazonaws.com/my_data_cla
config = {"user":"root",
          "password": "xxxxx",
          "driver":"org.postgresql.Driver"}
```

```
# Write review_df to to active_user table in RDS
review_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', mode=mode, properties=config)
```

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

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