```
# 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
```

```
--2020-08-07 00:32:58--  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::
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.2’

postgresql-42.2.9.j 100%[===================>] 892.61K  1.43MB/s    in 0.6s

2020-08-07 00:33:00 (1.43 MB/s) - ‘postgresql-42.2.9.jar.2’ saved [914037/914037]
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Level2").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_Lawn_and_Garden_v1_0
spark.sparkContext.addFile(url)
lawn_gard_df = spark.read.csv(SparkFiles.get("amazon_reviews_us_Lawn_and_Garden_v1_00.tsv.gz"
lawn_gard_df.show()
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|p
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   32787517| RED72VWWCOS7S|B008HDQYLQ|     348668413|Garden Weasel Gar...|
|         US|   16374060| RZHWQ208LTEPV|B005OBZBD6|     264704759|10 Foot Mc4 Solar...|
|         US|    9984817|R37LBC3XAVLYOO|B00RQL8U2G|      95173602|GE String A Long ...|
|         US|   12635190|R3L7XJMA0MVJWC|B0081SBO4Y|     835659279|Key Pair Lawn Wit...|
|         US|   43905102|R2I2GHSI7T1UBN|B008E6OK3U|     539243347|Zodiac R0502300 L...|
|         US|   52596997|R2GFFKHK4I6VMX|B00W6NTULY|     337446474|Hirts Gardens Swe...|
|         US|   43871104|R1R0UDX2XAN1S4|B00GXUMYKA|     468857193|AGPtEK 12 PCS Smo...|
|         US|   11346008|R22C8FMBSTFRY8|B005EIX8JS|     125753094|Design Toscano Ea...|
|         US|   49206471|R118NNIQ75XPGO|B000HJBKMQ|     834273114|TERRO T300 Liquid...|
|         US|   37596267|R30HYXHZQ49621|B004LY59V6|     612086079|BLACK+DECKER LBXR...|
|         US|   31554283|R3EMLKY0GF1E90|B00CAVM85M|     280334010|Reach 'n Spray Pe...|
|         US|   43211735|R23BX7EGJMGQJR|B00DP6X1LG|     233116679|Puro-Kleen Ultra-...|
|         US|   25705116|R2Z4B6SDEAZF6E|B00025H2PY|     592807498|Diatomaceous Eart...|
```

```
# Count rows - before cleanup
print("Total product review count before cleanup: ",lawn_gard_df.count() )
```

```
Total product review count before cleanup:  2557288
```

```
|         US|   10291713|R1TMSZWIT21A31|B000UJH6HQ|     228393894|Toro 53746 Drip B...|
```

```
# Drop null values
lawn_gard_df = lawn_gard_df.dropna()
lawn_gard_df.show()
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|p
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   32787517| RED72VWWCOS7S|B008HDQYLQ|     348668413|Garden Weasel Gar...|
|         US|   16374060| RZHWQ208LTEPV|B005OBZBD6|     264704759|10 Foot Mc4 Solar...|
|         US|    9984817|R37LBC3XAVLYOO|B00RQL8U2G|      95173602|GE String A Long ...|
|         US|   12635190|R3L7XJMA0MVJWC|B0081SBO4Y|     835659279|Key Pair Lawn Wit...|
|         US|   43905102|R2I2GHSI7T1UBN|B008E6OK3U|     539243347|Zodiac R0502300 L...|
|         US|   52596997|R2GFFKHK4I6VMX|B00W6NTULY|     337446474|Hirts Gardens Swe...|
|         US|   43871104|R1R0UDX2XAN1S4|B00GXUMYKA|     468857193|AGPtEK 12 PCS Smo...|
|         US|   11346008|R22C8FMBSTFRY8|B005EIX8JS|     125753094|Design Toscano Ea...|
|         US|   49206471|R118NNIQ75XPGO|B000HJBKMQ|     834273114|TERRO T300 Liquid...|
|         US|   37596267|R30HYXHZQ49621|B004LY59V6|     612086079|BLACK+DECKER LBXR...|
|         US|   31554283|R3EMLKY0GF1E90|B00CAVM85M|     280334010|Reach 'n Spray Pe...|
|         US|   43211735|R23BX7EGJMGQJR|B00DP6X1LG|     233116679|Puro-Kleen Ultra-...|
|         US|   25705116|R2Z4B6SDEAZF6E|B00025H2PY|     592807498|Diatomaceous Eart...|
|         US|   47041108|R35289PGJERP5J|B0079GHJXY|     408290044|Perky-Pet 312C Pa...|
|         US|    1534667|R39BPRMDKKIZL2|B004HFJ762|     404737140|Crossbow Dow Spec...|
|         US|   52287759| R6WFPPBS1DZMG|B00004RAGL|     773636542|Apex REM 15 15-Fo...|
|         US|   37010286| RK72M0ZBV9YLS|B010PWBNNK|     461072629|Elucto Electric B...|
|         US|   30576559| RX5G150AUWRDJ|B00T77AWY6|     365662076|Ohuhu® 100 Ft Exp...|
|         US|   10291713|R1TMSZWIT21A31|B000UJH6HQ|     228393894|Toro 53746 Drip B...|
|         US|   50656780|R2FURVPW763CIM|B000HJBKMQ|     834273114|TERRO T300 Liquid...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 20 rows
```

```
# Count rows - after cleanup
print("Number of product reviews after cleanup: ",lawn_gard_df.count())
print("Number of product reviews after cleanup: ",lawn_gard_df.count())
```

```
Number of product reviews after cleanup:  2557005
```

```
# Get Product vote info
product_voter_df = lawn_gard_df.select(["star_rating", "helpful_votes", "total_votes", "vine"
product_voter_df.show(10)
product_voter_df.count()
```

```
+-----------+-------------+-----------+----+-----------------+
|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+-----------+-------------+-----------+----+-----------------+
|          1|            2|          8|   N|                Y|
|          5|            0|          0|   N|                Y|
|          5|            4|          5|   N|                Y|
|          5|            0|          0|   N|                Y|
|          1|            5|          6|   N|                Y|
|          5|            0|          0|   N|                Y|
|          4|            0|          0|   N|                Y|
|          5|            2|          2|   N|                Y|
|          3|            0|          0|   N|                Y|
|          2|            0|          0|   N|                Y|
+-----------+-------------+-----------+----+-----------------+
only showing top 10 rows

2557005
```

```
# Get total verified and Unverified purchases
total_verified_df = lawn_gard_df.filter(lawn_gard_df['verified_purchase'] == 'Y')
total_unverified_df = lawn_gard_df.filter(lawn_gard_df['verified_purchase'] == 'N')

print("Total verified purchases: ",total_varified_df.count())
print("Total Unverified purchases: ",total_unvarified_df.count())
print("% Ratio of Unverified to verified purchase reviews: ",float(total_unvarified_df.count(
```

```
Total verified purchases:  2251611
Total Unverified purchases:  305394
% Ratio of Unverified to verified purchase reviews:  0.1356335530426881
```

```
# Get the products with 10 plus total votes
total_vote_df = lawn_gard_df.filter(lawn_gard_df['total_votes'] >= 10)
total_vote_df.show(10)
print("Number of products with reviews with at least 10 votes: ",total_vote_df.count())
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|p
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   33399595| R6J125A9S5H1G|B00OMSJ9WG|     583972217|10x10 V-Series 2 ...|
|         US|   12020967|R3SJT43TE6IY0O|B00S96Q2UO|     997068254|Heat Resistant Si...|
|         US|   42142471|R3CIYLO59XNDVJ|B002ASAB6I|      13220966|Timber Tuff TMB-4...|
|         US|   39823685| RQQ3KVTU5TJ4I|B00005A3L1|     436617004|Bounty Hunter TK4...|
|         US|   35875101|R3FELXWV9T5CWE|B00VQVPRH8|     733961147|Multi-Purpose Boo...|
|         US|   15323081| ROBYK6EZYK398|B00GOH6WVY|     746010001|Root Naturally Az...|
|         US|   13866645|R2RKCSAG6GBA4A|B00Q97CPIK|     860333862|Rid Tech Ultrason...|
```

```
# Percentage of products with 10 plus total votes for the review
print("Percentage of products with 10 plus reviews: ",float(total_vote_df.count()/lawn_gard_d
```

```
Percentage of products with 10 plus reviews:  0.049227123138202704 %
```

```
# Describe stats for paid and unpaid products
from pyspark.sql.functions import col, avg
paid_df = lawn_gard_df.filter(lawn_gard_df['vine']== 'Y')
unpaid_df = lawn_gard_df.filter(lawn_gard_df['vine']== 'N')

paid_df.describe().show()
unpaid_df.describe().show()
```

```
+-------+-----------+--------------------+--------------+----------+--------------------
|summary|marketplace|         customer_id|     review_id|product_id|     product_parent
+-------+-----------+--------------------+--------------+----------+--------------------
|  count|      13454|               13454|         13454|     13454|              13454
|   mean|       null|4.0954081251449384E7|          null|      null|5.074572449275308E8
| stddev|       null|1.2849717934906457E7|          null|      null|2.913488896656647E8
|    min|         US|            10044936|R100Q8WPKHEE17|B00004ZAVI|          100001885
|    max|         US|             9944883| RZZR0HC19L9HQ|B010QVQKJ2|          999221604
+-------+-----------+--------------------+--------------+----------+--------------------

+-------+-----------+--------------------+--------------+--------------------+----------
|summary|marketplace|         customer_id|     review_id|          product_id|      proc
+-------+-----------+--------------------+--------------+--------------------+----------
|  count|    2543551|             2543551|       2543551|             2543551|
|   mean|       null|2.8691685062479187E7|          null| 5.546940814737113E9|4.97965778
| stddev|       null|1.5269180832884334E7|          null|3.5584258942622313E9|2.88396352
|    min|         US|            10000009|R10001EZRM7QD7|          0618307354|
|    max|         US|             9999997| RZZZX5XP2S3X6|          B01JPMYFSG|
+-------+-----------+--------------------+--------------+--------------------+----------
```

```
#  Determine paid and five star paid reviews
paid_number = paid_df.count()
paid_five_star_number = paid_df[paid_df['star_rating']== 5].count()

print("Paid Review count: ",paid_number)
print("Paid Five Star Review count: ",paid_five_star_number)
```

```
Paid Review count:  13454
Paid Five Star Review count:  6006
```

```
#  Determine the percentage of five-star reviews among Vine reviews
percentage_five_star_vine = paid_five_star_number/paid_number

print("% of five-star reviews among Vine reviews",float(percentage_five_star_vine),"%")
```

```
% of five-star reviews among Vine reviews 0.4464099895941727 %
```

```
#  Determine the percentage of five-star reviews among non-Vine reviews.
unpaid_number = unpaid_df.count()
unpaid_five_star_number = unpaid_df[unpaid_df['star_rating']== 5].count()

print("Unpaid Review count: ",paid_number)
print("Unpaid Five Star Review count: ",paid_five_star_number)
```

```
Unpaid Review count:  13454
Unpaid Five Star Review count:  6006
```

```
# Determine the percentage of five-star reviews among non-Vine reviews.
percentage_five_star_non_vine = unpaid_five_star_number/unpaid_number
print("% of five-star unpaid reviews among non Vine reviews", float(percentage_five_star_non_
```

```
% of five-star unpaid reviews among non Vine reviews 0.605333645757447 %
```