```
import os
# Find the latest version of spark 3.0 from http://www.apache.org/dist/spark/ and enter as th
# 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 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 http://archive.ubuntu.com/ubuntu bionic InRelease
Get:2 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/ InRelease [3,626 B]
Get:3 http://archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]
Get:4 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Hit:5 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic InRelease
Get:6 http://archive.ubuntu.com/ubuntu bionic-backports InRelease [83.3 kB]
Ign:7 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x8
Hit:8 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  InRele
Hit:9 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x8
Hit:10 http://ppa.launchpad.net/cran/libgit2/ubuntu bionic InRelease
Hit:11 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic InRelease
Hit:12 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease
Get:13 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 Packages [3,472 kB]
Get:14 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 Packages [2,332 kB
Get:16 http://security.ubuntu.com/ubuntu bionic-security/universe amd64 Packages [1,554
Get:17 http://security.ubuntu.com/ubuntu bionic-security/main amd64 Packages [3,040 kB]
Fetched 10.7 MB in 5s (1,986 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
```

```
--2022-11-12 04:21:07--  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::2
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’

postgresql-42.2.16. 100%[===================>] 979.38K  --.-KB/s    in 0.09s

2022-11-12 04:21:08 (11.0 MB/s) - ‘postgresql-42.2.16.jar’ saved [1002883/1002883]
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("M16-Amazon-Challenge").config("spark.driver.extraClassP


from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Pet_Products_v1_00.t
spark.sparkContext.addFile(url)
df = spark.read.option("encoding", "UTF-8").csv(SparkFiles.get(""), sep="\t", header=True, in
df.show()
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|p
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   28794885| REAKC26P07MDN|B00Q0K9604|     510387886|(8-Pack) EZwhelp ...|
|         US|   11488901|R3NU7OMZ4HQIEG|B00MBW5O9W|     912374672|Warren Eckstein's...|
|         US|   43214993|R14QJW3XF8QO1P|B0084OHUIO|     902215727|Tyson's True Chew...|
|         US|   12835065|R2HB7AX0394ZGY|B001GS71K2|     568880110|Soft Side Pet Cra...|
|         US|   26334022| RGKMPDQGSAHR3|B004ABH1LG|     692846826|EliteField 3-Door...|
|         US|   22283621|R1DJCVPQGCV66E|B00AX0LFM4|     590674141|Carlson 68-Inch W...|
|         US|   14469895|R3V52EAWLPBFQG|B00DQFZGZ0|     688538603|Dog Seat Cover Wi...|
|         US|   50896354|R3DKO8J1J28QBI|B00DIRF9US|     742358789|The Bird Catcher ...|
|         US|   18440567| R764DBXGRNECG|B00JRCBFUG|     869798483|Cat Bed - Purrfec...|
|         US|   50502362| RW1853GAT0Z9F|B000L3XYZ4|     501118658|PetSafe Drinkwell...|
|         US|   33930128|R33GITXNUF1AD4|B00BOEXWFG|     454737777|Contech ZenDog Ca...|
|         US|   43534290|R1H7AVM81TAYRV|B001HBBQKY|     420905252|Wellness Crunchy ...|
|         US|   45555864|R2ZOYAQZNNZZWV|B007O1FHB0|     302588963|Rx Vitamins Essen...|
|         US|   11147406|R2FN1H3CGW6J8H|B001P3NU30|     525778264|Virbac C.E.T. Enz...|
|         US|    6495678| RJB41Q575XNG4|B00ZP6HS6S|     414117299|Kitty Shack - 2 i...|
|         US|    2019416|R28W8BM1587CPF|B00IP05CUA|     833937853|Wellness Kittles ...|
|         US|   40459386|R1II0M01NIG293|B001U8Y598|      85343577|OmniPet Anti-Ant ...|
|         US|   23126800| RMB8N0DBRH34O|B011AY4JWO|     499241195|K9KONNECTION [New...|
|         US|   30238476|R24WB6A6WVIPU6|B00DDSHE5A|     409532388|SUNSEED COMPANY 3...|
|         US|   35113999| ROCJSH0P9YSRW|B00PJW5OR8|     259271919|CXB1983(TM)Cute P...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 20 rows
```

```
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|   28794885| REAKC26P07MDN|B00Q0K9604|     510387886|(8-Pack) EZwhelp ...|
|         US|   11488901|R3NU7OMZ4HQIEG|B00MBW5O9W|     912374672|Warren Eckstein's...|
|         US|   43214993|R14QJW3XF8QO1P|B0084OHUIO|     902215727|Tyson's True Chew...|
|         US|   12835065|R2HB7AX0394ZGY|B001GS71K2|     568880110|Soft Side Pet Cra...|
|         US|   26334022| RGKMPDQGSAHR3|B004ABH1LG|     692846826|EliteField 3-Door...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 5 rows
```

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

```
vine_df.show()
```

```
+--------------+-----------+-------------+-----------+----+-----------------+
|     review_id|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+--------------+-----------+-------------+-----------+----+-----------------+
| REAKC26P07MDN|          5|            0|          0|   N|                Y|
|R3NU7OMZ4HQIEG|          2|            0|          1|   N|                Y|
|R14QJW3XF8QO1P|          5|            0|          0|   N|                Y|
|R2HB7AX0394ZGY|          5|            0|          0|   N|                Y|
| RGKMPDQGSAHR3|          5|            0|          0|   N|                Y|
|R1DJCVPQGCV66E|          5|            0|          0|   N|                Y|
|R3V52EAWLPBFQG|          3|            0|          0|   N|                Y|
|R3DKO8J1J28QBI|          2|            0|          0|   N|                Y|
| R764DBXGRNECG|          5|            1|          1|   N|                N|
| RW1853GAT0Z9F|          5|            0|          0|   N|                Y|
|R33GITXNUF1AD4|          2|            0|          0|   N|                Y|
|R1H7AVM81TAYRV|          1|            2|          2|   N|                Y|
|R2ZOYAQZNNZZWV|          5|            0|          0|   N|                Y|
|R2FN1H3CGW6J8H|          1|            0|          0|   N|                Y|
| RJB41Q575XNG4|          5|            0|          3|   N|                Y|
|R28W8BM1587CPF|          5|            0|          0|   N|                Y|
|R1II0M01NIG293|          2|            0|          0|   N|                N|
| RMB8N0DBRH34O|          5|            1|          1|   N|                Y|
|R24WB6A6WVIPU6|          5|            0|          0|   N|                Y|
| ROCJSH0P9YSRW|          5|            0|          0|   N|                Y|
+--------------+-----------+-------------+-----------+----+-----------------+
only showing top 20 rows
```

```
top_reviews_df = vine_df.filter("total_votes>=20").show(5)
```

```
+--------------+-----------+-------------+-----------+----+-----------------+
|     review_id|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+--------------+-----------+-------------+-----------+----+-----------------+
|R21KC552Y6HL8X|          1|           27|         31|   N|                Y|
| RX9WC9FTIR1XR|          5|           25|         25|   N|                Y|
| RGDCOU1KBHMNG|          3|           29|         31|   N|                Y|
| RVTYWID2TPMMY|          2|           35|         42|   N|                Y|
|R2CMPZ5VESGRLY|          4|           27|         28|   N|                Y|
+--------------+-----------+-------------+-----------+----+-----------------+
```

```
only showing top 5 rows
```

```
helpful_reviews_df = vine_df.filter("helpful_votes>=20").show(5)
```

```
+--------------+-----------+-------------+-----------+----+-----------------+
|     review_id|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+--------------+-----------+-------------+-----------+----+-----------------+
|R21KC552Y6HL8X|          1|           27|         31|   N|                Y|
| RX9WC9FTIR1XR|          5|           25|         25|   N|                Y|
| RGDCOU1KBHMNG|          3|           29|         31|   N|                Y|
| RVTYWID2TPMMY|          2|           35|         42|   N|                Y|
|R2CMPZ5VESGRLY|          4|           27|         28|   N|                Y|
+--------------+-----------+-------------+-----------+----+-----------------+
only showing top 5 rows
```

```
vine_review_df = vine_df.filter("vine == 'Y'").show(5)
```

```
+--------------+-----------+-------------+-----------+----+-----------------+
|     review_id|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+--------------+-----------+-------------+-----------+----+-----------------+
|R1CBOJMJAYL75C|          5|            0|          0|   Y|                N|
| R37IHP001XZVR|          5|            0|          1|   Y|                N|
|R175KT8QHRRK2G|          4|            0|          1|   Y|                N|
| RNWWD2B3X0CU2|          5|            0|          0|   Y|                N|
|R2FDITZFABSUN8|          3|            0|          1|   Y|                N|
+--------------+-----------+-------------+-----------+----+-----------------+
only showing top 5 rows
```

```
vine_review_df = vine_df.filter("vine == 'N'").show(5)
```

```
+--------------+-----------+-------------+-----------+----+-----------------+
|     review_id|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+--------------+-----------+-------------+-----------+----+-----------------+
| REAKC26P07MDN|          5|            0|          0|   N|                Y|
|R3NU7OMZ4HQIEG|          2|            0|          1|   N|                Y|
|R14QJW3XF8QO1P|          5|            0|          0|   N|                Y|
|R2HB7AX0394ZGY|          5|            0|          0|   N|                Y|
| RGKMPDQGSAHR3|          5|            0|          0|   N|                Y|
+--------------+-----------+-------------+-----------+----+-----------------+
only showing top 5 rows
```

Colab paid products - Cancel contracts here

✓ 0s completed at 9:57 PM