

```
%pyspark
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_Games_v1_00.tsv.gz"
spark.sparkContext.addfile(url)
video_games_df = spark.read.csv(SparkFiles.get("amazon_reviews_us_Video_Games_v1_00.tsv.gz"), sep="\t", header=True)
video_games_df.show()
```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

| marketplace | customer_id | review_id | product_id | product_parent | product_title | product_category | star_rating | helpful_votes | total_votes | vine | verified_purchase | re |
|-------------------------------------|-------------|----------------|-------------|----------------|----------------------|------------------|-------------|---------------|-------------|------|-------------------|-----------|
| view_headline | | review_body | review_date | | | | | | | | | |
| US | 12039526 | RTIS3L2M1F5SM | B001CXYMFS | 737716809 | Thrustmaster T-Fl... | Video Games | 5 | 0 | 0 | N | | Y an amaz |
| ing joystick... Used this for El... | | | | 2015-08-31 | | | | | | | | |
| US | 9636577 | R1ZV7R40OLHDK | B00M920ND6 | 569686175 | Tonsee 6 buttons ... | Video Games | 5 | 0 | 0 | N | | Y Definit |
| ely a sile... Loved it, I didn... | | | | 2015-08-31 | | | | | | | | |
| US | 2331478 | R3BH071QLH8QMC | B0029CSOD2 | 98937668 | Hidden Mysteries:... | Video Games | 1 | 0 | 1 | N | | Y |
| One Star poor quality work... | | | | 2015-08-31 | | | | | | | | |
| US | 52495923 | R127K9NTSXAYH | B00GOOSV98 | 23143350 | GelTabz Performan... | Video Games | 3 | 0 | 0 | N | | Y good, b |
| ut could b... nice, but tend to... | | | | 2015-08-31 | | | | | | | | |
| US | 14533949 | R3Z2WUXDJPW270 | B00Y074JOM | 821342511 | Zero Suit Samus a... | Video Games | 4 | 0 | 0 | N | | Y Grea |
| t but flawed. Great amiibo, gre... | | | | 2015-08-31 | | | | | | | | |
| US | 2377552 | R3AQ04YUKJWB6 | B002UBI6W6 | 328764615 | Psyclone Recharge... | Video Games | 1 | 0 | 0 | N | | Y |
| One Star The remote consta... | | | | 2015-08-31 | | | | | | | | |
| US | 17521011 | R2F0POU5K6F73F | B008XHCLFO | 24234603 | Protection for yo... | Video Games | 5 | 0 | 0 | N | | Y |
| A Must I have a 2012-201... | | | | 2015-08-31 | | | | | | | | |
| US | 19676307 | R3VNR804HYSMR6 | B00BRA9R6A | 682267517 | Nerf 3DS XL Armor | Video Games | 5 | 0 | 0 | N | | Y |
| Five Stars Perfect, kids lov... | | | | 2015-08-31 | | | | | | | | |
| US | 224068 | R3GZTM72WA2QH | B009EPWJLA | 435241890 | One Piece: Pirate... | Video Games | 5 | 0 | 0 | N | | Y |
| Five Stars Excellent | | | | 2015-08-31 | | | | | | | | |
| US | 48467989 | RNQOY62705W1K | B0000AV7GB | 256572651 | Playstation 2 Dan... | Video Games | 4 | 0 | 0 | N | | Y |
| Four Stars Slippery but expe... | | | | 2015-08-31 | | | | | | | | |
| US | 106569 | R1VTIA3JTYBY02 | B00008KTNN | 384411423 | Metal Arms: Glit... | Video Games | 5 | 0 | 0 | N | | N |
| Five Stars Love the game. Se... | | | | 2015-08-31 | | | | | | | | |
| US | 48269642 | R29DOU8791QZL8 | B000A3IA0Y | 472622859 | 72 Pin Connector ... | Video Games | 1 | 0 | 0 | N | | Y Game w |
| ill get stuck Does not fit prop... | | | | 2015-08-31 | | | | | | | | |

Interpreter: spark.pyspark. FINISHED Took 58 sec 551 millisec. Updated by ZGrinacoff on November 08 2019, 4:12:33 PM (PST)



```
%pyspark
# Count rows
video_games_df.count()
```

1785997

Interpreter: spark.pyspark. FINISHED Took 15 sec 883 millisec. Updated by ZGrinacoff on November 08 2019, 4:12:49 PM (PST)



```
# Clean up DataFrames to match tables
```

Interpreter: python. FINISHED Took 2 sec 9 millisec. Updated by ZGrinacoff on November 08 2019, 4:12:51 PM (PST)



```
%pyspark
from pyspark.sql.functions import to_date
# Review DataFrame
review_id_df = video_games_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 |
|-----------------|-------------|-------------|----------------|-------------|
| RTIS3L2M1F5SM | 12039526 | B001CXYMFS | 737716809 | 2015-08-31 |
| R1ZV7R40OLHDK | 9636577 | B00M920ND6 | 569686175 | 2015-08-31 |
| R3BH071QLH8QMC | 2331478 | B0029CSOD2 | 98937668 | 2015-08-31 |
| R127K9NTSXAYH | 52495923 | B00GOOSV98 | 23143350 | 2015-08-31 |
| R3Z2WUXDJPW270 | 14533949 | B00Y074JOM | 821342511 | 2015-08-31 |
| R3AQ04YUKJWB6 | 2377552 | B002UBI6W6 | 328764615 | 2015-08-31 |
| R2F0POU5K6F73F | 17521011 | B008XHCLFO | 24234603 | 2015-08-31 |
| R3VNR804HYSMR6 | 19676307 | B00BRA9R6A | 682267517 | 2015-08-31 |
| R3GZTM72WA2QH | 224068 | B009EPWJLA | 435241890 | 2015-08-31 |
| RNQOY62705W1K | 48467989 | B0000AV7GB | 256572651 | 2015-08-31 |
| R1VTIA3JTYBY02 | 106569 | B00008KTNN | 384411423 | 2015-08-31 |
| R29DOU8791QZL8 | 48269642 | B000A3IA0Y | 472622859 | 2015-08-31 |
| R15DUT1VIJ9RJZ | 52738710 | B0053BQN34 | 577628462 | 2015-08-31 |
| R3IMF2M0Q30U9ZM | 10556786 | B002IOHIMI | 988218515 | 2015-08-31 |
| R23H79DHOZTYAU | 2963837 | B0081EH12M | 770100932 | 2015-08-31 |
| RIV24EQAIXAX40 | 23092109 | B005FMLZQ0 | 24647669 | 2015-08-31 |
| R3UCNGYDVN24YB | 23091728 | B002BSA388 | 33706205 | 2015-08-31 |
| RUL4H4XTTN2DV | 10712640 | B000BUSLSAC | 829667834 | 2015-08-31 |
| R20JF7Z4DHTNX5 | 17455376 | B00KWF38AW | 110680188 | 2015-08-31 |
| R2T1AJ5MF12260 | 14754850 | B00BRQJYA8 | 616463426 | 2015-08-31 |

only showing top 20 rows

Interpreter: spark.pyspark. FINISHED Took 262 millisec. Updated by ZGrinacoff on November 08 2019, 4:12:51 PM (PST)



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

+-----+
|product_id| product_title|
+-----+
|B00CJ7IUI6|The Elder Scrolls...|
|B00DHF39KS|Wolfenstein: The ...|
|B00MUTAVH6|Under Night In-Bi...|
|B001AZSEUW| Peggle|
|B00KV0VBGM|PlayStation 4 Con...|
|B0009VGH4Y|USPRO® Headph...|
|B00400NZY4|Phineas and Ferb:...|
|B00ZLN9800|Donop seablue 2.4...|
|B002LB8W5V6|Topo Nintendo Ga...|
|B001J5MF12260|...

```
|B007A1Z5PY|Nyko Game Case fo...|
|B000AOE2K|Fire Emblem: Path...|
|B000HBBW7U|Taranus (PC) (Com...|
|B013RADQQQ|Suseystone® 2400D...|
|B00KQKKU2J|FIFA 15 (Ultimate...|
|B006W41X2C|Turtle Beach - Ea...|
|B000KCX9M4|Grand Theft Auto:...|
|B00YT90JNC|Red Wii Mini Cons...|
|B0096KG6A8|Wii U Super Mario...|
|B00L6AVLB0|World of Tanks-X3...|
|B000IMYKQ0|Wii Nunchuk Contr...|
+-----+
only showing top 20 rows
```

Interpreter: spark.pyspark. FINISHED Took 16 sec 488 millisec. Updated by ZGrinacoff on November 08 2019, 4:13:08 PM (PST)



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

+-----+-----+-----+
| review_id | review_headline | review_body |
+-----+-----+-----+
| RTIS3L2M1F5SM | an amazing joysti... | Used this for Eli... |
| R1ZV7R400LHKD | Definitely a sile... | Loved it, I didn... |
| R3BH071QLH8QMC | One Star | poor quality work... |
| R127K9NTXA2YH | good, but could b... | nice, but tend to... |
| R32ZWUXDJPW27Q | Great but flawed. | Great amiibo, gre... |
| R3AQQ4YUKJWBA6 | One Star | The remote consta... |
| R2F0POU5K6F73F | A Must | I have a 2012-201... |
| R3VNR804HYSMR6 | Five Stars | Perfect, kids lov... |
| R3GZTM72WA2QH | Five Stars | Excellent |
| RNQOY62705W1K | Four Stars | Slippery but expe... |
| R1VTIA3JTYBY02 | Five Stars | Love the game. Se... |
| R29DOU8791QZLB | Game will get stuck | Does not fit prop... |
| R15DUT1VIJ9RJZ | We have tried it ... | This was way too ... |
| R3IMF2M0Q3U9ZM | Four Stars | Works great good ... |
| R23H79DH0ZTYAU | Now i have to buy... | It did not fit th... |
| RIV24EQAIXA40 | Five Stars | perfect lightweig... |
| R3UCNGYDVN24YB | Five Stars | great |
| RUL4HAXTTN2DY | Five Stars | Works beautifully... |
| R20JF7Z4DHTNX5 | Five Stars | Kids loved the ga... |
| R2T1AJ5MFI2260 | Four Stars | Goodngame |
+-----+-----+-----+
only showing top 20 rows
```

Interpreter: spark.pyspark. FINISHED Took 161 millisec. Updated by ZGrinacoff on November 08 2019, 4:13:08 PM (PST)



```
%pyspark
# Customers DataFrame
customers_df = video_games_df.groupby("customer_id").agg({"customer_id": "count"}).withColumnRenamed("count(customer_id)", "customer_count")
customers_df.show()

+-----+-----+
|customer_id|customer_count|
+-----+-----+
| 690560 | 2 |
| 45046643 | 4 |
| 47822294 | 1 |
| 740955 | 1 |
| 12536186 | 1 |
| 3700565 | 1 |
| 2663941 | 2 |
| 49352379 | 1 |
| 7770381 | 2 |
| 711009 | 1 |
| 588545 | 1 |
| 7377220 | 4 |
| 24168883 | 2 |
| 35129371 | 6 |
| 16674821 | 1 |
| 8379448 | 2 |
| 1815445 | 1 |
| 24540860 | 2 |
| 26791558 | 3 |
| 36407303 | 1 |
+-----+-----+
only showing top 20 rows
```

Interpreter: spark.pyspark. FINISHED Took 16 sec 987 millisec. Updated by ZGrinacoff on November 08 2019, 4:13:25 PM (PST)



```
%pyspark
# Vine DataFrame
vine_df = video_games_df.select(["review_id", "star_rating", "helpful_votes", "total_votes", "vine"])
vine_df.show()

+-----+-----+-----+-----+-----+
| review_id | star_rating | helpful_votes | total_votes | vine |
+-----+-----+-----+-----+-----+
| RTIS3L2M1F5SM | 5 | 0 | 0 | N |
| R1ZV7R400LHKD | 5 | 0 | 0 | N |
| R3BH071QLH8QMC | 1 | 0 | 1 | N |
| R127K9NTXA2YH | 3 | 0 | 0 | N |
| R32ZWUXDJPW27Q | 4 | 0 | 0 | N |
| R3AQQ4YUKJWBA6 | 1 | 0 | 0 | N |
| R2F0POU5K6F73F | 5 | 0 | 0 | N |
| R3VNR804HYSMR6 | 5 | 0 | 0 | N |
| R3GZTM72WA2QH | 5 | 0 | 0 | N |
| RNQOY62705W1K | 4 | 0 | 0 | N |
| R1VTIA3JTYBY02 | 5 | 0 | 0 | N |
| R2T1AJ5MFI2260 | 1 | 0 | 0 | N |
+-----+-----+-----+-----+-----+
only showing top 20 rows
```

```
|RIV1P391710Z| 5| 0| 0| N|
|R29D0U8791QZL8| 1| 0| 0| N|
|R15DUT1VIJ9RJZ| 2| 0| 0| N|
|R3IMF2M3Q30U9ZM| 4| 0| 0| N|
|R23H79DH0ZTYAU| 1| 1| 1| N|
| RIV24EQAIXA40| 5| 0| 0| N|
|R3UCNGYDVN24YB| 5| 0| 0| N|
| RUL4H4XTTN2DV| 5| 0| 0| N|
|R20JF7Z4DHTNX5| 5| 0| 0| N|
|R2T1AJ5MFI2260| 4| 0| 0| N|
+-----+-----+-----+-----+
only showing top 20 rows
```

Interpreter: spark.pyspark. FINISHED Took 161 millisec. Updated by ZGrinacoff on November 08 2019, 4:13:25 PM (PST)



```
# Push to AWS RDS instance.
```

Interpreter: python. FINISHED Took 58 millisec. Updated by ZGrinacoff on November 08 2019, 4:13:25 PM (PST)



```
%pyspark
mode = "append"
jdbc_url = "jdbc:postgresql://mypostgresdb-hw.cb3cpmpbygch.us-east-1.rds.amazonaws.com:5432/postgres"
config = {"user": "root", "password": "Rangers123!", "driver": "org.postgresql.Driver"}
```

Interpreter: spark.pyspark. FINISHED Took 109 millisec. Updated by ZGrinacoff on November 08 2019, 4:20:10 PM (PST)



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

Interpreter: spark.pyspark. FINISHED Took 10 min 30 sec 4 millisec. Updated by ZGrinacoff on November 08 2019, 4:30:44 PM (PST)



```
%pyspark
# Write products_df to table in RDS
products_df.write.jdbc(url=jdbc_url, table='products', mode=mode, properties=config)
```

Interpreter: spark.pyspark. FINISHED Took 52 sec 285 millisec. Updated by ZGrinacoff on November 08 2019, 4:31:37 PM (PST)



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

Interpreter: spark.pyspark. FINISHED Took 2 min 11 sec 246 millisec. Updated by ZGrinacoff on November 08 2019, 4:33:48 PM (PST)



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

Interpreter: spark.pyspark. FINISHED Took 10 min 0 sec 133 millisec. Updated by ZGrinacoff on November 08 2019, 5:12:06 PM (PST)

