Read in CSV from S3

Interpreter: md. **FINISHED** Took 4 millisec. Updated by Diana on December 16 2019, 12:09:42 AM (EST)

```
%pyspark
from pyspark import SparkFiles
# Load in csv from S3 into a DataFrame
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Beauty_v1_00.tsv.gz"
spark.sparkContext.addFile(url)

df = spark.read.option('header', 'true').csv(SparkFiles.get("amazon_reviews_us_Beauty_v1_00.tsv.gz"), inferSchema=True, sep='\t', timestampFormat="yyyy/mm
    /dd")

df.show(10)
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----
+------+--------------------+--------------------+------------------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_p
urchase|     review_headline|         review_body|       review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----
+------+--------------------+--------------------+------------------+
|         US|    1797882|R3I2DHQBR577SS|B001ANOOOE|       2102612|The Naked Bee Vit...|          Beauty|          5|            0|          0|   N|
Y|          Five Stars|Love this, excell...|2015-08-31 00:00:00|
|         US|   18381298|R1QNE9NQFJC2Y4|B0016J22EQ|     106393691|Alba Botanica Sun...|          Beauty|          5|            0|          0|   N|
Y|Thank you Alba Bo...|The great thing a...|2015-08-31 00:00:00|
|         US|   19242472|R3LIDG2Q4LJBAO|B00HU6UQAG|     375449471|Elysee Infusion S...|          Beauty|          5|            0|          0|   N|
Y|          Five Stars|Great Product, I'...|2015-08-31 00:00:00|
|         US|   19551372|R3KSZHPAEVPEAL|B002HWS7RM|     255651889|Diane D722 Color,...|          Beauty|          5|            0|          0|   N|
Y|          GOOD DEAL!!|I use them as sho...|2015-08-31 00:00:00|
|         US|   14802407| RAI2OIG50KZ43|B00SM99KWU|     116158747|Biore UV Aqua Ric...|          Beauty|          5|            0|          0|   N|
Y|this soaks in qui...|This is my go-to ...|2015-08-31 00:00:00|
|         US|    2909389|R1R30FA4RB5P54|B000NYL1Z6|     166146615|Murad Clarifying ...|          Beauty|          4|            0|          0|   N|
Y|          Four Stars|                Good|2015-08-31 00:00:00|
|         US|   19397215|R30IJKCGJBGPJH|B001SYWTFG|     111742328|CoverGirl Queen C...|          Beauty|          5|            0|          0|   N|
Y|            Good buy|Great eyeliner, d...|2015-08-31 00:00:00|
|         US|    3195210|R18GLJJPVQ1OVH|B005F2EVMQ|     255803087|Bifesta Mandom Ey...|          Beauty|          5|            0|          0|   N|
Y|          Five Stars|Best makeup remover! |2015-08-31 00:00:00|
|         US|   52216383| R8TVYIJXLYJT0|B00M1SUW7K|     246816549|Can You Handlebar...|          Beauty|          5|            0|          0|   N|
Y|Tame the wild mus...|This is a great p...|2015-08-31 00:00:00|
|         US|   10278216|R1CJGF6M3PVHEZ|B001KYQA1S|       9612905|Maybelline Great ...|          Beauty|          1|            0|          2|   N|
Y|but it's like hav...|I thought it woul...|2015-08-31 00:00:00|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----
+------+--------------------+--------------------+------------------+
only showing top 10 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 2 min 43 sec 30 millisec. Updated by Diana on December 16 2019, 12:12:31 AM (EST)

Drop duplicates and incomplete rows

Interpreter: md. **FINISHED** Took 0 millisec. Updated by Diana on December 16 2019, 12:12:31 AM (EST)

```
%pyspark
print(df.count())
df = df.dropna()
print(df.count())
df = df.dropDuplicates()
print(df.count())
```

```
5115666
5114733
5114733
```

Interpreter: spark.pyspark. **FINISHED** Took 3 min 10 sec 456 millisec. Updated by Diana on December 16 2019, 12:15:41 AM (EST)

Examine the schema

Interpreter: md. **FINISHED** Took 1 millisec. Updated by Diana on December 16 2019, 12:15:41 AM (EST)

```
%pyspark
df.printSchema()
```

```
root
 |-- marketplace: string (nullable = true)
 |-- customer_id: integer (nullable = true)
 |-- review_id: string (nullable = true)
 |-- product_id: string (nullable = true)
 |-- product_parent: integer (nullable = true)
 |-- product_title: string (nullable = true)
 |-- product_category: string (nullable = true)
 |-- star_rating: string (nullable = true)
 |-- helpful_votes: integer (nullable = true)
 |-- total_votes: integer (nullable = true)
 |-- vine: string (nullable = true)
 |-- verified_purchase: string (nullable = true)
 |-- review_headline: string (nullable = true)
 |-- review_body: string (nullable = true)
 |-- review_date: timestamp (nullable = true)
```

Interpreter: spark.pyspark. **FINISHED** Took 112 millisec. Updated by Diana on December 16 2019, 12:15:41 AM (EST)

Create a new DataFrame for customers

Interpreter: md. **FINISHED** Took 0 millisec. Updated by Diana on December 16 2019, 12:15:41 AM (EST)

```
%pyspark
df_grouped = df.groupby("customer_id").count()
df_grouped.show(5)
```

```
+-----------+-----+
|customer_id|count|
+-----------+-----+
|   27596904|    4|
|   40338257|    1|
|   13138785|    1|
|   19846587|    2|
|   46800377|    1|
+-----------+-----+
only showing top 5 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 1 min 45 sec 677 millisec. Updated by Diana on December 16 2019, 12:17:27 AM (EST)

```
%pyspark
df_grouped=df_grouped.withColumnRenamed("count", "customer_count")
#df_grouped.show(5)
customers=df_grouped.select(["customer_id", "customer_count"])
print(customers.count())
customer = customers.dropna()
print(customers.count())
customers = customers.dropDuplicates(["customer_id"])
print(customers.count())
customers.show(5)
```

```
2815977
2815977
2815977
+-----------+--------------+
|customer_id|customer_count|
+-----------+--------------+
|   27596904|             4|
|   40338257|             1|
|   13138785|             1|
|   19846587|             2|
|   46800377|             1|
+-----------+--------------+
only showing top 5 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 7 min 0 sec 920 millisec. Updated by Diana on December 16 2019, 12:24:28 AM (EST)

Create a new DataFrame for products

Interpreter: md. **FINISHED** Took 0 millisec. Updated by Diana on December 16 2019, 12:24:28 AM (EST)

```
%pyspark
products = df.select(["product_id", "product_title"])
#products.show(5)
print(products.count())
products = products.dropna()
print(products.count())
products = products.dropDuplicates(["product_id"])
print(products.count())
products.show(5)
```

```
5114733
5114733
588771
+----------+--------------------+
|product_id|       product_title|
+----------+--------------------+
|1935682016|Henna Art, Tools ...|
|9788077757|Overnight Success...|
|9790770367|Dolce & Gabbana C...|
|9790773587|Calvin Klein Euph...|
|9790796927|Herve Leger Perfu...|
+----------+--------------------+
only showing top 5 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 7 min 3 sec 418 millisec. Updated by Diana on December 16 2019, 12:31:31 AM (EST)

Create a new DataFrame for vine_table

Interpreter: md. **FINISHED** Took 0 millisec. Updated by Diana on December 16 2019, 12:31:31 AM (EST)

```
%pyspark
vine_table = df.select(["review_id", "star_rating", "helpful_votes", "total_votes", "vine"])
#vine_table.show(5)
print(vine_table.count())
vine_table= vine_table.dropna()
print(vine_table.count())
vine_table = vine_table.dropDuplicates(["review_id"])
print(vine_table.count())
vine_table.show(5)
```

```
5114733
5114733
5114733
```

```
5114733
+--------------+-----------+-------------+-----------+----+
|     review_id|star_rating|helpful_votes|total_votes|vine|
+--------------+-----------+-------------+-----------+----+
|R1005LNVVR6FJB|          5|            0|          0|   N|
|R100MW88S1EM6X|          5|            0|          0|   N|
| R100NSSD3H3A8|          5|            0|          0|   N|
|R100X2VOGDRUWM|          4|            0|          0|   N|
|R1010AUDGGG8F7|          4|            0|          1|   N|
+--------------+-----------+-------------+-----------+----+
only showing top 5 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 7 min 2 sec 627 millisec. Updated by Diana on December 16 2019, 12:38:34 AM (EST)

Create a new DataFrame for review_id_table

Interpreter: md. **FINISHED** Took 1 millisec. Updated by Diana on December 16 2019, 12:38:34 AM (EST)

```
%pyspark
review_id_table  = df.select(["review_id", "customer_id", "product_id", "product_parent", "review_date"])
#review_id_table.show(5)
print(review_id_table.count())
review_id_table= review_id_table.dropna()
print(review_id_table.count())
review_id_table = review_id_table.dropDuplicates(["review_id"])
print(review_id_table.count())
review_id_table.show(5)
```

```
5114733
5114733
5114733
+--------------+-----------+----------+--------------+-------------------+
|     review_id|customer_id|product_id|product_parent|        review_date|
+--------------+-----------+----------+--------------+-------------------+
|R1005LNVVR6FJB|   28425726|B000C1Z4AI|     253474165|2014-03-29 00:00:00|
|R100MW88S1EM6X|   24144740|B00NZQ78OC|     886562127|2015-05-11 00:00:00|
| R100NSSD3H3A8|   39774520|B00P4G3Y6M|     442709301|2015-01-04 00:00:00|
|R100X2VOGDRUWM|   11896992|B00LGFKA3U|     182792021|2014-10-22 00:00:00|
|R1010AUDGGG8F7|   20522375|B000C1VX5S|     828677506|2014-12-29 00:00:00|
+--------------+-----------+----------+--------------+-------------------+
only showing top 5 rows
```

Interpreter: spark.pyspark. **FINISHED** Took 7 min 4 sec 735 millisec. Updated by Diana on December 16 2019, 12:45:39 AM (EST)

Write DataFrame to RDS

Interpreter: md. **FINISHED** Took 1 millisec. Updated by Diana on December 16 2019, 12:47:18 AM (EST)

```
%pyspark
# Configure settings for RDS
mode = "append"
jdbc_url="jdbc:postgresql://big-data-challenge.ceslmtkac6wc.us-east-2.rds.amazonaws.com:5432/beauty"
config = {"user":"root",
          "password": "postgres",
          "driver":"org.postgresql.Driver"}
```

Interpreter: spark.pyspark. **FINISHED** Took 109 millisec. Updated by Diana on December 16 2019, 12:49:20 AM (EST)

Interpreter: spark.pyspark. **FINISHED** Took 108 millisec. Updated by Diana on December 16 2019, 12:45:39 AM (EST)

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

Interpreter: spark.pyspark. **FINISHED** Took 5 min 52 sec 429 millisec. Updated by Diana on December 16 2019, 12:55:18 AM (EST)

```
%pyspark
# Write DataFrame to table

products.write.jdbc(url=jdbc_url, table='products', mode=mode, properties=config)
```

Interpreter: spark.pyspark. **FINISHED** Took 3 min 12 sec 658 millisec. Updated by Diana on December 16 2019, 12:59:41 AM (EST)

```
%pyspark
# Write DataFrame to table

review_id_table.write.jdbc(url=jdbc_url, table='review_id_table', mode=mode, properties=config)
```

Interpreter: spark.pyspark.

```
%pyspark
# Write DataFrame to table

vine_table.write.jdbc(url=jdbc_url, table='vine_table', mode=mode, properties=config)
```

Interpreter: spark.pyspark.

Interpreter: spark.

Interpreter: spark.