Read in CSV from S3

```
%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_Apparel_v1_00.tsv.gz"
spark.sparkContext.addFile(url)

df = spark.read.option('header', 'true').csv(SparkFiles.get("amazon_reviews_us_Apparel_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|vin
e|verified_purchase|     review_headline|         review_body|       review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----
+----------------+-------------------+-------------------+------------------+
|         US|   32158956|R1KKOXHNI8MSXU|B01KL6O72Y|      24485154|Easy Tool Stainle...|         Apparel|          4|            0|          0|
N|               Y|★ THESE REALLY DO...|These Really Do W...|2013-01-14 00:00:00|
|         US|    2714559|R26SP2OPDK4HT7|B01ID3ZS5W|     363128556|V28 Women Cowl Ne...|         Apparel|          5|            1|          2|
N|               Y|Favorite for wint...|I love this dress...|2014-03-04 00:00:00|
|         US|   12608825| RWQEDYAX373I1|B01I497BGY|     811958549|James Fiallo Men'...|         Apparel|          5|            0|          0|
N|               Y|Great Socks for t...|Nice socks, great...|2015-07-12 00:00:00|
|         US|   25482800|R231YI7R4GPF6J|B01HDXFZK6|     692205728|Belfry Gangster 1...|         Apparel|          5|            0|          0|
N|               Y|          Slick hat!|I bought this for...|2015-06-03 00:00:00|
|         US|    9310286|R3KO3W45DD0L1K|B01G6MBEBY|     431150422|JAEDEN Women's Be...|         Apparel|          5|            0|          0|
N|               Y|I would do it again!|Perfect dress and...|2015-06-12 00:00:00|
|         US|   26631939|R1C4QH63NFL5NJ|B01FWRXN0Y|     366144407|Levi's Boys' 514 ...|         Apparel|          5|            0|          0|
N|               Y|          Five Stars|Excellent for my ...|2014-04-22 00:00:00|
|         US|   48785098|R2GP65O1U9N7BP|B01EXNH1HE|     786052021|Minimalist Wallet...|         Apparel|          5|            0|          0|
N|               Y|            Love it!|Raw is the only w...|2015-07-28 00:00:00|
|         US|   39548589|R3O29CT5MQQ3XQ|B01E7OL09O|     108920964|Harriton Men's Ba...|         Apparel|          4|            0|          0|
N|               Y|         Three Stars|        A bit large.|2015-07-10 00:00:00|
|         US|   29355866|R1ZECD2AA8QFF6|B01DXHX81O|     317132458|Jockey Women's Un...|         Apparel|          5|            0|          0|
N|               Y|          Five Stars|          Great fit!|2015-08-09 00:00:00|
|         US|   27477484|R2S79GCF6J89OA|B01DDULIJK|     110598191|Alexander Del Ros...|         Apparel|          3|            0|          0|
N|               Y|    Not my favorite.|Shirt a bit too l...|2014-05-24 00:00:00|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----
+----------------+-------------------+-------------------+------------------+
only showing top 10 rows
```

Drop duplicates and incomplete rows

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

```
5906333
5905267
5905267
```

Examine the schema

```
%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: integer (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)
```

Create a new DataFrame for customers

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

```
+-----------+-----+
|customer_id|count|
+-----------+-----+
|   17309364|    1|
|    7064405|    1|
|   43447475|    2|
|    2496423|    1|
|   34630521|    3|
+-----------+-----+
only showing top 5 rows
```

```
%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)
```

```
3227834
3227834
3227834
+-----------+--------------+
|customer_id|customer_count|
+-----------+--------------+
|   17309364|             1|
|    7064405|             1|
|   43447475|             2|
|    2496423|             1|
|   34630521|             3|
+-----------+--------------+
only showing top 5 rows
```

Create a new DataFrame for products

```

```

Interpreter: spark.

```
%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)
```

```
5905267
5905267
2305345
+----------+--------------------+
|product_id|       product_title|
+----------+--------------------+
|B00001OGXK|Disguise Austin P...|
|B00006AMFF|Embroidered Stems...|
|B00006B7DV|Convertible Fleec...|
|B00006LZZA|Pinpoint Oxford B...|
|B00006WVZJ|OshKosh Wht/Blue ...|
+----------+--------------------+
only showing top 5 rows
```

```
%pyspark
products.printSchema()
```

```
root
 |-- product_id: string (nullable = true)
 |-- product_title: string (nullable = true)
```

|-- product_title: string (nullable = true)

Create a new DataFrame for vine table

```
%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)
```

```
5905267
5905267
5905267
+--------------+-----------+-------------+-----------+----+
|     review_id|star_rating|helpful_votes|total_votes|vine|
+--------------+-----------+-------------+-----------+----+
|R100ASCXUA4V0W|          5|            0|          0|   N|
|R100F54NKSHMAK|          5|            1|          1|   N|
|R100LQY8PNKPJ1|          3|            0|          0|   N|
|R100RHPNR94Z59|          5|            1|          1|   N|
|R100WXCZO6HX6M|          4|            0|          0|   N|
+--------------+-----------+-------------+-----------+----+
only showing top 5 rows
```

Create a new DataFrame for review_id_table

```
%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)
```

```
5905267
5905267
5905267
+--------------+-----------+----------+--------------+-------------------+
|     review_id|customer_id|product_id|product_parent|        review_date|
+--------------+-----------+----------+--------------+-------------------+
|R100ASCXUA4V0W|    1378574|B00HG6ZSCQ|     902112639|2014-12-29 00:00:00|
|R100F54NKSHMAK|   48586399|B0075FTSN8|     950472306|2013-10-01 00:00:00|
|R100LQY8PNKPJ1|   15367205|B00GM4B3GS|     512948810|2014-01-19 00:00:00|
|R100RHPNR94Z59|   15865080|B001HQRA1I|     263493850|2013-06-21 00:00:00|
|R100WXCZO6HX6M|   43391712|B00MY5UEZU|     775565320|2015-08-01 00:00:00|
+--------------+-----------+----------+--------------+-------------------+
only showing top 5 rows
```

Write DataFrame to RDS

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

```
%pyspark
# Write DataFrame to table

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

```
%pyspark
```

```
%pyspark
# Write DataFrame to table

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

```
%pyspark
# Write DataFrame to table

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

```
%pyspark
# Write DataFrame to table

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