I. ETL Amazon review data of Software

Interpreter: md. FINISHED Took 1 millisec. Updated by undefined on January 24 2020, 2:07:58 PM (EST)

Load Amazon Data into Spark DataFrame

Interpreter: md. FINISHED Took 0 millisec. Updated by undefined on January 24 2020, 1:21:37 PM (EST)

```
%pyspark
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Software_v1_00.tsv.gz"
spark.sparkContext.addFile(url)
Software_df = spark.read.option('header', 'true').csv(SparkFiles.get("amazon_reviews_us_Software_v1_00.tsv.gz"), sep="\t", header=True, inferSchema=True)
```

Interpreter: spark.pyspark. FINISHED Took 32 sec 970 millisec. Updated by undefined on January 24 2020, 12:23:16 PM (EST)

Check data

Interpreter: md. FINISHED Took 1 millisec. Updated by undefined on January 24 2020, 1:10:35 PM (EST)

```
%pyspark
Software_df.show(1)  # show one record to check data type of each field
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_headline|
review_body|       review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+
|         US|   42605767|R3EFW2STIYIY0I|B00MUTIDKI|     248732228|McAfee 2015 Inter...|        Software|          1|            2|          2|   N|                Y|I was very disapp...|
I was very disapp...|2015-08-31 00:00:00|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+-------------+-----------+----+-----------------+--------------------+
only showing top 1 row
```

Interpreter: spark.pyspark. FINISHED Took 267 millisec. Updated by undefined on January 24 2020, 12:23:28 PM (EST)

```
%pyspark
Software_df.count()
```

```
341931
```

Interpreter: spark.pyspark. FINISHED Took 3 sec 520 millisec. Updated by undefined on January 24 2020, 12:23:43 PM (EST)

```
%pyspark
len(Software_df.columns)
```

```
15
```

Interpreter: spark.pyspark. FINISHED Took 111 millisec. Updated by undefined on January 24 2020, 12:23:56 PM (EST)

```
%pyspark
Software_df.columns
```

```
['marketplace',
 'customer_id',
 'review_id',
 'product_id',
 'product_parent',
 'product_title',
 'product_category',
 'star_rating',
 'helpful_votes',
 'total_votes',
 'vine',
 'verified_purchase',
 'review_headline',
 'review_body',
 'review_date']
```

Interpreter: spark.pyspark. FINISHED Took 115 millisec. Updated by undefined on January 24 2020, 12:24:05 PM (EST)

```
%pyspark
Software_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)
```

Interpreter: spark.pyspark. FINISHED Took 113 millisec. Updated by undefined on January 24 2020, 12:24:24 PM (EST)

```
%pyspark
Software_df.cache()
```

```
DataFrame[marketplace: string, customer_id: int, review_id: string, product_id: string, product_parent: int, product_title: string, product_category: string, star_rating: int, helpful_vo
tes: int, total_votes: int, vine: string, verified_purchase: string, review_headline: string, review_body: string, review_date: timestamp]
```

Interpreter: spark.pyspark. FINISHED Took 110 millisec. Updated by undefined on January 24 2020, 12:24:32 PM (EST)

Interpreter: spark.pyspark. FINISHED Took 110 millisec. Updated by undefined on January 24 2020, 12:24:32 PM (EST)

```
%pyspark
Software_df.select(["vine","verified_purchase"]).describe().show()
```

```
+-------+------+-----------------+
|summary|  vine|verified_purchase|
+-------+------+-----------------+
|  count|341931|           341931|
|   mean|  null|             null|
| stddev|  null|             null|
|    min|     N|                N|
|    max|     Y|                Y|
+-------+------+-----------------+
```

Interpreter: spark.pyspark. FINISHED Took 12 sec 334 millisec. Updated by undefined on January 24 2020, 12:25:32 PM (EST)

```
%pyspark
Software_df.select("verified_purchase").filter("verified_purchase==null").count()
```

```
0
```

Interpreter: spark.pyspark. FINISHED Took 161 millisec. Updated by undefined on January 24 2020, 12:27:15 PM (EST)

```
%pyspark
Software_df.select("verified_purchase").filter("verified_purchase=='Y'").count()
```

```
195647
```

Interpreter: spark.pyspark. FINISHED Took 162 millisec. Updated by undefined on January 24 2020, 12:27:19 PM (EST)

Cleaned up DataFrames to match DB tables

Interpreter: md. FINISHED Took 1 millisec. Updated by undefined on January 24 2020, 1:10:19 PM (EST)

```
%pyspark
products_df = Software_df.select(["product_id", "product_title"]).drop_duplicates()
customers_df = Software_df.groupby("customer_id").agg({"customer_id": "count"}).withColumnRenamed("count(customer_id)", "customer_count")
from pyspark.sql.functions import to_date
reviews_df = Software_df.select(["review_id", "customer_id", "product_id", "product_parent", to_date("review_date").alias("review_date")])
vine_df = Software_df.select(["review_id", "star_rating", "helpful_votes", "total_votes", "vine"])
vine_df.show(10)
```

```
+--------------+-----------+-------------+-----------+----+
|     review_id|star_rating|helpful_votes|total_votes|vine|
+--------------+-----------+-------------+-----------+----+
|R3EFW2STIYIY0I|          1|            2|          2|   N|
|R12NR0R5A9F7FT|          5|            0|          0|   N|
|R1LSH74R9XAP59|          2|            0|          1|   N|
|R1QXUNTF76K7L6|          2|            0|          0|   N|
|R2F7DR75PS8NKT|          5|            0|          0|   N|
|R2C1DJSCC8UFS6|          3|            0|          0|   N|
|R1AXGS1W4YFXMX|          1|            0|          2|   N|
|R1XU1B93402SYJ|          1|            1|          1|   N|
|R2U432NB3OPVR0|          5|            0|          0|   N|
|R3R6FIMIOQ5SP9|          5|            0|          0|   N|
+--------------+-----------+-------------+-----------+----+
only showing top 10 rows
```

Interpreter: spark.pyspark. FINISHED Took 264 millisec. Updated by undefined on January 24 2020, 12:28:51 PM (EST)

Push to AWS RDS instance

Interpreter: md. FINISHED Took 2 millisec. Updated by undefined on January 24 2020, 1:18:29 PM (EST)

```
%pyspark
mode = "append"
jdbc_url="jdbc:postgresql://<EndPoint>:5432/<DBName>"
config = {"user":"root", "password": "<PW>", "driver":"org.postgresql.Driver"}
```

Interpreter: spark.pyspark. FINISHED Took 109 millisec. Updated by undefined on January 24 2020, 1:37:48 PM (EST) (outdated)

```
%pyspark
products_df.write.jdbc(url=jdbc_url, table='products', mode=mode, properties=config)
customers_df.write.jdbc(url=jdbc_url, table='customers', mode=mode, properties=config)
reviews_df.write.jdbc(url=jdbc_url, table='reviews', mode=mode, properties=config)
vine_df.write.jdbc(url=jdbc_url, table='vine', mode=mode, properties=config)
```

Interpreter: spark.pyspark. FINISHED Took 3 min 36 sec 447 millisec. Updated by undefined on January 24 2020, 1:49:49 PM (EST) (outdated)

II. Analyze Wine Program

Interpreter: md. FINISHED Took 0 millisec. Updated by undefined on January 24 2020, 1:23:29 PM (EST)

Filter by votes

Interpreter: md. FINISHED Took 0 millisec. Updated by undefined on January 24 2020, 1:19:05 PM (EST)

```
%pyspark
vine_df = Software_df.select(["star_rating", "helpful_votes", "total_votes", "vine", "verified_purchase"])
vine_filtertotalvotes_df = vine_df.filter(vine_df["total_votes"] >= 20)
vine_filtertotalvoteshelpfulvotes_df = vine_filtertotalvotes_df.filter(vine_filtertotalvotes_df["helpful_votes"]/vine_filtertotalvotes_df["total_votes"] >= 0.5)
```

Interpreter: spark.pyspark. FINISHED Took 160 millisec. Updated by undefined on January 24 2020, 1:54:23 PM (EST)

Describe Stats

Interpreter: md. FINISHED Took 0 millisec. Updated by undefined on January 24 2020, 1:19:38 PM (EST)

```
%pyspark
from pyspark.sql.functions import col, avg
paid_df = vine_filtertotalvoteshelpfulvotes_df.filter(vine_filtertotalvoteshelpfulvotes_df['vine']== 'Y')
unpaid_df = vine_filtertotalvoteshelpfulvotes_df.filter(vine_filtertotalvoteshelpfulvotes_df['vine']== 'N')

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

```
+-------+------------------+------------------+------------------+----+-----------------+
|summary|       star_rating|     helpful_votes|       total_votes|vine|verified_purchase|
+-------+------------------+------------------+------------------+----+-----------------+
|  count|               248|               248|               248| 248|              248|
|   mean|3.7943548387096775| 77.65725806451613| 81.91129032258064|null|             null|
| stddev| 1.304790452368455|142.24734497903623|145.08341937129128|null|             null|
|    min|                 1|                 15|                20|   Y|                N|
|    max|                 5|              1231|              1247|   Y|                Y|
+-------+------------------+------------------+------------------+----+-----------------+


+-------+------------------+-----------------+-----------------+-----+-----------------+
|summary|       star_rating|     helpful_votes|      total_votes| vine|verified_purchase|
+-------+------------------+-----------------+-----------------+-----+-----------------+
|  count|             17514|            17514|            17514|17514|            17514|
|   mean|2.8756423432682428|46.32185680027407|51.29542080621217| null|             null|
| stddev|1.6981277376160198|68.69833838338252|72.36000784694947| null|             null|
|    min|                 1|               10|               20|    N|                N|
|    max|                 5|             2243|             2394|    N|                Y|
+-------+------------------+-----------------+-----------------+-----+-----------------+
```

Interpreter: spark.pyspark. FINISHED Took 862 millisec. Updated by undefined on January 24 2020, 1:55:41 PM (EST)

Determine the percentage of five-star reviews among Vine reviews

Interpreter: md. FINISHED Took 1 millisec. Updated by undefined on January 24 2020, 1:20:12 PM (EST)

```
%pyspark
paid_five_star_number = paid_df[paid_df['star_rating']== 5].count()
paid_number = paid_df.count()
percentage_five_star_vine = float(paid_five_star_number) / float(paid_number)
print("paid_number",paid_number)
print("paid_five_star_number",paid_five_star_number)
print("percentage_five_star_vine",percentage_five_star_vine)
```

```
('paid_number', 248)
('paid_five_star_number', 102)
('percentage_five_star_vine', 0.4112903225806452)
```

Interpreter: spark.pyspark. FINISHED Took 360 millisec. Updated by undefined on January 24 2020, 1:55:51 PM (EST)

Determine the percentage of five-star reviews among non-Vine reviews

Interpreter: md. FINISHED Took 0 millisec. Updated by undefined on January 24 2020, 1:20:42 PM (EST)