Amazon_Watches_Reviews_EDA

FINISHED

settings

Took 59 sec. Last updated by anonymous at March 26 2019, 11:48:03 PM. (outdated)

SPARK JOB (http://ip-172-31-50-150.ec2.internal:4040/jobs/job?id=0) FINISHED

settings

Took 20 sec. Last updated by anonymous at March 26 2019, 11:53:47 PM. (outdated)

FINISHED

```
%sql
CREATE EXTERNAL TABLE IF NOT EXISTS amzn (
  marketplace string,
  customer_id string,
  review_id string,
  product_id string,
  product_parent string,
  product_title string,
  star_rating int,
  helpful_votes int,
  total_votes int,
  vine string,
  verified_purchase string,
  review_headline string,
  review_body string,
  review_date bigint,
  year int
)
PARTITIONED BY (product_category string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n'
STORED AS TEXTFILE
LOCATION 'hdfs:///user/hive/warehouse/amzn/'
```

settings

Amazon_Watches_Reviews_EDA

Took 1 sec. Last updated by anonymous at March 26 2019, 11:58:28 PM. (outdated)

```
%sql
MSCK REPAIR TABLE amzn
```

FINISHED

settings ▾

Took 1 sec. Last updated by anonymous at March 26 2019, 11:59:11 PM. (outdated)

```
%pyspark
spark.sql("""SET spark.hadoop.mapred.output.compress = true""")
spark.sql("""SET spark.hadoop.mapred.output.compression.codec = true""")
spark.sql("""SET spark.hadoop.mapred.output.compression.codec = org.apache.hadoop.io.compress.GzipCodec""")
spark.sql("""SET spark.hadoop.mapred.output.compression.type = BLOCK""")

spark.sql("""SET mapred.output.compress = true""")
spark.sql("""SET hive.exec.compress.output = true""")
spark.sql("""SET mapred.output.compression.codec = org.apache.hadoop.io.compress.GzipCodec""")
spark.sql("""SET io.compression.codecs = org.apache.hadoop.io.compress.GzipCodec""")

spark.sql("""SET spark.sql.shuffle.partitions = 500""")
spark.sql("""SET spark.yarn.executor.memoryOverhead = 4096""")
spark.sql("""SET hive.exec.dynamic.partition.mode=nonstrict""")
```

FINISHED

```
DataFrame[key: string, value: string]
```

Took 2 sec. Last updated by anonymous at March 27 2019, 12:11:54 AM.

SPARK JOB (http://ip-172-31-50-150.ec2.internal:4040/jobs/job?id=1) ERROR

```
%sql
INSERT OVERWRITE TABLE amzn PARTITION(product_category)
SELECT
    marketplace,
    customer_id,
    review_id,
    product_category,
    product_id,
    product_parent,
    product_title,
    star_rating,
    helpful_votes,
    total_votes,
    vine,
    verified_purchase,
    review_headline,
    review_body,
    review_date,
    year
FROM
    amazon_reviews_parquet
WHERE
    year = 2015
    AND product_category = 'Watches'
```

```
java.lang.UnsupportedOperationException: org.apache.parquet.column.values.dictionary.PlainValuesDictionary$PlainIntegerDictionary
        at org.apache.parquet.column.Dictionary.decodeToLong(Dictionary.java:49)
        at org.apache.spark.sql.execution.datasources.parquet.ParquetDictionary.decodeToLong(ParquetDictionary.java:36)
        at org.apache.spark.sql.execution.vectorized.OnHeapColumnVector.getLong(OnHeapColumnVector.java:364)
        at org.apache.spark.sql.catalyst.expressions.GeneratedClass$GeneratedIteratorForCodegenStage1.processNext(Unknown Source)
        at org.apache.spark.sql.execution.BufferedRowIterator.hasNext(BufferedRowIterator.java:43)
        at org.apache.spark.sql.execution.WholeStageCodegenExec$$anonfun$11$$anon$1.hasNext(WholeStageCodegenExec.scala:619)
        at org.apache.spark.sql.execution.UnsafeExternalRowSorter.sort(UnsafeExternalRowSorter.java:216)
        at org.apache.spark.sql.execution.SortExec$$anonfun$1.apply(SortExec.scala:108)
        at org.apache.spark.sql.execution.SortExec$$anonfun$1.apply(SortExec.scala:101)
        at org.apache.spark.rdd.RDD$$anonfun$mapPartitionsInternal$1$$anonfun$apply$24.apply(RDD.scala:836)
        at org.apache.spark.rdd.RDD$$anonfun$mapPartitionsInternal$1$$anonfun$apply$24.apply(RDD.scala:836)
        at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:52)
        at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:324)
```

Took 21 sec. Last updated by anonymous at March 27 2019, 12:12:18 AM.

```
%sql
```

READY