In [1]:
```
spark
```

```
Starting Spark application
```

ID	YARN Application ID	Kind	State	
0	application_1586188399015_0010	pyspark3	idle	Link (http:/63.ec2.internal:20888/proxy/application_158618839

```
SparkSession available as 'spark'.
<pyspark.sql.session.SparkSession object at 0x7f2659c0ab00>
```

In [2]:
```
from pyspark.sql.functions import *
from pyspark.sql.types import *
from pyspark.sql import *
```

In [3]:
```
spark.sql("select * from amazon_review.amazon_reviews_parquet limit 10").show(
n=100)
```

```
+-----------+-----------+--------------+----------+--------------+-----------
---------+-----------+-------------+-----------+----+-----------------+------
--------------+--------------------+-----------+----+----------------+
|marketplace|customer_id|     review_id|product_id|product_parent|       prod
uct_title|star_rating|helpful_votes|total_votes|vine|verified_purchase|     r
eview_headline|         review_body|review_date|year|product_category|
+-----------+-----------+--------------+----------+--------------+-----------
---------+-----------+-------------+-----------+----+-----------------+------
--------------+--------------------+-----------+----+----------------+
|         FR|   17853693|R3RYA47SSO16FD|B000089RV6|     473856475|Back in Bla
ck [Re...|          5|            0|          0|   N|                Y|
cd|livraison rapide ...| 2014-11-17|2014|           Music|
|         US|   14365559|R2EBKDCG4II0PI|B007I2BZIE|     285538061|    Clockwo
rk Angels|          3|            4|         12|   N|                Y|Latest
"meh" from...|This is a revisio...| 2012-06-29|2012|           Music|
|         FR|   27180553|R1P3RTS2ZTWIZJ|B00LW2KAFU|     851391217|Haendel : H
eroes ...|          5|            1|          2|   N|                Y|
magnifique|Nathalie Stutzman...| 2014-11-17|2014|           Music|
|         US|   49896047|R3QS4H9JAYCUE8|B005VR95M6|     943048926|Smooth Jazz
Hits:...|          4|            3|          3|   N|                Y|      C
alm your mind|Start gently with...| 2012-06-29|2012|           Music|
|         FR|   28449966|R3CIFFMNB9VKIO|B000002L61|     493231291|The Last In
Line ...|          5|            1|          1|   N|                N|
Un monument !|1984 second album...| 2014-11-17|2014|           Music|
|         US|   22542549|R3JM937EMOATQC|B000BP57BG|     333627898|A Light in
the Trees|          5|            1|          3|   N|                Y|
Amazing album|What can I say? B...| 2012-06-29|2012|           Music|
|         FR|   31154834|R1EMP2BCNX36S2|B00EMKV8A0|     233404951|Division Be
ll [Im...|          4|            0|          0|   N|                Y|
A vous de voir|Je le conseil à t...| 2014-11-17|2014|           Music|
|         US|   35027835|R3TUJ2S1UWTLN7|B000002W29|     515826178|Side By Sid
e By S...|          5|            0|          1|   N|                Y|
Perfect|I gave it as a gi...| 2012-06-29|2012|           Music|
|         FR|   37994294|R1VX091DL2E5GD|B00IWS73PW|     437944665|Corazon [Im
port a...|          5|            1|          3|   N|                Y|
Pur Jus !|Aucune surprise, ...| 2014-11-17|2014|           Music|
|         US|   12153181| ROLUFWYIEFE22|B007O3QC8K|     696357030|Bruckner: S
ymphon...|          5|            7|          9|   N|                N|
Finally|I think this is t...| 2012-06-29|2012|           Music|
+-----------+-----------+--------------+----------+--------------+-----------
---------+-----------+-------------+-----------+----+-----------------+------
--------------+--------------------+-----------+----+----------------+
```

In [4]:
```
df = spark.sql("select * from amazon_review.amazon_reviews_parquet limit 10")
```

In [5]:
```
df.printSchema()
```

```
root
 |-- marketplace: string (nullable = true)
 |-- customer_id: string (nullable = true)
 |-- review_id: string (nullable = true)
 |-- product_id: string (nullable = true)
 |-- product_parent: string (nullable = true)
 |-- product_title: 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: date (nullable = true)
 |-- year: integer (nullable = true)
 |-- product_category: string (nullable = true)
```

In [8]:
```
spark.sql("""select product_category,sum(COALESCE(helpful_votes,CAST(0 AS BIGI
NT)))
             from amazon_review.amazon_reviews_parquet
             where year>2004
             group by product_category
             order by product_category""").show(n=200)
```

```
+--------------------+--------------------------------------------------
--------+
|    product_category|sum(coalesce(CAST(helpful_votes AS BIGINT), CAST(0 AS B
IGINT)))|
+--------------------+--------------------------------------------------
--------+
|          Automotive|
3650619|
|Digital_Music_Pur...|
1032942|
| Digital_Video_Games|
215645|
|               Music|
9929328|
|              Sports|
7167820|
|                Toys|
6680296|
|         Video_Games|
3242435|
|            Wireless|
7934542|
+--------------------+--------------------------------------------------
--------+
```

In [10]:
```
spark.sql("""select product_category,year, sum(COALESCE(helpful_votes,CAST(0 A
S BIGINT))) as h_vote
              from amazon_review.amazon_reviews_parquet
              where year>2004
              group by product_category,year
              order by year,product_category
""").show(n=200)
```

```
+--------------------+----+-------+
|    product_category|year| h_vote|
+--------------------+----+-------+
|          Automotive|2005|   7844|
|Digital_Music_Pur...|2005|      3|
|               Music|2005|1681938|
|              Sports|2005|  47415|
|                Toys|2005| 263707|
|         Video_Games|2005| 178706|
|            Wireless|2005| 119956|
|          Automotive|2006|  25684|
|Digital_Music_Pur...|2006|     19|
| Digital_Video_Games|2006|      4|
|               Music|2006|1283713|
|              Sports|2006|  96116|
|                Toys|2006| 166478|
|         Video_Games|2006| 176944|
|            Wireless|2006| 154486|
|          Automotive|2007|  60132|
|Digital_Music_Pur...|2007|   4541|
|               Music|2007|1046390|
|              Sports|2007| 189238|
|                Toys|2007| 263156|
|         Video_Games|2007| 211143|
|            Wireless|2007| 197830|
|          Automotive|2008|  88546|
|Digital_Music_Pur...|2008|  44861|
| Digital_Video_Games|2008|     40|
|               Music|2008| 786220|
|              Sports|2008| 236770|
|                Toys|2008| 288045|
|         Video_Games|2008| 309154|
|            Wireless|2008| 203288|
|          Automotive|2009| 174620|
|Digital_Music_Pur...|2009|  78328|
| Digital_Video_Games|2009|   7782|
|               Music|2009| 791137|
|              Sports|2009| 386494|
|                Toys|2009| 379314|
|         Video_Games|2009| 286432|
|            Wireless|2009| 295974|
|          Automotive|2010| 229801|
|Digital_Music_Pur...|2010|  90914|
| Digital_Video_Games|2010|  10842|
|               Music|2010| 732403|
|              Sports|2010| 578909|
|                Toys|2010| 553568|
|         Video_Games|2010| 366979|
|            Wireless|2010| 432521|
|          Automotive|2011| 337297|
|Digital_Music_Pur...|2011| 133307|
| Digital_Video_Games|2011|  25536|
|               Music|2011| 763361|
|              Sports|2011| 817885|
|                Toys|2011| 703791|
|         Video_Games|2011| 352224|
|            Wireless|2011| 678208|
```

```
|          Automotive|2012| 455363|
|Digital_Music_Pur...|2012| 177468|
| Digital_Video_Games|2012|  40238|
|               Music|2012| 807251|
|              Sports|2012| 995098|
|                Toys|2012| 807897|
|         Video_Games|2012| 362084|
|            Wireless|2012| 984852|
|          Automotive|2013| 765769|
|Digital_Music_Pur...|2013| 233563|
| Digital_Video_Games|2013|  82081|
|               Music|2013| 891335|
|              Sports|2013|1359604|
|                Toys|2013|1191116|
|         Video_Games|2013| 433142|
|            Wireless|2013|1433440|
|          Automotive|2014| 808158|
|Digital_Music_Pur...|2014| 185545|
| Digital_Video_Games|2014|  34293|
|               Music|2014| 767337|
|              Sports|2014|1339794|
|                Toys|2014|1209069|
|         Video_Games|2014| 373430|
|            Wireless|2014|1875787|
|          Automotive|2015| 697405|
|Digital_Music_Pur...|2015|  84393|
| Digital_Video_Games|2015|  14829|
|               Music|2015| 378243|
|              Sports|2015|1120497|
|                Toys|2015| 854155|
|         Video_Games|2015| 192197|
|            Wireless|2015|1558200|
+--------------------+----+-------+
```

In [13]:
```
spark.sql("""select product_category,count(*)
    from amazon_review.amazon_reviews_parquet
    where year>2004
    group by product_category
    order by product_category
""").show(n=1000)
```

```
+--------------------+--------+
|    product_category|count(1)|
+--------------------+--------+
|          Automotive| 3515957|
|Digital_Music_Pur...| 1852161|
| Digital_Video_Games|  145431|
|               Music| 4478165|
|              Sports| 4856368|
|                Toys| 4912113|
|         Video_Games| 1627232|
|            Wireless| 9026949|
+--------------------+--------+
```

In [ ]:
```

```