Sparkify 6

```
%pyspark
sc
```

FINISHED

```
<SparkContext master=yarn appName=Zeppelin>
```

Took 32 sec. Last updated by anonymous at April 22 2020, 12:18:12 PM.

```
%pyspark
import json
log = sc.textFile("s3a://snell2-spark/input")
dat=log.collect()
```

FINISHED

Took 5 sec. Last updated by anonymous at April 22 2020, 12:18:20 PM. (outdated)

```
%pyspark
for dic in dat:
    data=json.loads(dic)
    #print(data['song'])
```

FINISHED

Took 0 sec. Last updated by anonymous at April 22 2020, 2:48:20 PM. (outdated)

```
%pyspark

#mapper code
s_lst=[]
a_lst=[]
for dic in dat:
    data=json.loads(dic)
    s=data['song']
    a=data['artist']
    s_lst.append([s,1])
    a_lst.append([a,1])
```

FINISHED

Took 0 sec. Last updated by anonymous at April 22 2020, 1:50:45 PM. (outdated)

```
%pyspark
#reducer code
def reducer(lst):
    dict={}
    list=[]
    for l in lst:
        if l[0] not in dict:
            dict[l[0]]=1
        else: dict[l[0]]+=1

    for key,values in dict.items():
        if key == None:
            continue
        else:
            list.append([key,values])
```

FINISHED

Sparkify 6

```
    list = sorted (list, key=lambda x:x[1],reverse=True)
    #list.pop(0)
    return list

song=[]
artist=[]
artist=reducer(a_lst)
song=reducer(s_lst)
```

Took 0 sec. Last updated by anonymous at April 22 2020, 2:05:11 PM. (outdated)

FINISHED

```
%pyspark
artist[:20]
```

```
[[u'Coldplay', 58], [u'Kings Of Leon', 55], [u'Dwight Yoakam', 38], [u'The Black Keys', 36], [u'Muse',
35], [u'Florence + The Machine', 35], [u'Jack Johnson', 35], [u'Bj\xc3\x83\xc2\xb6rk', 33], [u'John Ma
yer', 31], [u'The Killers', 31], [u'Radiohead', 30], [u'OneRepublic', 30], [u'Alliance Ethnik', 30],
[u'Justin Bieber', 29], [u'Linkin Park', 28], [u'Metallica', 27], [u'Eminem', 26], [u'Train', 25],
[u'Taylor Swift', 23], [u'Beyonc\xc3\x83\xc2\xa9', 23]]
```

Took 0 sec. Last updated by anonymous at April 22 2020, 4:12:59 PM.

FINISHED

```
%pyspark
song[:20]
```

```
[[u"You're The One", 37], [u'Undo', 28], [u'Revelry', 27], [u'Sehr kosmisch', 21], [u'Horn Concerto N
o. 4 in E flat K495: II. Romance (Andante cantabile)', 19], [u'Canada', 17], [u'Secrets', 17], [u'Dog
Days Are Over (Radio Edit)', 16], [u'Invalid', 14], [u'Fireflies', 14], [u'Repr\xc3\x83\xc2\xa9sente',
14], [u'Home', 13], [u'Somebody To Love', 13], [u'Hey_ Soul Sister', 12], [u'Yellow', 12], [u'Sincerit
\xc3\x83\xc2\xa9 Et Jalousie', 11], [u'Pursuit Of Happiness (nightmare)', 10], [u'The Gift', 10], [u'G
ears', 10], [u'OMG', 10]]
```

Took 0 sec. Last updated by anonymous at April 22 2020, 4:13:03 PM.

FINISHED

```
%pyspark
from pyspark.sql.types import *
sch = StructType([StructField("Songs", StringType()),StructField("Count", IntegerType())])
song_df = spark.createDataFrame(song,schema=sch)

sch1 = StructType([StructField("Artists", StringType()),StructField("Count", IntegerType())])
artist_df = spark.createDataFrame(artist,schema=sch1)
```

Took 0 sec. Last updated by anonymous at April 22 2020, 3:49:46 PM. (outdated)

FINISHED

```
%pyspark
song_df.show()
```

```
+--------------------+-----+
|               Songs|Count|
+--------------------+-----+
|      You're The One|   37|
|                Undo|   28|
|             Revelry|   27|
|       Sehr kosmisch|   21|
|Horn Concerto No....|   19|
```

Sparkify 6

```
|              Canada|   17|
|             Secrets|   17|
|Dog Days Are Over...|   16|
|             Invalid|   14|
|           Fireflies|   14|
|      ReprÃ□Â©sente|   14|
|                Home|   13|
|    Somebody To Love|   13|
|    Hey_ Soul Sister|   12|
```

Took 0 sec. Last updated by anonymous at April 22 2020, 3:49:29 PM.

```
%pyspark
artist_df.show()
```

FINISHED

```
|Florence + The Ma...|   35|
|        Jack Johnson|   35|
|            BjÃ□Â¶rk|   33|
|          John Mayer|   31|
|         The Killers|   31|
|           Radiohead|   30|
|         OneRepublic|   30|
|     Alliance Ethnik|   30|
|       Justin Bieber|   29|
|         Linkin Park|   28|
|           Metallica|   27|
|              Eminem|   26|
|               Train|   25|
|        Taylor Swift|   23|
|          BeyoncÃ□Â©|   23|
+--------------------+-----+
only showing top 20 rows
```

Took 0 sec. Last updated by anonymous at April 22 2020, 3:49:48 PM.

```
%pyspark
song_df.write.save("s3://snell2-spark/output/songs")
```

FINISHED

Took 1 sec. Last updated by anonymous at April 22 2020, 3:50:33 PM.

```
%pyspark
artist_df.write.save("s3://snell2-spark/output/artist")
```

FINISHED

Took 1 sec. Last updated by anonymous at April 22 2020, 3:50:35 PM.