test

```
%spark.pyspark
input_bucket = "s3a://dsci6007yshah/"

#you can use path as s3a://dsci6007yshah/2018-11-1*-events.json to use only subset
obj = spark.read.json(input_bucket)
```

☰ SPARK JOB (http://ip-172-31-21-103.ec2.internal:4040/jobs/job?id=16) FINISHED

Took 9 sec. Last updated by anonymous at April 14 2022, 9:41:42 AM. (outdated)

```
%spark.pyspark
#Top artists


# Using .map to make key-value pairs of (artist name, 1) or (None, 0) to avoid in counting
# Using .reduceByKey to aggregate all (artist name, 1) to (artist name, count)
counter = obj.rdd.map(lambda event : (event["artist"], 1) if event["artist"] else (None, 0)).\
          reduceByKey(lambda a,b:a +b)


# Sorting the key-values by value (i.e. count) in descending order
# Picking to top 10
counter = counter.sortBy(lambda a : -a[1]).take(10)

print('\nThe top 10 artists are:\nArtist\t\t\tCount\n')
for (artist, count) in counter:
    print(artist,'\t\t',count)
```

☰ SPARK JOB FINISHED

```
The top 10 artists are:
Artist                  Count


The Black Keys           12
Jack Johnson             11
Dwight Yoakam            10
Coldplay                 10
The Killers              9
```

```
Muse             9
Eminem           8
Linkin Park              8
Radiohead                8
Florence + The Machine           8
```

test

Took 4 sec. Last updated by anonymous at April 14 2022, 9:24:43 AM. (outdated)

SPARK JOB FINISHED

```
%spark.pyspark
#Top songs


#Following same 4 steps of map, reduceByKey, sort, and take as for top artists
counter = obj.rdd.map(lambda event : (event["song"], 1) if event["song"] else (None, 0)).\
          reduceByKey(lambda a,b:a + b)


print('\nThe top 10 songs are:\nSong\t\t\tCount\n')
for (song, count) in counter:
    print(song,'\t\t', count)
```

```
The top 10 songs are:
Song                    Count


You're The One           10
Undo             6
Invalid                  5
Bring Me To Life                 5
Secrets                  5
Horn Concerto No. 4 in E flat K495: II. Romance (Andante cantabile)              5
ReprÃ Â©sente            5
I CAN'T GET STARTED              4
Revelry                  4
Sehr kosmisch            4
```

Took 11 sec. Last updated by anonymous at April 14 2022, 9:32:00 AM.