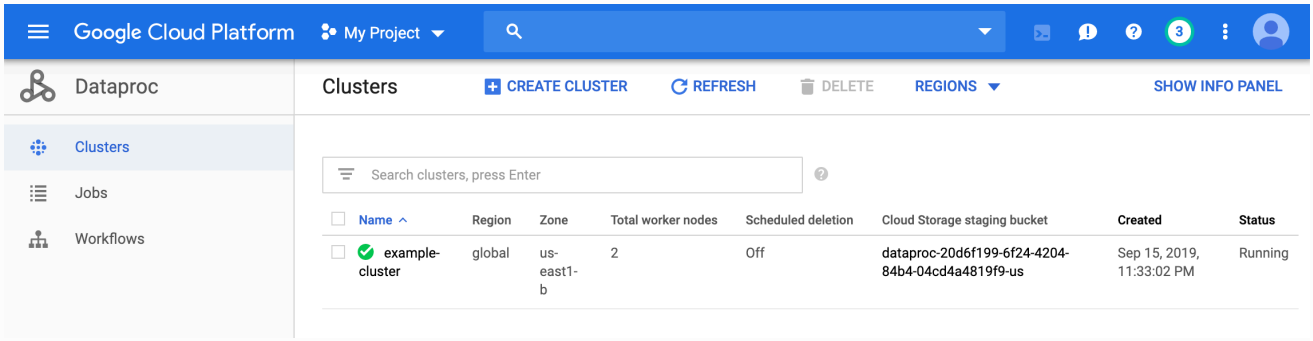
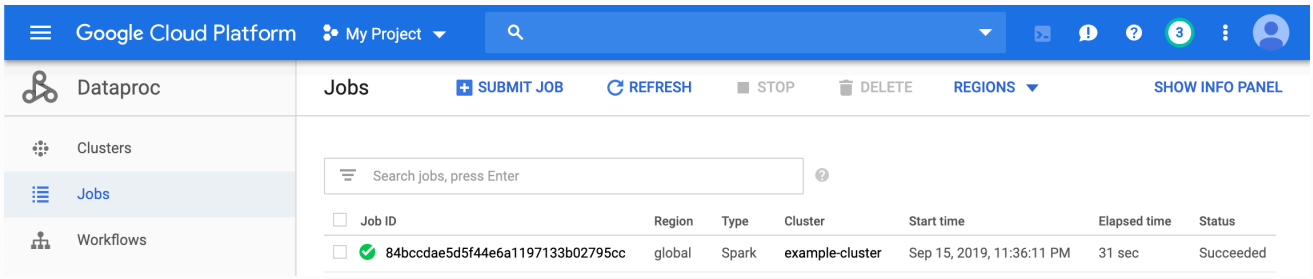
# HOMEWORK 0 (E6893)

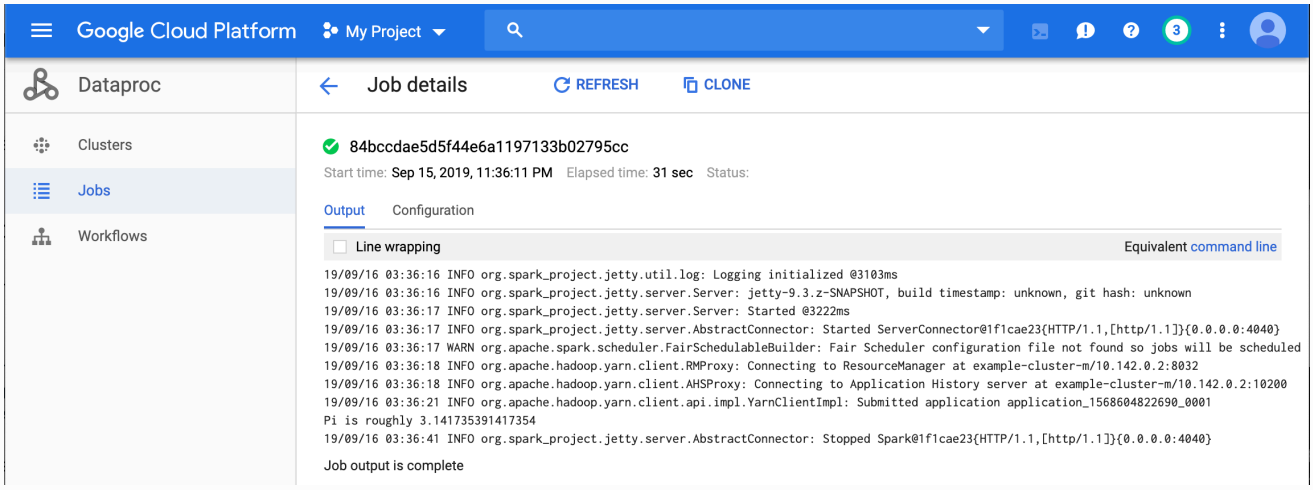
Jing Qian (jq2282)

## 1. Warm-up exercises

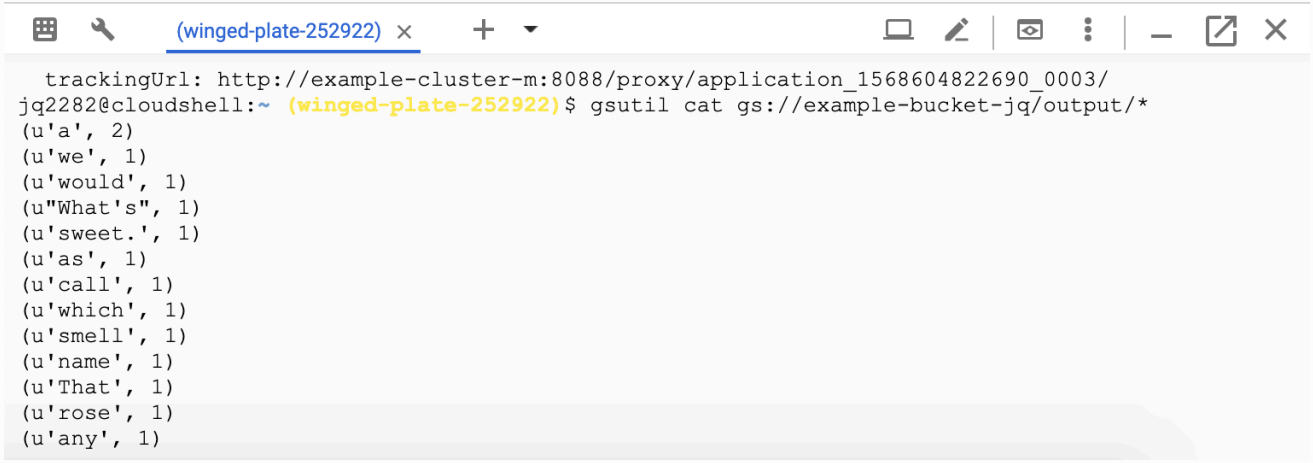
**1)** Screenshots of exercise 3:



Screenshots of exercise 4:




**2)**

Transformations in Exercise3: map.

Actions in Exercise3: reduce. The RDD operation that triggers the program to execute is the actions and hence "reduce".

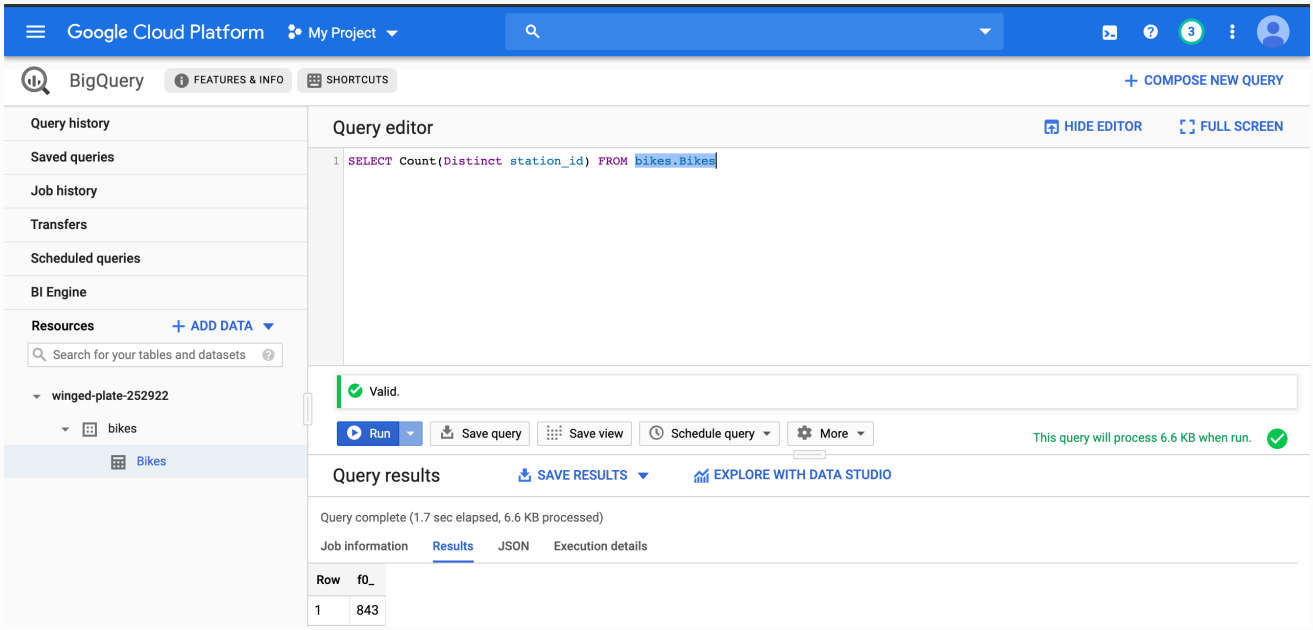
Transformations in Exercise4: flatMap, map, reduceByKey.

Actions in Exercise4: saveAsTextFile. The RDD operation that triggers the program to execute is the actions and hence "saveAsTextFile".

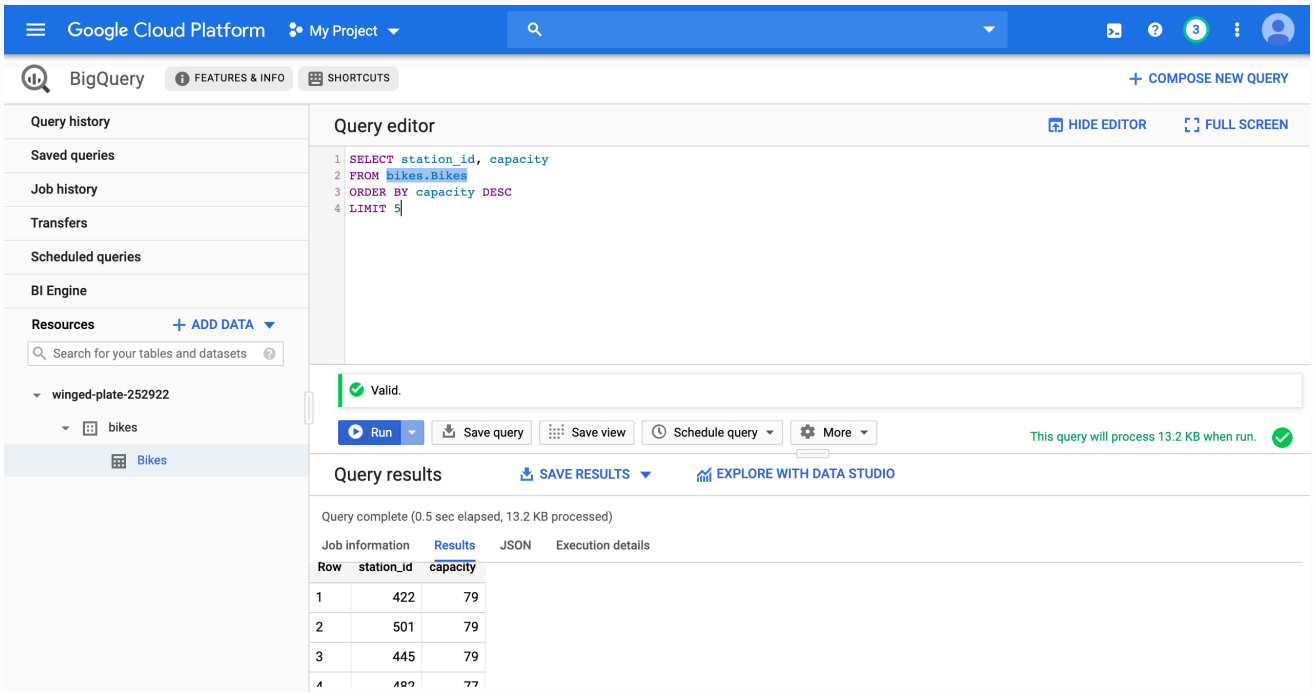
*Exercise 3 is an inside example and corresponding code is found at: https://github.com/apache/spark/blob/master/examples/src/main/java/org/apache/spark/examples/JavaWordCount.java. The code for Exercise 4 is provided in the given link: https://cloud.google.com/dataproc/docs/tutorials/gcs-connector-spark-tutorial.

## 2. NYC Bike expert

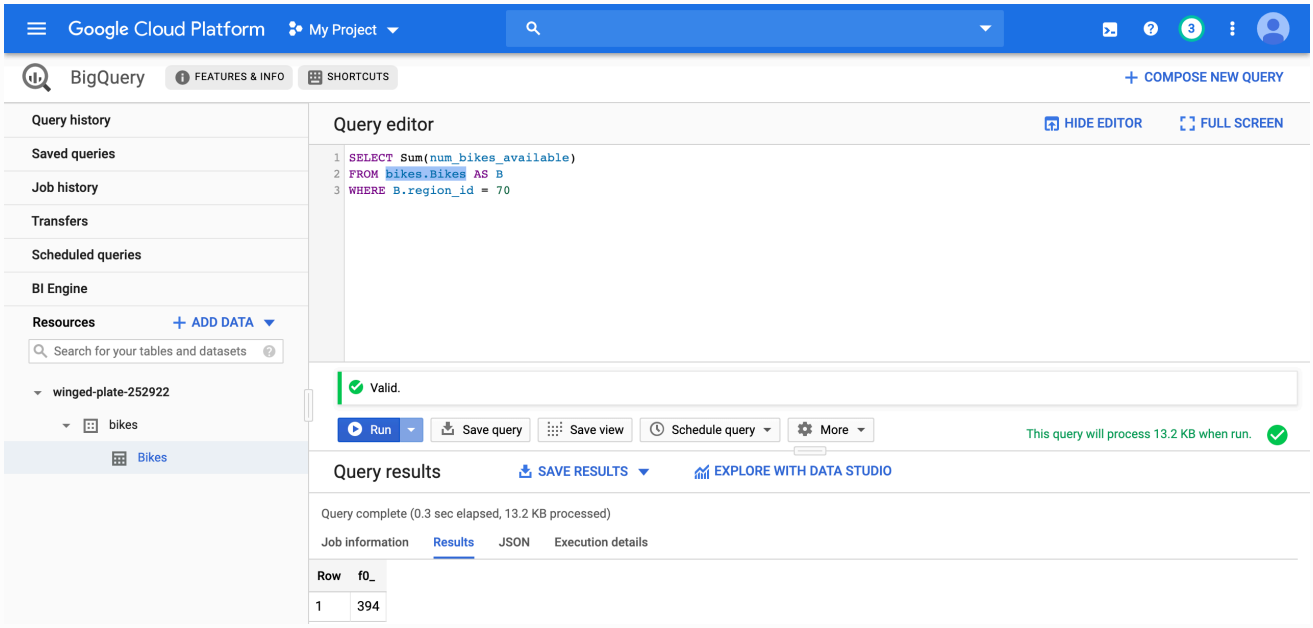
**1)** There are 843 unique station_ids in this dataset.




**2)** The largest capacity for a station is 79. The *station_id* of stations that have the largest capacity are: 445, 422, 501.

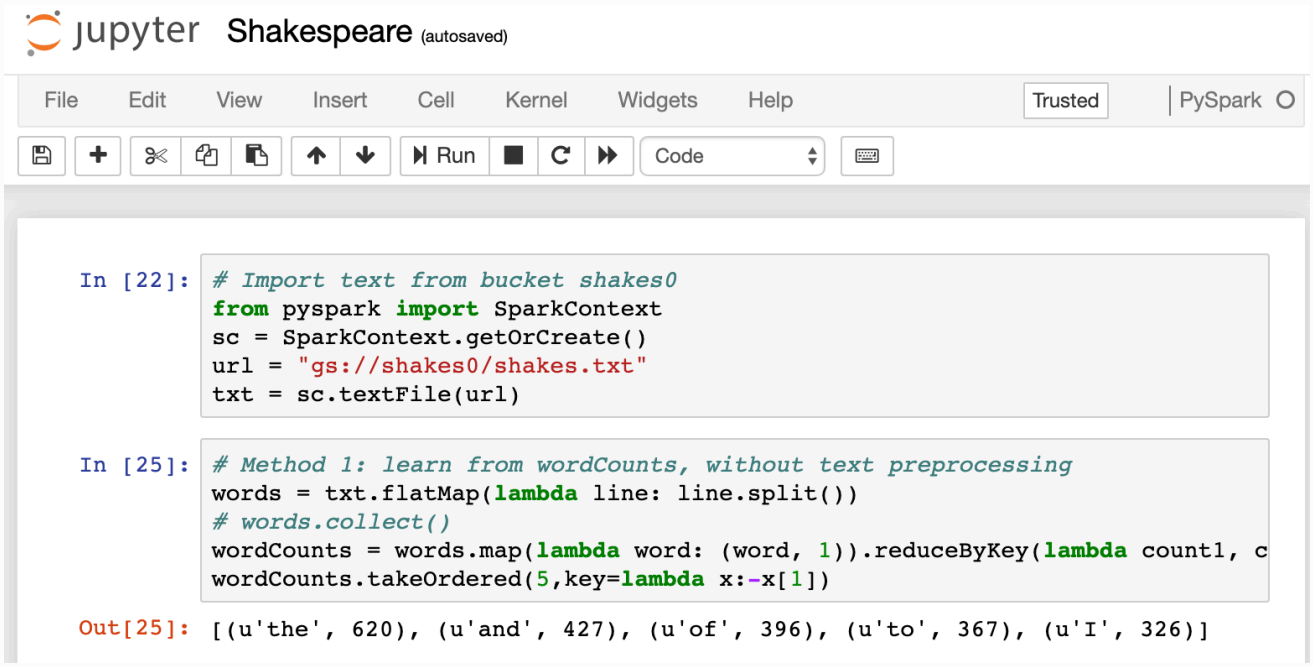



**3)** The total number of bikes available in region_id 70 is 394.




## 3. Understanding William Shakespeare

**1)** The top 5 frequent words without any text preprocessing are: (('the', 620), ('and', 427), ('of', 396), ('to', 367), ('I', 326)).



**2)** Top 5 frequent words by filtering out stop words provided by NLTK package are: (('I', 346), ('And', 170), ('not', 165), ('with', 143), ('be', 138)). Notice here I also removed the punctuations or the counted words will include punctuations.