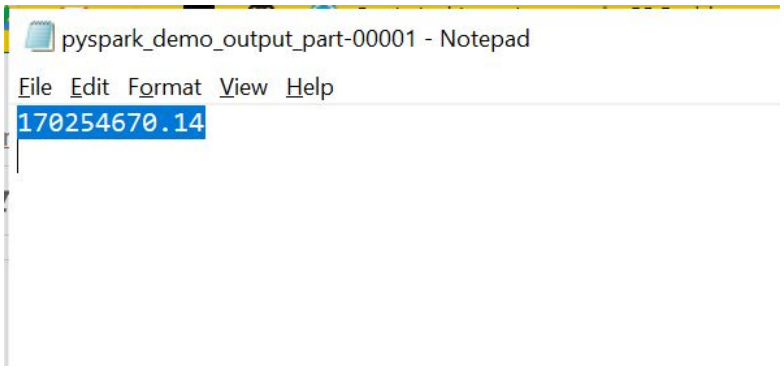
Nick Sun

Spark Assignment

The total distance of all flights I calculated using my Pyspark job was *170254670.14 km.* This value was printed out directly in stdout, but also output as a text file in the output storage bucket specified in the script.



The top ten flights were pretty printed directly into stdout (not a file):

```
(u'AD20C5', 4562016.42399164),
(u'406B4D', 3032985.116802839),
(u'AB8BA5', 2226580.2663955437),
(u'A7D68B', 2168124.8378791353),
(u'A234C0', 1865555.242918873),
(u'AB0E42', 1688773.5519092942),
(u'A01EB5', 1504094.906518219),
(u'AB4505', 992738.0206273091),
(u'4AB50B', 811225.8272344281)]
```

This was all the relevant output from the job I ran on dataproc.

While this script did run successfully, it could use a lot of work. I did not end up using any pyspark.sql.window functionality which Sean suggested last week since I was not able to get it to work so the job I submitted took over 14 minutes to complete. In the future, it is a good idea to get that working since I suspect that some shuffles might be going on between the partitions, causing the long processing time.

As a further note, I did very little preprocessing in dataprep that I know other folks discussed using such as lead and lag sql queries to get consecutive pings for the same ICAO. I am curious to try this approach later since it might lead to better performance.