Spark Assignment - Week 8

Nick Sun

This Spark assignment was pretty rough. The way that I ended up doing it was using `PySpark` all the way through without exporting any data into BigQuery to tackle in SQL.

The key `Dataprep` step was rerunning the recipe from last week, but in addition to ICAO, LAT, and LONG, we also need PosTime which is the epoch time of each ping received. I then loaded this new table into BigQuery so that I could access it from the Spark compute cluster. I used this information later to calculate the total distance travelled for each ICAO.

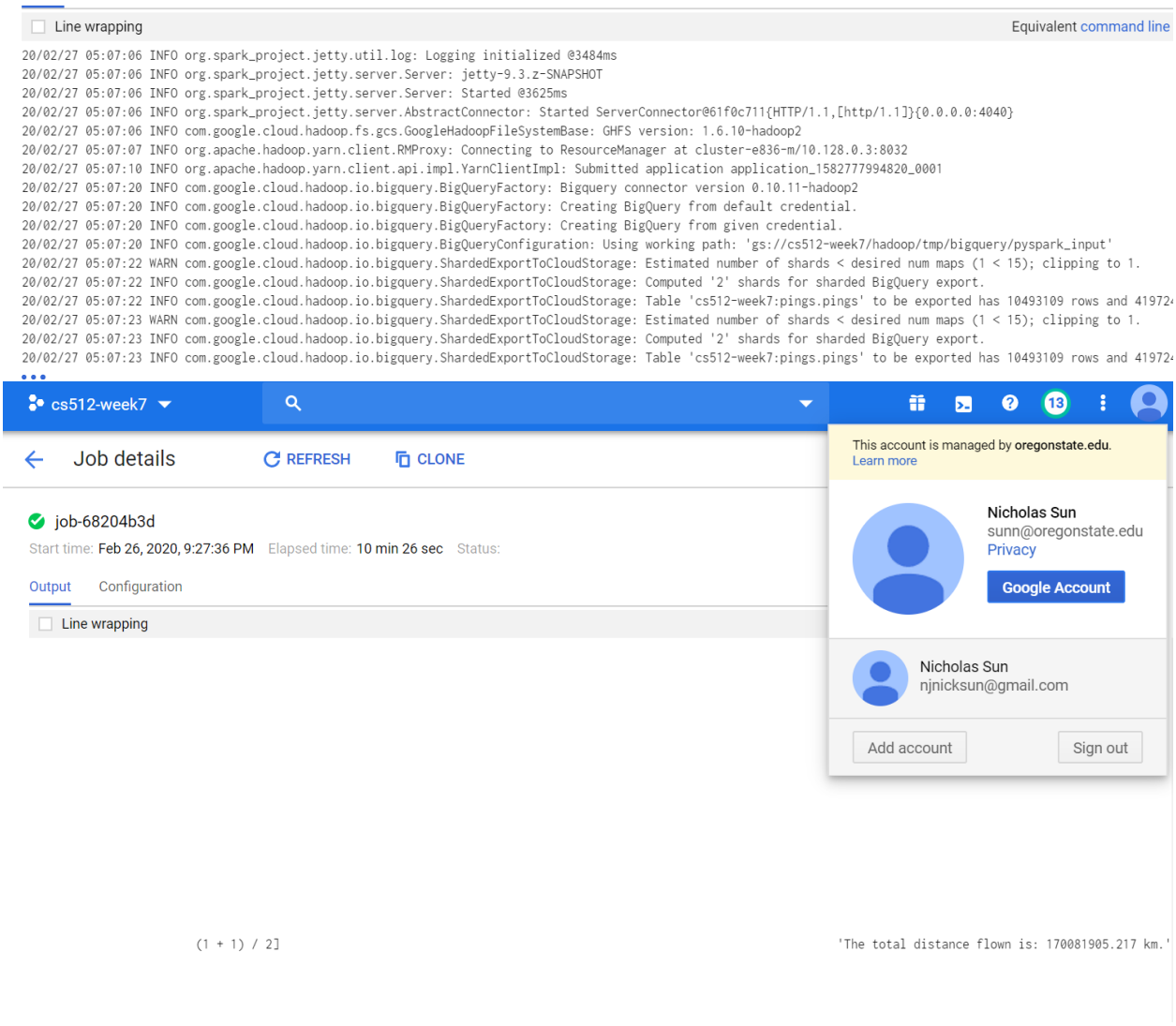
The relevant `PySpark` code is included below. The basic logic is that I have a function written which can calculate the Haversine distance for an entire flight:

```
def getHaversineDistance(pings):
    distance = 0
    for first_ping, second_ping in zip(pings, pings[1:]):
        distance += haversine(
            float(first_ping[1]['Long']), float(first_ping[1]['Lat']),
            float(second_ping[1]['Long']), float(second_ping[1]['Lat'])
        )
    return distance
```

I can then use this function on each flight and then sum over all flights using `PySpark`.

```
vals = table_data.values()
vals = vals.map(lambda line: json.loads(line))

key_pings = vals.map(lambda x: (x['Icao'], x))
sorted_key_pings = key_pings.sortBy(lambda x: x[1]['PosTime'])

grouped_key_pings = sorted_key_pings.groupBy(lambda x: x[0])
grouped_key_pings_lst = grouped_key_pings.map(lambda (x, y): (x, list(y)))

distances = grouped_key_pings_lst.map(lambda (x,y): (x, getHaversineDistance(y)))

totalDistance = distances.map(lambda (x,y): y).reduce(add)
```

My final answer was 170081905.217 km. Below are relevant screenshots:


Line wrapping
Equivalent command line
20/02/27 05:07:06 INFO org.spark_project.jetty.util.log: Logging initialized @3484ms
20/02/27 05:07:06 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT
20/02/27 05:07:06 INFO org.spark_project.jetty.server.Server: Started @3625ms
20/02/27 05:07:06 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@61f0c711{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
20/02/27 05:07:06 INFO com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
20/02/27 05:07:07 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-e836-m/10.128.0.3:8032
20/02/27 05:07:10 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1582777994820_0001
20/02/27 05:07:20 INFO com.google.cloud.hadoop.io.bigquery.BigQueryFactory: Bigquery connector version 0.10.11-hadoop2
20/02/27 05:07:20 INFO com.google.cloud.hadoop.io.bigquery.BigQueryFactory: Creating BigQuery from default credential.
20/02/27 05:07:20 INFO com.google.cloud.hadoop.io.bigquery.BigQueryFactory: Creating BigQuery from given credential.
20/02/27 05:07:20 INFO com.google.cloud.hadoop.io.bigquery.BigQueryConfiguration: Using working path: 'gs://cs512-week7/hadoop/tmp/bigquery/pyspark_input'
20/02/27 05:07:22 WARN com.google.cloud.hadoop.io.bigquery.ShardedExportToCloudStorage: Estimated number of shards < desired num maps (1 < 15); clipping to 1.
20/02/27 05:07:22 INFO com.google.cloud.hadoop.io.bigquery.ShardedExportToCloudStorage: Computed '2' shards for sharded BigQuery export.
20/02/27 05:07:22 INFO com.google.cloud.hadoop.io.bigquery.ShardedExportToCloudStorage: Table 'cs512-week7:pings.pings' to be exported has 10493109 rows and 41972
20/02/27 05:07:23 WARN com.google.cloud.hadoop.io.bigquery.ShardedExportToCloudStorage: Estimated number of shards < desired num maps (1 < 15); clipping to 1.
20/02/27 05:07:23 INFO com.google.cloud.hadoop.io.bigquery.ShardedExportToCloudStorage: Computed '2' shards for sharded BigQuery export.
20/02/27 05:07:23 INFO com.google.cloud.hadoop.io.bigquery.ShardedExportToCloudStorage: Table 'cs512-week7:pings.pings' to be exported has 10493109 rows and 41972
cs512-week7
13
Job details
REFRESH
CLONE
This account is managed by oregonstate.edu.
Learn more
Nicholas Sun
sunn@oregonstate.edu
Privacy
Google Account
Nicholas Sun
njnicksun@gmail.com
Add account
Sign out
job-68204b3d
Start time: Feb 26, 2020, 9:27:36 PM
Elapsed time: 10 min 26 sec
Status:
Output
Configuration
Line wrapping
(1 + 1) / 2]
'The total distance flown is: 170081905.217 km.'