Ingest Data in GCP BigQuery Table

1. Create new project, make sure you’re in it
2. Open cloud shell, confirm you’re in the correct project
3. Clone repo
   1. <https://github.com/GoogleCloudPlatform/data-science-on-gcp.git>
4. Cd into the ch 2 directory
5. “mkdir data” directory
6. “cp download.sh data” to copy the .sh file (shell script) into the new directory
   1. This particular one will populate the directories with data
7. “cd data”
8. “for MONTH in `seq 1 12`; do bash download.sh 2015 $MONTH; done”
9. Invoke bash and download 1 file:
   1. “bash download.sh 2015 1”
   2. “cp 201501.csv 201501.bck”
10. Then remove the file you downloaded:
    1. “rm \*.csv”
11. Invoke python via the command line, then read the lines of the file in command line:
    1. “python”
    2. “f=open(“201501.bck”).readlines()”
12. Then create a write file to write it out to csv
    1. “fout=open(“201501.bck”,”w”)
13. For the first 5000 lines, write the lines out:
    1. “for i in range(5000):

Fout(write(f(i))”

1. You should now have the .csv & .bck in your directory
2. python demo.py
3. create and name a bucket
   1. Via GUI:
      1. Search the word “bucket”
      2. Create a new one, name it, then click create button
4. Then load the data on the command line:
   1. Example via the command line: “gsutil -m cp \*.csv gs://alfrednewbucket”
5. Create new dataset
   1. “gcloud config set project <project id>”
   2. “bq mk dsongcpal”
6. Load everything into BigQuery:
   1. “bq load --autodetect --source\_format=CSV dsongcpal.flight\_auto gs://alfrednewbucket/201501.csv”
7. Go to Big Query and do some test SELECT statements
   1. Search for big query
   2. Your project should have the container listed in a directory
   3. Open subfiles & you should see the schema, etc.
   4. Click the query button & write a select statement