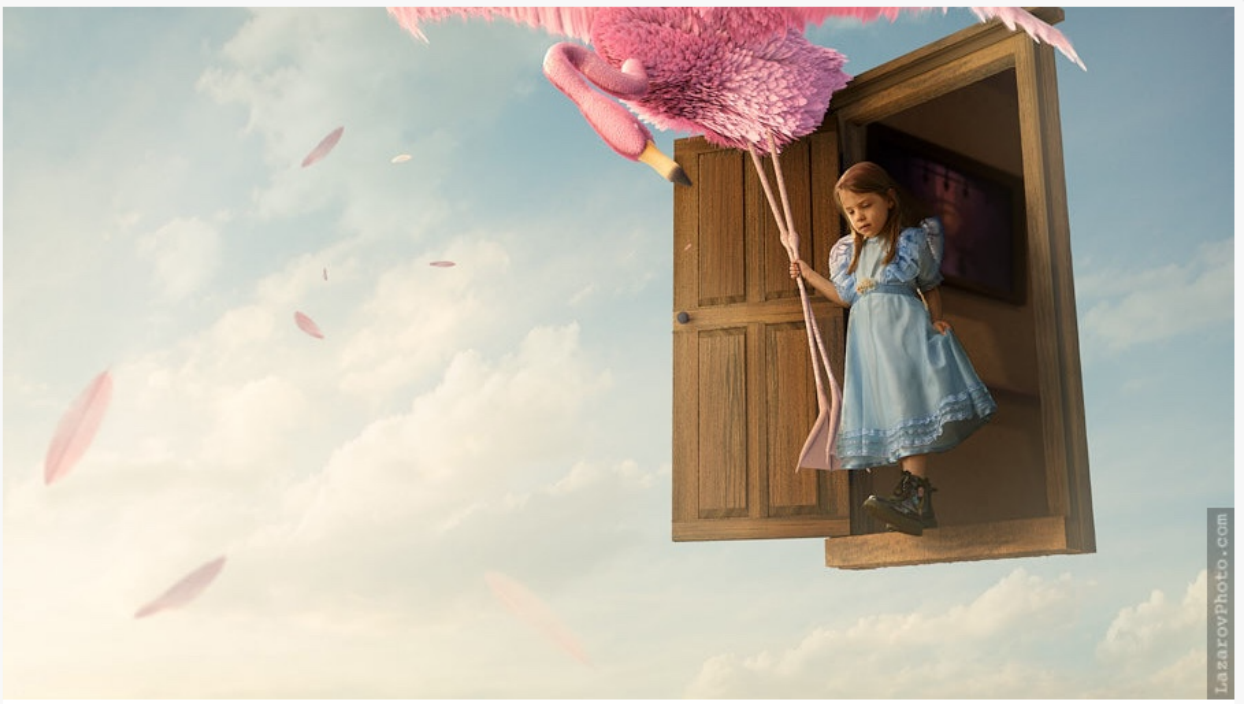
The Opioid Files with BigQuery



Alice steps into the clouds!

In a new chapter of the story, Alice steps into the clouds to discover the joys of floating weightlessly while investigating some Honking Big Data. This Opioid Files exercise follows the previous exercise.

Background

Over the summer of 2019, The Washington Post published a series of reports on the distribution of opioids in the US. Their data team published a portion of a database that tracks the path of every opioid pain pill, from manufacturer to pharmacy, in the United States between 2006 and 2012.

Details of how the data may be accessed are available here. This file represents the national data, it is about 7GB *compressed.* Uncompressed, it is about 75 GB in size. The downloaded file is available on the eecs Linux servers[1] as `/comp/119/arcos_all_washpost.tsv.gz`.

[1] Visit http://systems.eecs.tufts.edu/managing-your-password/ to obtain a login and password, then ssh linux.eecs.tufts.edu to access your linux account.

Problems[2]

This quiz assumes that you have already obtained Google Cloud credits.

1. [10 points] Activate your Google account, create a project, then create a Google Storage bucket named similar to `jsingh-bigdata-private`. The instructions for doing so from the command line are available here[3]. If you prefer, it can also be done using the Google Storage console, which is a web UI. While you're doing that,
    a. Set storage class to Standard,
    b. Set uniform bucket-level access setting to off,
    c. Set your bucket location to the us-central-1 region and
    d. Allow all other parameters to be defaults.

2. [10 points] Extract 5000 entries for the year 2007 from `/comp/119/arcos_all_washpost.tsv.gz`. Write the resulting lines to `~/arcos_2007.tsv`.

3. [5 points] Upload the generated file `~/arcos_2007.tsv` to your shiny new bucket and verify that it got there. The command to verify is your equivalent of `ls -la gs://jsingh-bigdata-private`. Please paste the output of the command into your submission.

4. [5 points] Navigate over to the BigQuery Console. Create a dataset named alice_2022 also located in us-central-1. Create a table arcos_2007 within this dataset. The columns of this table should correspond to the fields in arcos.tsv, as we will load that file into the table. Set the schema of the table as follows (most attributes are strings except for a handful).

```
REPORTER_DEA_NO:STRING,
REPORTER_BUS_ACT:STRING,
REPORTER_NAME:STRING,
REPORTER_ADDL_CO_INFO:STRING,
REPORTER_ADDRESS1:STRING,
REPORTER_ADDRESS2:STRING,
REPORTER_CITY:STRING,
REPORTER_STATE:STRING,
REPORTER_ZIP:STRING,
REPORTER_COUNTY:STRING,
BUYER_DEA_NO:STRING,
BUYER_BUS_ACT:STRING,
BUYER_NAME:STRING,
BUYER_ADDL_CO_INFO:STRING,
BUYER_ADDRESS1:STRING,
```

[2] Total out of 40 points.
[3] The Google Storage utility `gsutil` is preinstalled on the EECS Linux servers. It's not hard to install if you're using your own laptop instead.

```
BUYER_ADDRESS2:STRING,
BUYER_CITY:STRING,
BUYER_STATE:STRING,
BUYER_ZIP:STRING,
BUYER_COUNTY:STRING,
TRANSACTION_CODE:STRING,
DRUG_CODE:STRING,
NDC_NO:STRING,
DRUG_NAME:STRING,
QUANTITY:STRING,
UNIT:STRING,
ACTION_INDICATOR:STRING,
ORDER_FORM_NO:STRING,
CORRECTION_NO:STRING,
STRENGTH:STRING,
TRANSACTION_DATE:STRING,
CALC_BASE_WT_IN_GM:FLOAT,
DOSAGE_UNIT:STRING,
TRANSACTION_ID:INTEGER,
Product_Name:STRING,
Ingredient_Name:STRING,
Measure:STRING,
MME_Conversion_Factor:FLOAT,
Combined_Labeler_Name:STRING,
Revised_Company_Name:STRING,
Reporter_family:STRING,
dos_str:STRING
```

5. [5 points] Upload the arcos_2007.tsv file into BigQuery. The command for data loading is `bq load`. *Since `bq` isn't installed on the Linux machines, we can use the cloud shell for bq load.* The shell is accessible from the Cloud Console (top ribbon, right ).
    a. Authorize the shell as prompted,
    b. Set your project ID using `gcloud config set project [PROJECT_ID]` ,
    c. Upload the file to the table: `bq load --field_delimiter=tab --source_format=CSV [PROJECT_ID]:alice_2022.arcos_2007 gs://[BUCKET_NAME]/arcos.tsv`

6. [5 points] Use the data in BigQuery to find the names of all the drugs that were delivered during 2007. (Submit your query answers in JSON format).