


Using Google Cloud For DS:

A Data Science in Context Presentation

Prepared and Presented by William Jasmine


Set Up

- Download the Google Cloud CLI
  - This will enable to perform operations in Google Cloud directly from you command line.
  - However, Google still offers a lot of cloud function from its UI.
- Create a Gmail account if you don't already have one.
- Authenticate your Gmail account with your newly downloaded Google Cloud CLI via typing the following into your command line:
  - `gcloud auth login`
- Create a Google Cloud project that you will use to store, query, and analyze data. Can be done via the following in your command line:
  - `gcloud projects create [YOUR_PROJECT_NAME]`
- Set your project in the command line:
  - `gcloud config set project [YOUR_PROJECT_NAME]`
- What mine looks like

Loading Data Into BigQuery

- BigQuery is Google Cloud's database management platform.
- It has its own SQL language (similar to PostgreSQL) that can be used to query any stored data.
- BigQuery's hierarchy for storing data: Project ➔ Dataset ➔ Table
  - Referenced as `myproject.mydataset.mytable` when querying data
- Load data from a csv file into a BigQuery table
  - `bq load myDataset.newTable [PATH_TO_DATA_SOURCE] schema`

Example :

```
bq load movie_survey.movies movies.csv \
movie_name:string,category:string,release_year:string
```

Result

Querying Data in BigQuery

- UI is actually pretty good for this, and provides a decent environment for writing SQL code.
- Can also be done directly from command line (either method works):
  - bq query --use_legacy_sql=false < query.sql
  - bq query --use_legacy_sql=false “[INSERT_QUERY]”
- BigQuery’s main benefit is its performance on extremely large datasets.
- Extremely scalable/elastic: can allocate computing resources in correlation to the amount of data/operations that a query will use
- *The result:* BigQuery can query tables containing terabytes worth of data on the scale of seconds.

Using Colab

- Google Colaboratory is essentially Google's response to Jupyter Notebooks
- *Main benefit:* it is very easy to connect them to other Google tools, i.e. Bigquery and Google Sheets, and Google Drive.
- Can be used to query data and store as a pandas dataframe.
- Once data is analyzed, results can be uploaded directly as a Google Sheet or into Google Storage.

Example that shows all of the above

What is Google Cloud Storage?

- Google cloud is a file storage manager hosted on the Google Cloud Platform infrastructure.
- Essentially allows you to be able to store unlimited data (for a price) for any kind of file type.
- Different from Google Drive which focuses more on personal file storage and simplicity of file sharing.

Resources/Appendix

- How to install the Google Cloud CLI
- BigQuery command line library documentation
- General BigQuery documentation
- Google Colaboratory introduction and documentation
- Google Cloud Storage documentation
- My Colab example (send me a request if you'd like access)


Questions?