Fundamentals of Data Engineering

Week 03 - sync session

datascience@berkeley

Overview

- Dive into command line tools for figuring out what you have in datasets
- Setting up BigQuery from the command line
- What’s up next?

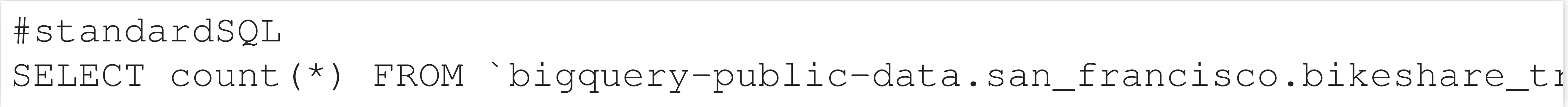
Project Breakout

Share queries

- Questions?
- Problems?
- Queries that you wanted to do but you couldn’t figure out how?

What's the size of this dataset? (i.e., how many trips)

983648



```
#standardSQL
SELECT count(*) FROM `bigquery-public-data.san_francisco.bikeshare_tr
```

What is the earliest start time and latest end time for a trip?

```
2013-08-29 09:08:00 2016-08-31 23:48:00
```

```
#standardSQL
SELECT min(start_date)
FROM `bigquery-public-data.san_francisco.bikeshare_trips`

#standardSQL
SELECT max(end_date)
FROM `bigquery-public-data.san_francisco.bikeshare_trips`
```

How many bikes are there?

700

```
#standardSQL
SELECT count(distinct bike_number)
FROM `bigquery-public-data.san_francisco.bikeshare_trips`
```

Activities: async content

- This week, videos about various tools you can use to work with your data
- Working with files: json, csv etc
- So for today's class activity, we're going to work with some command line tools to manipulate data in some handy ways.

Finding stuff out about your data

Download Datasets

Save data into your `w205` directory

```
cd ~/w205
curl -L -o annot_fpid.json https://goo.gl/qWiu7d
curl -L -o lp_data.csv https://goo.gl/FDFPYB
```

What’s in this file?

```
head lp_data.csv

tail lp_data.csv
```

What are variables in here?

```
head -n1 lp_data.csv
```

How many entries?

```
cat lp_data.csv | wc -l
```

How about sorting?

```
cat lp_data.csv | sort
```

Take a look at what options there are for sort

```
man sort
```

fix so sorting correctly

```
cat lp_data.csv | sort -g

cat lp_data.csv | sort -n
```

Find out which topics are more popular

What have we got in this file?

```
head annot_fpid.json
```

Hmmm, what now? jq

pretty print the json

```
cat annot_fpid.json | jq .
```

Just the terms

```
cat annot_fpid.json | jq '.[][]'
```

Remove the “”s

```
cat annot_fpid.json | jq '.[][]' -r
```

Can we sort that?

```
cat annot_fpid.json | jq '.[][]' -r | sort
```

Unique values only

```
cat annot_fpid.json | jq '.[][]' -r |
sort | uniq
```

How could I find out how many of each of those unique values there are?

```
cat annot_fpid.json | jq '.[][]' -r |
sort | uniq -c
```

Now, how could I sort by that?

```
cat annot_fpid.json | jq '.[][]' -r |
sort | uniq -c | sort -g
```

Ascending

```
cat annot_fpid.json | jq '.[][]' -r |
sort | uniq -c | sort -gr
```

Descending

So, what are the top ten terms?

```
cat annot_fpid.json | jq '.[][]' -r |
sort | uniq -c | sort -gr | head -10
```

bq cli

```
bq query --use_legacy_sql=false '
SELECT count(*)
FROM `bigquery-public-data.san_francisco.bikeshare_status`'
```

How many stations are there?

```
bq query --use_legacy_sql=false '
SELECT count(distinct station_id)
FROM `bigquery-public-data.san_francisco.bikeshare_status`'
```

How long a time period do these data cover?

```
bq query --use_legacy_sql=false '
SELECT min(time), max(time)
FROM `bigquery-public-data.san_francisco.bikeshare_status`'
```

Generate Ideas

- What do you know?
- What will you need to find out?

Summary

- Command line tools and jq to dive into your data
- BigQuery from the command line

Onward

- This week’s videos may seem like kind of a mixed bag.
- Starting to transition between Project 1 & Project 2.
- Working with query tools (bigquery and Athena)
- Getting a glimpse of how to use jupyter notebooks for Project 1
- Looking ahead at some bits of docker manipulation from the command line, getting ready for Project 2.
- Also you’ll see your first supplemental tag

Extras

Resources

sed and awk

http://www.catonmat.net/blog/awk-one-liners-explained-part-one/ http://www.catonmat.net/blog/sed-one-liners-explained-part-one/

jq

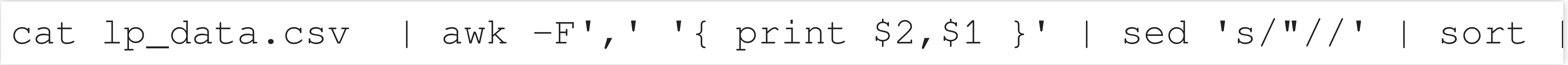
https://stedolan.github.io/jq/tutorial/

Advanced options

Sort by ‘product_name’

```
cat lp_data.csv | awk -F',' '{ print $2,$1 }' | sort
```

Fix the “”s issue



```
cat lp_data.csv  | awk -F',' '{ print $2,$1 }' | sed 's/"//' | sort |
```

Berkeley SCHOOL OF INFORMATION