Fundamentals of Data Engineering

Week 02 - sync session

datascience@berkeley

Assignment 1

- We will usually do a breakout to share solutions and ask questions, but this week's was pretty straightforward.
- Questions on process?

Your droplet set up

- repos cloned:
  - `course-content`
  - `assignment-01-<user-name>`

How to do a PR

- Review process from gui

Due tomorrow morning

Some things about this class

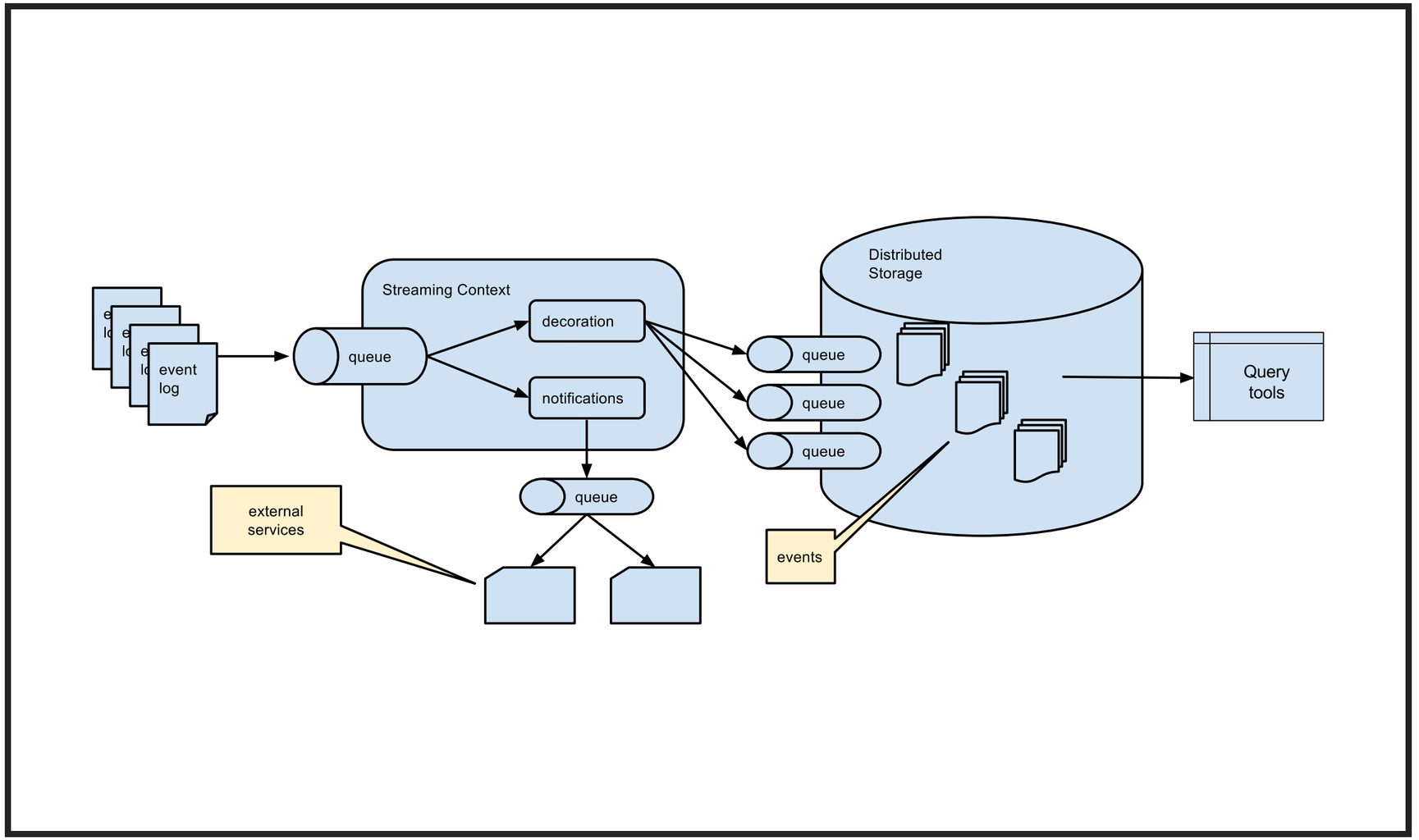
How to read

- Two screens (or devices)
- Reading
- Following along with coding

Pacing

- What you can do
- What you can understand

Where are we in the pipeline



Events

- What sort of events feed this pipeline?
- How were these events captured?

Main thing to pay attention to

- Pipeline is provided for this example
- We're *using* it to answer business questions

Big Ideas

Business Decisions

- All about the business
- Data-Driven Business Decisions …are queries

Translation

- SQL queries are really pretty easy
- How to get to the queries from the questions, sometimes not so much

Query Project

- In the Query Project, you will get practice with SQL while learning about Google Cloud Platform (GCP) and BiqQuery. You'll answer business-driven questions using public datasets housed in GCP. To give you experience with different ways to use those datasets, you will use the web UI (BiqQuery) and the command-line tools, and work with them in jupyter notebooks.
- We will be using the Bay Area Bike Share Trips Data (https://cloud.google.com/bigquery/public-data/bay-bike-share).

Problem Statement

- You're a data scientist at Ford GoBike (https://www.fordgobike.com/), the company running Bay Area Bikeshare. You are trying to increase ridership, and you want to offer deals through the mobile app to do so. What deals do you offer though? Currently, your company has three options: a flat price for a single one-way trip, a day pass that allows unlimited 30-minute rides for 24 hours and an annual membership.

Questions

- Through this project, you will answer these questions:
  - What are the 5 most popular trips that you would call “commuter trips”?
  - What are your recommendations for offers (justify based on your findings)?

Get Going: Google account

- Go to https://cloud.google.com/bigquery/
- Click on “Try it Free”
- It asks for credit card, but you get $300 free and it does not autorenew after the $300 credit is used,

Working with BQ gui

https://bigquery.cloud.google.com/table/bigquery-public-data:san_francisco.bikeshare_status

Tutorial

https://www.w3schools.com/sql/default.asp

Some annoying specific stuff about BQ

the ;

```
SELECT *
FROM Customers;
```

VS

```
SELECT *
FROM Customers
```

Legacy vs Standard SQL

```
SELECT *
FROM [bigquery-public-data:san_francisco.bikeshare_trips]
```

VS

```
#standardSQL
SELECT *
FROM `bigquery-public-data.san_francisco.bikeshare_trips`
```

The Big Difference

```
SELECT distinct(bikes_available)
FROM [bigquery-public-data:san_francisco.bikeshare_status]
```

NO

```
#standardSQL
SELECT distinct(bikes_available)
FROM `bigquery-public-data.san_francisco.bikeshare_status`
```

YES

For this class

```
#standardSQL
SELECT *
FROM `bigquery-public-data.san_francisco.bikeshare_status`
```

- More similar to command line bq
- More like most other SQL implementations

Querying Data

How many events are there?

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

How many stations are there?

```
#standardSQL
SELECT count(distinct station_id)
FROM `bigquery-public-data.san_francisco.bikeshare_status`
```

How long a time period do these data cover?

```
#standardSQL
SELECT min(time), max(time)
FROM `bigquery-public-data.san_francisco.bikeshare_status`
```

How many bikes does station 90 have?

```
#standardSQL
SELECT station_id,
(docks_available + bikes_available) as total_bikes
FROM `bigquery-public-data.san_francisco.bikeshare_status`
WHERE station_id = 90
```

What’s up with that?

```
#standardSQL
SELECT station_id, docks_available, bikes_available, time,
(docks_available + bikes_available) as total_bikes
FROM `bigquery-public-data.san_francisco.bikeshare_status`
WHERE station_id = 90
ORDER BY total_bikes
```

Get a table with `total_bikes` in it

```
#standardSQL
SELECT station_id, docks_available, bikes_available, time,
(docks_available + bikes_available) as total_bikes
FROM `bigquery-public-data.san_francisco.bikeshare_status`
```

```
#standardSQL
SELECT distinct (station_id), total_bikes
 FROM `ambient-cubist-185918.bike_trips_data.total_bikes`
```

```
#standardSQL
SELECT distinct station_id, total_bikes
FROM `ambient-cubist-185918.bike_trips_data.total_bikes`
WHERE station_id = 22
```

Independent Queries

https://www.w3schools.com/sql/default.asp

SecureShell (SSH)

remote terminal connections

```
ssh science@xxx.xxx.xxx.xxx
```

copying files

On your laptop, run

```
scp some_file science@xxx.xxx.xxx.xxx:
```

or

```
scp some_file science@xxx.xxx.xxx.xxx:/tmp/
```

On your laptop, run

```
scp science@xxx.xxx.xxx.xxx:~/w205/a_file.py .
```

Summary

- Business questions
- Answered using empirical data
- By running queries against (raw?) events
- Need a pipeline in place to capture these raw events
- SSH

Berkeley SCHOOL OF INFORMATION