Fundamentals of Data Engineering

Week 02 - sync session

datascience@berkeley

Your cloud instance set up

- repos cloned:
  - `course-content`

How to create a new branch, commit to that and do a PR

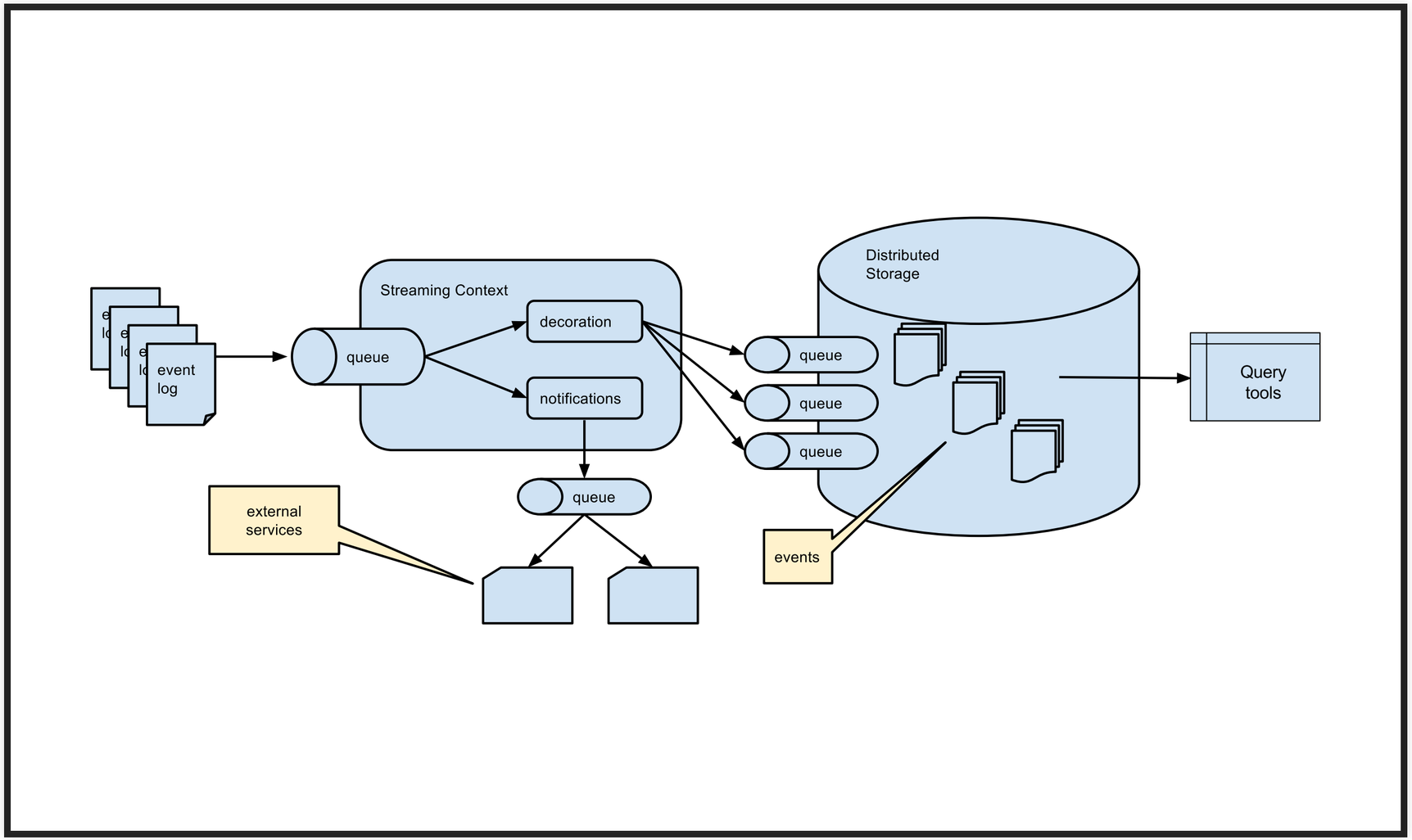
- Process from commandline and giu this time

Some things about this class

Pacing

- What you can do
- What you can understand

Where are we in the pipeline



Main thing to pay attention to

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

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, follow the class walk through to find the data set.

Problem Statement

- You’re a data scientist at a company formerly known Ford GoBike, now Lyft bay wheels (https://www.lyft.com/bikes/bay-wheels). 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)?

Working with BQ gui

https://console.cloud.google.com/bigquery

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`
```

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

Events

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

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
```

Independent Queries

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

Summary

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

Berkeley SCHOOL OF INFORMATION