Fundamentals of Data Engineering

Week 01 - sync session

datascience@berkeley

Week 1 - Overview

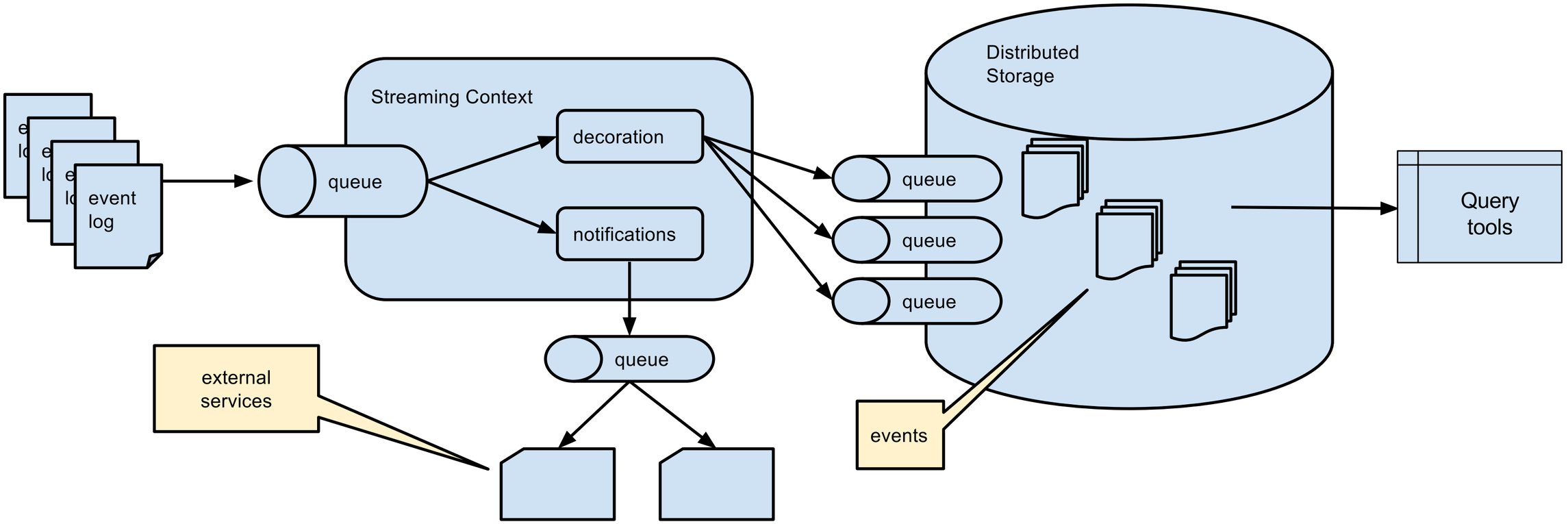
- Introductions
- Set up your working environment for this class
- Review syllabus, course goals, processes & tools ...

Introductions

In this class, you will

- Gain exposure to basic problems associated with data and data-driven decision-making
- Develop a working knowledge of some tools/techniques used to solve these problems
- Learn where to go for help and more info

Just enough


event log
queue
Streaming Context
decoration
notifications
queue
external services
queue
queue
queue
Distributed Storage
events
Query tools

Process/Procedures

- Good practices
- Appropriate tools
- Getting used to

How this class works

Syllabus

https://mids-w205-fund-of-data-eng.github.io/course-content/

Asyncronous Content

```
https://github.com/mids-w205-martin-mims/course-content/ \
blob/master/01-Introduction/async-videos.md
```

- Same as in ISVC, but you can access it all in one place here.

Readings

- No one textbook available for this course.
- Using subscription service to cover the range of topics.
- `https://www.safaribooksonline.com/pricing/`
- Individual option: $39/month (can stop whenever you want)
- Quick note: Get the mobile apps.

Prerequisites

- Resources listed under prereqs
- Safari has tons of other materials you can help yourself with.

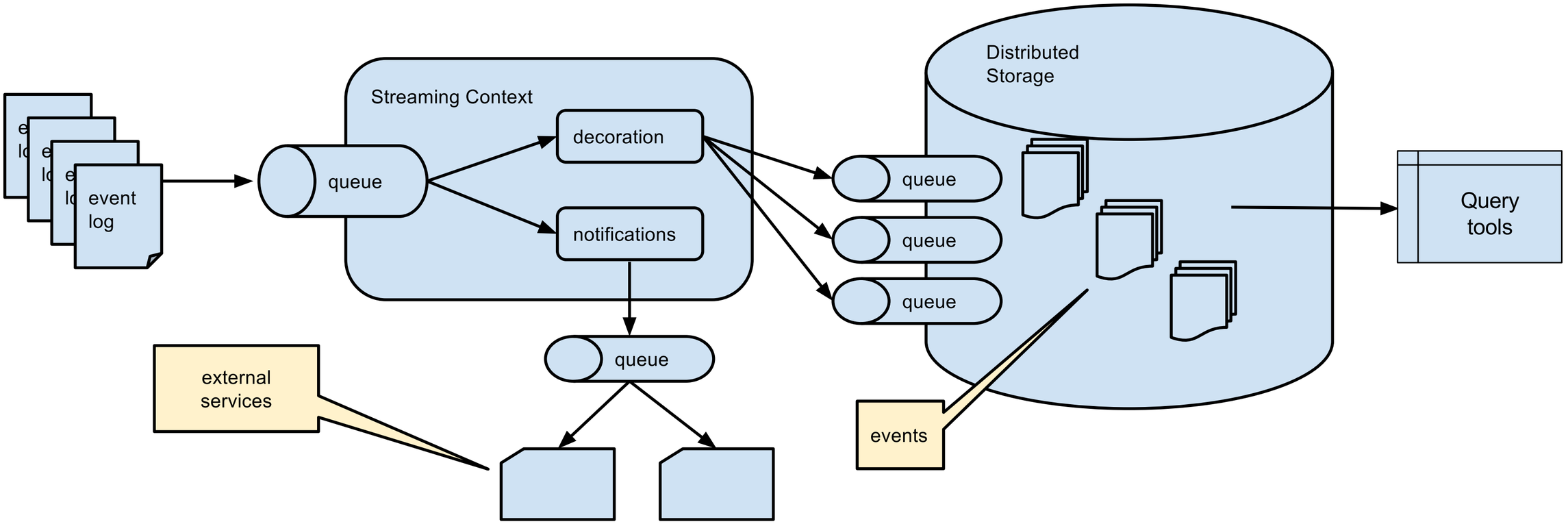
Course Outline

- 4 sections:
- 3-week Introduction
- 5-week Basics section
- 4-week Streaming Data section
- Putting it All Together

Student Projects

Student Projects

1. Querying Data
2. Tracking User Activity
3. Understanding User Behavior


event log
queue
Streaming Context
decoration
notifications
queue
external services
queue
queue
queue
Distributed Storage
events
Query tools

Querying Data

- Use existing tools/pipeline/dataset
- Answer basic business questions

Tracking User Activity

- Use provided pipeline components
- Transform/store data
- Answer business questions
- Bonus:
  - Trigger notifications

Understanding User Behavior

- Assemble an end-to-end pipeline
- Ingest/transform/store data
- Answer comprehensive business questions
- Bonus:
  - Manage sessionization / state

Levels of Expertise

Async

- To get ready for project 1, videos - how events are generated

Activities

- Let’s get going!

Slack

Cloud Instances

1. Sign up for GCP account
2. Create your instance
3. Access your instance

Create your instance

Access JupyterHub

AI Platform -> Notebooks -> “Open Jupyterlab”

Access your instance

- Compute Engine -> VM Instances -> “SSH”
- Settings -> Change Linux Login -> “jupyter”

Docker


DON'T PANIC

- pull the image:

```
docker pull midsw205/base
```

- create your mids-w205 workspace:

```
mkdir w205
```

- run (set *your* home directory for "`-v`")

```
docker run \
  -it \
  --rm \
  -v ~/w205:/w205 \
  midsw205/base:latest \
  bash
```

- `exit` (or `ctrl-d`)

Git

Git set up

Clone the `course-content` repo

- `cd w205`
- Clone the `course-content` repo into your mids-w205 workspace:

```
git clone https://github.com/mids-w205- \
  <instructor-last-name>/course-content
```

Signup Assignment

Clone the repo

- `cd w205`
- Clone the repo into your mids-w205 workspace:

```
git clone https://github.com/mids-w205- \
  <instructor-last-name>/ \
  signup-<git-user-name>
```

Open, Change, Close `README.md`

- `nano README.md`
- change line
- `ctrl-o`
- return
- `ctrl-x`
- Now you're out of nano.

Git: commit changes

- `git status`
- `git add README.md`
- `git commit -m 'my new readme'`
- The first time you commit, it doesn't know who you are.

```
git config --global user.email "you@example.com"
```

```
git config --global user.name "Your Name"
```

- `git commit -m 'my new readme'`
- `git push`

Git: submit a PR

- All projects submitted as PRs

```
https://github.com/mids-w205-martin-mims/signup-<user-name>
```

- Click on `README.md`
- Click on edit button (pencil icon)
- Make a change
- “Commit changes” section, select “Create a new branch for this commit…”
- Enter PR name & description
- Click “Propose file change” button
- Assign instructors as reviewers
- Click “Create pull request” button

Berkeley SCHOOL OF INFORMATION