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

In this class, you will

- Gain exposure to basic problems associated with data and data-driven decision-making

In this class, you will

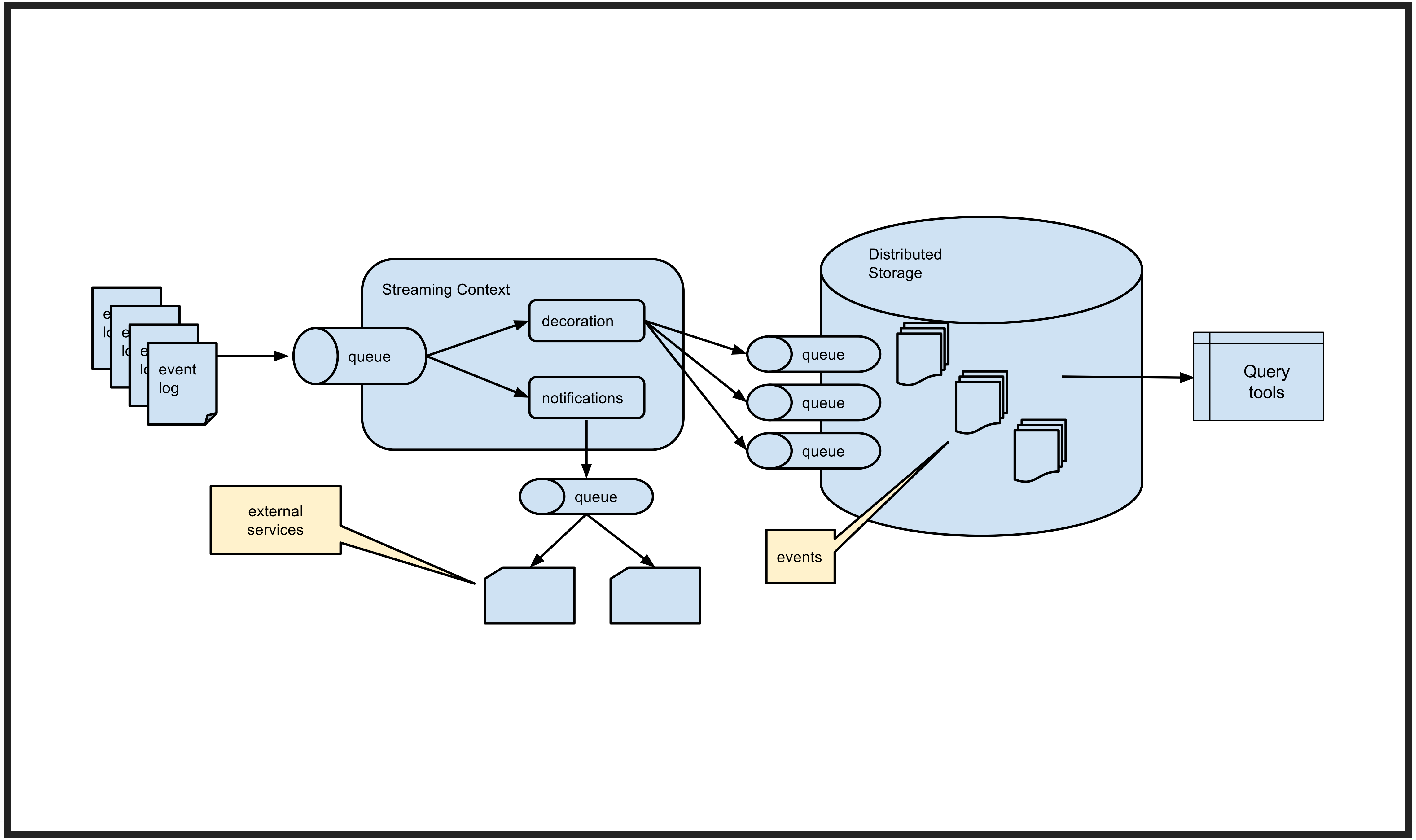
- Gain exposure to basic problems associated with data and data-driven decision-making
- Develop a working knowledge of some tools/techniques used to solve these problems

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

Pipeline: Contextual Anchor



Process/Procedures

Process/Procedures

- Good practices

Process/Procedures

- Good practices
- Appropriate tools

Process/Procedures

- Good practices
- Appropriate tools
- Getting used to

Approach

Approach

- Github-centric content

Approach

- Github-centric content
- Cloud accounts

Approach

- Github-centric content
- Cloud accounts
- Dockerhub

Approach

- Github-centric content
- Cloud accounts
- Dockerhub
- Activity submissions

Activities

- Let’s get going!

Docker

- What is docker?

Install docker

- Windows

  https://store.docker.com/editions/community/docker-ce-desktop-windows

- Mac

  https://store.docker.com/editions/community/docker-ce-desktop-mac

DON'T PANIC

Docker set up (from a terminal)

- 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 /Users/<user>/w205:/w205 \
  midsw205/base:latest \
  bash
```

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

git

- What is git?

Git set up

Get started

- If working on Mac or Linux, or have git installed, go to w205 folder.
- If windows or no git,

```
docker run \
  -it \
  --rm \
  -v /Users/<user>/w205:/w205 \
  midsw205/base:latest \
  bash
```

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`

Open, Change, Close `README.md`

- `nano README.md`

Open, Change, Close `README.md`

- `nano README.md`
- change line

Open, Change, Close `README.md`

- `nano README.md`
- change line
- `ctrl-o`

Open, Change, Close `README.md`

- `nano README.md`
- change line
- `ctrl-o`
- return

Open, Change, Close `README.md`

- `nano README.md`
- change line
- `ctrl-o`
- return
- `ctrl-x`

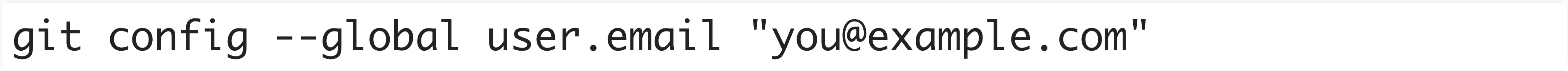
Open, Change, Close `README.md`

- `nano README.md`
- change line
- `ctrl-o`
- return
- `ctrl-x`
- Now you’re out of nano, but still in the container.

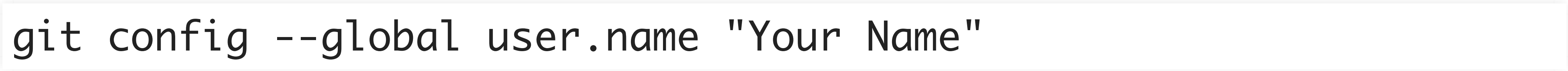
Git: commit changes

- `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`

After all that,

- Mac & Linux users
- Windows users
- for today, you used docker,
- What do we need to do going forward…

Git: submit a PR

- All assignments 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

What is Data Engineering?

Things are changing quickly

https://www.coursera.org/learn/gcp-big-data-ml-fundamentals

What surprised you about the points made?

- Enter 2 things on chat that you noticed.

Virtualization

GCP

DO

AWS

How this class works

Syllabus

```
https://github.com/mids-w205-martin-mims/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

Readings

- No one textbook available for this course.

Readings

- No one textbook available for this course.
- Using subscription service to cover the range of topics.

Readings

- No one textbook available for this course.
- Using subscription service to cover the range of topics.
- `https://www.safaribooksonline.com/pricing/`

Readings

- No one textbook available for this course.
- Using subscription service to cover the range of topics.
- https://www.safaribooksonline.com/pricing/
- Two Options:

Readings

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

Readings

- No one textbook available for this course.
- Using subscription service to cover the range of topics.
- `https://www.safaribooksonline.com/pricing/`
- Two Options:
- Individual option: $39/month (can stop whenever you want)
- Team option: (up to 25 people) - 1 year subscription for $399

Readings

- No one textbook available for this course.
- Using subscription service to cover the range of topics.
- https://www.safaribooksonline.com/pricing/
- Two Options:
- Individual option: $39/month (can stop whenever you want)
- Team option: (up to 25 people) - 1 year subscription for $399
- Quick note: Get the mobile apps.

Prerequisites

Prerequisites

- Resources listed under prereqs

Prerequisites

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

Course Outline

Course Outline

- 4 sections:

Course Outline

- 4 sections:
- 3-week Introduction

Course Outline

- 4 sections:
- 3-week Introduction
- 5-week Basics section

Course Outline

- 4 sections:
- 3-week Introduction
- 5-week Basics section
- 4-week Streaming Data section

Course Outline

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

Class flow

Class 1

Class 1

- Preview, discussion, walkthrough set up for github for Assignment 1

Between Class 1 & Class 2

Between Class 1 & Class 2

- async material for Week 1

Between Class 1 & Class 2

- async material for Week 1
- Readings for Week 1

Between Class 1 & Class 2

- async material for Week 1
- Readings for Week 1
- Assignment 01

Class 2

Class 2

- Review Assignment 01, questions, where did you hit a wall?

Class 2

- Review Assignment 01, questions, where did you hit a wall?
- Some lectur-ish stuff on this week’s topic

Class 2

- Review Assignment 01, questions, where did you hit a wall?
- Some lectur-ish stuff on this week’s topic
- Preview Query Project (spans Assignments 2-5)

Class 2

- Review Assignment 01, questions, where did you hit a wall?
- Some lectur-ish stuff on this week’s topic
- Preview Query Project (spans Assignments 2-5)
- Preview, discussion, SQL query activities to prepare for Assignment 2

Class 2

- Review Assignment 01, questions, where did you hit a wall?
- Some lectur-ish stuff on this week’s topic
- Preview Query Project (spans Assignments 2-5)
- Preview, discussion, SQL query activities to prepare for Assignment 2
- Final Assignment 01 due on Friday

Between Class 2 & Class 3

Between Class 2 & Class 3

- async material for Week 2

Between Class 2 & Class 3

- async material for Week 2
- Readings for Week 2

Between Class 2 & Class 3

- async material for Week 2
- Readings for Week 2
- Assignment 02

Class 3

Class 3

- Review Assignment 02, questions, where did you hit a wall?

Class 3

- Review Assignment 02, questions, where did you hit a wall?
- Some lectur-ish stuff on this week’s topic

Class 3

- Review Assignment 02, questions, where did you hit a wall?
- Some lectur-ish stuff on this week's topic
- Preview, discussion, do google cloud platform setup and sql statements for Assignment 03

Class 3

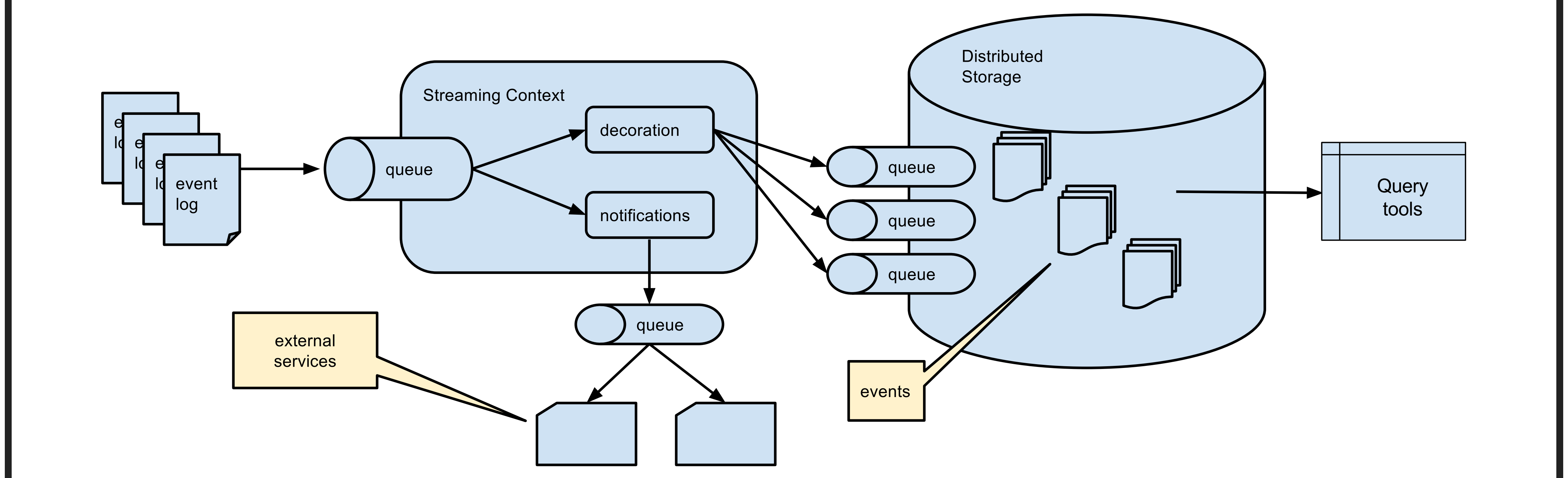
- Review Assignment 02, questions, where did you hit a wall?
- Some lectur-ish stuff on this week's topic
- Preview, discussion, do google cloud platform setup and sql statements for Assignment 03
- Final Assignment 02 due on Friday

Student Projects

Student Projects

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

Pipeline


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

Berkeley SCHOOL OF INFORMATION