University of California, Berkeley
Master of Information and Data Science (MIDS)
W205 – Fundamentals of Data Engineering

Week 1 – Introduction to Data Engineering

Agenda for Today's Class

- Welcome
- Introductions
- Course Overview
- Attendance and Participation
- Grading, Projects, etc
- Useful information
- Breakouts
- Summary

Introductions

- Name
- Location (with timezone)
- Optional: fun fact or something about you

Overview of w205

- What is data engineering?
- Terminology
- Tools
- General principles beyond the current tool

Overview of w205

Where is what?

- GitHub => https://github.com/kevin-crook-ucb/ucb_mids_w205_repo
- Other sources, ways: GDrive, Slack, etc

- Slack
- Schedule and Due Dates
- Office Hours
- Grading
- Grading Scale
- Attendance and Participation
- Asynch Assessments
- Projects
- Readings
- Cloud

Prerequisite Knowledge for w205

- Python programming
- SQL
- Linux CLI
- GitHub Git CLI

See Syllabus for links to tutorials and details

- NOT mentioned but assumed: basic OS knowledge, basic networking (IPs, Ports, Firewalls,...)

See course 1D computing basics (can ask me later)

Attendance and Participation (5% of grade)

Please record your attendance and participation for today's class:

- Find in class GitHub repo README
- or use the link pinned to my slack channel
- or found in the FAQ.

Grading

- 3 projects: 90% - 2 IC, 3rd group
- Attendance recording: 5%
- Assessments (found in ISVC): 5%

Debugging Skills

- Work through problems
  - Calmly, systematically, logically, orderly
- Optimized debugging
  - Narrow down search space as much as possible
  - Reach successful resolution as soon as possible
- Working Independently
  - Learn to do as much as we can before involving others
  - Solve most issues without involving others
- The art of asking for help

Debugging Skills cont.

The art of asking for help

- Timing: Try “a while”, then ask
- Who: NOT the instructor! Ask peers first on the channel, e.g., datasci-205 or datasci-205-schioberg
- How:
  - Send code as a **code snippet** in slack (or the whole file).
  - NO screenshots unless website content!!
  - Say what you were trying to do with some details. Give context
  - Copy paste the whole error message into a **code snippet** in Slack

Reminder on general rules

- Treat your peers like coworkers:
  - Respect boundaries!
  - Your project submissions have to be your own work!
  - Work together but submit your own solution!

Data Engineers

- Design, create, and deploy data pipelines.
- Design, create, and deploy data serving layers

Data Scientists

Process the data using sophisticated mathematical/statistical models, artificial intelligence, machine learning, deep learning, etc.

Full Stack Data Scientists

Data scientists who can do their own data engineering

Breakout: Google jobs in the data science/data engineering field

- What do you notice?
  - Skills asked
  - Size of the company
  - Separation of data science vs engineering vs full stack

Degree of Separation of Roles

- Data engineers separate from data scientists
  - Data scientists need to work effectively with data engineers
  - Breakout exercise later today

- Full stack data scientists
  - Typically smaller companies and startups
  - Although becoming popular at larger companies

Learning How to Learn is the Best IT Skill to Have

- Technology and products come and go at fast pace
- Must constantly learn new skills
  - Videos, tutorials, books, etc.
- **Bleeding edge technologies** are more challenging to learn than established technologies
- Data science often requires bleeding edge
- Frustration and confusion in learning are not always a bad thing
- Breakout exercise later today

Cloud

Brief overview (we will spend more time in coming weeks)

- VM: emulation of physical hardware computer
- Containers: emulation of OS
- Object store: scale up storage, AWS S3, GCP Store, Swift, etc.
- Elastic computing: increase / decrease computing resources on an as needed basis
- Managed services: ready to use software, everything handled for us, pay as you go

Breakouts

GitHub => ucb_mids_w205_repo => breakouts

(time permitting, we may not get to all of them)

Summary

- Find your way around the class material
- Get your VM set up -> labs, OHs, Slack
- Grading
- Debugging
- General information about data engineering