Fall 2017 - W205 – Storing and Retrieving Data
Week 1 Live Class Session Agenda
Kevin R. Crook

- Introductions
  - Instructor
  - Students (round robin)
    - Name
    - Where you are from
      - Now
      - Originally
    - Professional Background
      - Education history
      - Work history
      - Currently working in data science or a related area now?

- KevinCrook.com

- Asynchronous Material
  - Videos
    - Most important
    - Student slides are provided on my website
    - My notes are provided on my website
  - Readings
    - Importance varies
      - Some are landmark, but dated
      - Some are theoretical, videos and/or labs may explain them better
    - Links are provided on my website – matches the syllabus – ISVC out of date
  - Quizzes – least important, some not working

- 2 Safari books that don't appear to be available in the Berkeley online library of electronic books (Safari has multiple levels)
  - Marz, N. & Warren, J. (2015). Big data: Principles and best practices of scalable real-time data systems. Manning, Sections 1.4-1.10. (Safari)
  - Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufman,
    - Chapter 1, pp. 1-35. (Safari)
    - Chapter 3, pp. 83-120.Read the following sections: 3.1, 3.2, 3.3.1, 3.4.8-3.4.9 Optional : 3.3.2-3.4.7, 3.5 (Safari)
    - Chapter 5, pp. 187-194, 210-218.Read the following sections: 5.1 Optional : 5.2- 5.5 (Safari)

- Review accounts and software:
    - Berkeley Library – connection to Safari
    - Slack.com
    - Amazon Web Services (AWS)
    - Windows users
        - PuTTY
    - MacIntosh users
        - Terminal or similar
    - GitHub.com
        - Videos are now on KevinCrook.com
        - Education discount
            - github.com/academic
        - This next week, please try to do the following:
            - Create a GitHub repo called "w205_2017_fall" as spelled out in the instructions on KevinCrook.com
    - Tableau
        - Students will need an academic version of Tableau in a few weeks. It would be best to get it now so it's ready to go when you need it.
        - Don't get the 14 day free trial – get the academic version. Check the expiration date and make sure it's a year out.

- Things you will need to come up to speed quickly for success in this class. All are required for Exercise 1, so please be sure you come up to speed in time to complete Exercise 1.
    - Python Programming
        - Should have had the Python class or equivalent
    - Amazon Web Services
        - Should be comfortable with launching instances, connecting to instances using keys, safe shutdowns, safe startups, creating security groups aka firewall rules, attaching EBS volumes, creating AMIs as a save point
    - Linux Command Line
        - At this point should be comfortable with:
            - Files: creating, deleting, copying, moving, renaming, finding permissions, changing permissions
            - Directories: creating, deleting, copying, moving, renaming, finding permissions, changing permissions
            - Users: changing to w205 with login shell, exiting w205, differences between root and w205
            - Multiple login sessions to the same instance
    - Bash Shell Programming
        - May want to go through the chapters in my recommended book
    - Data Modeling using Entity-Relationship Diagrams (ERDs) in Third Normal Form (3NF)

- SQL
  - We are moving up Unit 7 and Lab 5 to give students SQL earlier
- HDFS
  - Command Line:
    - Directories: listing, creating, deleting
    - Files: copying files into HDFS, deleting
  - Web interface:
    - Starting, viewing directories and files

- Review asynchronous material – Unit 1 – Course Introduction and Architecture

- Schedule
  - 1 page PDF on KevinCrook.com

- Lab 1
  - Due Tuesday, September 12 at 11:59 pm PT
  - Want to keep all 3 sections in synch
    - Weird schedule
      - 5 & 6 Thursday
      - 4 Tuesday
      - Start with Thursday before Tuesday

- Live Class Sessions
  - Suggestions needed
  - MIDS guidelines
    - Student should all actively participate
    - No student should be sitting idle and not saying anything
    - Breakouts are the recommended way to achieve students participating
      - If some students are more knowledgeable than other students, the more knowledgeable students can teach the others. Often teaching others is a great way to enhance your own knowledge
  - My thoughts:
    - Not enough time to rehash 2 - 3 hours of the asynch material in 1.5 hours
    - Not enough time to work a lab in 1.5 hours
    - Use the live class session
      - "Soft Skills"
        - Students frequently complain that the asynch material is too theoretical
        - Breakout session discussing how to apply the theoretical material from the asynch material to real world problems
      - "Hard Skills"
        - Instructor showing a technical skill while students follow and ask questions –
        - Not use the breakout format - however, it is not participatory, by everyone, we will have to use the breakout format

- Office hours after class