Summer 2017 - W205 – Storing and Retrieving Data
Week 3 Live Class Session Agenda
Kevin R. Crook

- Review the schedule
  - Please note we are covering "Unit 7 – Querying Data" out of order to give students some exposure to SQL and ERDs earlier so they will be prepared for Exercise 1 which is coming up.
  - This coming week:
    - Lab 5 – Working with Relational Databases (using PostgreSQL)
    - Unit 3 – Structure and Organization
  - Labs 3, 4, 5 are all distributed in GitHub as students asked if they could work ahead during the holiday weekend. Please remember that current week's assignments get priority for slack and office hours.

- KevinCrook.com
  - Added the following:
    - My personal opinion notes from the asynchronous material
      - Units 1, 2, 3, 7 are completed, others are just outlines, will repost when more are completed
      - Feel free to disagree
      - Let me know if you feel something is incorrect or misleading, but remember I want to keep them to 2 or 3 pages max
    - Student slides from the asynchronous material
    - Links to the readings
    - Videos for Lab 2
    - Videos for Labs 9 and 10 coming. Lab 9 being reworked.

- 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    - Links are provided on my website – matches the syllabus – ISVC out of date
    - 2 Safari books
      - May want to go ahead and read or copy these if you don't have easy access to Safari:
        - 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)
    - Quizzes – least important, some not working

- 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
        - Was everyone able to get create an account and add the educational / academic discount?
        - This next week, please try to do the following:
            - Create a GitHub repo called "w205_2017_summer" 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.

- ***Reminder:***
  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)
        - Cover this week and next week
    - SQL
        - Cover this week and next week
        - Practice now that you have PostgreSQL – consider the PostgreSQL tutorial if you are new to SQL after completing Lab 5
        - Next week is Hive SQL – consider the tutorial after completing Lab 3
    - HDFS
        - Command Line:
            - Directories: listing, creating, deleting
            - Files: copying files into HDFS, deleting
        - Web interface:
            - Starting, viewing directories and files

- Review asynchronous material – Unit 7 – Querying Data

- Break Out Exercise
    - An non-profit area where Data Science can make the world a better place
    - Give an example of an old school company, can be Fortune 500, that has been around for at least 30 years, that has become a Data-Driven Organization

- Lab 2 – any questions?

- Lab 5 – go over (time permitting)

- Office hours after class