Summer 2017 - W205 – Storing and Retrieving Data
Week 4 Live Class Session Agenda
Kevin R. Crook

- Review the schedule
  - This coming week:
    - Lab 3 – Defining Schema and Basic Queries with Hive and Spark
    - Unit 4 – Data Lakes: Storage and Maintenance
    - Exercise 1 – Data Science Study on the Quality of Care for Medicare Patients
      - Get started after you finish Lab 3
      - To be on track this week you need to complete the following at a minimum:
        - Synch GitHub with the instance you will be using
        - Create the directory loading_and_modeling in GitHub as specified
        - Be sure to frequently check code into GitHub so you don't lose work
        - Write a first cut of the Bash script load_data_lake.sh
          - Bring down the zip file
          - Unzip it
          - Rename and remove the first line of the required files
          - Create HDFS directories for the files and copy the files to HDFS
        - Write a first cut of hive_base_ddl.sql
          - Create schema-on-read specifications in Hive for each of the files
        - Be able to run queries against the tables – verify the data against the files to make sure you have everything loaded correctly
        - Read the data dictionary PDF, run queries against the tables to gain an understanding of the data
        - First cut of your ERD and associated analytics
  - Back on sequence going forward, labs and unit in sequential order

- Review accounts and software
  - Last time to review – assume everyone has had 3 weeks to get access to everything
  - Slack.com
  - Amazon Web Services (AWS)
  - Windows users
    - PuTTY
  - MacIntosh users
    - Terminal or similar
  - GitHub.com
  - Tableau

- Coming up to Speed
    - Last time to review this
    - 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
        - Will be needed this week
    - Data Modeling using Entity-Relationship Diagrams (ERDs) in Third Normal Form (3NF)
        - Cover this week and next week
    - SQL
        - Should be up to speed after practicing with PostgreSQL
        - Hive SQL this week's 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 3 – Structure and Organization

- Lab 5 – any questions?

- Lab 3 – go over

- Exercise 1 – go over

- Office hours after class