Summer 2017 - W205 – Storing and Retrieving Data
Week 5 Live Class Session Agenda
Kevin R. Crook

- Change of direction for the live class sessions
  - Weeks 1, 2, 3, 4
    - Focus on getting software loaded
    - Focus on getting everyone up to speed with AWS, GitHub, Linux command line, HDFS, Bash shell programming, SQL, ERDs, etc.
    - More coverage of asynchronous material as it's usually the heaviest learning curve and foundational to the remaining material in the course
  - Week 5 going forward
    - Focus more on interactive content during class
    - Focus more on exercise 1, exercise 2, and the project
    - Less coverage of asynchronous material

- This coming week:
  - Lab 4 – An Introduction to Apache Spark and Spark SQL
  - Unit 5 – Data Ingestion: Storage and Maintenance

- Exercise 1
  - To be on track, everyone should have already completed the following at a minimum (otherwise you are behind).
    These steps are needed so you can start your analytics:
    - Synch GitHub with the instance you will be using
    - Create the directory loading_and _modeling in GitHub as specified
    - Frequently checking code into GitHub so you don't lose work
    - Have a first cut of the Bash script load_data_lake.sh
      - Brings down the zip file
      - Unzips it
      - Renames and removes the first line of the required files
      - Creates HDFS directories and copies the files to HDFS
    - Have a first cut of hive_base_ddl.sql
      - Creates schema-on-read specifications in Hive for each of the files
    - Be able to run queries against the tables – verified that data with schema applied matches data in the files – verified that it is correctly loaded
      (for today's break out session, you will be running Hive queries against these tables)
  - To stay on track, before next week, at a minimum, you should have gone through and understand all of the tables from the data dictionary, have explored the data through SQL queries and have enough understanding to start designing your analytics.

- Break out exercise
    - From exercise 1 data
        - each group will take a file
            - group 1 – hospitals
            - group 2 – effective_care
            - group 3 – readmissions
            - group 4 – measures
            - group 5 – survey_responses
        - read the data dictionary to understand what the file layout is
        - explore the data using Hive SQL on the schema on read tables, you may also want to create some schema on write parquet tables as it will be easier to explore the data
            - primary key for the table?
            - Any foreign keys to other tables?
            - Any tables that appear to be child tables of this table?
            - How might it relate to other tables in other ways?
            - How can it be used to answer the analytical questions? Can it be used as is or should it be filtered? Aggregated? Joined? Correlated to other data?
        - present a summary about the file and its data and how it might be used for the analytics for exercise 1

- Review asynchronous material – Unit 4 – Data Lakes: Storage and Maintenance

- Lab 3 – any questions?

- Lab 4 – go over

- Exercise 1 – any questions?

- Office hours after class