Fall 2017 - W205 – Storing and Retrieving Data
Week 9 Live Class Session Agenda
Kevin R. Crook

- Schedule
  - Lab 6 – due Tuesday, 11/7/2017 at 11:59 pm
  - Asynchronous for next week
    - Unit 10 – Cleaning Data, Entity Linkage, String Clustering and Fuzzy Methods for Merging Datasets
  - Project
    - Milestones
      - Proposal Report
        - In class - 11/2/2017 or 11/7/2017
        - 10 minute presentation per project team
      - Progress Report
        - In class – 11/16/2017 or 11/28/2017
        - 10 minute presentation per project team
      - Final Presentations
        - In class – 12/14/2017 or 12/19/2017
        - All materials must be checked into GitHub repo prior to class time
      - Weeks in which we do not have a formal milestone – we will spend 3 to 5 minutes going round robin through the project teams
    - Focus
      - Covering 2 or 3 of the V's
        - Volume, Variety, Velocity
      - Steel Thread
        - Minimal thread working end to end as soon as you can
        - Add to functionality as incrementally as possible
      - Scale Out
        - Decide how much scale out you plan to build in your prototype
        - Balance
          - Build as much scale out as you can
          - But, don't jeopardize having a working prototype at the end of the semester
        - Address any scale out that you didn't build
          - Show a path to get there

        - No Scale Out
            - Storage layer – PostgreSQL, etc.
            - Processing layer – SQL, Python, etc.
            - Machine Learning – Python with Scikit-learn, R, etc.
            - Streaming media – Python single threaded API, etc.
            - Data Visualization – freebee as Tableau has a scale out server!
        - Full scale out
            - Storage layer – Hive, Redshift, etc.
            - Processing layer – Spark, Hadoop MapReduce, etc.
            - Machine Learning – Spark MLlib, Mahout, etc.
            - Streaming Media – Storm, Heron, Spark Streaming, Kafka, etc.
            - Data Visualization – freebee as Tableau has a scale out server!
    - Machine Learning
        - Not required, but most teams have at least one member who has experience
        - Steel thread – predictions can be random selection, median, mean, etc.
        - Simple algorithms and enhance time permitting
    - Data Visualizations
        - Easy to add with Tableau

- Today in class
    - Project
        - Milestone – Proposal Report
    - Exercise 2
        - Walkthrough
    - Streaming Media
        - Storm
        - Twitter Heron
        - Spark Streaming
        - Kafka – putting Kafka in front of streaming media

- Next Week in Class
    - Tableau exercise – all the way through Tableau – much more detail than lab 7