Fall 2017 - W205 – Storing and Retrieving Data
Week 7 Live Class Session Agenda
Kevin R. Crook

- Schedule
  - Exercise 1 - Due Tuesday, 10/31/2017 at 11:59 pm
  - Lab 6 – due Tuesday, 11/7/2017 at 11:59 pm
  - Asynchronous for next week
    - Unit 8 – Exploring Data
  - 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

(next page)

- Today in class
    - Instructor led exercise
        - Message Queuing (MQ) Architectures
            - Vendors: IBM MQ Series, MSMQ, Tibco, etc.
            - Open Source: Kafka – for Big Data Speed Layer
                - Classic MQ Architecture - Topics, topics of topics, publisher, subscriber, etc.
                - Front end for streaming data
        - Massively Parallel Processing
            - Lists
            - Lambda
            - DAGs
                - Old school Massively Parallel Processing (LISP – circa late 1980's, early 1990's) – open ended DAG, but write it yourself
                - Hadoop MapReduce – fixed DAG – can't do everything – have to chain DAGs
                - Spark – open ended DAG, Spark does it for you
        - Hadoop MapReduce
            - Hadoop Essential book
                - Great, short, easy to read overview of Hadoop Ecosystem
            - Hadoop 2 Quick-Start Guide… book
                - Chapter 5 – Hadoop MapReduce Framework
                    - Fixed DAG – take it or leave it
                    - Can combine – but have to load memory each time we DAG
                    - DAG Steps:
                        - Input
                        - Split
                        - Map Step
                        - Combiner Step
                        - Shuffle Step
                        - Reduce Step
                - Chapter 6 –
                    - Example of word count using Java / Scala
                    - Python not able to see Java data structures – only streams interface

- Spark
    - Example of word count – 1 line
    - Sam's Teach Yourself Apache Spark book
        - Hour 9 – Functional Programming with Python
        - Hour 10 – Working with the Spark API (Transformations and Actions)
            - Expands on all the functions we learned in lab 4
            - Bike Rental Example
        - Hour 16 – Machine Learning with Spark
            - Classification – Decision Trees, Naïve Bayes
            - Clustering – k-means
    - Spark created sophisticated DAGs for you:
        - https://databricks.com/blog/2015/06/22/understanding-your-spark-application-through-visualization.html
        - Keeps data in memory, does not keep reading back into memory like Hadoop MapReduce
- Data Algorithms book
    - "real world examples"
        - in case you are tired of counting words and classifying Irises by petal lengths
    - Hadoop MapReduce
    - Spark
        - Mix of Scala and Python
- Advanced Analytics with Spark, 2nd Edition
    - More "real world examples"
    - Scala – not much Python