CSE 6331 Cloud Computing
Summer 2016, © DL, UTA, 2016

Programming Assignment 5,6,7
Data Science, Analytics

Introduction to Map-Reduce (Hadoop),
Machine Learning (Clustering), and Visualization

Description:

The following are interesting data sets:
http://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php (all earthquakes)
https://www.ncdc.noaa.gov/cdo-web/ (various detailed weather)
https://github.com/fivethirtyeight/data/blob/master/pollster-ratings/pollster-stats-full.xlsx (small but interesting)
https://research.stlouisfed.org/fred2/ (economic data sets)

You should use several, both smaller and larger, of different types.

Map - Reduce

1. Get, install, try Hadoop.
   (downloads: http://www.apache.org/dyn/closer.cgi/hadoop/common/,
   more at: http://hadoop.apache.org/)
   Or, use a prebuilt image or, use the AWS service
   But, you will need to use Hadoop on a cloud service provider (Google, AWS)
2. Interesting data sets have at least 100 thousand tuples up to a few million tuples. At these web-sites there are schema/meta-data describing the data.
3. Using earthquakes as an example, we would like to know: are magnitude 1 or 2 (or others) increasing, week-by-week? Day-by-day? Is there a relationship between magnitude and depth? Location and magnitude? Week-by-week?
   We want to take large amounts of data and categorize into groups (ranges), for example magnitude groups (1-2, 2-3, 3-4,…) or latitude groups (20-25, 25-30, …) using Hadoop.
4. Try with different numbers of mappers and reducers. (1 mapper, 1 reducer (1,1), then: (2,1), (2,2), (10,1), (10,2)
   Run with 1, 2, and 3 instances.
5. "Instrument" (time) running.

Machine Learning

1. Using Weka, Python libraries, or Java libraries, which contain k-means cluster methods, try clustering your Map-Reduce output (using two dimensional data, or only two of the dimensions) into clusters of 3, 7, 10, and 20.
   Which give the best (most meaningful) result?

Visualizing

1. Using D3.JS (or similar) packages
2. Import the output from the clustering above, and show scatter carts, bar charts, bubble charts, with annotated color, through your web browser on your screen.