CS 6343: CLOUD COMPUTING
Project #1

- Hadoop MapReduce
  - http://hadoop.apache.org/docs/r1.0.4/mapred_tutorial.html
  - You will need to install Hadoop system on your own PC or on Lab PCs.
- Input files
  - A sanitized crime database from http://utdallas.edu/~ilyen/course/cloud/for15f/data.zip
    - The source of this crime database was from http://www.police.uk/data
  - An example input file: http://utdallas.edu/~ilyen/course/cloud/for15f/proj1-input.txt
- Write a MapReduce program to compute the total crime incidents of each crime type in each region
  - Region definition
    - Crime location is defined on a coordinate system (East, North)
    - East and North are defined by a 5-digit numerical value
    - Region definition 1: use the first digit of the coordinates only to define a region
      - (5xxxx, 7xxxx), (5xxxx, 3xxxx), (8xxxx, 6xxxx), each is one region
      - Supposedly there are 100 regions, but not all the numbers appear in the files
    - Region definition 2: use the first three digits of the coordinates to define a region
      - (535xx, 726xx) is one region
    - Consider other region definitions
  - Crime types include: Anti-social behavior, Burglary, Criminal damage and arson, Drugs, Other theft, Public disorder and weapons, Robbery, Shoplifting, Vehicle crime, Violent crime, Other crime
- Study Hadoop behaviors from its logs
    - How the file(s) are distributed over the nodes
      - One large input file or many small input files
    - How the mapper and reducer tasks are distributed over the nodes
    - The performance of map, reduce, and shuffling&sorting phases, including execution time, memory usage, etc.
    - How the system handle errors
    - …
- Submission
  - Submit through e-learning
    - Your source code
    - Your report in a “doc” file and name it as “report.doc”
      - Discuss your findings about Hadoop behaviors from the logs