CS 6301: Special Topics in Computer Science
CLOUD COMPUTING
Project #1

- Hadoop MapReduce
  - http://hadoop.apache.org/docs/r1.0.4/mapred_tutorial.html
- Download input files and copy them to HDFS
  - Try out different inputs
    - Provide a single file with a very large data set and see how Hadoop partition the file and place them on the slave nodes
    - Provide a large number of smaller files
  - Input file available at http://www.utdallas.edu/~ilyen/course/cloud/for14s/data.zip
  - Acknowledgement: Dataset is from http://www.police.uk/data
- 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
- Under different settings, study Hadoop behaviors from its logs
  - Settings
    - One large input file or many small input files
    - Different number of virtual nodes (same as or much larger than the number of physical nodes)
    - Different number of mapper tasks (defined through InputFormat)
    - Different number of reducer tasks (JobConf.setNumReduceTasks(int))
    - Introduce errors in a couple of input files
    - …
  - Study
    - How the file(s) are distributed over the nodes
    - 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 procedure
  - Go to university systems, such as apache, cs1, cs2
  - Prepare a project directory which includes
    - Your source code
    - Your report in a “doc” file and name it as “report.doc”
      - Discuss your findings about Hadoop solutions and their impacts on performance
      - Discuss your comments

- Go to your project directory and issue the command: ~ilyen/handin/handin.cloud
- Now you are done with the submission
- Please do not wait till last minute to submit your program. You can make as many submissions as you wish and the later submission overwrites the earlier submission. You can try out way in advance to see whether the submission script works for you. If there is any problem, please email TA to resolve the submission problem