Big Data Mining: HW#1

By J. H. Wang

Mar. 9, 2018

Programming Exercise: Hadoop/Spark Setup & the First MapReduce Program

- Goal: Setting up your Distributed Environment and Writing your first MapReduce program
- Input: Numeric data (to be detailed later)
- Output: Results of simple statistics (to be detailed later)

Tasks and Data

- Tasks
  - Some simple statistics on numeric data (to be detailed later)
  - You have to do it in a distributed way
- Data: to be announced later
- You have to submit the generated output
- You also have to output the efficiency (running time) of each task

Note on Programming Exercises

- Programming exercises can be done as a team (at most two persons per team)

- You can use any programming language in Hadoop or Spark to implement
  - Java, Scala, Python, or R

Homework Submission

- For implementation projects, please submit a compressed file containing:
  - Your cluster environment setup
    - How many PCs, what spec, network setup, …
  - Your source codes
  - The generated output
  - Documentation on how to compile, install, or configure the environment, and also the detailed responsibility of each member
- Due: 2 weeks (Mar. 23, 2018)

Evaluation of Results

- In completion of each of the tasks, you get part of the scores
- Correctness of Output
- Efficiency
- Please specify the environment setup of your (virtual) machines
- You might need to demo if your program was unable to run

Questions or Comments?