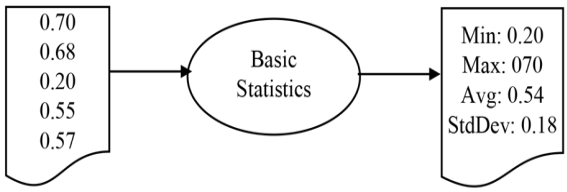
Project 1: Basic Statistics with Hadoop
Cloud Computing
Spring 2017

Professor Judy Qiu

Problem statement

The idea of this project is to get you started with Hadoop and the MapReduce concept. You may have already looked at the WordCount example, both serial and Hadoop implementations. This problem is similar to WordCount except that you will be computing the basic statistics such as min, max, average, and standard deviation of a given data set.

The input to the program will be a text file carrying exactly one floating point number per line. The output should include **min, max, average, and standard deviation** of these numbers.



Files

A test input file is available as a separate attachment. The statistics values for this input are **Min: 0.01 Max: 0.99 Avg: 0.50 StdDev: 0.2817**

Deliverables

You will need to complete the source code and write a report. Zip your work into a file with the name username_project1.zip (replace 'username' with your own) and submit the following:

- Complete source code
- A document with the following details:
  - Transformation of data during the computations, i.e. data type of key, value
  - The data structure used to transfer between Map and Reduce phases
  - How the data flow happens through disk and memory during the computation

Evaluation

The point total for this project is 5.

- Correctness of the source code (2 points)
- Completeness of the report (3 points)