BdiF Homework 1
The two problems were complimented in one single cpp file (HW.cpp). The code was tested in local machine with openmpi, and also at HPC.

The steps I tested with HPC:

a) `<xiaobo.he@penzias:BigDataHW1> mpic++ -o scrub HW.cpp`
b) `<xiaobo.he@penzias:BigDataHW1> mpirun -np 16 scrub "data10k.txt" 2>&1`
c) Then get four output files: signal.txt, noise.txt, log.txt, NormalityTest.txt

(1) Scrub data:

The basic idea for separating the original data into signal and noise parts is to treat the raw data as noise if it is far away from the distribution center in the tails.

The approach:

a) Load the input file, equally separate to the nodes.
b) The input record data string stream is converted to a vector of records. The first line of every data buffer is thrown out because the possible broken data during cutting to different nodes.
c) Choose a slide window with size of 50 records, compare the time tick, and sort the data in the small window.
d) Move the slide window to next data record, and sort the new data to right position. Go through the whole vector.
e) Go through the vector, calculate the mean and standard error for record prices. During the process, set the data as noise if the recorded price jumps too much in one tick, which I set 10 times, comparing to one tick before and after.
f) Adjust proper parameter for how much data to treat as noise by calculate how far sway the data price from the price mean. This major contribution for separating signal and noise. The implement is realized in function ScrubRecord(), with default parameter as 0.30. The default parameter was calibrated with sample data file data10k.txt and check with plotting in R. The noise is about 10% of the original data.
g) Get two separating record vectors: signal and noise. Output to two files: signal.txt and noise.txt.
h) The operation time was record in log.txt.

(2) Test normality.

Jarque-Bera method is used to test data normality. The formula comes from these two website:
http://www.itl.nist.gov/div898/software/dataplot.html/refman1/auxillar/jarqbera.htm
https://en.wikipedia.org/wiki/Jarque–Bera_test

a) I only use one node to test normality. Once separate original data into noise and signal, take the signal vector do JB test.
b) Go through the signal vector, and calculate the return vector.
c) Calculate the forth moments of the data: mean, variance, skewness, and kurtosis.
d) Calculate the p value according to JB test formula.
e) If the p value is small than 9.21, we will accept the normal hypothesis with confidence level 99%. This threshed value was checked in R.
f) This function is implemented in function NormalTest( ).
g) The result is recorded into file NormalityTest.txt