A3

Abhishek Ravi Chandran, Chinmayee Vaidya

February 6, 2016

Design

In this assignment we measure the performance difference between different type of jobs. We have 3 solutions, single threaded job, multi threaded job and then a map-reduce job. The map reduce job is tested both in pseudo distributed mode and distributed mode.

All the jobs read all the input files and process the data sequentially. For the Multi threaded solution, a new thread is created for each file that is present in the folder.

In the map-reduce program we have used a custom writable class(CarrierMonthWritable) to group the data with the composite key month and carrier

For the fast median calculation we have implemented quick select algorithm, which uses randomization to find the median without having to sort the whole data. it has a worst case of O(nˆ2), but the chances of getting this performance is greatly reduced as we choose the pivot randomly. It has a best case performance of O(n) and an average of O(n)

The program takes a mode input, where -mn is to calculate mean, -md is to calculate median and -fm uses quick select to calculate the median.

All configurations are run with 25 files as input. We can see that the multi threaded program is much faster than the single threaded one. In the same way the distributed version is faster that the pseudo distributed version.

The graph below shows the performance of each of the configurations for 25 files(~0.6 GB) of data:

Graph:

Key:

i-fastmed : Single thread solution calculating median using quick select

i-mean : Single thread solution calculating mean

i-median : Single thread solution calculating median

ii-fastmed : multi threaded solution calculating median using quick select

ii-mean : multi threaded solution calculating mean

ii-median : multi threaded solution calculating median

iii-fastmed : pseudo distributed solution calculating median using quick select

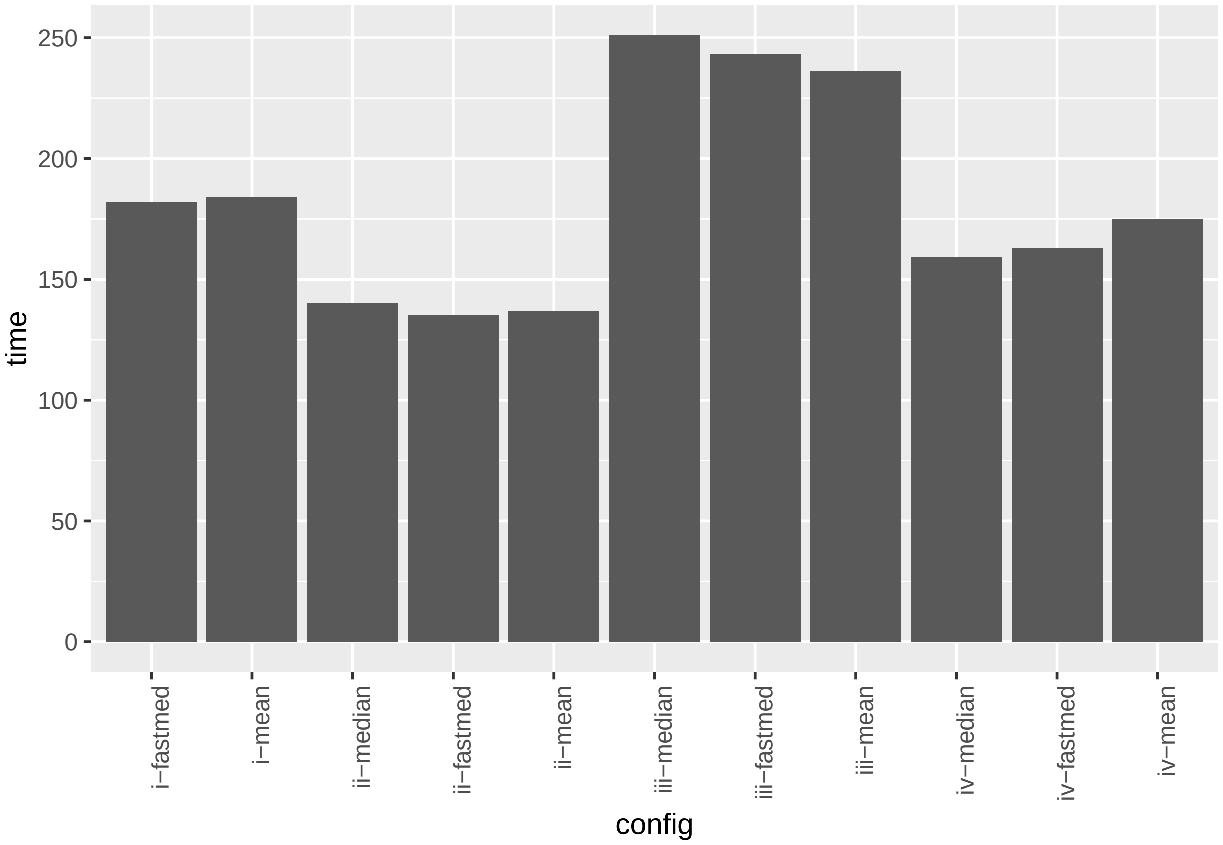
iii-mean : pseudo distributed solution calculating mean

iii-median : pseudo distributed solution calculating median

iv-fastmed : distributed solution calculating median using quick select

iv-mean : distributed solution calculating mean

iv-median : distributed solution calculating median



time
250
200
150
100
50
0
i−fastmed
i−mean
ii−median
ii−fastmed
ii−mean
iii−median
iii−fastmed
iii−mean
iv−median
iv−fastmed
iv−mean
config