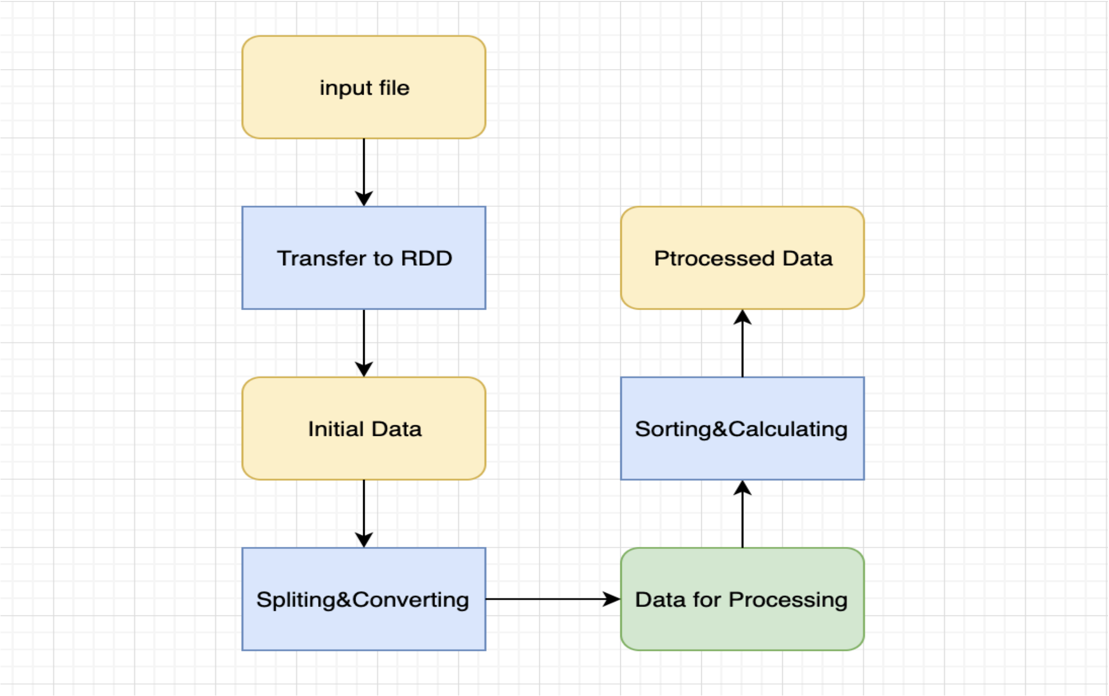
Comp6452 Ass2

Student ID: z5184142

Student Name: Jiachen Li

For this assignment, the solution will be described into two parts. First is data integration. After load the file, the program will take useful information from the file. First split every row of the RDD and take the first element and the last element after splitting. Every last element which is data size will be put into bitstranfer function, converted from string to int, and transfer different unit of measurement to Bytes.

After that, the program will begin data process. First using groupbykey to combine the values (data size) together. Then sorted them in ascending order. Next is calculating the mean of the data size. The program will sum all the data size together and divide the number of lines to get the result, and then the program will use the mean to calculate the variance of data size. Finally, the program will integrate all the request data to one String and then write to the file.


input file
Transfer to RDD
Initial Data
Spliting&Converting
Data for Processing
Sorting&Calculating
Ptrocessed Data