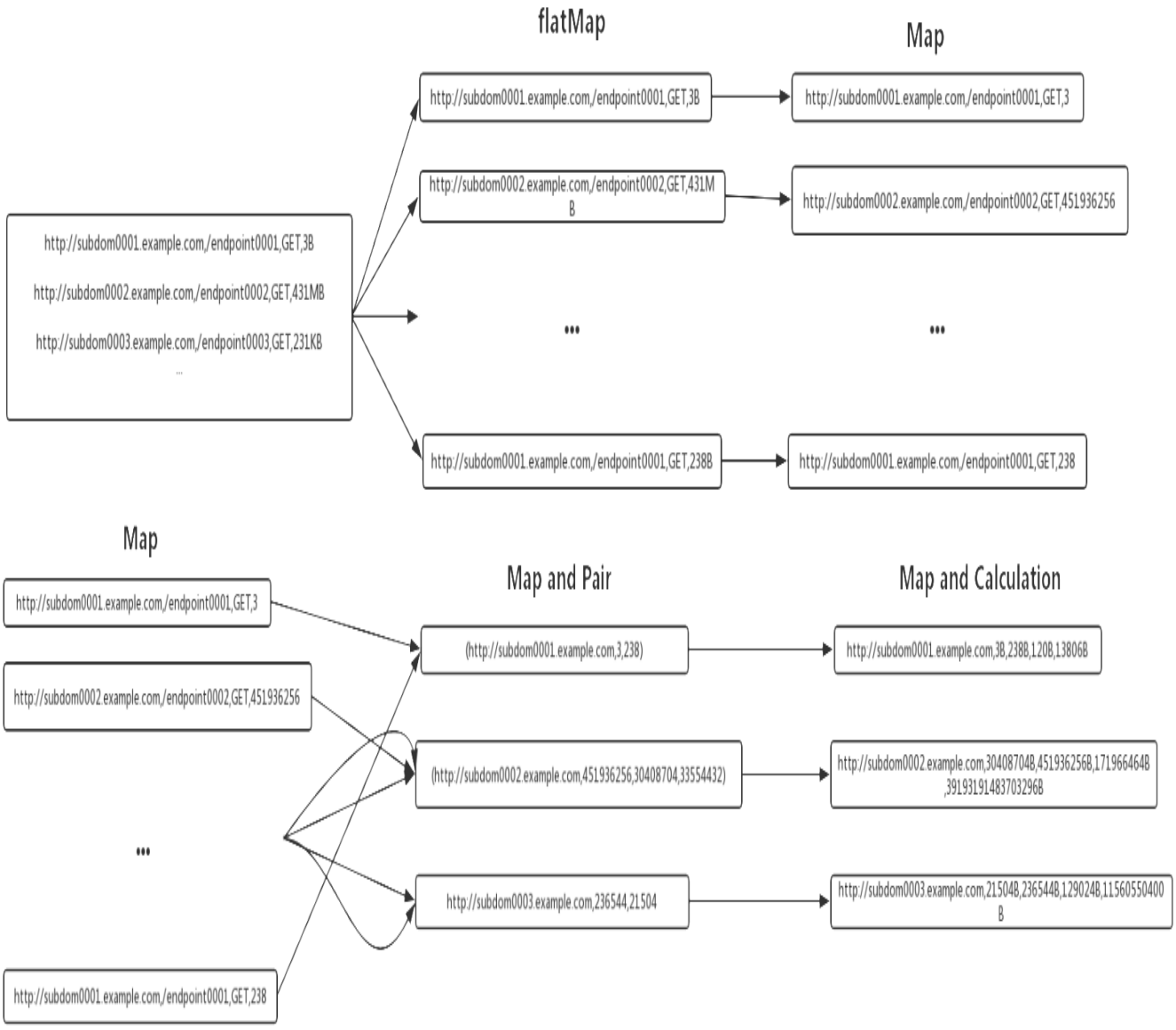
COMP9313 - 2019T2 Assignment #2 (Spark)

Student Number: z5195349 Student Name: Wenxun Peng

The solution of this assignment is as follow:



First, I used flatMap to split the whole txt file to lines and each row contains a https record. Secondly, I used map and other function to convert all “MB” and “KB” to “B” (MB = 1024*1024 B, KB = 1024 B). Third, using map again and the groupByKey to get the pair (The last part in the first diagram (Map part) is the same as the first part in the second diagram). Finally, using the map and doing some calculation (max, min, mean and variance value).

The RDD operation

The RDD I used in this assignment includes flatMap, map, groupByKey and other some input and output function in RDD. The map operation is the most important part to classification and calculation.