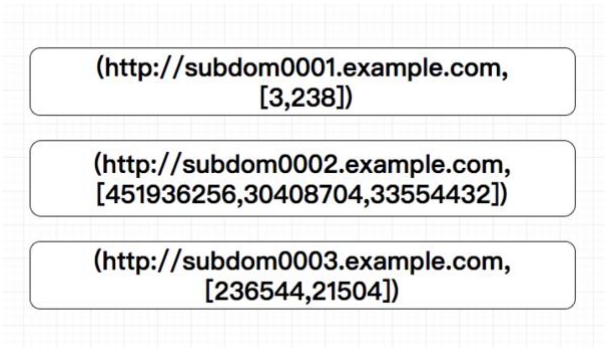
Assignment 2 Solution

This assignment is asking us to calculate some numbers of different website with multi-payload.

I looked lecture's slides and the book: Learning Spark (Chapter3 and 4), they are very helpful, especially in Learning Spark chapter 4 (page 42 – 49), which tells me how to do with a pair RDD.

First, I create Rdd from outside file, and do some pre-processing work like remove the blank line and remove some invalid information like "endpoint" and "GET", turning MB and KB to Bytes number. Now I have turned my Rdd to pairRDD.

Then I transform it to a new pairRDD, the same key coutains different value, by using groupBy().



Then, for each key, I could calculate its min-value, max-value, mean-value and variance-value.

Finally, put it into a file by using saveAsTextFile().