COMP 9313 - 2019T2 Assignment 2
Student: z5146092

The Implementation of Data Integration

At first, each line of the content in provided data source is extracted and split by commas, then the first element (URL) and the last element (data amount) are extracted and used only. A data unit conversion function is used for converting all data units (e.g. KB, MB) to Bytes (B), for normalization. Then the key-value pairs are grouped by applying groupByKey function. As a result, URLs with all of the data amount are stored in a map for further calculation. The whole process is described in Fig 1.

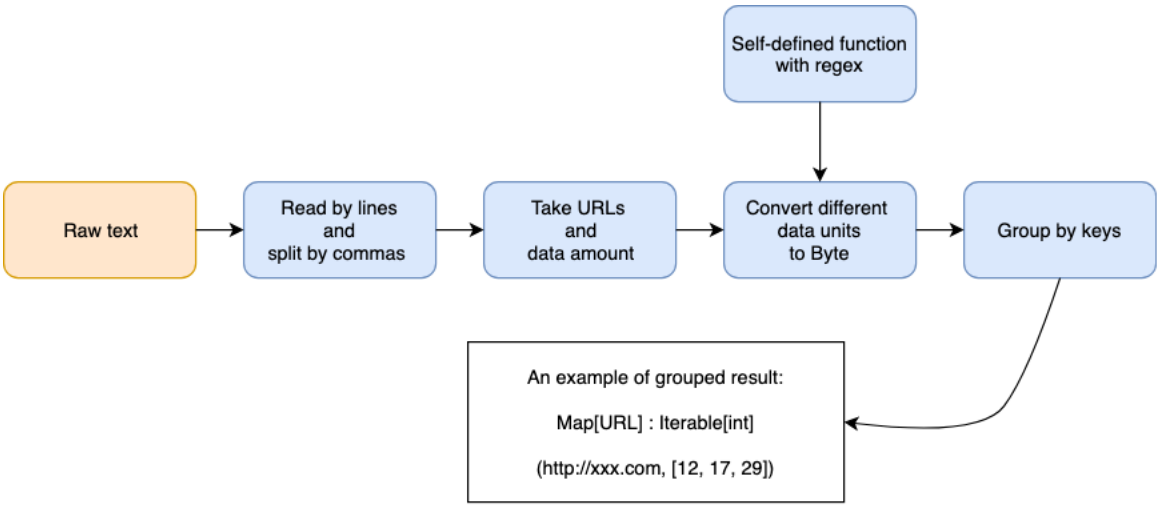

Fig1. Data Integration

The Implementation of Calculation of Min, Max, Mean, Variance

As the data are grouped to key with iterable int array pairs, the first task is to sort them in an ascending sequence with using sortWith function. After sorting, it is easy to find the minimum and the maximum value by calling line.head and line.last values. The mean value is calculated by sum of all the values of the same key, dividing by the length of value array. Then variance is calculated by the sum of all the squares of values minus the mean, then dividing by the length of value array. The whole process is described in Fig 2.

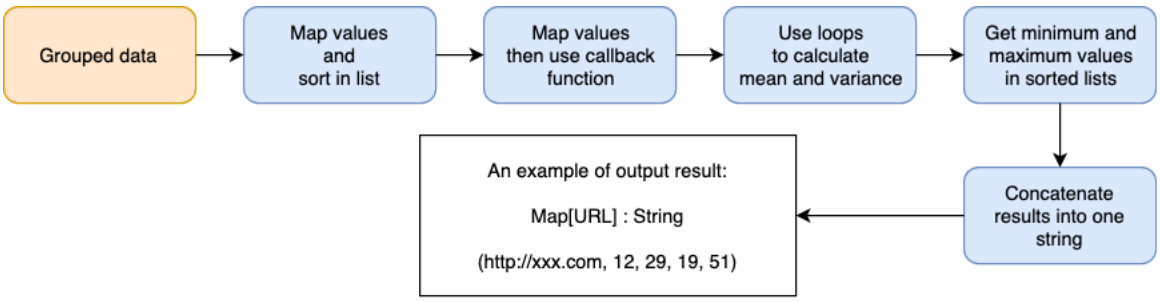

Fig2. Calculation Values

Other Implementation

For the result generation, the result string is generated by concatenating minimum, maximum, mean and variance, rather than return a tuple of four values with the key (URL). Therefore, the final output is created by such key value pairs. After that, another map function is used to remove redundant parentheses, then coalesce function is used to instruct that the output result should not be in several partitions. Finally, the output result is generated by saveAsText function.