COMP9313 - 2019T2 Assignment #2 (Spark)

z5147810 Meng Lingxu

Environment: CSE Lab environment, Spark-shell, Scala
Working period: 2019.07.08 - 2019.07.15

1. Input File Processing

Each line in the input file contains four arguments, which are domain URL, endpoint, HTTP method and payload size. Only the first and the last arguments are needed for the assignment. Firstly, use *sc.textFile* to read the input file and then, map the lines and split each line by a comma, as it is a csv file. Finally, map the split lines and group them by key. We get a map of data that contains domain URL and all payload size under this domain. The format is:

("http://xxxxxx.example.com", CompatBuffer(xx,xx,xx))

2. Helper Functions

Payload size is String format ending with byte size ("GB", "MB", "KB", "B")... The function foo() is used to help get payload size in Byte size, as the number of order is 1024, we just need to distinguish last two chars of the string and multiply them in respect.
As we are only allowed to use RDD, some other functions are needed to calculate the minimum, maximum, mean and variance. Those are basic math work.
One thing need to notice is that we should use Long data structure because the payload size could be really huge.

3. Create Output

After all the calculation that we need, we just use key and value in the map we produced in the first part and get all numbers that we need.
And we use *saveAsTextFile* to write the result into a file, with commas to separate each value. To make sure all result lines are written into one single file, *coalesce(1)* function is added before to make sure this point.