COMP9313 – Assignment2

Han YANG – z5140181

Spark

In my solution, there are two main RDD operations, Transformation and Action. In Transformation functions, data in RDD is not evaluated until the action is executed, it just takes an RDD and returns a new RDD. In Action function, all data processes are computed when it is called, and returns a new value.

Firstly, I create an RDD from the file. Then using Transformation function 'map' to extract base URL and Payload as key-value pairs, saving as a new RDD. Next, designing an if-else structure with 'map' to transform the different units (KB, MB) of digits to bytes (B) for each payload. Then I use a Transformation function 'groupByKey' to group all payloads for each URL. After that, I define two functions called mean and variance to calculate the mean and variance for a List, and using 'map' to calculate the minimum, maximum, mean and variance payload for each base URL. Finally, I use an Action function 'saveAsTextFile' to evaluate the data in RDD and output it into the storage system (HDFS).

The whole program process can be seen as below: