UNSW
AUSTRALIA

COMP9313

Assignment2-report

Name: Luyao Zhang
ZID: z5151973

Assignment 2 is an implementation of apache spark which aims to get a file that include given keys' minimum, maximum, mean and variance values. I use the apache spark with Scala language to write my program.

First, I consider to create a list that include the useful data like URLs(as a key), and each data payload as a value.

```
var urls_list:List[(String,Long)]=List()
//clean the useful data
for (it <- items) {
  if(it(0)!="") {

    val size_num = it(3).replaceAll("[A-Za-z]","").toLong
    val size_type = it(3).replaceAll("[0-9]", "")

    if(size_type=="KB"){

      urls_list = urls_list :+ ((it(0),size_num*1024))
    }
    else if(size_type=="MB"){

      urls_list = urls_list :+ ((it(0),size_num*1024*1024))
    }
    else if(size_type=="B"){


      urls_list = urls_list :+ ((it(0),size_num))
    }

  }

}
```

When we get the clean data, we create a RDD and start to calculate the final result.

```
val result_List = Urls.groupByKey().map(x=>{val res_min=(x._2).min+"B";val res_max=(x._2).max+"B";
  var sum:Long=0;var num1=0;var sum1:Long=0;for(i<-x._2){sum=sum+i;num1=num1+1};val avg=sum/num1;val res_mean=avg.floor.toLong;
  for(e<-x._2){val diff=e-avg;val diff_squa=diff*diff;sum1=sum1+diff_squa};val res1=sum1/num1;val res_var=res1.floor.toLong;(x._1, res_min, res_max, res_mean+"B", res_var+"B")})
```

This part I use URLs to group the data, and use min and max method to get the min value and max value. Then, I compute the mean and variance value use the given formula. After this step, save result in a new RDD.

The final part is build a CSV format data. So I rewrite the data.

```
val lines = result_List.map(x=>{val res=x._1+","+x._2+","+x._3+","+x._4+","+x._5;res})
//output the result
lines.repartition(1).saveAsTextFile(outputDirPath)
```

The last step is output the result using saveAsTextFile method.