Assignment – 2

BIG DATA ANALYSIS

Hadoop MapReduce for Climate Data Analytics

NAME - SAHIL
ROLL NO. -107121086
ELECTRICAL AND ELECTRONICS ENGINEERING

Task 1

(a) What will the output pairs for map phase look like?

< 956, "At dawn shey(1) departed" >
< 8, "My mind tried to console me –" >
< 57, " Everything is Maya(2)" >
< 5, "Angrily I replied:" >
< 89, ""Here's this sewing box on the table," >
< 146, "that flower-pot on the terrace," >

Input to map phase

(< "At", 1 > < "dawn", 1 > <"shey", 1> <"1", 1> <"departed", 1>)
(< "My", 1 > < "mind", 1 > <"tried", 1> <"to", 1> <"console", 1><"me",1>)
(< "Everything", 1 > < "is", 1 > <"Maya", 1> <"2", 1>)
(< "Angerily", 1 > < "I", 1 > <"replied", 1> <"1", 1>)
(< "Here's", 1 > < "this", 1 > <"sewing", 1> <"box", 1> <"on", 1> <"the", 1> <"table", 1>)
(< "that", 1 > < "flower", 1 > <"plot", 1> <"on", 1> <"the", 1> <"terrace", 1>)

Output of map phase

(b) What will be the types of keys and values of the input and output pairs in the Map phase?

- Input key : Integer or IntWritable
- Input Value : String (sentence) or Text
- Output key : String (single word) or Text
- Input Value : Integer or IntWritable

(c) What will the input pairs for reduce phase look like?

< "At", { 1 } > < "dawn", { 1 } > <"shey", { 1 }> <"1", { 1, 1 }> <"departed", { 1 }>
"My", { 1 } > < "mind", { 1 } > <"tried", { 1 }> <"to", { 1 }> <"console", { 1 }><"me",{ 1 }>
< "Everything", { 1 } > < "is", { 1 }> <"Maya", { 1 }> <"2", { 1 }>
< "Angerily", { 1 } > < "I", { 1 } > <"replied", { 1 }>
(< "Here's", { 1 } > < "this", { 1 } > <"sewing", { 1 }> <"box", { 1 }> <"on", { 1, 1}> <"the", { 1, 1}> <"table", { 1 }>
< "that", { 1 } > < "flower", { 1 } > <"plot", { 1 }><"terrace", { 1 }>

(d) What will be the types of keys and values of the input and output pairs in the Reduce phase?

- Input key : String (single word) or Text
- Input Value : list/array of Integers or Iterable<IntWritable>
- Output key : String (single word) or Text
- Input Value : Integer or IntWritable

(e) Write map () function for Questions a and b.

```
public static class TokenizerMapper extends Mapper<Object, Text, Text, IntWritable> {

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
                StringTokenizer itr = new StringTokenizer(value.toString());
                while (itr.hasMoreTokens()) {
                        word.set(itr.nextToken());
                        context.write(word, one);
                }
        }
}
```

(f) Write reduce () function for Questions c and d.

```
public static class IntSumReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
        private IntWritable result = new IntWritable();

        public void reduce(Text key, Iterable<IntWritable> values, Context context)
                        throws IOException, InterruptedException {
                int sum = 0;
                for (IntWritable val : values) {
                        sum += val.get();
                }
                result.set(sum);
                context.write(key, result);
        }
}
```

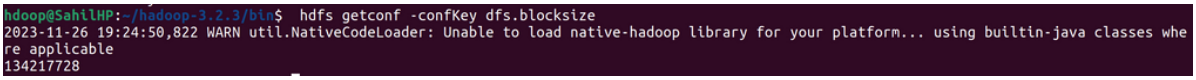
Task 2

(b) How many Map and Reduce tasks did running Word Count on Gberg-100M.txt produce? Run it again on Gberg-200M.txt and Gberg-500M.txt and write your observations. Additionally, run the following command on the cluster: $ hdfs getconf -confKey dfs.blocksize

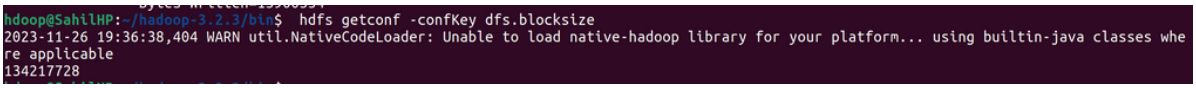
Text file	Map Tasks	Reduce Tasks	Input file Size	Block size
Gberg-100M.txt	1	1	104.3 Mb	134.21 Mb
Gberg-200M.txt	3	1	209.7 Mb	134.21 Mb
Gberg-500M.txt	4	1	524.3 Mb	134.21 Mb

After running given command for each Cluster :

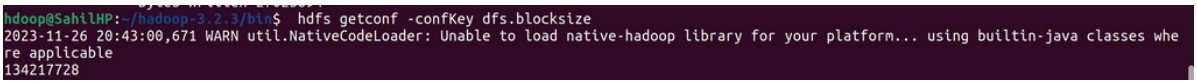
1. Gberg-100.txt



2. Gberg-200.txt



3. Gberg-500.txt



(c) What is the link between the input size, the number of Map tasks, and the size of a block on HDFS?

In above table in part (b), we can clearly see that as input size increases map tasks increases but block size on HDFS remain same for all three input files.