

Assignment No. 2

Subject: Data Science & Big Data Analytics Lab

Name: Yash Dhavale

Class: TE 09 L9

Roll No: 33122

Log File Input

Problem Statement:

Design a distributed application using MapReduce(Using Java) which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the Internet and process it using a pseudo distribution mode on Hadoop platform.

Code:

```
import java.io.IOException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
//import org.apache.hadoop.util.GenericOptionsParser;

public class logfile
{
    public static void main(String [] args) throws Exception
    {
        Configuration conf = new Configuration();
        Job j = Job.getInstance(conf, "word count");
        j.setJarByClass(logfile.class);
        j.setMapperClass(MapForWordCount.class);
        j.setReducerClass(ReduceForWordCount.class);
        j.setOutputKeyClass(Text.class);
        j.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(j, new
Path(args[0]));
        FileOutputFormat.setOutputPath(j, new
Path(args[1]));
        System.exit(j.waitForCompletion(true)?0:1);
    }
    public static class MapForWordCount extends Mapper<LongWritable,
Text, Text, IntWritable>
```

```

        {
            public void map(LongWritable key, Text value, Context
con) throws IOException, InterruptedException
            {
                String line = value.toString();
                String[] words=line.split(" ");

                Text outputKey = new Text(words[0]);
                int time = Integer.parseInt(words[1]);
                IntWritable outputValue = new IntWritable(time);
                con.write(outputKey, outputValue);
            }
        }

        public static class ReduceForWordCount extends Reducer<Text,
IntWritable, Text, IntWritable>
        {
            public void reduce(Text word, Iterable<IntWritable>
values, Context con) throws IOException, InterruptedException
            {
                int sum = 0;
                for(IntWritable value : values)
                {
                    sum += value.get();
                }
                con.write(word, new IntWritable(sum));
            }
        }
    }
}

```

OUTPUT:

1. Display as it is

```

[cloudera@quickstart ~]$ pwd
/home/cloudera
[cloudera@quickstart ~]$ cat > /home/cloudera/ProcessFile2.txt
10.0.0.5
10.0.0.1
10.0.0.5
20.0.0.0
20.0.0.0
10.0.0.5
10.0.0.1
^Z
[1]+  Stopped                  cat > /home/cloudera/ProcessFile2.txt
[cloudera@quickstart ~]$ cat /home/cloudera/ProcessFile2.txt
10.0.0.5
10.0.0.1
10.0.0.5
20.0.0.0
20.0.0.0
10.0.0.5
10.0.0.1

```

```

[cloudera@quickstart ~]$ hdfs dfs -ls
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 5 items
drwxr-xr-x   - hbase supergroup          0 2023-01-30 22:09 /hbase
drwxr-xr-x   - solr  solr                  0 2015-06-09 03:38 /solr
drwxrwxrwx   - hdfs supergroup            0 2023-01-29 22:26 /tmp
drwxr-xr-x   - hdfs supergroup            0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs supergroup            0 2015-06-09 03:36 /var
[cloudera@quickstart ~]$ hdfs dfs -put
/home/cloudera/ProcessFile2.txt /inputfolder2/
[cloudera@quickstart ~]$ hdfs dfs -ls /out1
Found 2 items
-rw-r--r--    1 cloudera supergroup          0 2023-01-30 23:27
/out1/_SUCCESS
-rw-r--r--    1 cloudera supergroup          21 2023-01-30 23:27
/out1/part-r-00000
[cloudera@quickstart ~]$ hdfs dfs -cat /out1/part-r-00000
10.0.0.5      3
10.0.0.1      2
20.0.0.0      2
[cloudera@quickstart ~]$

```

2. Sum of time of a particular IP

```

[cloudera@quickstart ~]$ pwd
/home/cloudera
[cloudera@quickstart ~]$ cat > /home/cloudera/ProcessFile3.txt
10.0.0.5 10
10.0.0.1 12
10.0.0.5 15
20.0.0.0 10
20.0.0.0 20
10.0.0.5 12
10.0.0.1 15
^Z
[1]+  Stopped                  cat >
/home/cloudera/ProcessFile3.txt
[cloudera@quickstart ~]$ cat /home/cloudera/ProcessFile3.txt
10.0.0.5 10
10.0.0.1 12
10.0.0.5 15
20.0.0.0 10
20.0.0.0 20
10.0.0.5 12
10.0.0.1 15
[cloudera@quickstart ~]$ hdfs dfs -ls
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 5 items
drwxr-xr-x   - hbase supergroup          0 2023-01-30 22:09
/hbase
drwxr-xr-x   - solr  solr                  0 2015-06-09 03:38 /solr
drwxrwxrwx   - hdfs supergroup            0 2023-01-29 22:26 /tmp
drwxr-xr-x   - hdfs supergroup            0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs supergroup            0 2015-06-09 03:36 /var

```

```
[cloudera@quickstart ~]$ hdfs dfs -put
/home/cloudera/ProcessFile3.txt /inputfolder3/
[cloudera@quickstart ~]$ hdfs dfs -cat
/inputfolder3/ProcessFile3.txt
10.0.0.5 10
10.0.0.1 12
10.0.0.5 15
20.0.0.0 10
20.0.0.0 20
10.0.0.5 12
10.0.0.1 15
[cloudera@quickstart ~]$ hdfs dfs -ls /out1
Found 2 items
-rw-r--r-- 1 cloudera supergroup          0 2023-01-30 23:27
/out1/_SUCCESS
-rw-r--r-- 1 cloudera supergroup        21 2023-01-30 23:27
/out1/part-r-00000
[cloudera@quickstart ~]$ hdfs dfs -cat /out1/part-r-00000
10.0.0.5 37
10.0.0.1 27
20.0.0.0 30
[cloudera@quickstart ~]$
```