Lab Task-4: Weather data analysis for analysing hot and cold days using MapReduce.

Submitted by:

Asish Karthikeya

CSE_B

Code
Mapper

```java
package weather_data;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Mapper;

public class weather_mapper extends Mapper<LongWritable, Text, Text, IntWritable>
{
private static final int MISSING = 9999;

@Override

public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException
{
String line = value.toString();

String year = line.substring(15, 19);

int airTemperature;

if (line.charAt(87) == '+')

{ // parseInt doesn't like leading plus signs

airTemperature = Integer.parseInt(line.substring(88, 92));

}
else
{
airTemperature = Integer.parseInt(line.substring(87, 92));

}
```

```java
String quality = line.substring(92, 93);

if (airTemperature != MISSING && quality.matches("[01459]"))

{

context.write(new Text(year), new IntWritable(airTemperature));

}

}

}
```

Reducer

```java
package weather_data;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;

public class weather_reducer extends Reducer<Text, IntWritable, Text, IntWritable>

{

@Override

public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException,
InterruptedException

{

int maxValue = Integer.MIN_VALUE;

for (IntWritable value : values)

{

maxValue = Math.max(maxValue, value.get());

}

context.write(key, new IntWritable(maxValue));

}

}
```

Driver program

```java
package weather_data;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
```

```java
public class weather_main
{
public static void main(String[] args) throws Exception
{
if (args.length != 2)
{
System.err.println("Usage: MaxTemperature <input path> <output path>");
System.exit(-1);
}
@SuppressWarnings("deprecation")
Job job = new Job();
job.setJarByClass(weather_main.class);
job.setJobName("Max temperature");
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
job.setMapperClass(weather_mapper.class);
job.setReducerClass(weather_reducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
```

Dataset Used:

Index of /pub/data/noaa/2022

Name	Last modified	Size	Description
Parent Directory		-	
010010-99999-2022.gz	2023-05-18 18:08	122K	
010014-99999-2022.gz	2023-05-18 18:08	183K	

Input:

File Preview
/weather_analysis/input/010010-99999-2022

```
0093010010999992022010101004+70939-008669FM-12+001099999V0203581N0139199999999999999999-00871-01111101401ADDKA1010M-007
0093010010999992022010102004+70939-008669FM-12+001099999V0203531N0127199999999999999999-00901-01141101411ADDKA1010M-008
0093010010999992022010105004+70939-008669FM-12+001099999V0203501N0105199999999999999999-00991-01271101381ADDKA1010M-009
0093010010999992022010107004+70939-008669FM-12+001099999V0203511N0085199999999999999999-01051-01411101341ADDKA1010M-010
0093010010999992022010108004+70939-008669FM-12+001099999V0203491N0079199999999999999999-01051-01581101321ADDKA1010M-010
0192010010999992022010109004+70939-008669FM-12+001099999V0200021N00681006001C9020000199-01071-01731101321ADDAY171031AY2
0093010010999992022010110004+70939-008669FM-12+001099999V0203431N0089199999999999999999-01171-01581101341ADDKA1010M-010
0093010010999992022010111004+70939-008669FM-12+001099999V0203441N0060199999999999999999-01151-01621101331ADDKA1010M-011
0230010010999992022010112004+70939-008669FM-12+001099999V0203401N00821003001C9008000199-01241-01761101251ADDAY181061AY2
0093010010999992022010113004+70939-008669FM-12+001099999V0203401N0074199999999999999999-01251-01931101191ADDKA1010M-012
0093010010999992022010114004+70939-008669FM-12+001099999V0203391N0075199999999999999999-01271-01881101071ADDKA1010M-012
0224010010999992022010115004+70939-008669FM-12+001099999V0203411N00721006001C9018000199-01251-01911100991ADDAY181031AY2
0093010010999992022010116004+70939-008669FM-12+001099999V0203561N0064199999999999999999-01241-01791100871ADDKA1010M-012
0093010010999992022010117004+70939-008669FM-12+001099999V0203481N0076199999999999999999-01201-01791100761ADDKA1010M-012
0241010010999992022010118004+70939-008669FM-12+001099999V0203531N00671006001C9018000199-01171-01831100671ADDAA112000231
0093010010999992022010119004+70939-008669FM-12+001099999V0203411N0068199999999999999999-01171-01771100611ADDKA1010M-011
0093010010999992022010120004+70939-008669FM-12+001099999V0203411N0089199999999999999999-01101-01451100481ADDKA1010M-011
```

✖ Cancel   ⬇ Download

Jar File Path & Location:

Path:

C:\PBDM\Lab

Location:

Name	Date modified
📁 New folder	05-09-2023 22:49
☕ hadoop-common-2.5.0	05-09-2023 22:16
☕ hadoop-mapreduce-client-core-2.5.0	05-09-2023 22:16
☕ MatrixMul	05-09-2023 22:18
☕ Weather_analysis	18-09-2023 20:44
☕ Word_Count_Program	04-09-2023 21:07

Execution:

```
[root@sandbox ~]# hadoop fs -get /weather_analysis/Weather_analysis.jar
get: `Weather_analysis.jar': File exists
[root@sandbox ~]# hadoop jar Weather_analysis.jar /weather_analysis/input /weather_
23/09/18 10:35:23 INFO impl.TimelineClientImpl: Timeline service address: http://sa
23/09/18 10:35:23 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hor
23/09/18 10:35:24 INFO client.AHSProxy: Connecting to Application History server at
23/09/18 10:35:25 WARN mapreduce.JobResourceUploader: Hadoop command-line option pa
oolRunner to remedy this.
23/09/18 10:35:26 INFO input.FileInputFormat: Total input paths to process : 1
23/09/18 10:35:26 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
23/09/18 10:35:26 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo l
23/09/18 10:35:27 INFO mapreduce.JobSubmitter: number of splits:1
23/09/18 10:35:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16945
23/09/18 10:35:28 INFO impl.YarnClientImpl: Submitted application application_16945
23/09/18 10:35:29 INFO mapreduce.Job: The url to track the job: http://sandbox.hort
23/09/18 10:35:29 INFO mapreduce.Job: Running job: job_1694535283099_0010
23/09/18 10:36:00 INFO mapreduce.Job: Job job_1694535283099_0010 running in uber mo
23/09/18 10:36:00 INFO mapreduce.Job:  map 0% reduce 0%
23/09/18 10:36:27 INFO mapreduce.Job:  map 100% reduce 0%
```

```
23/09/18 10:37:16 INFO mapreduce.Job:  map 100% reduce 100%
23/09/18 10:37:18 INFO mapreduce.Job: Job job_1694535283099_0010 failed with state
Job failed as tasks failed. failedMaps:1 failedReduces:0

23/09/18 10:37:18 INFO mapreduce.Job: Counters: 13
        Job Counters
                Failed map tasks=4
                Killed reduce tasks=1
                Launched map tasks=4
                Other local map tasks=3
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=63769
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=63769
                Total time spent by all reduce tasks (ms)=0
                Total vcore-milliseconds taken by all map tasks=63769
                Total vcore-milliseconds taken by all reduce tasks=0
                Total megabyte-milliseconds taken by all map tasks=15942250
                Total megabyte-milliseconds taken by all reduce tasks=0
[root@sandbox ~]# hadoop jar Weather_analysis.jar /weather_analysis/input /weather_
23/09/18 10:39:51 INFO impl.TimelineClientImpl: Timeline service address: http://sa
23/09/18 10:39:51 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hor
23/09/18 10:39:52 INFO client.AHSProxy: Connecting to Application History server at
23/09/18 10:39:53 WARN mapreduce.JobResourceUploader: Hadoop command-line option pa
oolRunner to remedy this.
23/09/18 10:39:54 INFO input.FileInputFormat: Total input paths to process : 1
23/09/18 10:39:54 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
23/09/18 10:39:54 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo l
23/09/18 10:39:56 INFO mapreduce.JobSubmitter: number of splits:1
23/09/18 10:39:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16945
23/09/18 10:39:57 INFO impl.YarnClientImpl: Submitted application application_16945
23/09/18 10:39:58 INFO mapreduce.Job: The url to track the job: http://sandbox.hort
23/09/18 10:39:58 INFO mapreduce.Job: Running job: job_1694535283099_0011
23/09/18 10:40:22 INFO mapreduce.Job: Job job_1694535283099_0011 running in uber mo
23/09/18 10:40:22 INFO mapreduce.Job:  map 0% reduce 0%
23/09/18 10:40:44 INFO mapreduce.Job:  map 100% reduce 0%
23/09/18 10:41:04 INFO mapreduce.Job:  map 100% reduce 100%
23/09/18 10:41:06 INFO mapreduce.Job: Job job_1694535283099_0011 completed successf
ully
23/09/18 10:41:06 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=44248
```

```
23/09/18 10:41:06 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=44248
                FILE: Number of bytes written=376859
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=918721
                HDFS: Number of bytes written=8
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=19604
                Total time spent by all reduces in occupied slots (ms)=16484
                Total time spent by all map tasks (ms)=19604
                Total time spent by all reduce tasks (ms)=16484
                Total vcore-milliseconds taken by all map tasks=19604
                Total vcore-milliseconds taken by all reduce tasks=16484
                Total megabyte-milliseconds taken by all map tasks=4901000
                Total megabyte-milliseconds taken by all reduce tasks=4121000
        Map-Reduce Framework
                Map input records=4022
                Map output records=4022
                Map output bytes=36198
                Map output materialized bytes=44248
                Input split bytes=141
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=44248
                Reduce input records=4022
                Reduce output records=1
                Spilled Records=8044
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=627
                CPU time spent (ms)=6220
```

```
                Map-Reduce Framework
                        Map input records=4022
                        Map output records=4022
                        Map output bytes=36198
                        Map output materialized bytes=44248
                        Input split bytes=141
                        Combine input records=0
                        Combine output records=0
                        Reduce input groups=1
                        Reduce shuffle bytes=44248
                        Reduce input records=4022
                        Reduce output records=1
                        Spilled Records=8044
                        Shuffled Maps =1
                        Failed Shuffles=0
                        Merged Map outputs=1
                        GC time elapsed (ms)=627
                        CPU time spent (ms)=6220
                        Physical memory (bytes) snapshot=297906176
                        Virtual memory (bytes) snapshot=3891314688
                        Total committed heap usage (bytes)=139984896
                Shuffle Errors
                        BAD_ID=0
                        CONNECTION=0
                        IO_ERROR=0
                        WRONG_LENGTH=0
                        WRONG_MAP=0
                        WRONG_REDUCE=0
                File Input Format Counters
                        Bytes Read=918580
                File Output Format Counters
                        Bytes Written=8
[root@sandbox ~]# hadoop fs -cat /weather_analysis/output/part-r-00000
2022    92
[root@sandbox ~]#
```

Output:

File Preview

/weather_analysis/output/part-r-00000

```
2022    92
```