



TOPIC:

Project on NCDC (National Climate Data Center)

Course: CPSC-651-11-Big Data Systems & Analysis – Fall 2019

Submission to: Prof. Jeongkyu Lee

Submitted by: Srikanth Dabbiru (UB ID 1046112) & Aditya Lavu (UB ID 1066404)

- Testing the Project with Sample Data on VM:

1. To test the project and the codes, we have uploaded a sample of year 2015 from the NOAA dataset.
2. The idea was to find out if it was a hot or a cold day depending on the temperatures recorded.
3. Days of the year crossing 35.0 were considered to be the hottest days from the sample.
4. Snapshot of the sample .csv used in the VM is as below:

23907 20150101 2.423 -98.08 30.62 2.2 -0.6 0.8 0.9 7.0 1.47 C 3.7 1.1 2.5 99.9 85.4 97.2 0.369 0.308 -99.000 -99.000 -99.000 7.0 8.1 -9999.0 -9999.0 -9999.0																			
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q		
1	23907 20150101	2.423	-98.08	30.62	2.2	-0.6	0.8	0.9	7.0	1.47 C	3.7	1.1	2.5	99.9	85.4	97.2	0.369	0.308	-99.000 -99.000 -99.000 7.0 8.1 -9999.0 -9999.0 -9999.0
2	23907 20150102	2.423	-98.08	30.62	3.5	1.3	2.4	2.2	10.2	1.43 C	4.9	2.3	3.1	100.0	98.8	99.8	0.391	0.327	-99.000 -99.000 -99.000 7.1 7.9 -9999.0 -9999.0 -9999.0
3	23907 20150103	2.423	-98.08	30.62	15.9	2.3	9.1	7.5	3.1	11.00 C	16.4	2.9	7.3	100.0	34.8	73.7	0.450	0.397	-99.000 -99.000 -99.000 7.6 7.9 -9999.0 -9999.0 -9999.0
4	23907 20150104	2.423	-98.08	30.62	9.2	-1.3	3.9	4.2	0.0	13.24 C	12.4	-0.5	4.9	82.0	40.6	61.7	0.414	0.352	-99.000 -99.000 -99.000 7.3 7.9 -9999.0 -9999.0 -9999.0
5	23907 20150105	2.423	-98.08	30.62	10.9	-3.7	3.6	2.6	0.0	13.37 C	14.7	-3.0	3.8	77.9	33.3	57.4	0.399	0.340	-99.000 -99.000 -99.000 6.3 7.0 -9999.0 -9999.0 -9999.0
6	23907 20150106	2.423	-98.08	30.62	20.2	2.9	11.6	10.9	0.0	12.90 C	22.0	1.6	9.9	67.7	30.2	49.3	0.395	0.335	-99.000 -99.000 -99.000 8.0 8.0 -9999.0 -9999.0 -9999.0
7	23907 20150107	2.423	-98.08	30.62	10.9	-3.4	3.8	4.5	0.0	12.68 C	12.4	-2.1	5.5	82.7	36.5	55.7	0.387	0.328	-99.000 -99.000 -99.000 7.6 8.3 -9999.0 -9999.0 -9999.0
8	23907 20150108	2.423	-98.08	30.62	0.6	-7.9	-3.6	-3.3	0.0	4.98 C	3.9	-4.8	-0.5	57.7	37.6	48.1	0.372	0.316	-99.000 -99.000 -99.000 4.7 6.1 -9999.0 -9999.0 -9999.0
9	23907 20150109	2.423	-98.08	30.62	2.0	0.1	1.0	0.8	0.0	2.52 C	4.1	1.2	2.5	87.8	48.9	64.4	0.368	0.312	-99.000 -99.000 -99.000 5.4 6.2 -9999.0 -9999.0 -9999.0
10	23907 20150110	2.423	-98.08	30.62	0.5	-2.0	-0.8	-0.6	3.9	2.11 C	2.5	-0.1	1.4	99.9	47.7	85.8	0.373	0.314	-99.000 -99.000 -99.000 5.1 6.0 -9999.0 -9999.0 -9999.0
11	23907 20150111	2.423	-98.08	30.62	10.9	0.0	5.4	4.4	2.6	6.38 C	12.7	1.3	5.8	100.0	77.8	97.1	0.420	0.362	-99.000 -99.000 -99.000 6.5 6.7 -9999.0 -9999.0 -9999.0
12	23907 20150112	2.423	-98.08	30.62	6.5	1.4	4.0	4.3	0.0	1.55 C	6.9	2.7	5.1	100.0	89.4	97.8	0.412	0.350	-99.000 -99.000 -99.000 7.3 7.5 -9999.0 -9999.0 -9999.0
13	23907 20150113	2.423	-98.08	30.62	3.0	-0.7	1.1	1.2	0.0	3.26 C	5.6	0.7	2.9	99.7	80.7	90.7	0.401	0.337	-99.000 -99.000 -99.000 6.1 6.8 -9999.0 -9999.0 -9999.0
14	23907 20150114	2.423	-98.08	30.62	2.9	0.9	1.9	1.8	0.7	1.88 C	4.7	2.0	3.1	99.6	90.8	97.9	0.395	0.331	-99.000 -99.000 -99.000 6.1 6.7 -9999.0 -9999.0 -9999.0
15	23907 20150115	2.423	-98.08	30.62	13.2	1.2	7.2	6.4	0.0	13.37 C	16.4	1.4	6.7	98.9	46.7	73.4	0.395	0.333	-99.000 -99.000 -99.000 6.7 7.0 -9999.0 -9999.0 -9999.0
16	23907 20150116	2.423	-98.08	30.62	16.7	3.5	10.1	9.9	0.0	13.68 C	19.2	1.3	8.7	80.2	38.1	58.2	0.391	0.330	-99.000 -99.000 -99.000 7.3 7.4 -9999.0 -9999.0 -9999.0
17	23907 20150117	2.423	-98.08	30.62	19.5	5.0	12.2	12.3	0.0	10.96 C	20.9	3.3	10.6	87.7	30.4	55.7	0.388	0.327	-99.000 -99.000 -99.000 8.7 8.4 -9999.0 -9999.0 -9999.0
18	23907 20150118	2.423	-98.08	30.62	20.9	7.6	14.3	13.7	0.0	15.03 C	23.4	3.5	11.9	45.9	14.6	31.4	0.383	0.325	-99.000 -99.000 -99.000 9.5 9.2 -9999.0 -9999.0 -9999.0
19	23907 20150119	2.423	-98.08	30.62	23.9	6.7	15.3	14.3	0.0	14.10 C	25.6	3.8	12.6	65.3	26.8	45.6	0.376	0.321	-99.000 -99.000 -99.000 9.9 9.5 -9999.0 -9999.0 -9999.0
20	23907 20150120	2.423	-98.08	30.62	26.0	9.5	17.8	15.9	0.0	14.57 C	27.9	6.5	14.5	88.4	16.1	50.2	0.373	0.320	-99.000 -99.000 -99.000 10.9 10.4 -9999.0 -9999.0 -9999.0
21	23907 20150121	2.423	-98.08	30.62	11.0	6.9	8.9	8.9	1.7	2.71 C	13.1	6.8	9.7	99.2	68.0	88.1	0.369	0.317	-99.000 -99.000 -99.000 10.7 10.6 -9999.0 -9999.0 -9999.0
22	23907 20150122	2.423	-98.08	30.62	8.6	3.5	6.1	5.6	40.0	1.28 C	9.1	4.1	6.3	99.6	95.2	98.0	0.546	0.418	-99.000 -99.000 -99.000 9.0 9.3 -9999.0 -9999.0 -9999.0
23	23907 20150123	2.423	-98.08	30.62	9.4	2.2	5.8	4.2	7.5	6.58 C	11.1	2.0	4.8	98.4	58.8	86.5	0.554	0.409	-99.000 -99.000 -99.000 7.6 8.1 -9999.0 -9999.0 -9999.0
24	23907 20150124	2.423	-98.08	30.62	16.0	1.4	8.7	8.0	0.0	14.26 C	18.8	0.4	7.7	92.0	33.0	63.0	0.494	0.381	-99.000 -99.000 -99.000 7.7 7.9 -9999.0 -9999.0 -9999.0
25	23907 20150125	2.423	-98.08	30.62	20.2	6.4	13.3	12.7	0.0	14.99 C	22.0	4.4	11.0	69.2	18.9	43.8	0.456	0.357	-99.000 -99.000 -99.000 9.1 8.9 -9999.0 -9999.0 -9999.0
26	23907 20150126	2.423	-98.08	30.62	21.5	7.2	14.4	14.1	0.0	12.01 C	22.9	5.5	12.2	56.8	23.7	40.6	0.433	0.349	-99.000 -99.000 -99.000 10.0 9.7 -9999.0 -9999.0 -9999.0
27	23907 20150127	2.423	-98.08	30.62	26.5	10.7	18.6	17.5	0.0	15.18 C	28.9	8.1	15.5	52.2	21.4	38.8	0.420	0.344	-99.000 -99.000 -99.000 11.4 10.8 -9999.0 -9999.0 -9999.0
28	23907 20150128	2.423	-98.08	30.62	26.3	13.3	19.8	19.1	0.0	15.11 C	28.1	7.9	16.3	54.9	19.4	35.5	0.410	0.339	-99.000 -99.000 -99.000 12.1 11.5 -9999.0 -9999.0 -9999.0
29	23907 20150129	2.423	-98.08	30.62	23.1	9.8	16.5	16.4	0.0	13.74 C	27.4	9.7	16.4	87.0	34.2	55.6	0.402	0.334	-99.000 -99.000 -99.000 13.1 12.4 -9999.0 -9999.0 -9999.0
30	23907 20150130	2.423	-98.08	30.62	13.0	6.9	10.0	9.0	0.2	7.19 C	19.2	8.3	11.0	67.6	48.4	58.8	0.389	0.328	-99.000 -99.000 -99.000 11.6 11.8 -9999.0 -9999.0 -9999.0
31	23907 20150131	2.423	-98.08	30.62	15.1	7.4	11.3	10.2	8.5	1.18 C	14.5	8.4	10.7	100.0	63.1	90.3	0.401	0.337	-99.000 -99.000 -99.000 11.2 11.3 -9999.0 -9999.0 -9999.0
32	23907 20150201	2.423	-98.08	30.62	18.3	3.9	11.1	13.3	0.0	8.69 C	22.1	4.1	13.8	98.8	53.6	79.1	0.450	0.386	-99.000 -99.000 -99.000 12.9 12.6 -9999.0 -9999.0 -9999.0
33	23907 20150202	2.423	-98.08	30.62	8.0	-1.9	3.1	3.3	0.0	12.48 C	15.2	-0.6	5.8	69.4	34.8	54.2	0.420	0.354	-99.000 -99.000 -99.000 9.3 10.1 -9999.0 -9999.0 -9999.0
34	23907 20150203	2.423	-98.08	30.62	5.3	2.3	3.8	3.8	0.8	2.69 C	8.3	3.9	5.7	100.0	65.1	82.6	0.409	0.343	-99.000 -99.000 -99.000 8.7 9.3 -9999.0 -9999.0 -9999.0
35	23907 20150204	2.423	-98.08	30.62	11.8	4.3	8.1	7.9	0.3	4.41 C	13.8	5.5	8.8	100.0	80.6	94.1	0.410	0.344	-99.000 -99.000 -99.000 9.6 9.8 -9999.0 -9999.0 -9999.0

- **Mapper-Reducer Code:**

For the mapper, the max temperature class is static, this method takes the input as text data type. Leaving the first five tokens, the 6th token is taken as the temp_max and the 7th as temp_min.

Now temp_max value is set to be >35.0 and the temp_min is set to be <10.0 and are now passed to the reducer step.

If the temp values for the day are >35.0 output as Hot Day and if <10.0 output as a Cold Day.

For the Reducer method, it takes the input as key and the pairs would be the list of values from the Mapper.

Now **Aggregation** is applied, and it produce the next result.

For the main method, it is used for setting up all the configuration properties. This will be acting as the driver for our Map Reduce code.

- Below is the Complete Source Code used:

```
MyMaxMin.java > M main(String[] args)
1  import java.io.IOException;
2  import java.util.Iterator;
3
4  import org.apache.hadoop.fs.Path;
5  import org.apache.hadoop.io.LongWritable;
6  import org.apache.hadoop.io.Text;
7
8  import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
9  import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
10 import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
11 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
12 import org.apache.hadoop.mapreduce.Job;
13 import org.apache.hadoop.mapreduce.Mapper;
14 import org.apache.hadoop.mapreduce.Reducer;
15 import org.apache.hadoop.conf.Configuration;
16
17 public class MyMaxMin {
18
19
20     //Mapper
21
22
23
24     public static class MaxTemperatureMapper extends
25         Mapper<LongWritable, Text, Text, Text> {
26
27
28         @Override
29         public void map(LongWritable arg0, Text Value, Context context)
30             throws IOException, InterruptedException {
31
32             //To Convert the record (single line) to String and storing it
33
34             String line = Value.toString();
35
36             //To Check if the line is not empty
37
38             if (!(line.length() == 0)) {
39
40                 //date
41
42                 String date = line.substring(6, 14);
43
44                 //maximum temperature
45
46                 float temp_Max = Float
47                     .parseFloat(line.substring(39, 45).trim());
48
49                 //minimum temperature
50
51                 float temp_Min = Float
52                     .parseFloat(line.substring(47, 53).trim());
53
54                 //if maximum temperature is greater than 35.0 , its a h
55
56                 if (temp_Max > 35.0) {
57                     // Hot day
58                     context.write(new Text("Hot Day " + date),
59                         new Text(String.valueOf(temp_Max)));
60                 }
61             }
62         }
63     }
64 }
```

```

60         }
61
62         //if minimum temperature is less than 10.0 , its a cold day
63
64         if (temp_Min < 10) {
65             // Cold day
66             context.write(new Text("Cold Day " + date),
67                 new Text(String.valueOf(temp_Min)));
68         }
69     }
70 }
71
72 }
73
74 //Reducer
75
76
77 public static class MaxTemperatureReducer extends
78     Reducer<Text, Text, Text, Text> {
79
80
81     public void reduce(Text Key, Iterator<Text> Values, Context context)
82         throws IOException, InterruptedException {
83
84
85         //To put all the values in temperature variable of type String
86
87         String temperature = Values.next().toString();
88         context.write(Key, new Text(temperature));
89     }
90 }
91
92
93
94
95 public static void main(String[] args) throws Exception {
96
97     //reads the default configuration of cluster from the configuration xml files
98     Configuration conf = new Configuration();
99
100    //Initializing the job with the default configuration of the cluster
101    Job job = new Job(conf, "weather example");
102
103    //Assigning the driver class name
104    job.setJarByClass(MyMaxMin.class);
105
106    //Key type coming out of mapper
107    job.setMapOutputKeyClass(Text.class);
108
109    //value type coming out of mapper
110    job.setMapOutputValueClass(Text.class);
111
112    //Defining the mapper class name
113    job.setMapperClass(MaxTemperatureMapper.class);
114
115    //Defining the reducer class name
116    job.setReducerClass(MaxTemperatureReducer.class);
117
118    //Defining input Format class which is responsible to parse the dataset into a key value pair
119    job.setInputFormatClass(TextInputFormat.class);

```



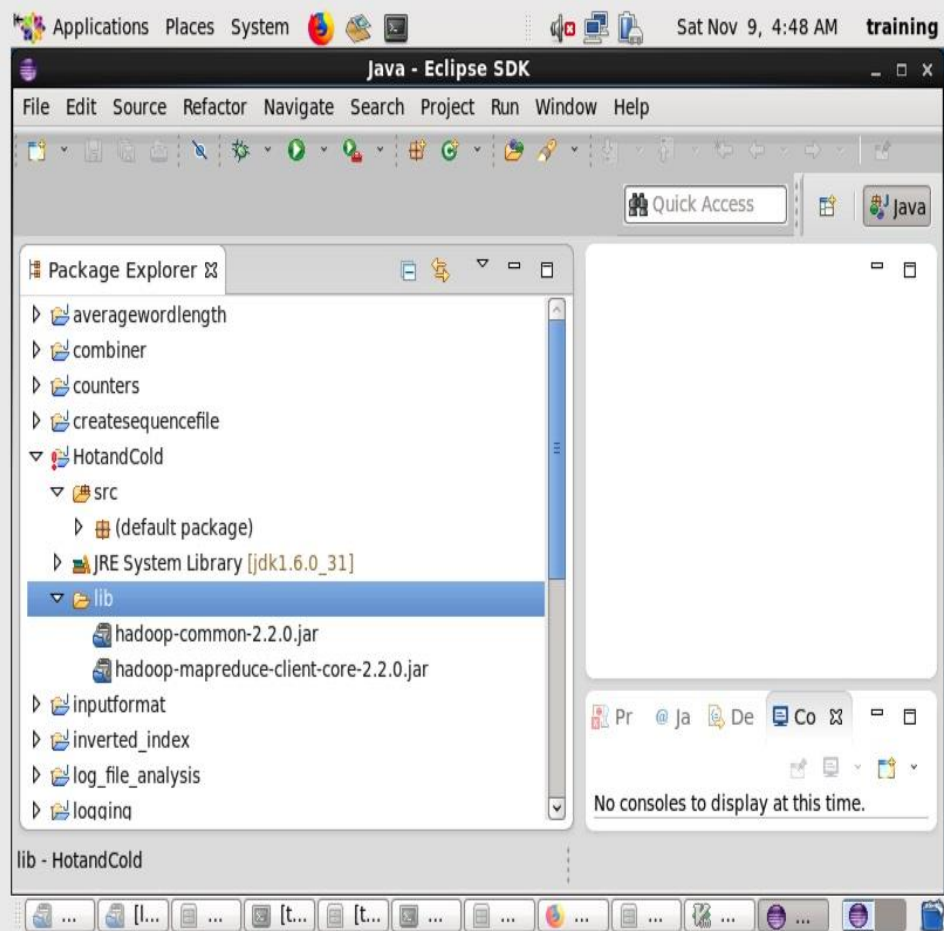
```
MyMaxMin.java > M main(String[] args)

113     job.setMapperClass(MaxTemperatureMapper.class);
114
115     //Defining the reducer class name
116     job.setReducerClass(MaxTemperatureReducer.class);
117
118     //Defining input Format class which is responsible to parse the dataset into a key value pair
119     job.setInputFormatClass(TextInputFormat.class);
120
121     //Defining output Format class which is responsible to parse the dataset into a key value pair
122     job.setOutputFormatClass(TextOutputFormat.class);
123
124     //setting the second argument as a path in a path variable
125     Path outputPath = new Path(args[1]);
126
127     //Configuring the input path from the filesystem into the job
128     FileInputFormat.addInputPath(job, new Path(args[0]));
129
130     //Configuring the output path from the filesystem into the job
131     FileOutputFormat.setOutputPath(job, new Path(args[1]));
132
133     //deleting the context path automatically from hdfs so that we don't have delete it
134     outputPath.getFileSystem(conf).delete(outputPath);
135
136     //exiting the job only if the flag value becomes false
137     System.exit(job.waitForCompletion(true) ? 0 : 1);
138
139 }
140 }
141
142
```

- **Eclipse IDE:**

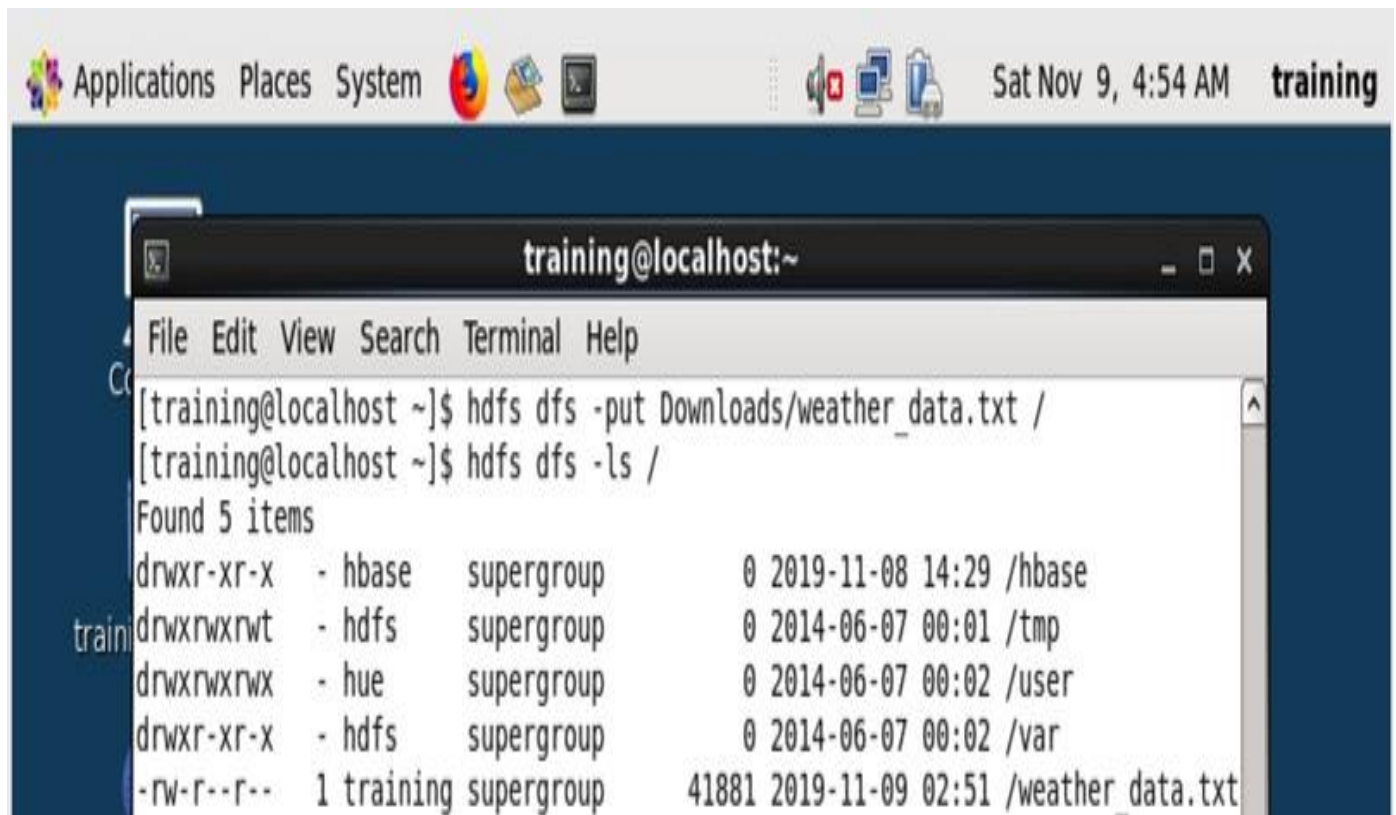
Now the project is created in the Eclipse IDE to analyze the sample dataset.

The jar file is then exported after having no issues with the project files.



The next step was to send the sample dataset onto HDFS.

Hdfs dfs -put Downloads/Noaa_Weather_data.txt /

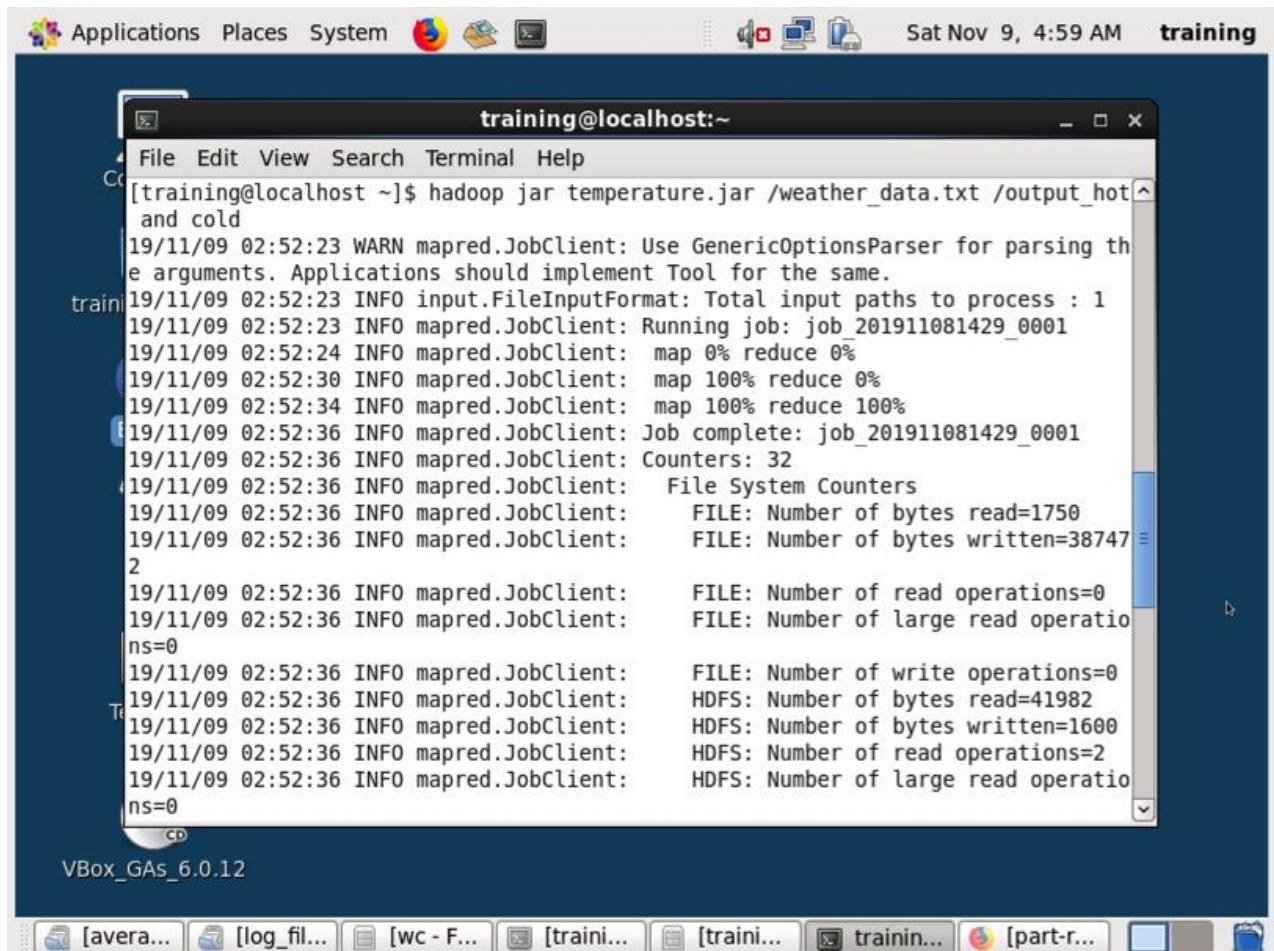


The screenshot shows a Linux desktop with a top bar containing icons for Applications, Places, System, and a clock showing 'Sat Nov 9, 4:54 AM' with the username 'training'. A terminal window titled 'training@localhost:~' is open, displaying the following commands and output:

```
training@localhost:~$ hdfs dfs -put Downloads/weather_data.txt /
training@localhost:~$ hdfs dfs -ls /
Found 5 items
drwxr-xr-x  - hbase  supergroup          0 2019-11-08 14:29 /hbase
drwxrwxrwt  - hdfs   supergroup          0 2014-06-07 00:01 /tmp
drwxrwxrwx  - hue    supergroup          0 2014-06-07 00:02 /user
drwxr-xr-x  - hdfs   supergroup          0 2014-06-07 00:02 /var
-rw-r--r--  1 training supergroup    41881 2019-11-09 02:51 /weather_data.txt
```


Run the Jar file for output.

**Hadoop jar temperature.jar /Noaa_weather_data.txt
/output_hotandcold**



The screenshot shows a terminal window titled "training@localhost:~" with a menu bar (File, Edit, View, Search, Terminal, Help). The command executed is `hadoop jar temperature.jar /weather_data.txt /output_hotandcold`. The output shows a warning about `GenericOptionsParser` and then progress information: "Total input paths to process : 1", "Running job: job_201911081429_0001", and progress updates for map and reduce tasks. The job completes with "Job complete: job_201911081429_0001" and "Counters: 32". A detailed list of counters follows, including File System Counters (bytes read/written, read/write operations) and HDFS Counters (bytes read/written, read operations).

```
training@localhost:~$ hadoop jar temperature.jar /weather_data.txt /output_hotandcold
19/11/09 02:52:23 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
19/11/09 02:52:23 INFO input.FileInputFormat: Total input paths to process : 1
19/11/09 02:52:23 INFO mapred.JobClient: Running job: job_201911081429_0001
19/11/09 02:52:24 INFO mapred.JobClient: map 0% reduce 0%
19/11/09 02:52:30 INFO mapred.JobClient: map 100% reduce 0%
19/11/09 02:52:34 INFO mapred.JobClient: map 100% reduce 100%
19/11/09 02:52:36 INFO mapred.JobClient: Job complete: job_201911081429_0001
19/11/09 02:52:36 INFO mapred.JobClient: Counters: 32
19/11/09 02:52:36 INFO mapred.JobClient:   File System Counters
19/11/09 02:52:36 INFO mapred.JobClient:     FILE: Number of bytes read=1750
19/11/09 02:52:36 INFO mapred.JobClient:     FILE: Number of bytes written=38747
19/11/09 02:52:36 INFO mapred.JobClient:     FILE: Number of read operations=0
19/11/09 02:52:36 INFO mapred.JobClient:     FILE: Number of large read operations=0
19/11/09 02:52:36 INFO mapred.JobClient:     FILE: Number of write operations=0
19/11/09 02:52:36 INFO mapred.JobClient:   HDFS: Number of bytes read=41982
19/11/09 02:52:36 INFO mapred.JobClient:   HDFS: Number of bytes written=1600
19/11/09 02:52:36 INFO mapred.JobClient:   HDFS: Number of read operations=2
19/11/09 02:52:36 INFO mapred.JobClient:   HDFS: Number of large read operations=0
```

Check the Output directory in the HDFS.

The image consists of two screenshots of the Hue web interface, a web-based tool for managing Hadoop clusters. The top screenshot shows a file viewer for a specific file, while the bottom screenshot shows a directory listing for the 'output_hot' directory.

Top Screenshot: File Viewer

The browser window is titled 'part-r-00000 - File Viewer - Mozilla Firefox'. The address bar shows 'localhost:8888/filebrowser/view//output_hot'. The interface displays a table of file details for 'part-r-00000'.

File	Download	View	File	Location	Refresh
Cold Day 20150103	2.3				
Cold Day 20150104	-1.3				
Cold Day 20150105	-3.7				
Cold Day 20150106	2.9				
Cold Day 20150107	-3.4				
Cold Day 20150108	-7.9				
Cold Day 20150109	0.1				
Cold Day 20150110	-2.0				
Cold Day 20150111	0.0				
Cold Day 20150112	1.4				
Cold Day 20150113	-0.7				
Cold Day 20150114	0.9				
Cold Day 20150115	1.2				
Cold Day 20150116	3.5				
Cold Day 20150117	5.0				
Cold Day 20150118	7.6				
Cold Day 20150119	6.7				
Cold Day 20150120	9.5				
Cold Day 20150121	6.9				
Cold Day 20150122	3.5				

Bottom Screenshot: File Browser

The browser window is titled 'File Browser - Mozilla Firefox'. The address bar shows 'localhost:8888/filebrowser/view/user/training#/ou'. The interface displays a directory listing for the 'output_hot' directory.

Search for file name:

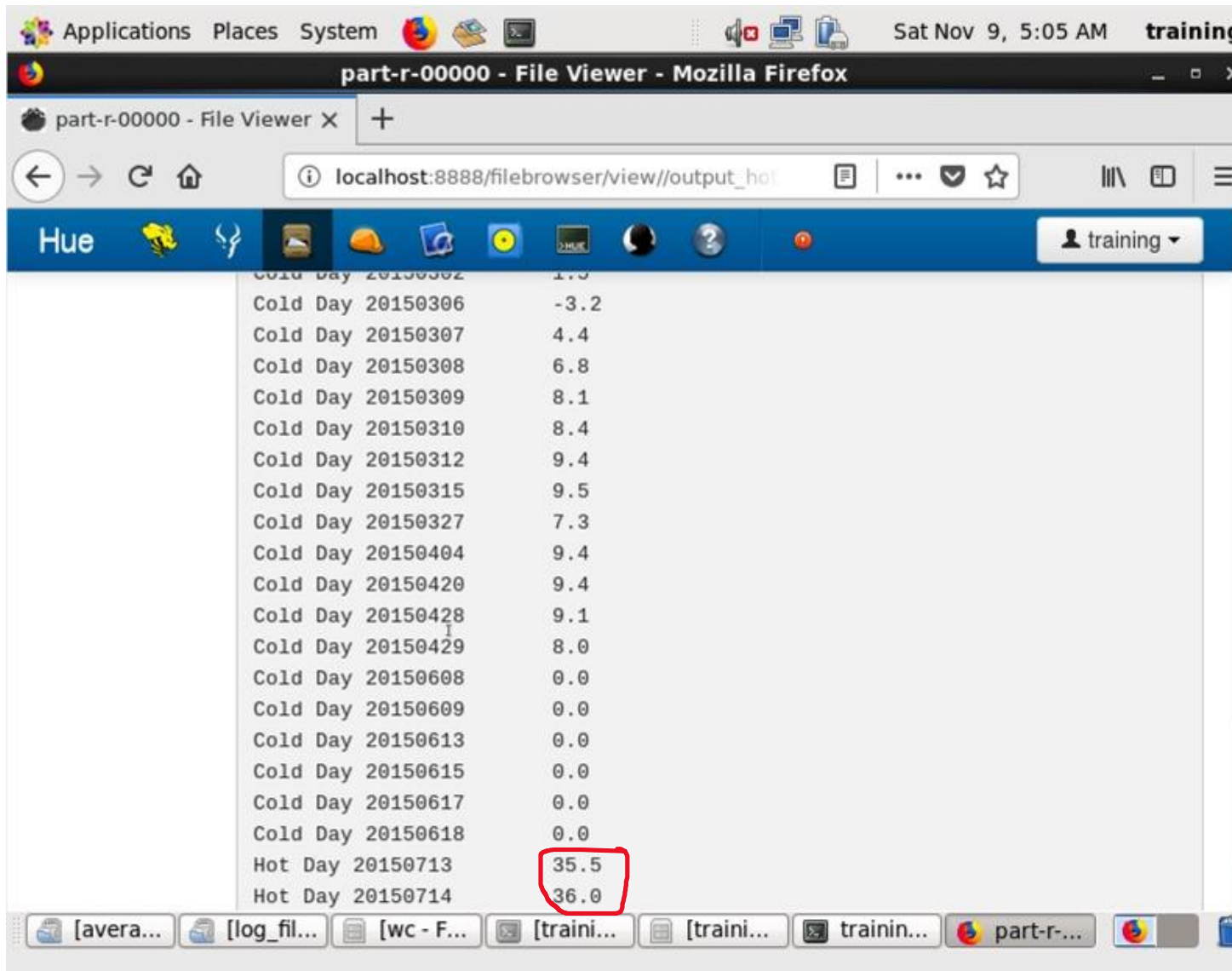
Home / output_hot

Type	Name	Size	User	Group	Permissions	Date
Folder	..		hdfs	supergroup	drwxr-xr-x	November 09, 2019 02:52 am
File	_SUCCESS	0 bytes	training	supergroup	-rw-r--r--	November 09, 2019 02:52 am
Folder	_logs		training	supergroup	drwxr-xr-x	November 09, 2019 02:52 am
File	part-r-00000	1.6 KB	training	supergroup	-rw-r--r--	November 09, 2019 02:52 am

Show 45 items per page. Showing 1 to 3 of 3 items, page 1 of 1

Results analysis:

Depending on the 2015 sample dataset only two days above 35.0 recorded.



Cold Day 20150302	1.9
Cold Day 20150306	-3.2
Cold Day 20150307	4.4
Cold Day 20150308	6.8
Cold Day 20150309	8.1
Cold Day 20150310	8.4
Cold Day 20150312	9.4
Cold Day 20150315	9.5
Cold Day 20150327	7.3
Cold Day 20150404	9.4
Cold Day 20150420	9.4
Cold Day 20150428	9.1
Cold Day 20150429	8.0
Cold Day 20150608	0.0
Cold Day 20150609	0.0
Cold Day 20150613	0.0
Cold Day 20150615	0.0
Cold Day 20150617	0.0
Cold Day 20150618	0.0
Hot Day 20150713	35.5
Hot Day 20150714	36.0