Bharathwin MA
205229105

Exercise 06: Earth Quack data analytics using MapReduce (Java & Python)

This exercise's MapReduce process is doing Earth Quack data analysis. This analysis is used to find maximum magnitude earth quack in each region. In this exercise students try to create Mapper and Reducer process using Java and Python.

Prerequisites

Ensure that Hadoop is installed, configured and is running. More details:

Single Node Setup for first-time users.

Cluster Setup for large, distributed clusters.

Inputs and Outputs

i. Input file should be in : /earth/in/

WAData.txt

Copy the content text from earth.csv, Which is attached in Google classroom.

ii. Output file should be in /earth/out/

Step 1

Create and Compile `EarthQuack.java` and create an EarthQuack.jar:

(i) Create `EarthQuack`.java project.

```
import org.apache.hadoop.fs.Path; import
org.apache.hadoop.io.DoubleWritable; 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;
```

```
public class EarthQuake {

    public static void main(String[] args) throws Exception {
if (args.length != 2) {

            System.err.println("Usage: hadoopex <input path> <output path>");

            System.exit(-1);

        }

// Create the job specification object
        Job job = new Job();
        job.setJarByClass(EarthQuake.class);
        job.setJobName("Earthquake Measurment");

        // Setup input and output paths
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        // Set the Mapper and Reducer classes
        job.setMapperClass(EarthQuakeMapper.class);
        job.setReducerClass(EarthQuakeReducer.class);

        // Specify the type of output keys and values
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(DoubleWritable.class);

       // Wait for the job to finish before terminating          // Wait for the job to finish before
terminating

        System.exit(job.waitForCompletion(true) ? 0 : 1);

    }
}
```

(ii) Create `EarthquakeMapper`.java project.

```
import org.apache.hadoop.io.DoubleWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text; import

org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;


public class EarthquakeMapper extends

        Mapper<LongWritable, Text, Text, DoubleWritable>

{



    @Override

    public void map(LongWritable key, Text value, Context context) throws

            IOException, InterruptedException {
```

```
        String[] line = value.toString().split(",", 12);

        // Ignore invalid lines
    if (line.length != 12) {
    System.out.println("- " +
    line.length);            return;
        }
```


```
String[] line = value.toString().split(",", 12);

    // Ignore invalid lines
    if (line.length != 12) {
    System.out.println("- " + line.length);
    return;
    }
```

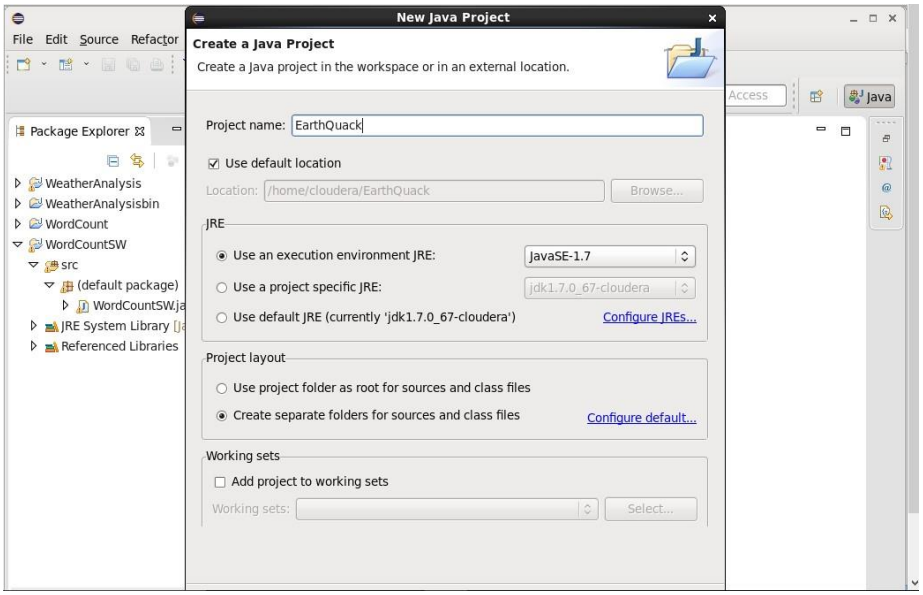

```
    // The output `key` is the name of the region
    String outputKey = line[1];

    // The output `value` is the magnitude of the earthquake
    double outputValue = Double.parseDouble(line[9]);

   // Record the output in the Context object
   context.write(new Text(outputKey), new DoubleWritable(outputValue));

    }
}
```

(iii) Create `EarthquakeMapper`.java project.

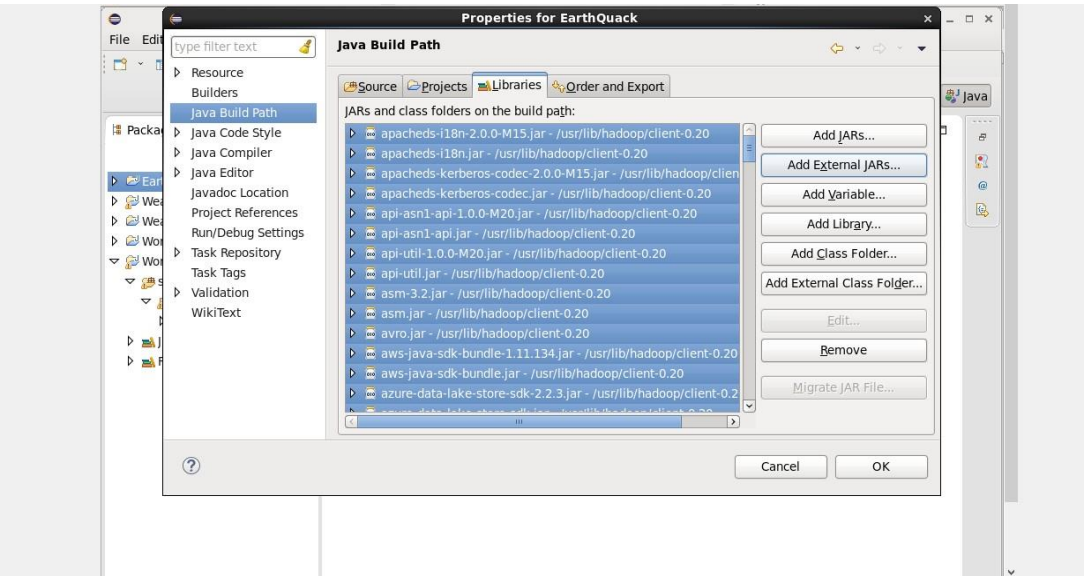
```
import org.apache.hadoop.io.DoubleWritable;

import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException; import

org.apache.hadoop.io.Text;


public class EarthquakeReducer extends

        Reducer<Text, DoubleWritable, Text, DoubleWritable>

{

    @Override
```

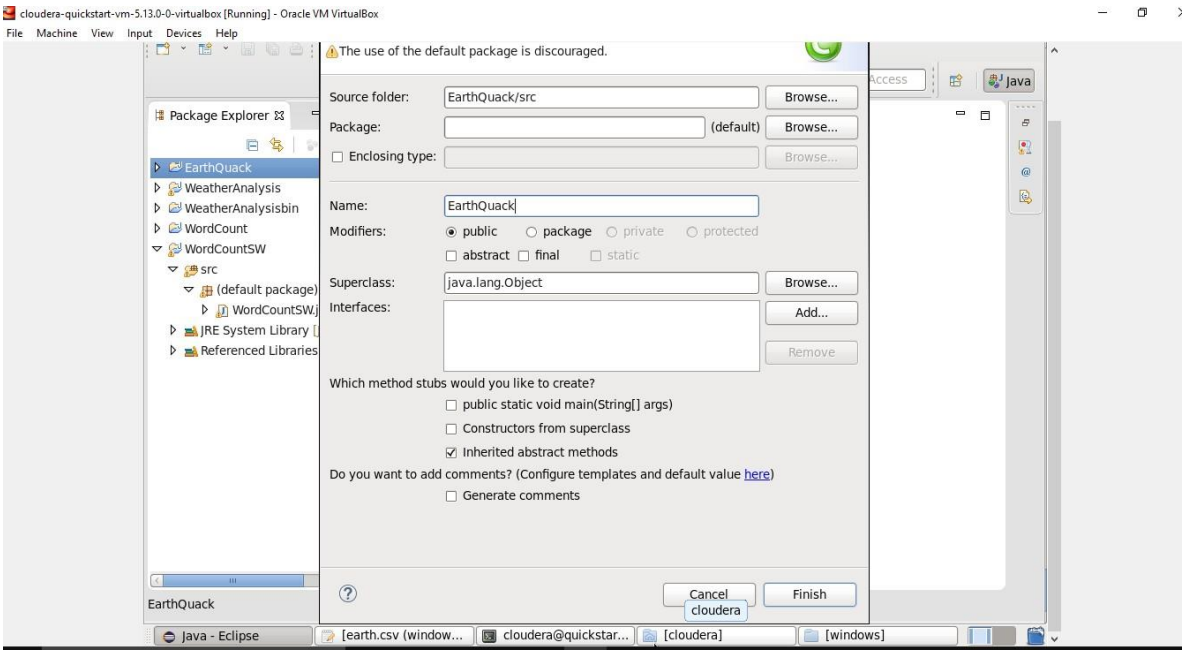
```
public void reduce(Text key, Iterable<DoubleWritable> values,

        Context context) throws IOException, InterruptedException {



    // Standard algorithm for finding the max value

double maxMagnitude = Double.MIN_VALUE;
    for (DoubleWritable value : values) {
    maxMagnitude = Math.max(maxMagnitude, value.get());
    }

    context.write(key, new DoubleWritable(maxMagnitude));


  }


      context.write(key, new DoubleWritable(maxMagnitude));

    }

  }
```

(iv) Import external .jar files

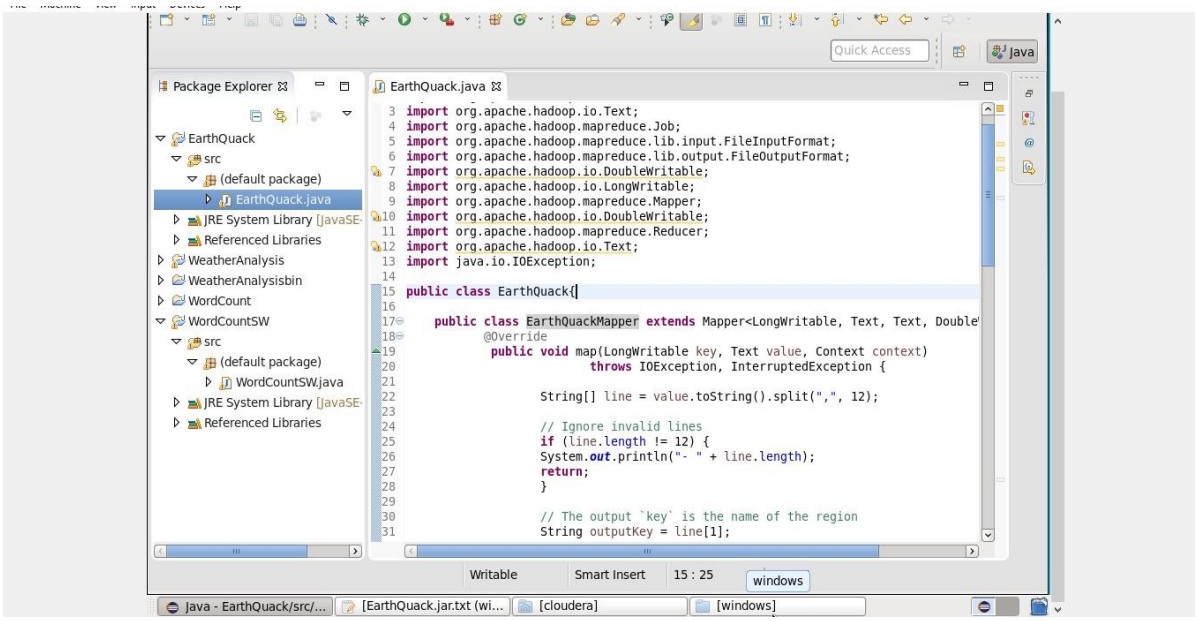


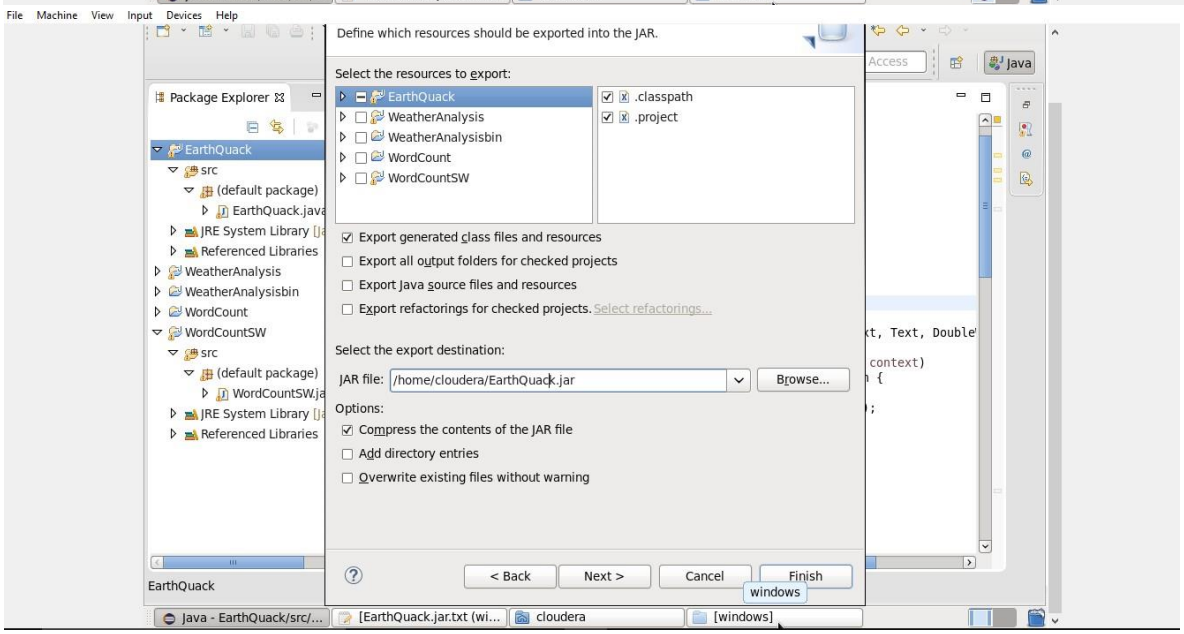
(v)





(vi) Create EarthQuake.jar file

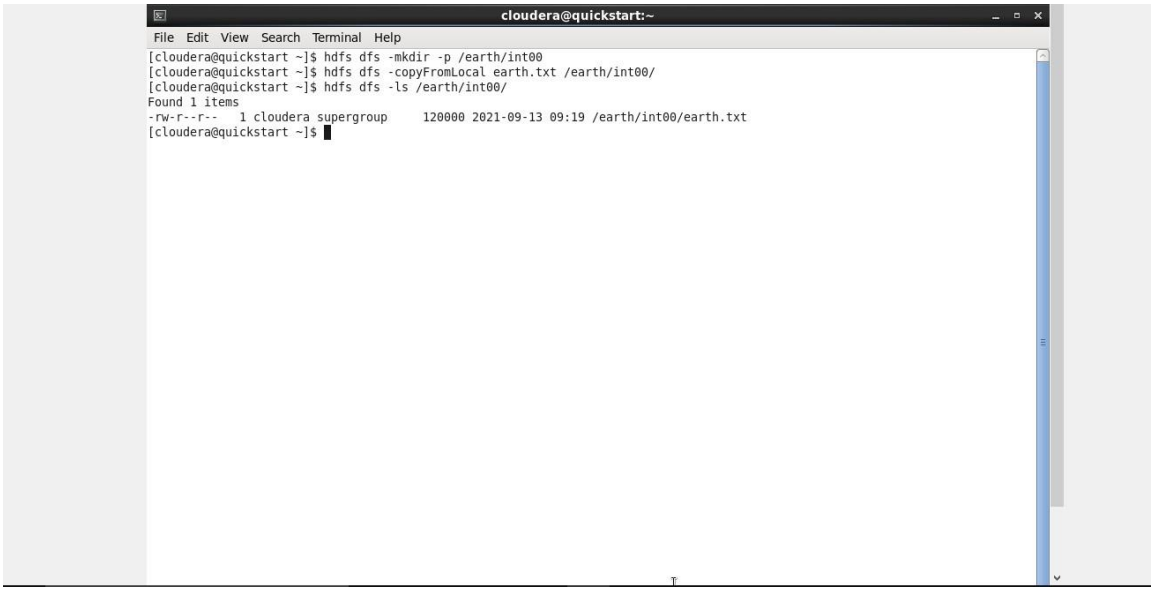




Step 2

Create following folders in HDFS:

- `/earth/in` - input directory in HDFS
- `/earth/out` - output directory in HDFS



Step 3

Create and copy earth.txt-files into input folder:

cloudera@quickstart:~

File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ cat > earth.txt
```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

```
ak,off the coast of Southeastern Alaska,10636696,2,"Saturday, January  5, 2013 21:56:05 UTC",55.5883,-135.6547,3.2,20,12

ak,Central Alaska,10631969,3,"Saturday, January  5, 2013 21:49:34 UTC",62.2735,-149.5358,1.4,53.2,16

ak,"Andreanof Islands, Aleutian Islands, Alaska",10631794,3,"Saturday, January  5, 2013 21:44:03 UTC",51.4817,-176.1561,2,26.
8,15

ak,"Kenai Peninsula, Alaska",10636687,2,"Saturday, January  5, 2013 21:34:25 UTC",60.4062,-148.3471,1.5,4.8,10

ak,Central Alaska,10631789,2,"Saturday, January  5, 2013 21:27:45 UTC",64.5252,-151.3103,1.5,13,10

us,Southeastern Alaska,c000ek0i,6,"Saturday, January  5, 2013 21:26:56 UTC",55.3116,-134.7276,4,9.9,81

ak,"Fox Islands, Aleutian Islands, Alaska",10636684,2,"Saturday, January  5, 2013 21:06:53 UTC",53.2942,-167.2919,2,55,12

ak,Southeastern Alaska,10636683,2,"Saturday, January  5, 2013 21:01:57 UTC",55.4375,-134.7142,2.7,20,8

ak,Southeastern Alaska,10636682,2,"Saturday, January  5, 2013 20:54:57 UTC",55.8668,-134.9697,2.1,10,5

us,"off the east coast of Honshu, Japan",c000ejzq,7,"Saturday, January  5, 2013 20:46:03 UTC",36.3921,142.9928,4.6,30.7,92

ak,off the coast of Southeastern Alaska,10636681,2,"Saturday, January  5, 2013 20:46:03 UTC",55.4689,-135.311,2.7,10,4

ak,off the coast of Southeastern Alaska,10631776,3,"Saturday, January  5, 2013 20:43:18 UTC",55.4611,-135.3817,4.2,10.4,22

ak,Southeastern Alaska,10636679,2,"Saturday, January  5, 2013 20:41:48 UTC",55.8116,-134.9622,2.9,10,7

ak,Southeastern Alaska,10636678,2,"Saturday, January  5, 2013 20:38:32 UTC",55.6674,-134.7293,2.4,2.3,7

ak,off the coast of Southeastern Alaska,10636677,2,"Saturday, January  5, 2013 20:32:57 UTC",55.5115,-135.3832,2.4,1,7

ci,Southern California,15270569,2,"Saturday, January  5, 2013 20:31:45 UTC",33.3873,-116.3855,1,5.5,52

ak,"Andreanof Islands, Aleutian Islands, Alaska",10636675,2,"Saturday, January  5, 2013 20:11:13 UTC",51.7622,-175.8648,1.9,5
9.4,9

ak,"Acme Islands, FOR TEST PURPOSES",10636675,2,"Saturday, January  5, 2013 20:11:13 UTC",51.7622,-175.8648,8.9,59.4,9
^C
[cloudera@quickstart ~]$
```

[earth.csv (windows) -... | cloudera@quickstart:~ | [windows]

cloudera@quickstart:~

File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir -p /earth/int00
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal earth.txt /earth/int00/
[cloudera@quickstart ~]$ hdfs dfs -ls /earth/int00/
Found 1 items
-rw-r--r--   1 cloudera supergroup     120000 2021-09-13 09:19 /earth/int00/earth.txt
[cloudera@quickstart ~]$
```

[cloudera@quickstart ~]$ hdfs dfs -ls /earth/in00/

Found 1 items

-rw-r--r-- 1 cloudera supergroup 12054 2021-08-26 15:48 /earth/in/earth.txt

Step 4

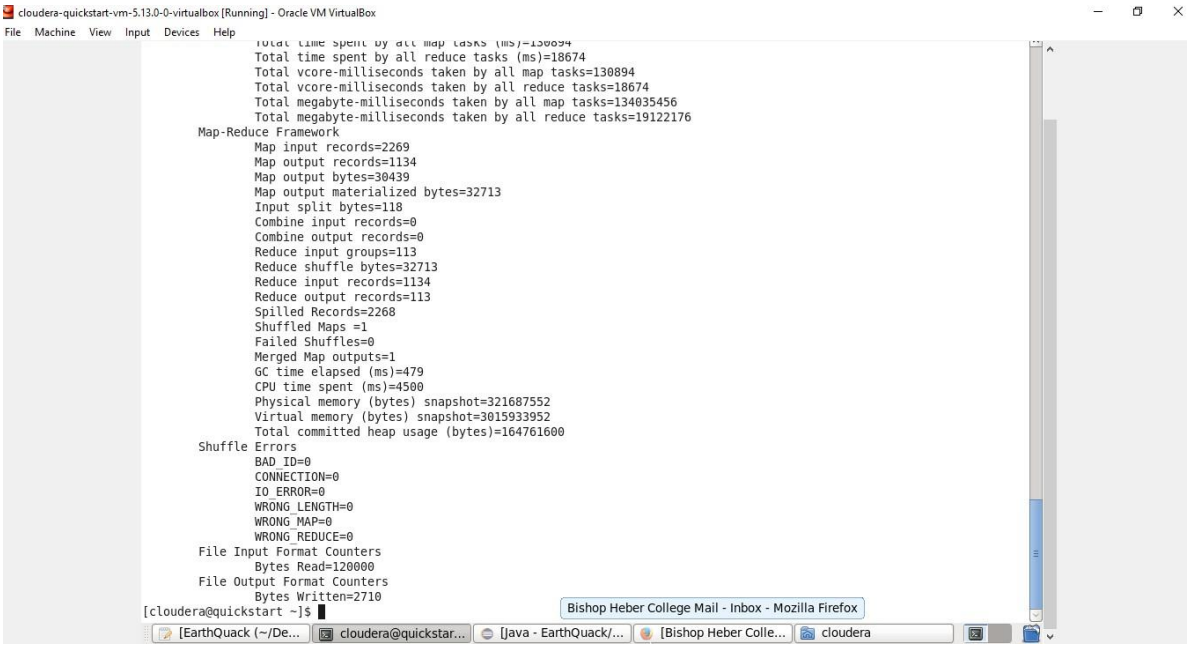
Run the MapReduce application：

Java

[cloudera@quickstart ~]$ hadoop jar EarthQuake.jar EarthQuake /earth/in/earth.txt /earth/out/

Python

[cloudera@quickstart ~]$ hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoopstreaming-2.6.0-mr1-cdh5.13.0.jar -file /home/cloudera/map.py /home/cloudera/reduce.py mapper "python map.py" -reducer "python reduce.py" -input /earth/in/earth.txt -output /earth/out Show MapReduce Framework



Step 5

Output:
[cloudera@quickstart ~]$ hdfs dfs -ls /earth/out/

Found 2 items

-rw-r--r-- 1 cloudera supergroup 0 2021-08-26 15:50 /weather/out00/_SUCCESS
-rw-r--r-- 1 cloudera supergroup 228 2021-08-26 15:50 /weather/out00/part-r-00000

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

cloudera@quickstart:~

File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hdfs dfs -ls /earth/out00/
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2021-09-13 09:39 /earth/out00/_SUCCESS
-rw-r--r--   1 cloudera supergroup       2710 2021-09-13 09:39 /earth/out00/part-r-00000
[cloudera@quickstart ~]$ hdfs dfs -cat /earth/out00/part-r-00000
"Acme Islands   4.9E-324
"Andaman Islands        92.3832
"Andreanof Islands      52.8796
"Anguilla region        4.9E-324
"Antofagasta    4.9E-324
"Arunachal Pradesh      94.3088
"Babuyan Islands region 121.2571
"Baja California        4.9E-324
"British Columbia       4.9E-324
"Channel Islands region 4.9E-324
"Fox Islands    54.115
"Greater Los Angeles area       4.9E-324
"Gulf of Santa Catalina 4.9E-324
"Halmahera      127.4821
"Hawaii region  4.9E-324
"Island of Hawaii       4.9E-324
"Izu Islands    141.5995
"Jujuy  4.9E-324
"Kenai Peninsula        4.9E-324
"Kodiak Island region   4.9E-324
"Lassen Peak area       4.9E-324
"Mona Passage   4.9E-324
"New Britain region     152.7111
"New Guinea     141.9851
"Newberry Caldera area  4.9E-324
"Oaxaca 4.9E-324
"Oklahoma City urban area       4.9E-324
"Olympic Peninsula      4.9E-324
"Papua  138.859
"Puget Sound region     4.9E-324
"Rat Islands    51.8203
"Salta  4.9E-324
```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

```
Solomon Islands 5.2
South Atlantic Ocean    5.1
Southeastern Alaska     4.7
Southern Alaska 3.5
Southern California     2.8
Spain   3.5
Strait of Hormuz        4.3
Taiwan region   5.2
Tennessee       2.6
Tonga   4.6
Utah    1.8
Vanuatu 4.7
Virgin Islands region   4.9
Virginia        2.0
Washington      2.9
central Turkey  4.5
east of the Kuril Islands       4.9
eastern Tennessee       2.2
near the coast of southern Peru 4.7
north of the Virgin Islands     3.4
northern Alaska 3.2
northern Idaho  1.5
off the coast of Oregon 4.3
off the coast of Southeastern Alaska    4.5
off the west coast of northern Sumatra  5.0
off the west coast of the North Island of New Zealand   4.6
offshore Central California     1.9
offshore Honduras       4.5
offshore Northern California    3.1
south of Alaska 3.4
south of the Aleutian Islands   3.1
south of the Fiji Islands       5.1
south of the Kermadec Islands   5.4
southeast of Taiwan     4.3
southern Iran   4.1
southern Mid-Atlantic Ridge     5.0
western Iran    5.4
western Montana 2.3
[cloudera@quickstart ~]$
```

Click to switch to "Workspace 2"

[EarthQuack (~/De... | cloudera@quickstar... | [Java - EarthQuack/... | [Bishop Heber Colle... | cloudera

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

```
Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/18CYB7kJ5XYbAIuwC0AlWmzJEqS__9DOa
"""

import re
import sys

for line in sys.stdin:
  val[] = line.strip()
  (reg, mag) = (val[1], val[9])
    print "%s\t%s" % (reg,mag)

[cloudera@quickstart Desktop]$ cat reduce.py
# -*- coding: utf-8 -*-
"""max_temperature_reduce.ipynb

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/18CYB7kJ5XYbAIuwC0AlWmzJEqS__9DOa
"""

import sys

(last_key, max_val) = (None, -sys.maxint)
for line in sys.stdin:
  (key, val) = line.strip().split("\t")
  if last_key and last_key != key:
    print "%s\t%s" % (last_key, max_val)
    (last_key, max_val) = (key, int(val))
  else:
    (last_key, max_val) = (key, max(max_val, int(val)))

if last_key:
  print "%s\t%s" % (last_key, max_val)

[cloudera@quickstart Desktop]$ hadoop  jar /usr/lib/hadoop-0.20-mapreduce/contrib/streming/hadoop-streaming-2.6.0-mrl-cdh5.13
```

cloudera@quickstart:...