Athens University of Economics and Business

Department of Management Science and Technology

MSc in Business Analytics—Big Data Systems

Hadoop Assignment

Associate Professor P. Louridas

Instructions

Your submission will consist of a zipped archive of all the source files that you will write. The zipped archive will have a name of the form:

`STUDENTNUMBER_SURNAME_FIRSTNAME.zip`.

So, if you have written the files `Foo.java`, `Bar.java`, and `FooBar.java`, then the zip archive will contain exactly these three Java files.

The Problem

You will write a Hadoop program that reads a dataset containing measurements from the National Climatic Data Center (NCDC, http://www.ncdc.noaa.gov/). The dataset consists of files, one file per year. Each file consist of lines. Each line is a distinct measurement of meteorological data. The exact data format is explained at ftp://ftp.ncdc.noaa.gov/pub/data/noaa/ish-format-document.pdf.

You will get the input dataset with:

```
> wget -O ncdc_data.tar.gz
↪  https://pithos.okeanos.grnet.gr/public/Pq6yWKjrCLwHmIVkrRuxk3
```

This will download a tar archive, which you will then uncompress and untar:

```
> tar zxvf ncdc_data.tar.gz
```

When this is done, the yearly data files will be contained in a directory called `ncdc_data`.

The output of the program will be a sequence of lines of the form:

```
1901 030650_62167_-333_022267_60450_317
1902 022617_61183_-328_030650_62167_244
1903 030650_62167_-306_030650_62167_289
1904 022617_61183_-294_022267_60450_256
1905 030650_62167_-328_030650_62167_283
```

The first part of the line is the year. The second part contains:

- the latitude of the minimum temperature of the year
- the longitude of the minimum temperature of the year
- the minimum temperature of the year
- the latitude of the maximum temperature of the year
- the longitude of the maximum temperature of the year
- the maximum temperature of the year

Notes

- You may use Hadoop in stand-alone or pseudocluster mode. Remember that if you setup Hadoop in pseudocluster mode then to run it in stand-alone mode you have to undo the changes in the configuration files.
- If you are not familiar with Linux commands, you can have a look at http://linuxcommand.org/tlcl.php.
- If in your solution the key and value of the mapper class are not the same with the key and the value of the reducer class, you can specify which is what with calls like:

  ```
  job.setMapOutputKeyClass(MapperKey.class);
  job.setMapOutputValueClass(MapperValue.class);
  job.setOutputKeyClass(ReducerKey.class);
  job.setOutputValueClass(ReducerValue.class);
  ```

  where of course you substitute the real class names in your applications. It is not a bad idea to use these four calls even if the keys and the values are the same.

Go for it!