

**CS4545/CS6545 – Hands-on 3**  
**Winter 2021, UNB, Fredericton**  
**Due: February 25, 2020 at 11 am**

In this hands-on, you'll become familiar with Hadoop and MapReduce programming paradigm.

**INSTRUCTIONS: launching Hadoop and running the demo program**

1. Remotely connect to the FCS lab (see *RemoteDesktopToALabMachine.pdf*)  
Launch and login to the BigDataSystems Master VM and the 3 Slave VMs.
2. On BigDataSystemsMaster VM type the command on a terminal:  
start-all.sh  
This will start the Hadoop cluster (on all 4 VMs).
3. On BigDataSystemsMaster VM, go to /home/bigdata/eclipse/java-mars/eclipse and launch Eclipse.
4. Once Eclipse is started, create a new java project named *hadoopdemo* as: File> New> Java Project
5. Get the hadoopdemo.zip file by typing the command below:  
\$ wget https://www.cs.unb.ca/~sray/teaching/bds/handson/hadoopdemo.zip  
  
Unzip it (command: unzip *filename*) and copy and overwrite the folders *src*, *lib* and *input* to the Eclipse project you just created.
6. Include the jar files in the *lib* folder in your Eclipse project.
7. Once the *.java* files are compiled, take the *.class* files in the *bin* folder and create a jar file *maxt.jar*  
(Note: the *.class* files are in *bin* folder of the workspace. Inside it type command: jar cvf *maxt.jar \*.class*)
8. You can run the demo by running the commands:  
hadoop dfs -mkdir /input  
cd /home/bigdata/eclipse/java-mars/eclipse/workspace/hadoopdemo  
hadoop fs -copyFromLocal input/data /input/.  
hadoop jar ./maxt.jar MaxTemperatureJobRunner /input /output  
hadoop dfs -lsr /output  
hadoop dfs -cat /output/part-r-00000  
hadoop dfs -rmr /output (note: output folder must be deleted before the next run)
9. Finally, to stop the Hadoop cluster (on all 4 VMs) on BigDataSystemsMaster VM type the command:  
stop-all.sh

**INSTRUCTIONS: Task and deliverables**

1. Download the dataset *worldcities.csv*  
\$ wget https://www.cs.unb.ca/~sray/teaching/bds/handson/worldcities.csv

It contains information about many cities of the countries of the world. The first line of the data file contains information about the fields, which are self-descriptive, as follows:

*"city","city\_ascii","lat","lng","country","iso2","iso3","admin\_name","capital","population","id"*

2. Write a Map-Reduce application using Hadoop APIs (as shown in the demo) to show the largest and the smallest city and their population for each country.
3. Run your code with the Hadoop cluster in the lab VMs and copy/rename the output into a plain text file called *output.txt*

In the output (*output.txt*) file there will be two lines of text for each country, the first line showing the smallest city (by population) and the second line showing the largest city (by population) and their population. An example is shown below:

```
...  
Canada    Antigonish    5002  
Canada    Toronto      5429524  
...
```

4. **TIPS:**

You may want to debug your code by running it locally within Eclipse. For that, please refer to the lecture slides (slide # 39) titled “Write the driver class to run/debug locally” in the lecture notes L05\_CS4545\_CS6545\_Hadoop.pdf

The code from the textbook (Hadoop: The Definitive Guide. By: Tom White) is available at: <https://github.com/tomwhite/hadoop-book/>

## INSTRUCTIONS: D2L Submission

1. Submit the via Desire To Learn (D2L) with the following:
  - a. \*.java files in a single zip file: h3\_<your\_student\_id>.zip
  - b. A plain text file containing the output: *output.txt*
  - c. A screenshot (.png or .pdf file) showing the execution/run of your program in the Cluster mode in the lab VMs.
2. Hands-on not submitted electronically via D2L or submitted after the due date will NOT be marked.

## NOTES ABOUT PLAGIARISM

Please note that the handsons are meant to be done individually. Any submission that appears to be in violation of an academic offence (plagiarism) may be reported to the Registrar’s Office as per UNB regulations (See section VII of UNB Undergraduate Calendar).