Exercise 19

For the next couple of weeks we need a working HADOOP installation. However, installing HADOOP itself on your computer is more than tricky. Therefor we will use a LINUX image that already has everything we need installed. In order to run this image properly I recommend running it virtually. For that you need a software called VirtualBox. Please go here:

https://www.virtualbox.org/

and install VirtualBox on your computer. Virtual box already is quite large, so please do this soon.

After you installed VB, you also need an “image”. This image is a snapshot of a working operating system that in our case also has Hadoop installed. We will use the Cloudera Quickstart VM, which you can download here:

http://www.cloudera.com/content/support/en/downloads/download-components/download-products.html?productID=F6mO278Rvo

please choose the one for VirtualBox. In case you want to use a different virtualization software like VMWare, feel free to do so, but be aware that you should only use this option if you know what you are doing.

The Cloudera Quickstart VM is a 2.6 GB zip file, you need a little time to download it, do this early and not in the last minute.

The Cloudera Image is compressed, I recommend unpacking it already since this also takes a couple of minutes.

I will show you how to run the image on the VirtualBox in class.

1) Open Cloudera-quickstart-vm-4.4.0.-1-virtualbox (takes a couple of minutes to load)

2) start the hadoop service using the firefox interface, username cloudera password cloudera, start the console (little black icon in the title bar of the screen)

We are going to navigate and play around with the HDFS for a couple of moments.
HDFS commands to know
ls command on the HDFS, lists the content of the hdfs:
3) hadoop fs -ls

We are going to make a folder for our word count example
4) hadoop fs -mkdir /user/cloudera/wordcount

And within this folder we make another folder where we will drop the input files
5) hadoop fs -mkdir /user/cloudera/wordcount/input

Let us make another folder with a wrong name

6) hadoop fs -mkdir /user/cloudera/wordcount/impult

Removing this folder using:
hadoop fs -rm /user/cloudera/wordcount/impult
simply gives us an error saying this is a folder … -r was missing
7) hadoop fs -rm -r /user/cloudera/wordcount/impult

make a regular folder called wordcount and change into it
8) mkdir wordcount
9) cd wordcount

create files with words as content, we need them later
10)
echo “the lazy fox jumped over another lazy fox” > file0
echo “what does the fox say” > file1

Let us move both files onto hdfs into the input folder
11) hadoop fs -put file* /user/cloudera/wordcount/input

Did it arrive?
12) hadoop fs -ls /user/cloudera/wordcount/input

Let us remove them, just to show how files are removed
13) hadoop fs -rm /user/cloudera/wordcount/input/*

Check using 12) if they are gone, and just put them there again using 11)

Now we need to get the word count example file, goto google “cloudera word count” go to the first link, click on the word count tutorial v1.0 and then goto source code, copy the entire example, go back to the command line and execute emacs on WordCount.java

14) emacs WordCount.java

15) Paste the entire content into this file, save, and exit emacs

Check if you actually filled stuff into this file:
16) less WordCount.java
Now we need to create a folder that will contain our java project
17) mkdir wordcount-classes

Wordcount.java now will be compiled using the java compiler
18) javac -cp <classpath> -d wordcount-classes/ WordCount.java
Here the class path has to be replaced by the actual path, which in vm4 is this: /usr/lib/hadoop/*:/usr/lib/hadoop/client-0.20/*

We need to make a jar package using jar
19) jar -cvf wordcount.jar -C wordcount-classes/ .

This can now be executed on hadoop:

20) hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/wordcount/input /user/cloudera/wordcount/output
You will see a couple or updates on the screen

Let us check the results in the output folder:
21) hadoop fs -ls /user/cloudera/wordcount/output

there are three files: SUCCESS,_log, part-00000
you can either look at them while they remain of the hdsf:
22) hadoop fs -cat /user/cloudera/wordcount/output/part-00000

Or you retrieve the file first and then inspect it using less
23) hadoop fs -get /user/cloudera/wordcount/output/part-00000 .
24) less part-00000

Handin a screenshot of the console showing your successful execution of the hadoop wordcount job