**Hadoop streaming with R tutorial**

In this tutorial we will work through steps to run the word count example with mappers and reducers written in R and show you how to use hadoop streaming to perform the ordinary map-reduce.

You need a machine with hadoop and R installed on it. If you use the Ubuntu image that we provided you need to install R via this command:

sudo apt-get install r-base

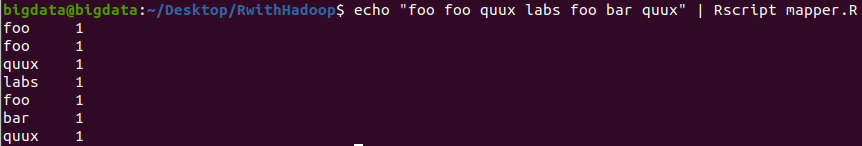
Steps to run a map-reduce program in R with Hadoop:

1- Write your map and reduce logic in two separate files (e.g. mapper.R & reducer.R)

2- Both mapper and reducer reads input via stdin and output via stdout to work with Hadoop streaming → refer to mapper.R and reducer.R

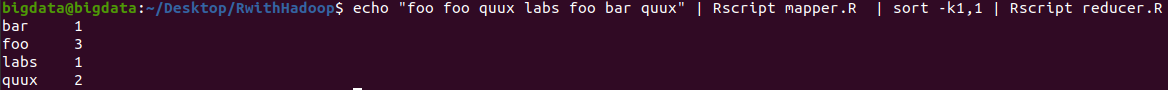
3- Let’s test locally the map logic

echo "foo foo quux labs foo bar quux" | Rscript mapper.R



4- Let’s test locally the reduce logic

echo "foo foo quux labs foo bar quux" | Rscript mapper.R | sort -k1,1 | Rscript reducer.R



At this point we are sure that map and reduce logic works as expected, Let’s move to Hadoop

5- Start hadoop

start-all.sh

6- Create directories on HDFS for this tutorial

hdfs dfs -mkdir /RwithHadoop

hdfs dfs -mkdir /RwithHadoop/Input

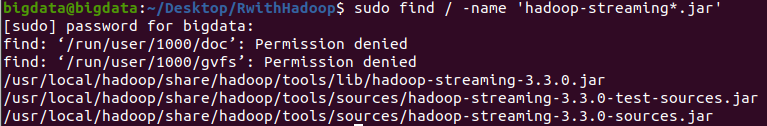
7- move input.txt from local FS to HDFS (make sure that the terminal current directory is where input.txt exists)

hdfs dfs -put input.txt /RwithHadoop/Input

8- Now, we need to run Hadoop streaming with mapper.R and reduce.R and for this to work we need to find hadoop-streaming.jar via this command:

sudo find / -name 'hadoop-streaming\*.jar'

This will ask you for the password (if you use our machine the password is: 123456)



The path that we need here is:

/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.0.jar

Where 3.3.0 is the hadoop version.

This jar file is the hadoop streaming where it can take external mappers and reducers with your favourable language (R, Python, …)

9- Run the map reduce job via this command:

hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.0.jar -file /home/bigdata/Desktop/RwithHadoop/mapper.R -mapper /home/bigdata/Desktop/RwithHadoop/mapper.R -file /home/bigdata/Desktop/RwithHadoop/reducer.R -reducer /home/bigdata/Desktop/RwithHadoop/reducer.R -input /RwithHadoop/Input -output /RwithHadoop/Output

10- Print the map reduce output:

hdfs dfs -cat /RwithHadoop/Output/\*

