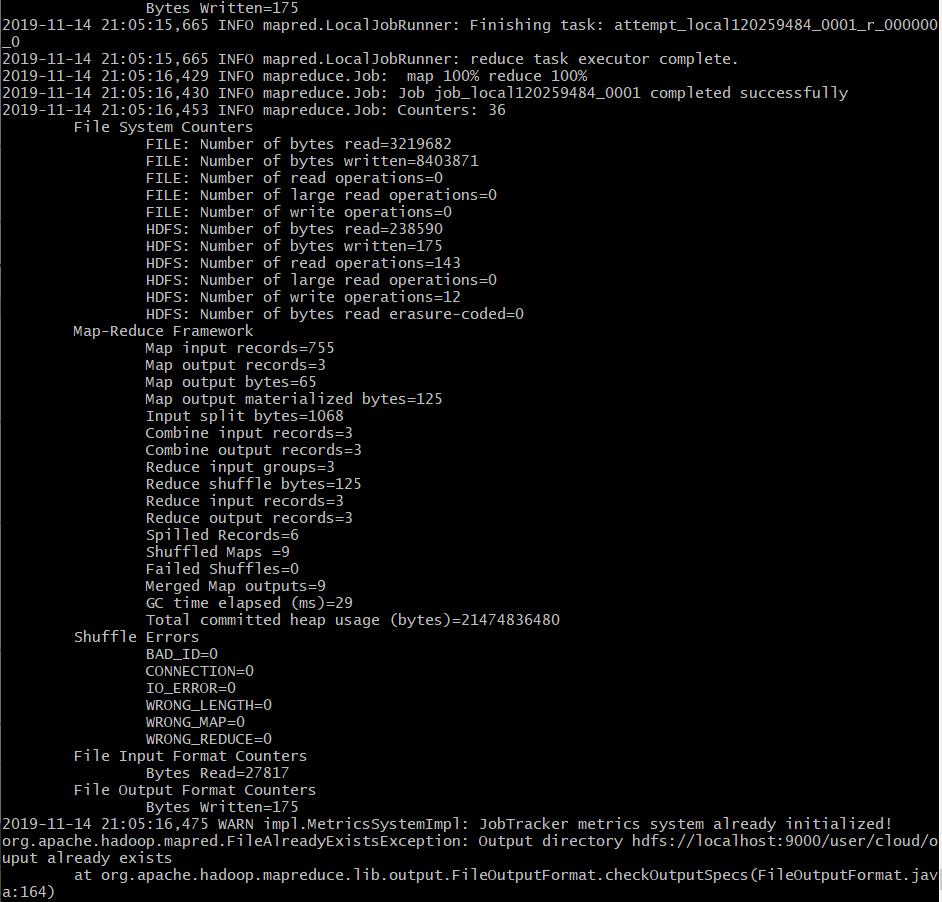
This is a very meaningful problem. Using Hadoop is very interesting. The operation in the HDFS is quite similar to what we do with the normal Linux system. The differences focus on that we should upload our operation to the DFS and retrieve the result throw the network.

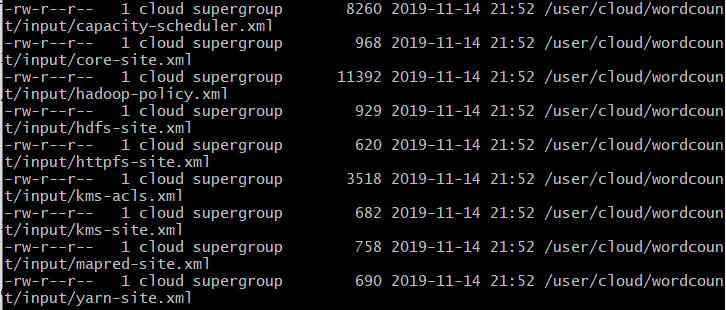
To use Hadoop on a server, I first set the standalone mode on a server and then change it to pseudo parallel mode. The tutorial for the pseudo parallel mode is not complete actually. It still requests some other settings like HADOOP\_CLASSPATH on our own server while the temporary folder is also missing in the set-up tutorial.

For deploying on the EMR as fully parallel mode, it is also quite different. The JAVA version on the EMR is OpenJDK 1.8, the JRE and JDK are separated. Therefore, to deploy the demo, we need to find the JDK in some other folder but not in the JAVA\_HOME path. The environment in EMR can still get improved for better user experience.

Following shows some of the screenshot when running Hadoop.

This is the result of the example (running in pseudo-parallel mode) in the share folder:

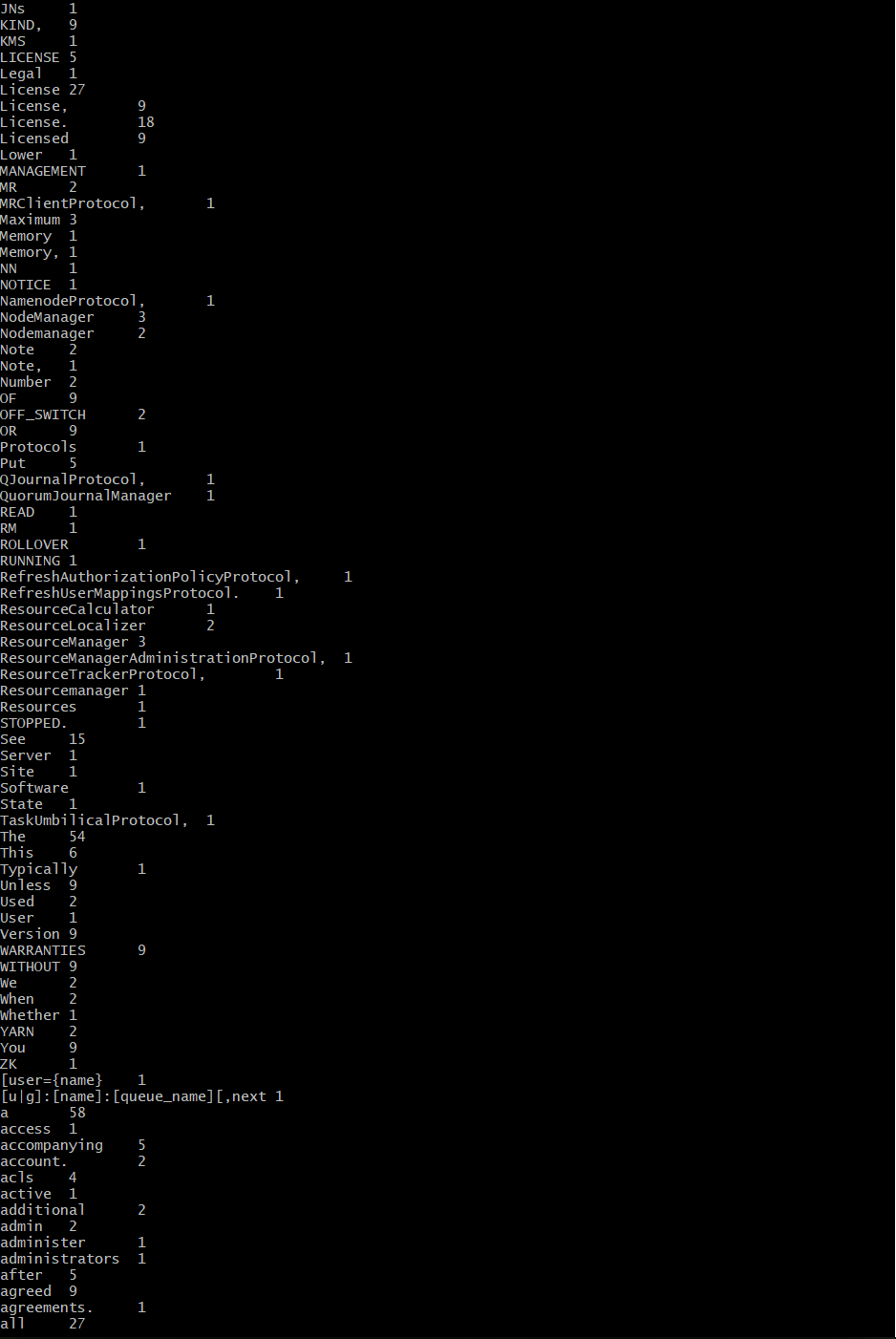
These are the input files:

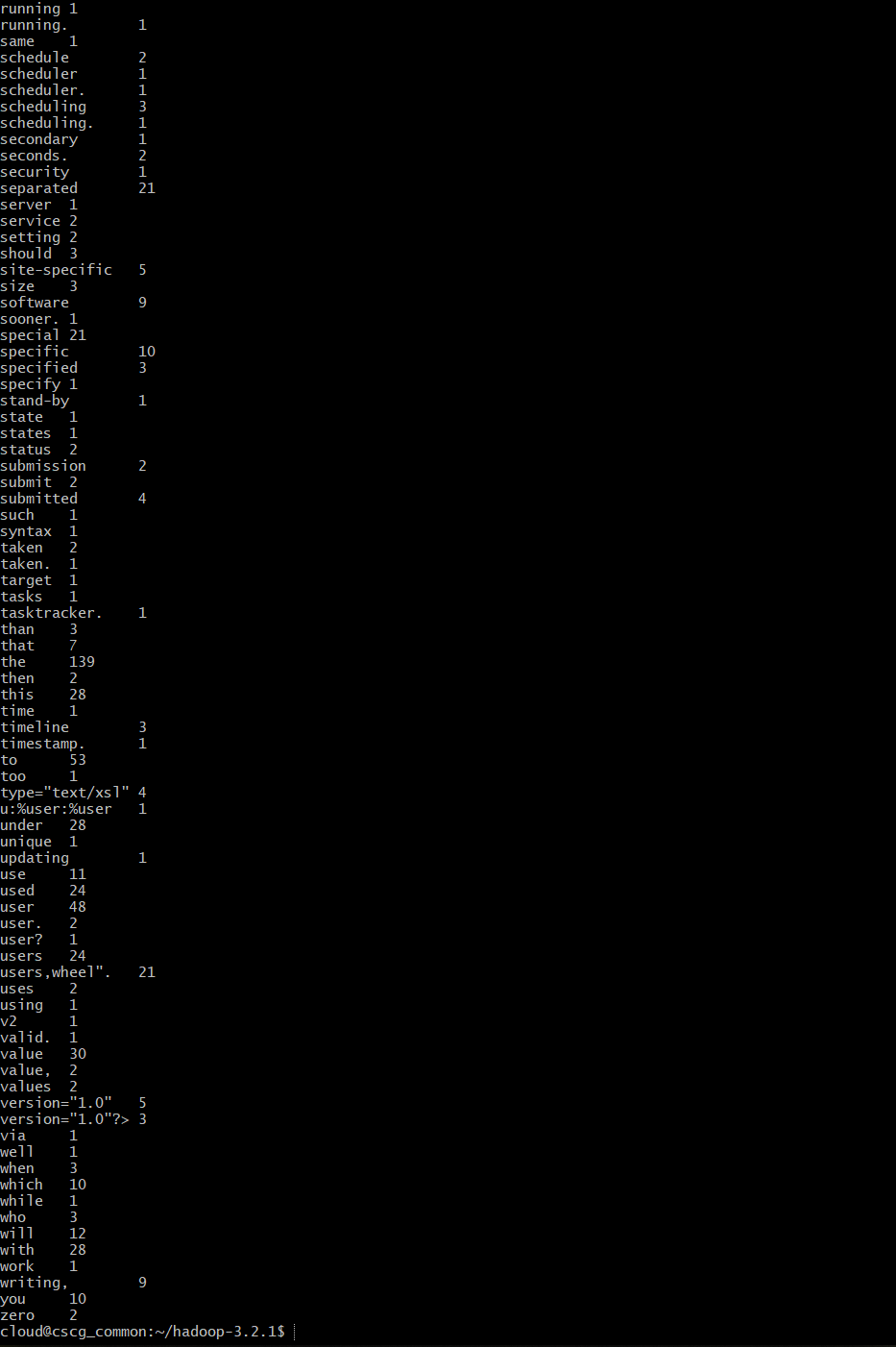


Output for the wordcount program are shown below:



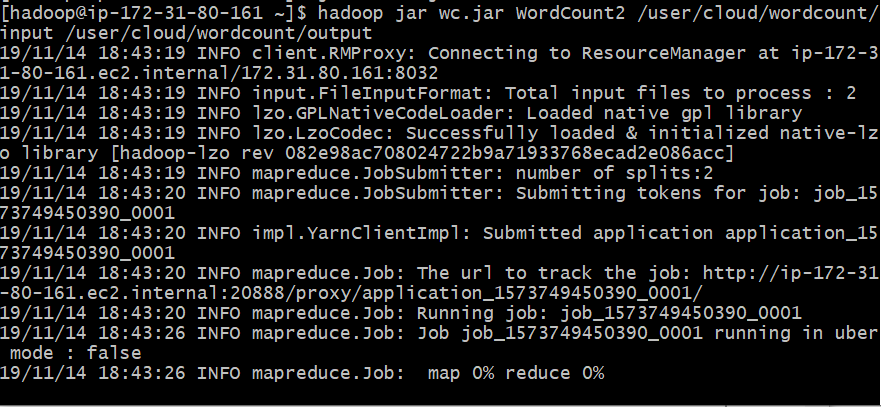




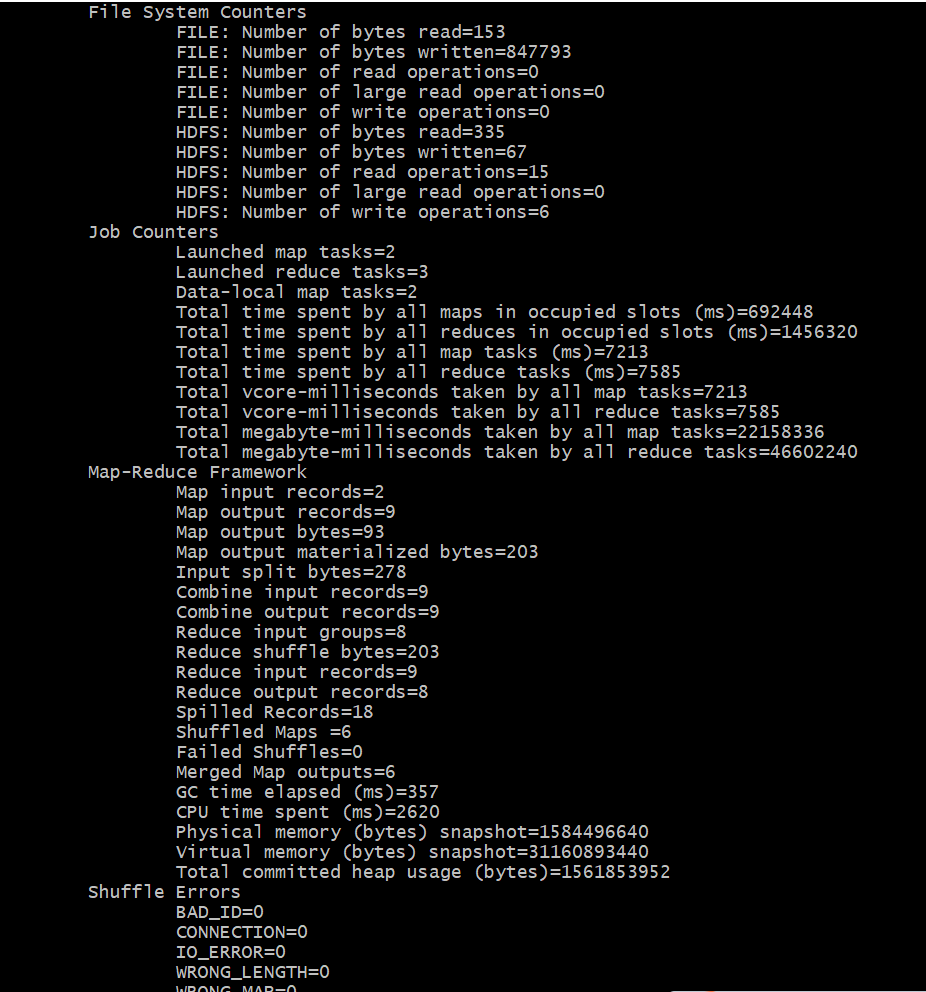


All the output above are part of the result but not the total output.

Down here is the screenshot when running the map reduce in EMR



Down here is part of the running end of the mapreduce sample:



This is the result of the map reduce running in EMR

