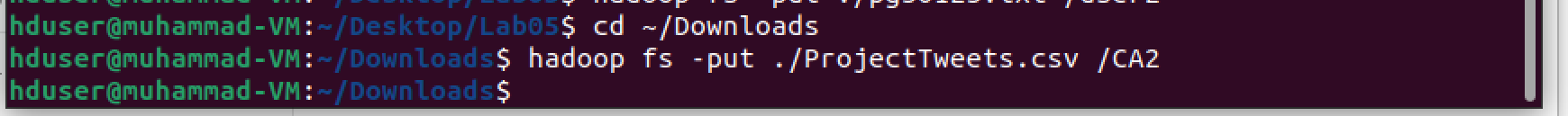
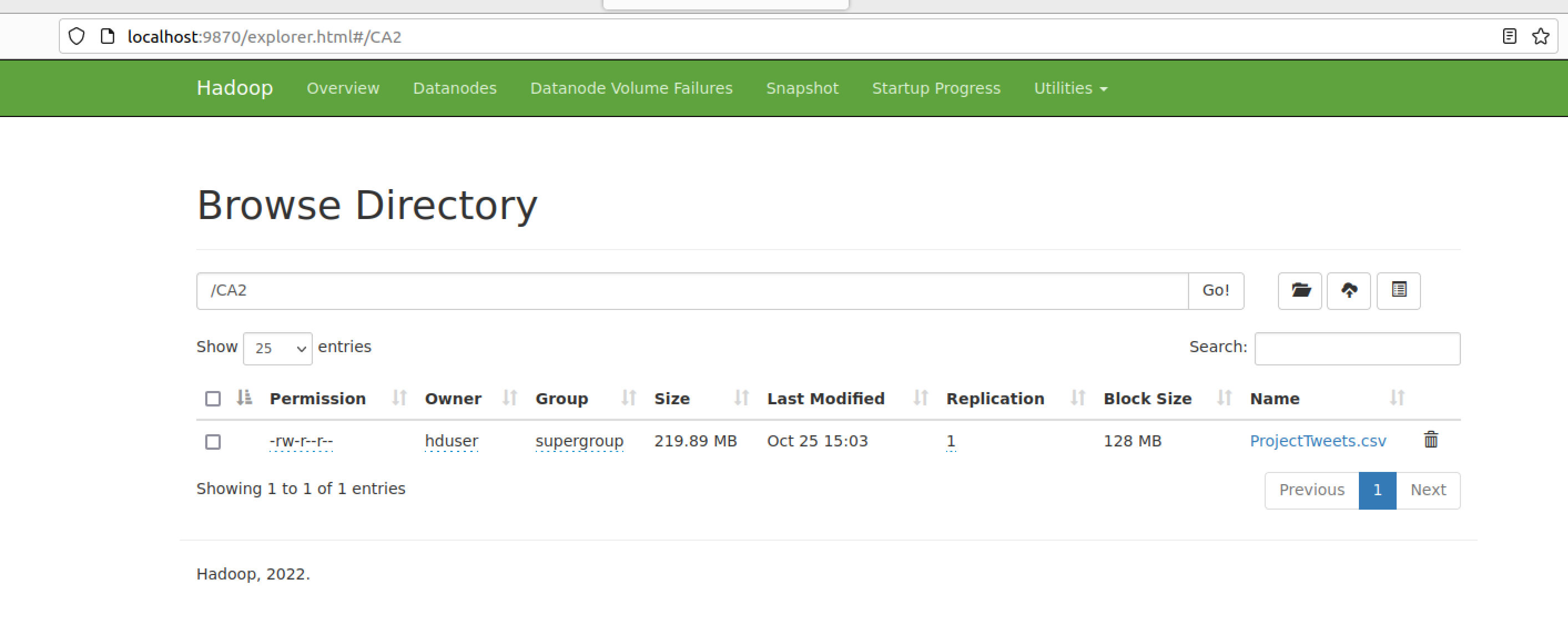
First I put the file on HDFS in a folder called CA2:



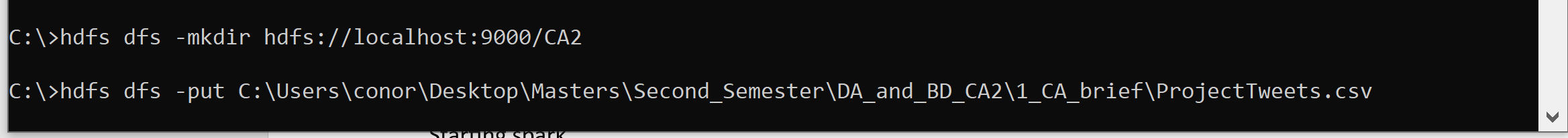
Then I checked to make sure it was present in the HDFS:

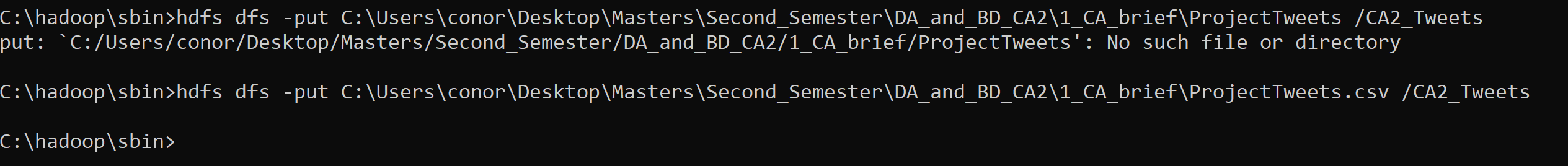


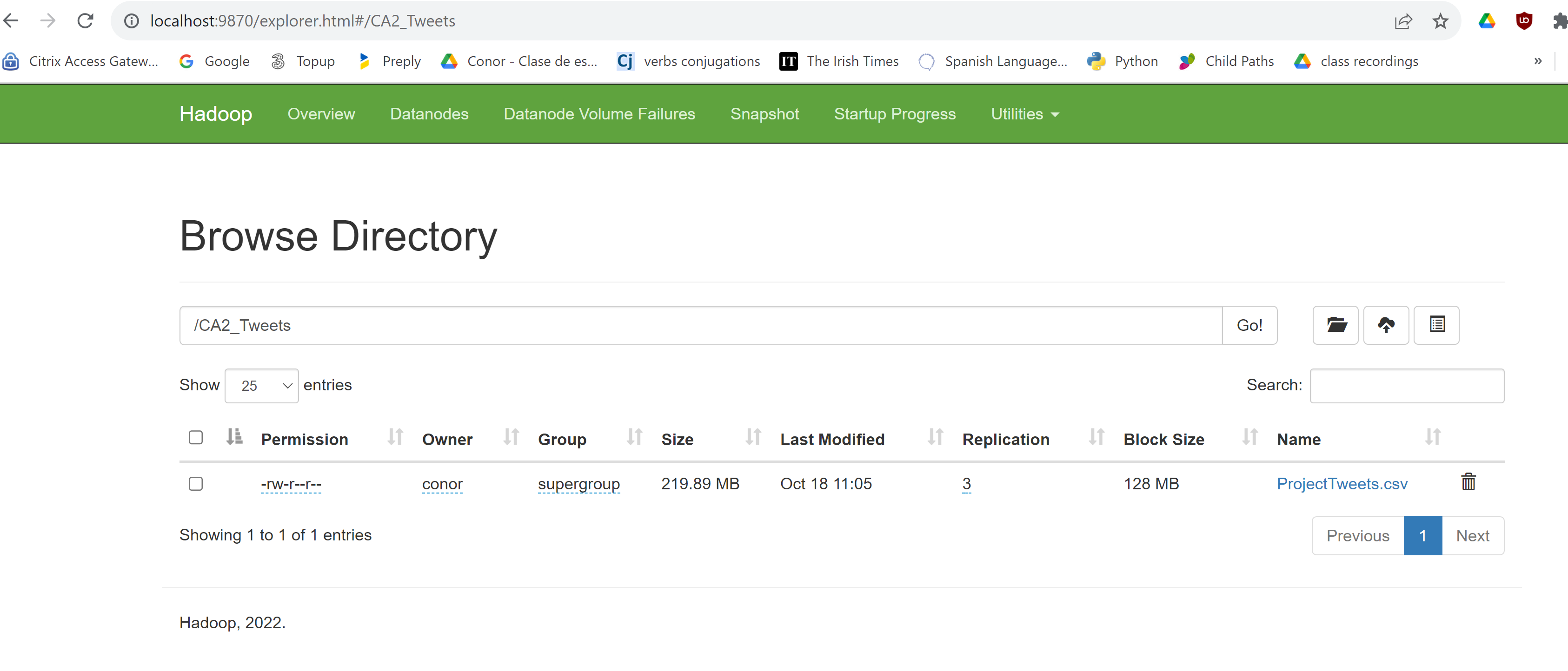
Loading the file into HDFS

I opened HDFS using the CMD prompt (ran as an administrator) and I uploaded the ProjectTweets dataset using the put function in the CMD line:

Initially there were issues because the file path had spaces within it but after those spaces were removed, the file successfully uploaded into HDFS:



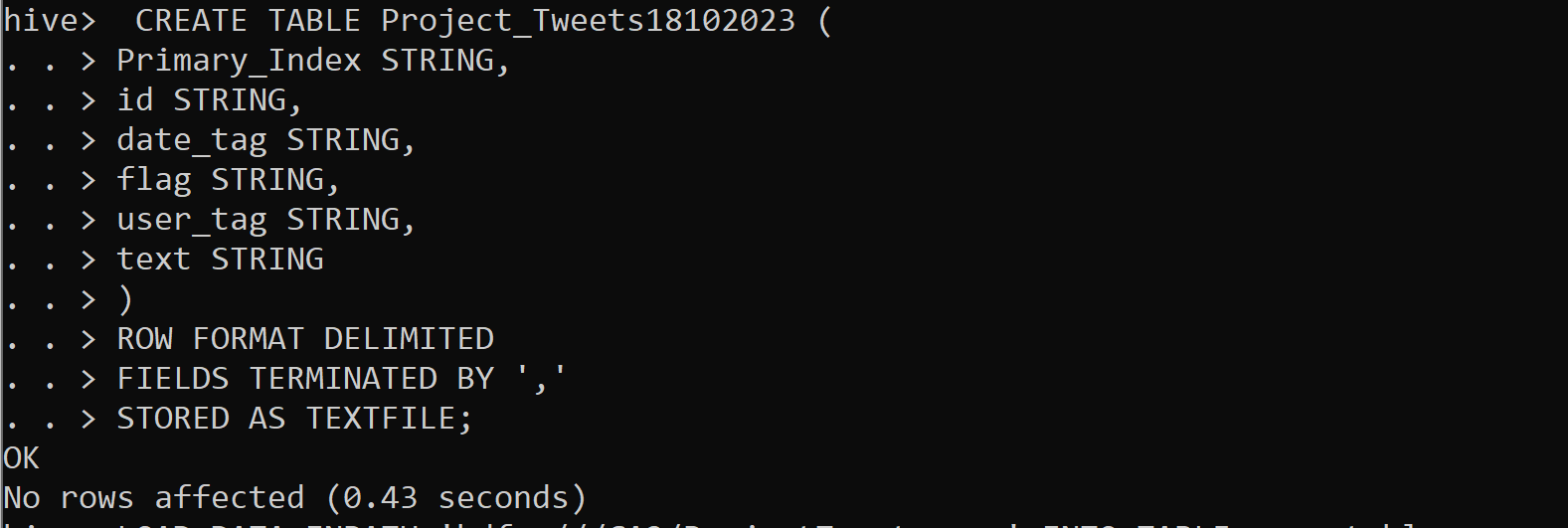


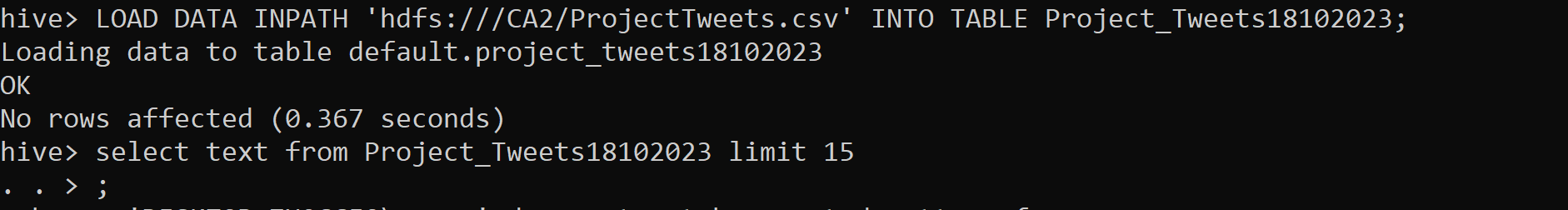


We are using a standalone Hadoop distribution. But it’s possible to check the cluster detail, so it’s possible to check how many clusters on the standalone machine Hadoop is developing.

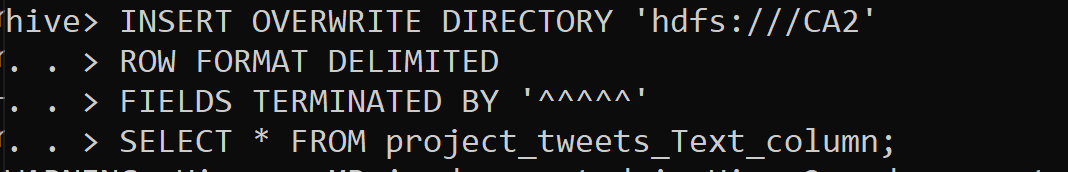
If we had different machines we could decide how to distribute the data, Hadoop automatically takes responsibility that but there are. Hadoop does it all internally.

HIVE





Send back the one column for the map reduce job



Mapreduce:

C:\>hadoop jar C:\hadoop\share\hadoop\mapreduce\hadoop-mapreduce-examples-2.10.2.jar wordcount /CA2/000000\_0 /output/part\_1

