Read Me

Here, I use Amazon EMR and S3 to run hadoop mapreduce programs. The instances is 3, so I got 3 pieces of result of each wordcount task. The input file bible, jar files, the file for distributed cache and outputs are all stored in Amazon S3, and also HDFS file system. First space of each line (bible.gz) is not calculate for the result.

Also, all the result txt files are in the result package.(one wordcount--wcresult, double wordcount--dcresult, distributed cache--distributedcacheResult)

The link is <https://console.aws.amazon.com/s3/buckets/wordcountj> (may need to log in)

Result link:

Onewordcount: <https://s3.us-east-2.amazonaws.com/wordcountj/output/wcresult/part-r-00000>

<https://s3.us-east-2.amazonaws.com/wordcountj/output/wcresult/part-r-00001>

<https://s3.us-east-2.amazonaws.com/wordcountj/output/wcresult/part-r-00002>

Doublewordcount:

<https://s3.us-east-2.amazonaws.com/wordcountj/doubleoutput/dcresult/part-r-00000>

[https://s3.us-east-2.amazonaws.com/wordcountj/doubleoutput/dcresult/part-r-00001](https://s3.us-east-2.amazonaws.com/wordcountj/doubleoutput/dcresult/part-r-00000)

<https://s3.us-east-2.amazonaws.com/wordcountj/doubleoutput/dcresult/part-r-00002>

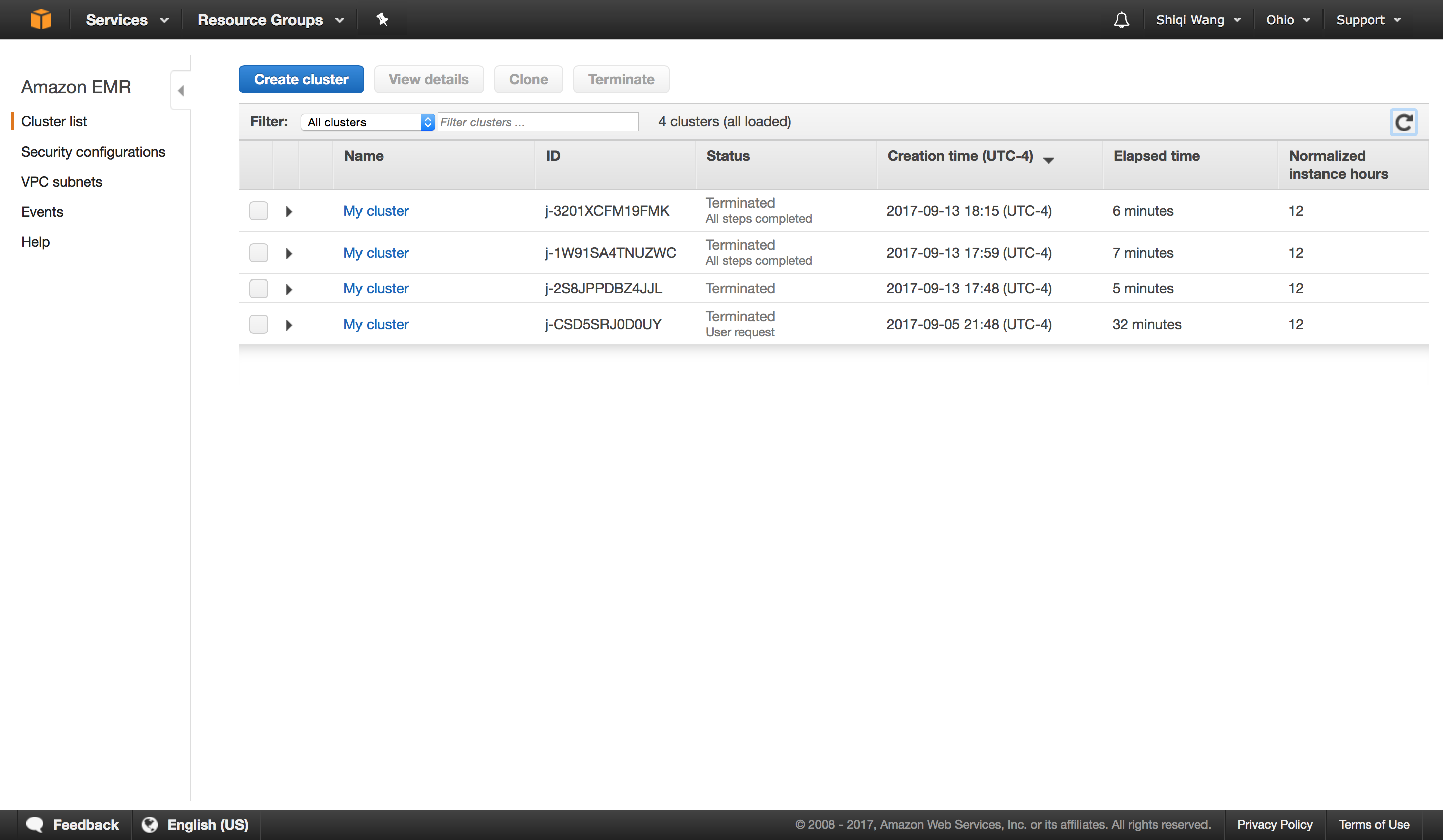
Distributedcache:

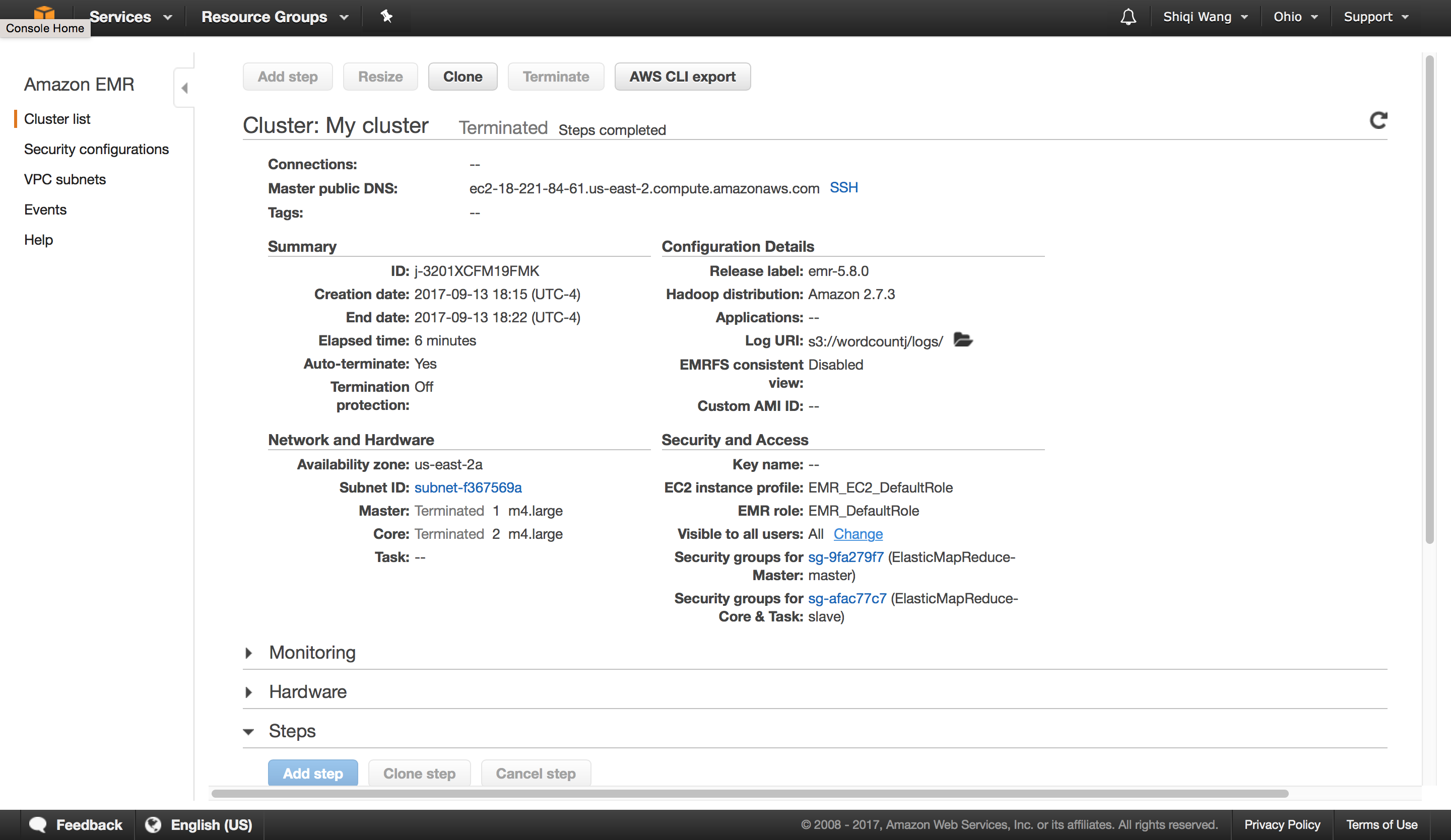
<https://s3.us-east-2.amazonaws.com/wordcountj/discoutput/result/part-r-00000>

<https://s3.us-east-2.amazonaws.com/wordcountj/discoutput/result/part-r-00001>

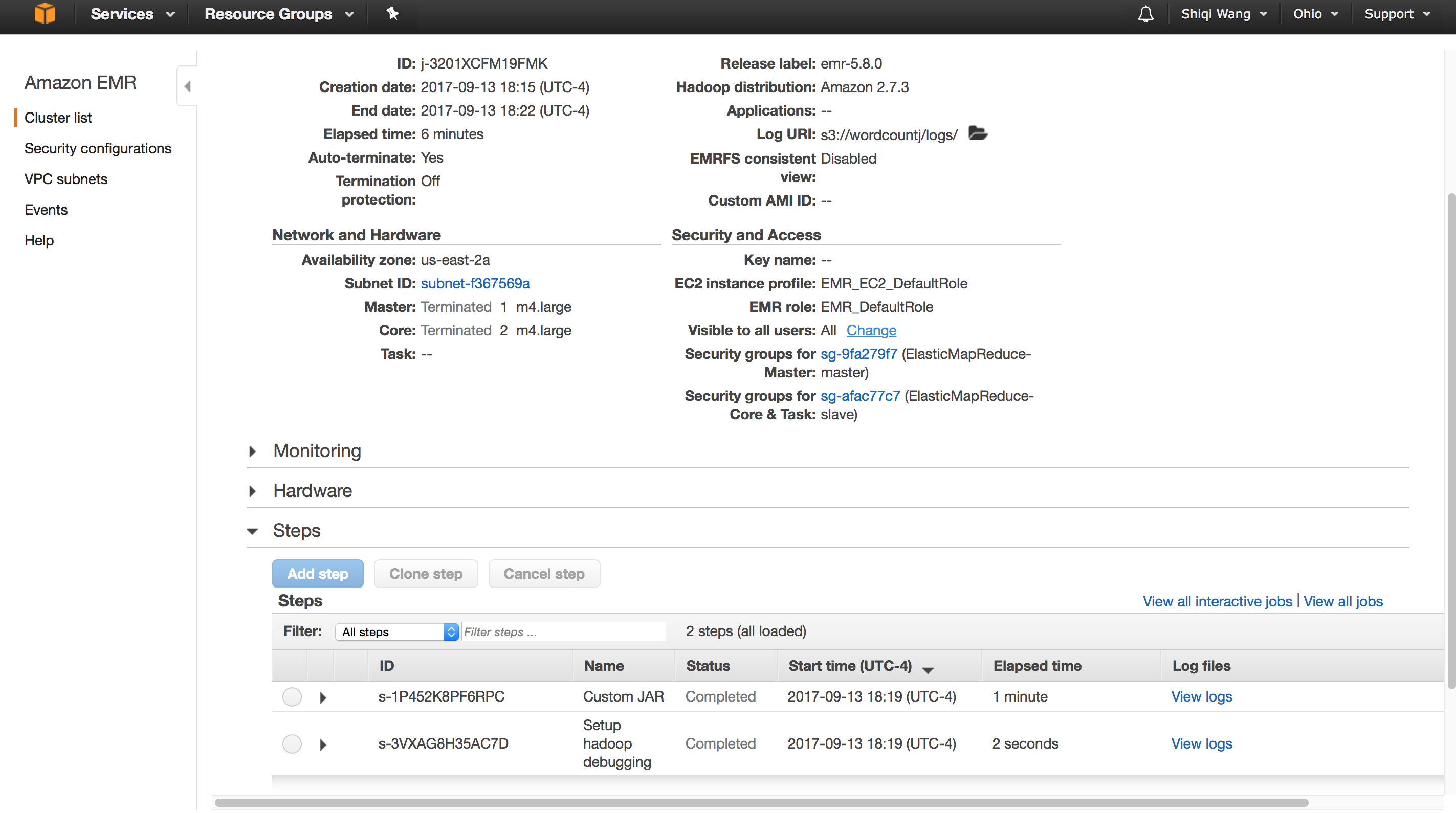
<https://s3.us-east-2.amazonaws.com/wordcountj/discoutput/result/part-r-00002>

Screenshots for AWS:

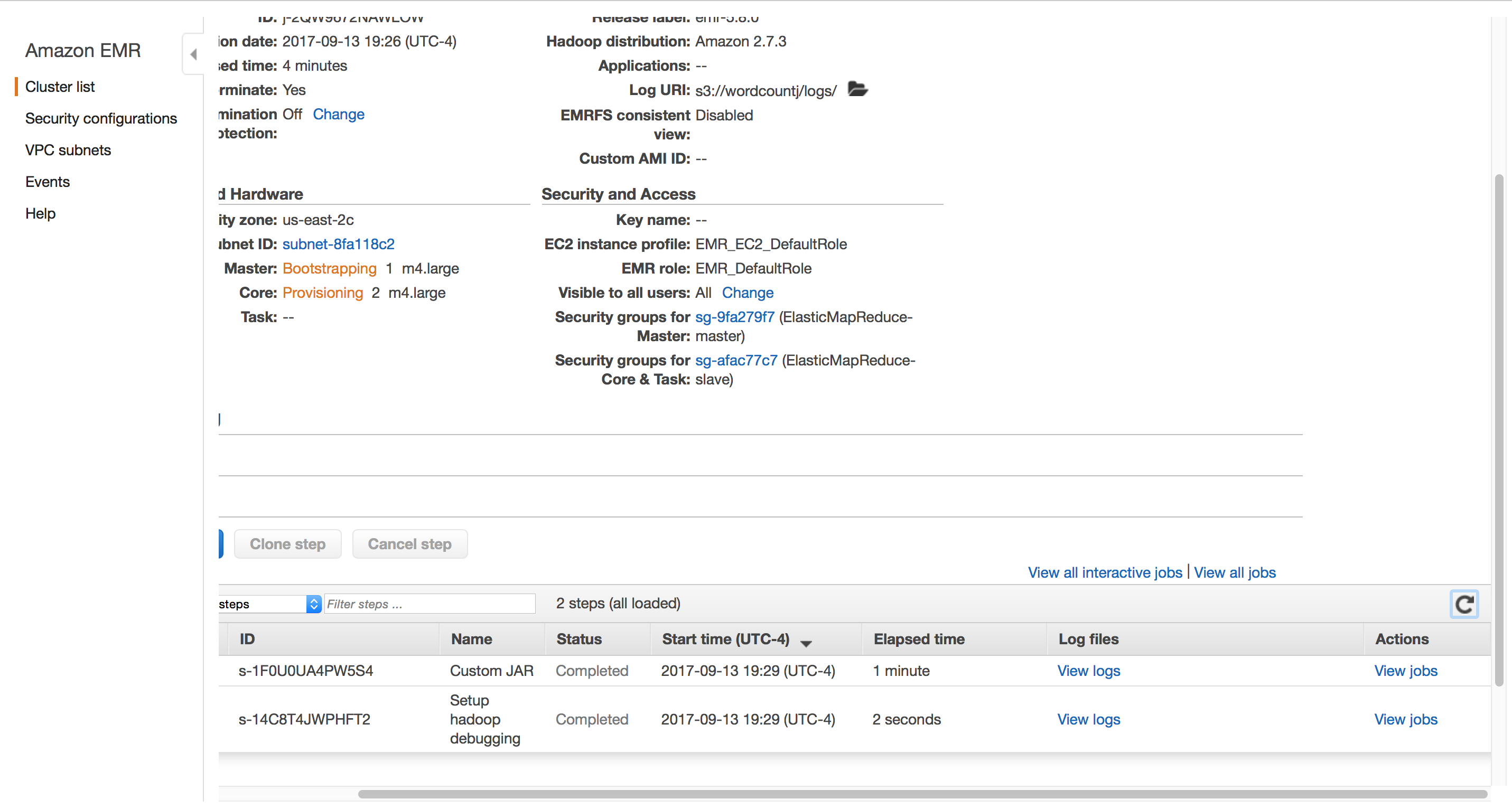




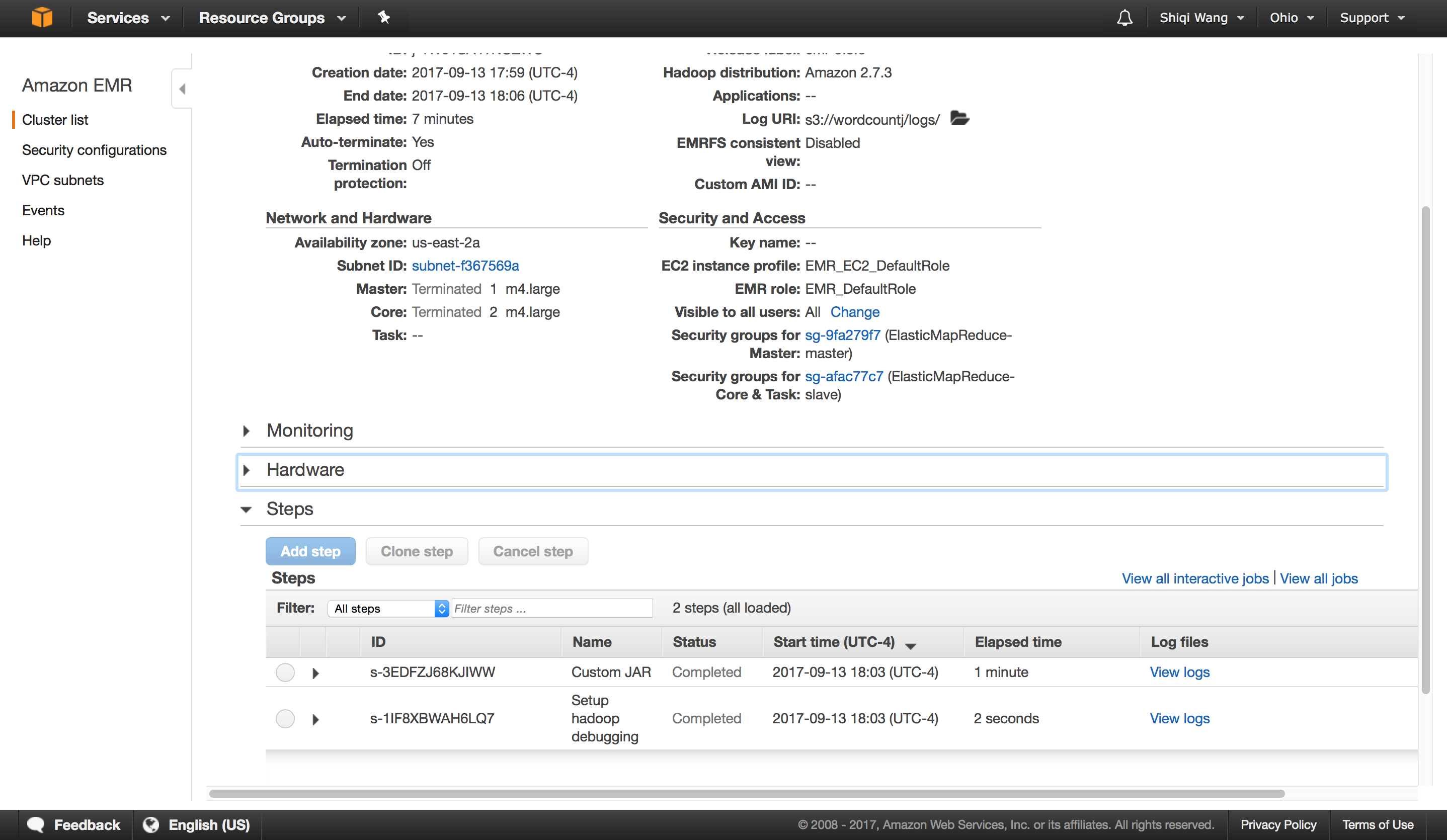
one wordcount



double wordcount



distributed cache



Amazon S3 result

