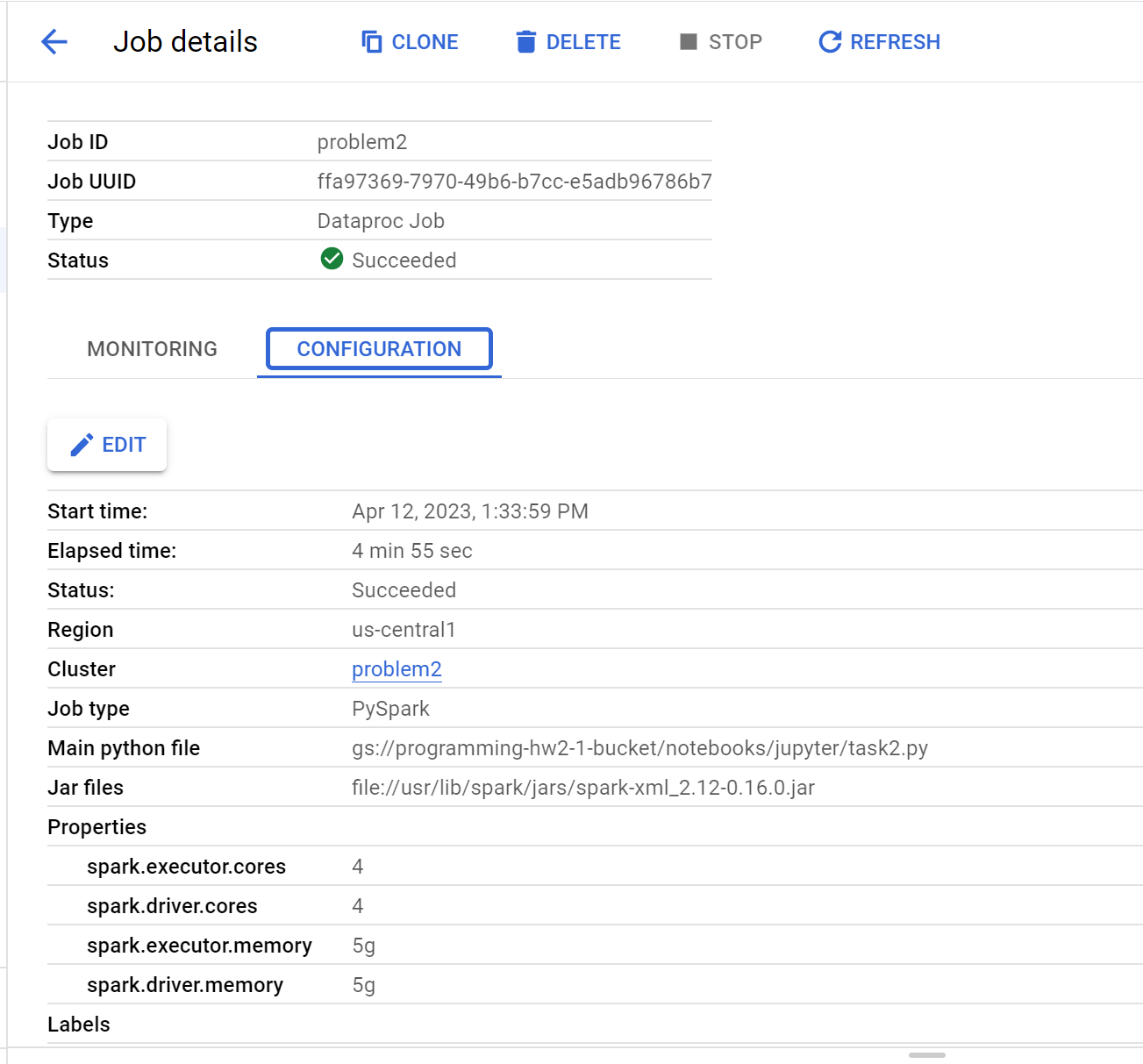
Question 1:

The default block size on HDFS is 128 MB.

The default replication factor on Dataproc is 2.

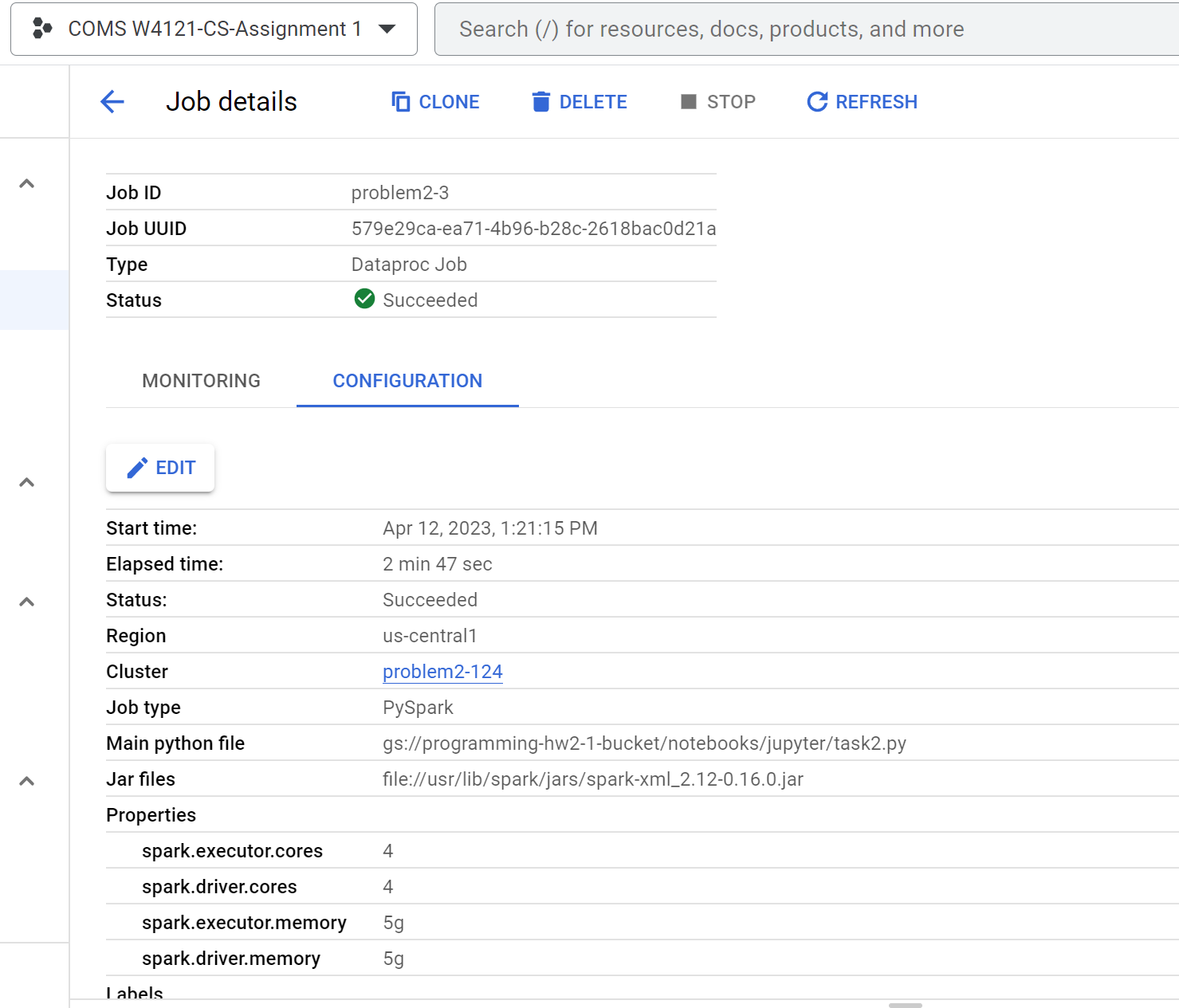
Question 2:

The execution time is 4min 55sec



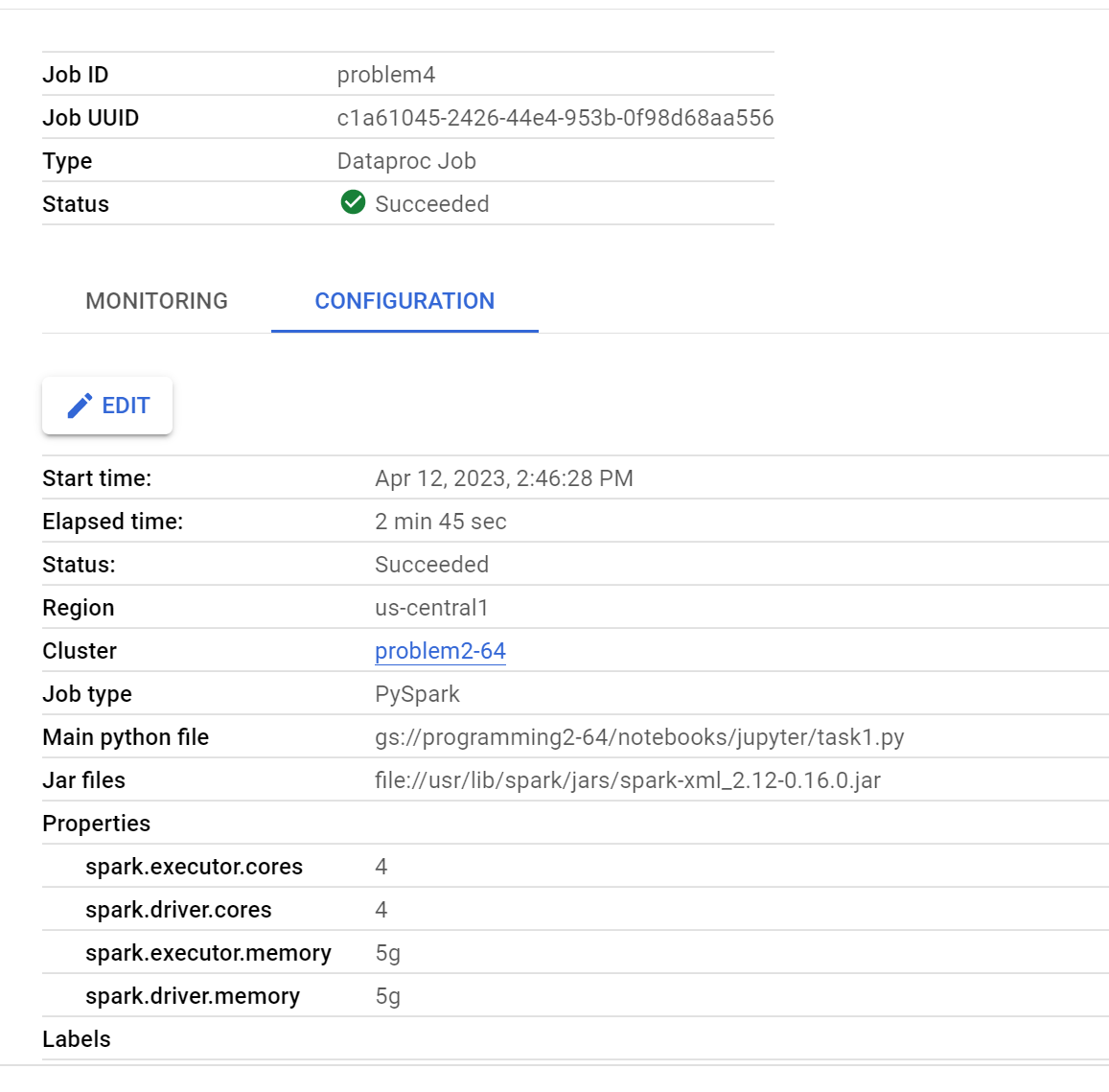
Question 3:

The execution time is 2min 47sec. It is much faster than a single node cluster, the performance is much better than single node cluster. Basically, because we have two more worker nodes which can help to run the tasks. It will do the tasks simultaneously which will speed up the total process. See the screenshot:



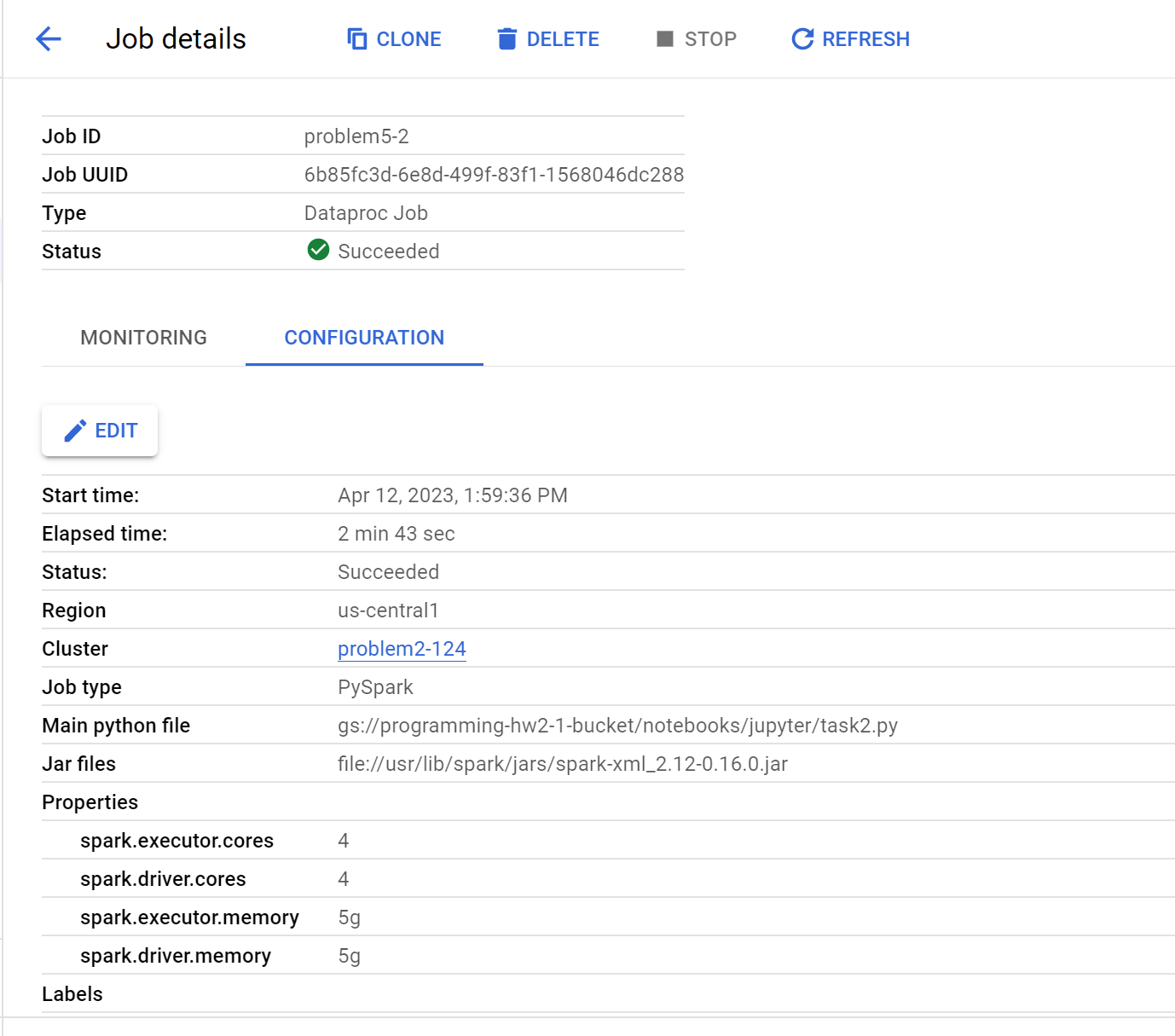
Question 4:

The execution time is 2min 45sec under block size is 64MB. It is almost the same as question 3 (a little faster than question 3). See the screenshot. This makes sense since block size does not have a large impact on running time.

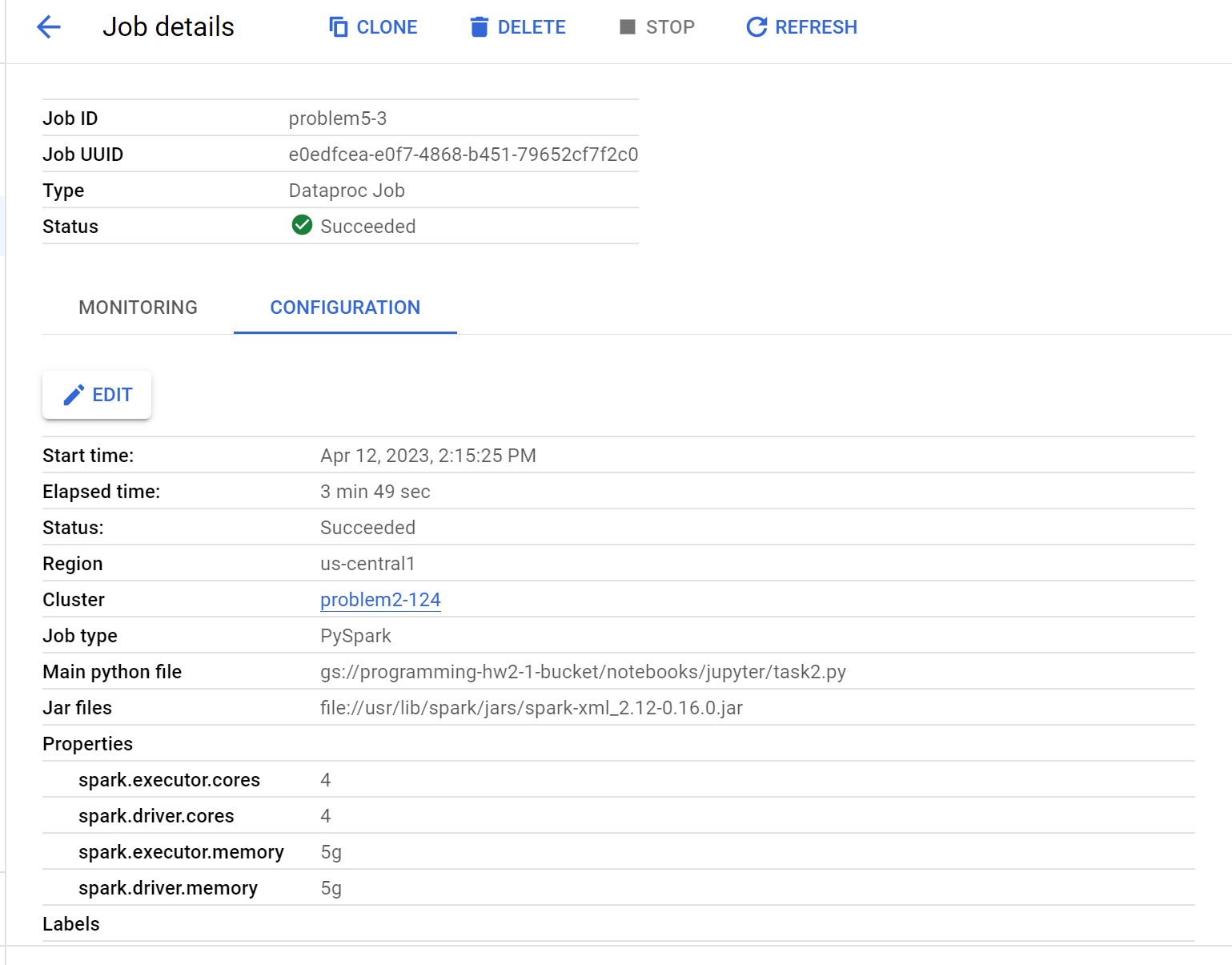


Question 5:

The job with two worker nodes takes 2min 43sec to complete it. See the screenshot:

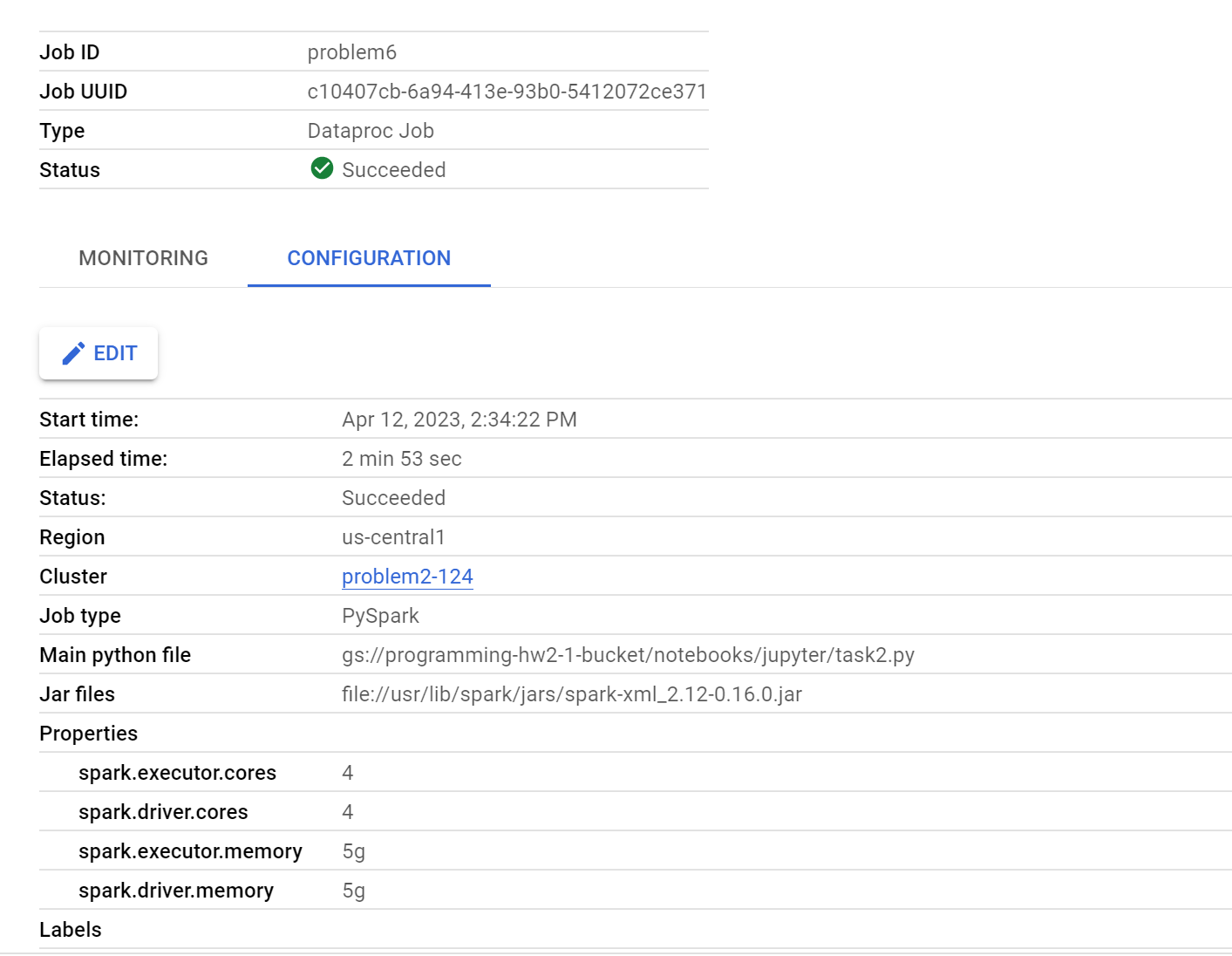


While when I stop one of the workers, the job still executed successfully in the end. The job with only one worker node takes 3min 49sec to complete it, one more minute for the execution time compare with the original task. This makes sense as we now only have half of the workers to work, which, theoretically will double the execution time. See the screenshot:



Question 6:

We can see from screenshot that when change replication factor to 1, the execution time almost the same (only a little slower than question 5) compare with question 5. Which is 2min 53sec. This makes sense because when clusters run successfully, the master does not need to access replicated data to fix some data failures, so the whole execution processes are not affected by replication factor, this task supposed to be the same as problem 5.



Question 7:

When set block size to 64MB and repeat question 5, we can see from the screenshot that the execution time almost unchanged, still 2min 43sec, same as the question 5. We can conclude that the block size has very limited influence to the total execution time. This makes sense as block size only control each block’s size, doesn’t prolong or influence the whole working processes.

