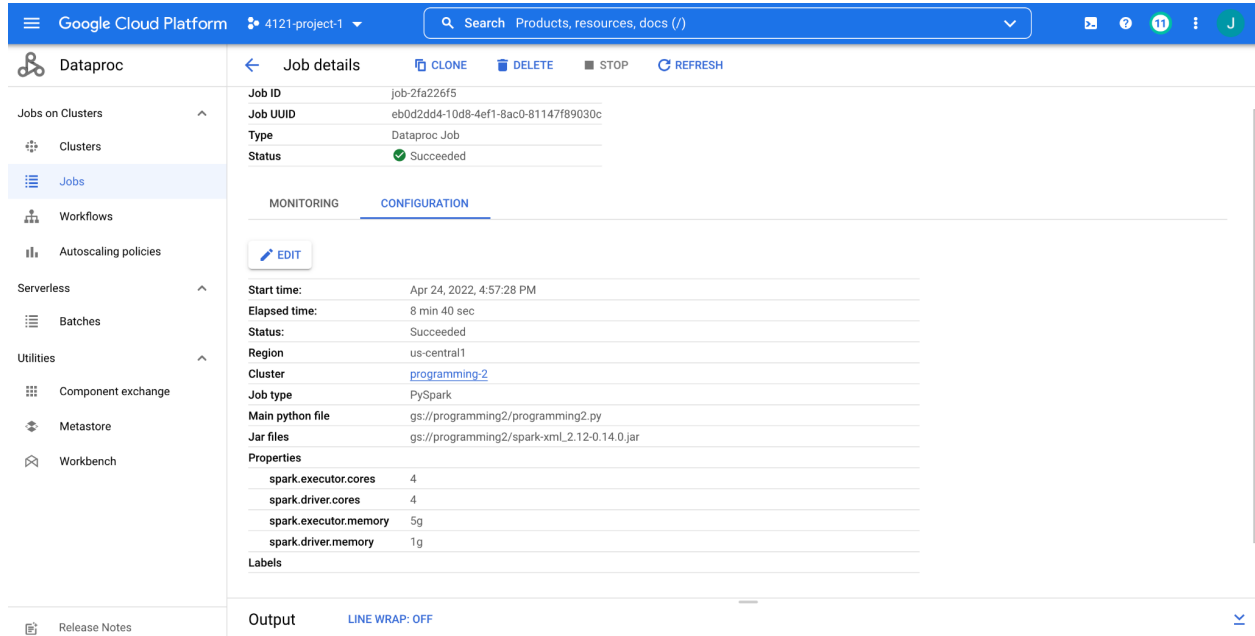# Computer System Programming 2 README
Junhao Zhang jz3430 and Yuning Di yd2617

Question 1: The default block size is 128MB and the default replication factor is 2

Question 2: The completion time is shown in the below graph: 8 min 40 sec

Question 3: The completion time is 4 min 41 sec. The performance is getting better in terms of the completion time. This is because we are using 3 node clusters with 2 works and thus it's more efficient than the previous problem.

Question 4: The completion time is 4 min 44 sec. We can see that even though we decreased the block size from 128MB to 64 MB, the completion time has not increased too much, only a few seconds longer than before. This makes sense because the decrease in block size will increase the completion time. On the other hand, the change in block size is not big enough to generate a significant effect on the performance and that's why the time difference between question 4 and question 3 is not too big.

Question 5: We can see from the below two graphs that the job still finishes even though 1 worker is killed. The completion time of the normal cluster is around 1 hour and the completion time of the other is around 2 hour 2 min. This makes sense because the normal cluster has 2 workers. The other has only 1 worker. When we kill 1 worker, the efficiency will decrease and the time it takes to finish the job should be 2 times longer than the normal one, which matches the result.

1. Normal cluster with 2 workers

2. Cluster with 1 worker killed

Question 6: The completion time is 1 hour 3 min and it's slower than question 5 but the difference in completion time is not too big. Recall from the lecture that by increasing the replication factor, the resource manager is more likely to find a node that stores the data. So if we decrease the replication factor from 2 to 1, then the resource manager is less likely to find a

node that stores the data and that's why the performance gets slower in terms of completion time.

Question 7: The performance is getting a little bit worse after changing the block size 64MB since the completion time is longer than question 5. As mentioned in the previous problem, decreasing the block size will decrease the efficiency and slow the performance and that's why the time to complete question 7 is a little bit longer than question 5. Since the change in block size is not too big, the time difference between question 7 and 5 is not too big as well.

Question 8: The completion time is 34 min 17 sec.

Question 9:

There are 1167144 articles in the database that have a rank greater than 0.5 and we got the same result in both cases.

Question 10: It is feasible and efficient. Using TCP sockets, we can guarantee that the streaming data we received is complete. And we can also identify different data in the same directory (folder). For file streams, the official document says "Once moved, the files must not be changed. So if the files are being continuously appended, the new data will not be read". However, the TCP socket can be always listened to receive continuous appended data. Sadly, we have to bear flow control of TCP which means that the throughput might be not very large.

Question 11: For this extra credit assignment, we spent one night on it. We spent three days on this whole assignment,