LAB2 report

Hu Zhengdong, Sun Yan, Wu Jiayao, Yang Ziqi

Notes

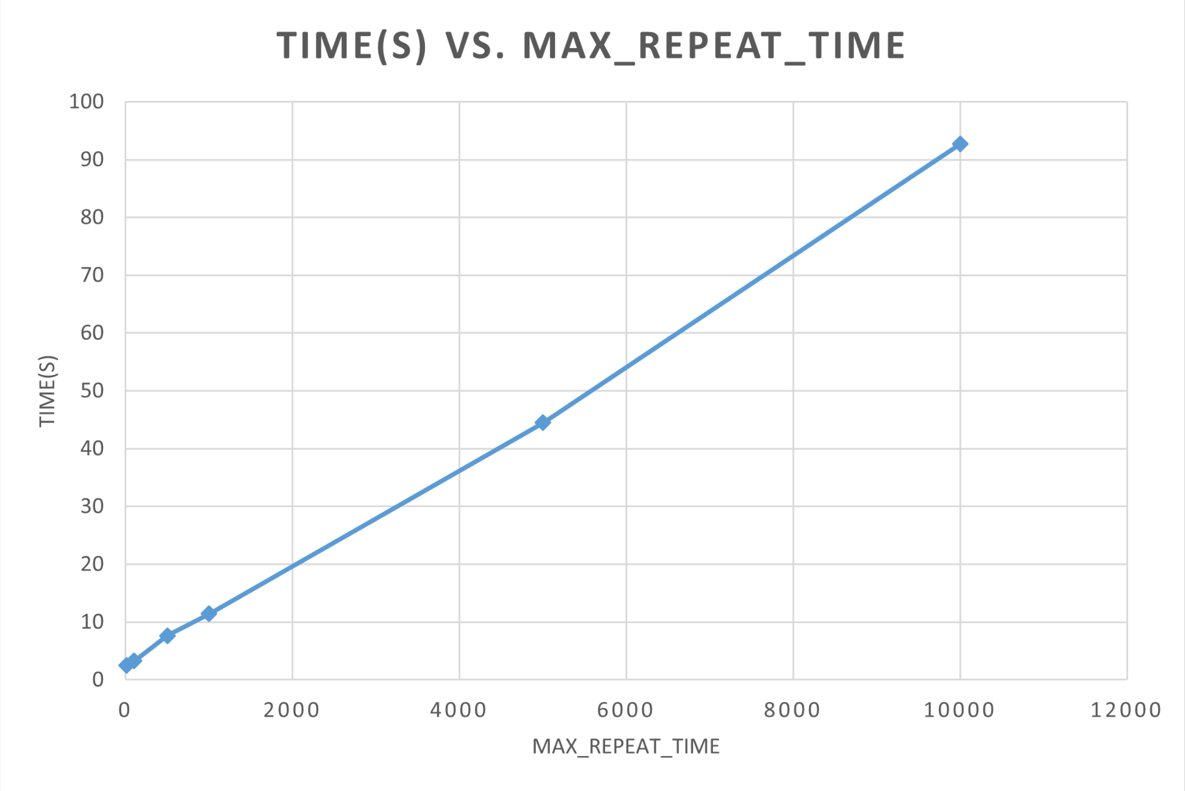
Using the provided files, we generate 5494 students.

When generating grades, we repeat each student for `[0, MAX_REPEAT_INTERVAL]` times, which means in the grade file, each student will appear for a random frequency between 0 times or `MAX_REPEAT_INTERVAL` times.

When on a single computer

Virtual Machine: Ubuntu 20.04 LTS, 2 Core 4 Threads, 4G

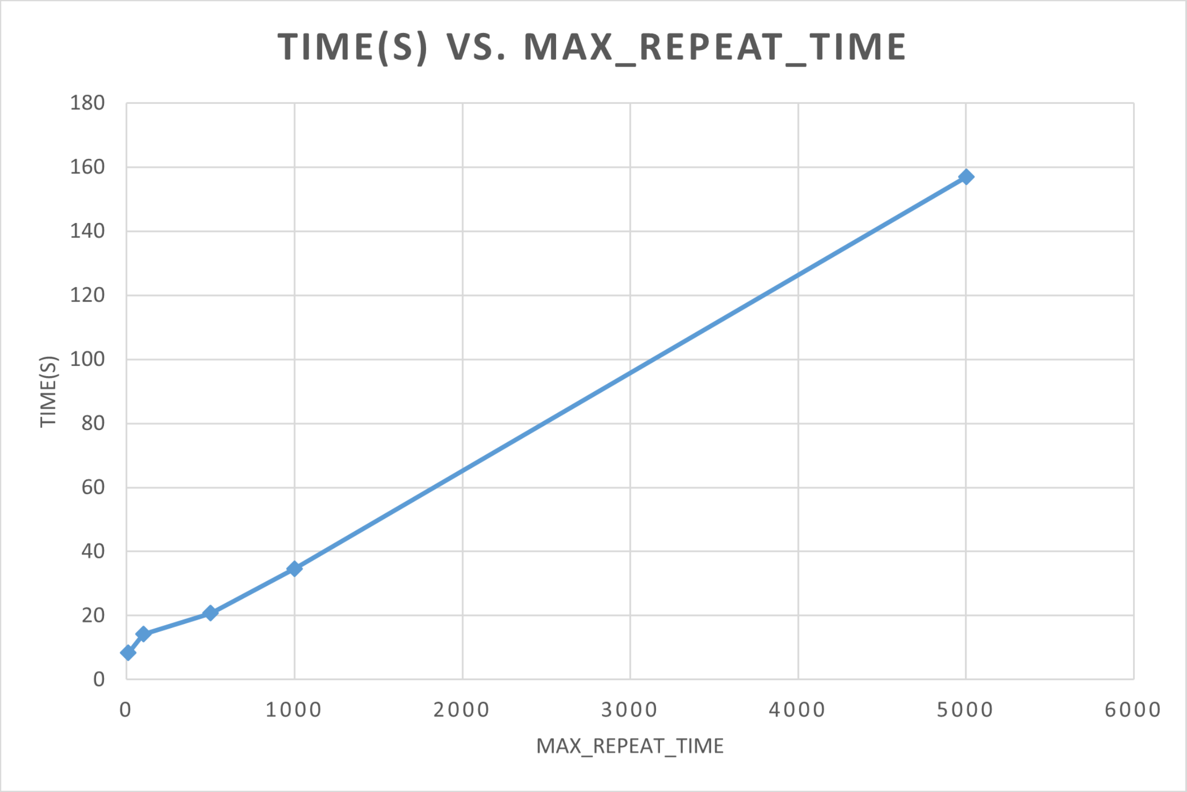
We calculate with the real time, rather than with user time or sys time.



When on a cluster

Virtual Machine: Ubuntu 20.04 LTS, 2 Core 4 Threads, 4G * 3

We calculate with the real time, rather than with user time or sys time.



Comparison

(When doing `MAX_REPEAT_TIME=10000`, cluster fails because that hadoop itself fails. The reason, indicated from logs and error info, may be that our VMs are not powerful enough to handle such a large file.)

The result is kind of out of expectation. We suppose that cluster will be faster than single.The fact is that calculation on a cluster is slower than on a single machine.

The time for cluster may be something rather than linear increasing. But, both single and cluster turn out to be linear.

Some factors may contribute to the outcomes:

1. Network speed limits. We are using local network with WiFi rather than LAN. The slow transmission speed leads to delay in receiving files, which leads to waste of time during calculations.
2. Data size limits. Since hadoop itself has to do extra works to handle things that only happens during parallel work, if the data size is rather small, handling such things (race condition, network, namenode/datanode failure, etc.) may actually take more time than calculations itself. But on the other hand, the bigger the data set, the harder our computer can deal with it without problems...