CS5630 Programming Project-1

Due: Sep 30, 2022 midnight
Upload to ICON a single zip file containing all materials

Part-A. Understanding your workload **40 points**

The goal of the this part is to familiarize yourself with the workload that you would be eventually running in the cloud. In this task, you will learn to install a database system and then run YCSB to benchmark its performance. The choice of the database system is up to you but please make sure that YCSB supports your system. For more details on YCSB, please visit: https://ycsb.site/

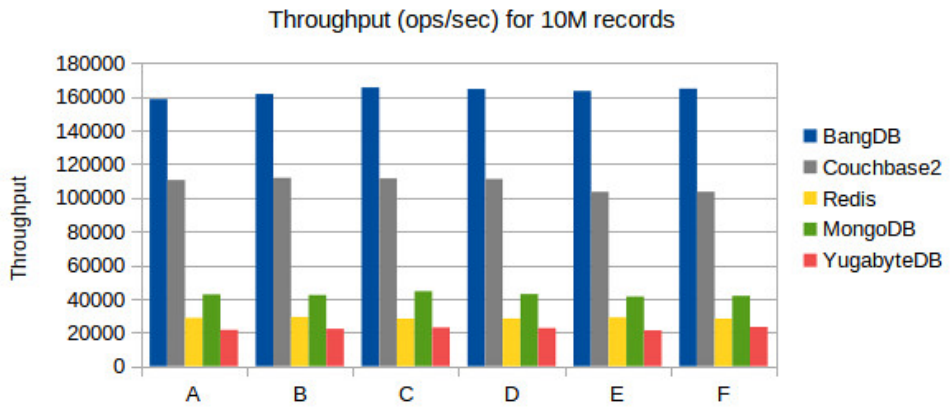
This part is to be completed on your own machine (e.g., laptop). Note that benchmark results will depend on the choice of your database as well as the specifications of your machine. If you do not have access to computing resources, please contact your teaching staff.

YCSB Configuration: YCSB has six core workloads (https://github.com/brianfrankcooper/YCSB/wiki/Core-Workloads). Please understand these workloads, and learn how to run them (https://github.com/brianfrankcooper/YCSB/wiki/Running-a-Workload). Configure YCSB to load 1M records and perform 1M operations under each workload. Make sure to set YCSB thread count to one.

Benchmarking: Benchmark your database system by running all six of the YCSB workloads using the configuration above. Next, repeat the whole process in two additional configurations of YCSB (using thread counts of two and four).

Submission: You should submit a report that contains the following three items:

1. A brief description of your database system, its configuration as well as your machine specification including its CPU, RAM, and disk.
2. Final summary from each of the YCSB runs. You should have 6 workloads * 3 thread configurations = 18 outputs.
3. A single graph that helps visualize the complete benchmark performance. Below is an example graph† that plots the YCSB workloads on X-axis and throughputs on Y-axis. It is comparing the YCSB performance of several database systems against the core workloads. Identify and explain any two salient observations/trends from your graph.



† ref: http://highscalability.com/blog/2021/2/17/benchmark-ycsb-numbers-for-redis-mongodb-couchbase2-yugabyte.html

Part-B. Setting up the AWS environment **20 points**

The second part is to familiarize yourself with the cloud environment. You will get a notification from AWS Academy to register for a lab course, which in turn will give you access to AWS resources. Resources provided to each student is capped at $100, so you should closely monitor your usage and aim to remain within that budget.

For this project, you will need to set up two virtual machines: one for running the database system and another for YCSB. Create two identical VMs of medium type (e.g., A1.medium, or c6g.medium). Please ensure that your selected VM has 1 vCPU and 2 GiB of memory (https://aws.amazon.com/ec2/pricing/on-demand/). Boot up the VMs using EBS as the root drive, and configure your AWS network such that they can communicate with each other. Please refer to AWS guides on how to perform these steps.

Characterizing the Inter-VM Network: Since we will be running the database and YCSB on two different VMs, it is important to know about the network connecting these VMs. While ping is useful for checking connectivity, it cannot holistically characterize the network performance. So, use a tool such as iperf (https://iperf.fr/) on your VMs to characterize your network. Here is an example guide: https://aws.amazon.com/premiumsupport/knowledge-center/network-throughput-benchmark-linux-ec2/

Submission: You should submit a report that contains the following two items:

1. A brief description of the process of creating and configuring VMs. Include a screenshot of your AWS academy dashboard showing two VMs.
2. Output summary of iperf from both VMs.

Part-C. Benchmarking in the cloud **40 points**

The final part requires you to run the benchmark in the cloud and compare its performance across two sets VMs (i.e., two medium VMs, then two large VMs). Using your experiences from Part-A, install and configure your database system on one cloud server and YCSB on the second cloud server.

Benchmarking: Once the VMs and software are set up, benchmark your database by running the six core workloads of YCSB. Here, you are only required to run YCSB in one configuration (1M records, 1M operations, 4 threads). Next, procure two VMs of large type (e.g., A1.large, or c6g.large), and repeat the benchmarking on them. Please ensure that these VMs have 2 vCPU and 4 GiB of memory.

Submission: You should submit a report that contains the following three items:

1. A table comparing your laptop and VM specifications including CPU, RAM, and disk. Also, mention your total cloud cost (i.e., how much of your allocated money was spent).
2. Final summary from each of the YCSB runs. You should have 6 workloads * 2 VM configurations = 12 outputs.
3. A single graph that helps visualize the two benchmark runs (following the same graph format that you used in Part-A). Please explain any two salient observations from this graph. *For example, (i) how does the performance change from <medium> to <large> VM, (ii) how does the 2-node cloud setup perform vis-a-vis the single node laptop setup, (iii) which cloud resources (cpu, memory, disk, network) was the bottleneck, and so on.*

Note: (1) we will not accept late submissions. However, you are allowed to submit partial solutions and will receive a grade proportional to your work. (2) If you took any help from external sources such as StackOverflow, blogs, videos, etc, make sure to cite them in your report.