Advances in Social Science, Education and Humanities Research, volume 514
Proceedings of the First International Conference on Social Science, Humanity, and Public Health (ICOSHIP 2020)

The Implementation of Big Data Technology in Virtual Machines for Mapping 2019-nCoV Pandemic on the Students of Information Technology

Surateno*
Politeknik Negeri Jember
Jember, Indonesia
ratno@polije.ac.id

Ery Setiyawan Jullev Atmadji
Politeknik Negeri Jember
Jember, Indonesia
ery@polije.ac.id

***Abstract*—The spread of Covid-19 has influenced changes in human behavior and paradigm including in teaching and learning process. The teaching and learning process has now been replaced by E-Learning models which offer flexibility for both teachers and students. Current learning process can be accessed by students from different places and not necessarily gather in one room. However, this raises a new problem, namely the difficulty of tracking the scattered students when they return to campus. One way to overcome this problem is to use a big data approach combined with a virtual machine to recognize and detect these students. In this approach, the results of tracking students with server computer resources are up to 15% smaller than using non virtual machines. This is a new approach to data processing.**

Keywords— covid19, virtual machine, e-learning, machine learning

I. INTRODUCTION

The spread of the new generation of Corona Virus or commonly known as Covid-19 is increasingly widespread it has been identified as a pandemic by WHO. This virus has also entered Indonesia and implies the Teaching and Learning Activities process in several educational institutions. Some of them has been closed and the teaching and learning process is replaced by an E-Learning model. However, this turned out to create a new problem, namely the massive spread of Covid-19 that had an impact in the form of exposure to students who were outside their homes or boarding houses.

One way to be able to detect students who are likely to be exposed to the Covid-19 virus, an approach is needed to be able to detect this possibility, one of the way is to use and implement big data, but the current use of big data is still not fully implemented as long as the infrastructure cost still expensive to build one of them.

The new resource in the realm of Big Data is the use of Cloud Computing, Cloud Computing as a computing model which embodies the concept of providing IT resources (CPU, RAM, storage media) by utilizing and making other computers to provide services. Cloud Computing is influential in server virtualization technology, which is one and a combination of computing resources originating from the physical infrastructure [1].

To achieve of cloud technology itself, an infrastructure that cannot be said to be small is needed, several physical computers will be made into one unit by utilizing virtualization. Virtualization itself in the computer world is a process of implementing software in an intangible computer hardware function where its performance resembles actual hardware or even more.

Virtualization technology has a disadvantage, that's the large overhead of virtualization technology, because in this technology the use of hardware layer and also using parent kernel so that virtual systems that are in it will more or less experience a decrease in performance because automatically the virtual system must provide its own "embedded" kernel in the abstraction from the results of virtualization, it is not natively "embedded" in the parent's hardware and kernel.

Meanwhile, the technology that is currently developing rapidly is containerization. Container based operating system is a container technology that treats containers inside as a whole system as if it were an isolated operating system.

One of the advantages of this technology is that it offers features that are similar to virtualization but with a significant increase in performance because there is no overhead from the virtualization system and natively utilizing the kernel and hardware from the parent [2]. While the weaknesses of this technology are also the same as the weaknesses of Linux container technology in general, namely that it cannot be run on operating systems other than Linux because of differences in kernels, so virtualization is still needed when running containers on operating systems other than Linux [3].

II. LITERATURE REVIEW

A. Openstack and Docker

Because Docker is opensource developers and system administrators usually using it as a deployment test which allows them to create, deploy, and run micro-service based applications. It has a lightweight and portable packaging tool, and also powerful Docker Hub, as service that enables easily distribution of the Docker containers. Docker allows applications to be rapidly packed from different components and easily deployed. This method

will eliminate the back and forwards between development, QA, and production environments.

What distinguishes between VM and Docker is the deployment approach. Where unlike a VM that requires a full operating system when it will download an application, Docker only requires a few resources while the operating system still uses the main operating system where the docker is installer. As a result, software developer can deliver application faster and also run the same application on single machine, without changing it, on notebooks, baremetal virtual machines (VM), or and any cloud platform [4].

III. METHODS

The data used were data from the tracking results of the academic community in the Politeknik Negeri Jember environment. The respondents were 350 users with amount of data is 10800 data containing of the location, health status results from the tracking system and activity logs from the users.

The database engine used was a machine based on a non-relational database, in this case, mongoDB that runs on a docker machine that works on the main operating system compared to machines using VMs. Here is the proposed model.

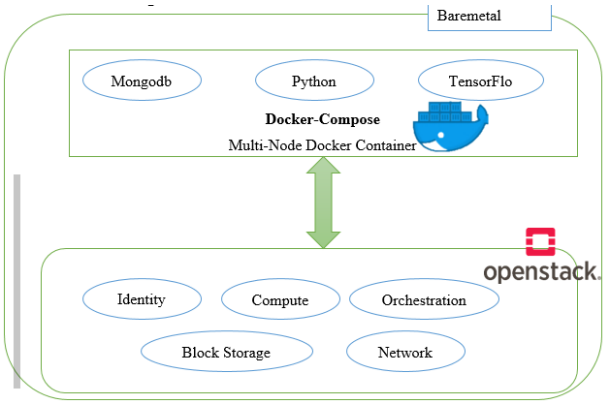

Fig. 1. System Model

A. Data preprocessing

The data were taken at the beginning of the pandemic using the help of Google Form. Then the respondent's address data was carried out by a forward geocoding process to produce location data that was closer to the actual location. Respondents came from students as well as lecturers and administrative staff who are carrying out work from home.

The data were then processed into csv form so it will be easier to enter into the database engine that has been previously provided.

B. Data Store

After the data is prepared in such a way, the next is the process of entering data into the database, in this case using the help of the Python programming language.

The data that will be entered is processed first using a preprocessing data approach, which is to remove noise from the data itself so that the data entered in the database is clean data.

C. Docker Environment

Docker is a virtualization that have light-weighted solution that has been attract attentions from the research community to explore its potential in different applications.

Several cases have made performance comparisons between dockers and virtual machines [5] [6], the result of this comparison is that Docker is able to provide more efficient output on the performance of CPU performance.

Docker also has better scalability than VMs. As the number of VM instances and docker containers increases, docker can keep its performance without dramatically losing computing ability due to large overhead like VMs do [2].

Boettiger complements it by stating that Docker is capable of virtualization at the system level operation, deploying containers in a portable manner though cross platform, provides reuse features components, sharing, archiving, and versioning of the image container. In the industry, container technology has been implemented as one of the trending cloud computing models. Some of the advantages of containerization with Docker are.

- Docker images are read-only templates for running containers. An Image can consist of an operating system and several applications that have been installed. The Web application that has installed the Image is used to run the container.
- Docker Container is a directory that stores everything needed so that the application can run smoothly. Each container is run from a docker image that has been determined that the container can be run, stopped and removed. Docker uses the union-file system as the back-end file system for its container, so any changes saved to the container will cause a new layer to be created on top of the base image, which makes the container a layer where applications can run normally. Each container that runs isolated in one environment and a secure application platform, does not conflict with other applications on the same host.
- Docker registry is a repository (public or private) that provides thousands of docker images. Public docker registries are called Docker hubs. Users can push commands via docker client to the docker registry for storage and sharing. And other users can do a pull command to download and run it directly
- Docker file is an automation script (builder) that builds an image, a Docker file is a text document or script that contains all the commands that are usually performed manually to build an image, using the docker build command from the terminal.
- Docker uses a model similar to that used on Github and other source control systems, but with a different type, the Repository is an ID for each image stored in the registry. When running the docker commit command, the image will be named in the format username / name image. Giving this name implies when doing the docker push command, the index will see the image name and make sure there is no same repository name, otherwise index will check whether it has access to the repository, then it is allowed to upload a new version of the image to the repository.

- The Docker index is linked to the Hub Registry, although they have different functions, the Index manages user accounts, permissions, search, tagging and other things stored on the public web interface. When executing the docker run command to run a docker image, it is used to find data in the index instead of the registry. When running docker pull or docker push commands, index will determine whether it is allowed to access or modify the image, and then the registry is the part that will store the image after getting access rights from the index.

IV. Results

In this study, the testing environment was performed on two baremetal servers focus on the CPU workload, GPU workload, disk I / O workload and deep learning tools side, the test model was carried out using performance evaluation of some work on the docker container side compared to vm. In order to perform the same test, several test variables will be equalized, such as the compiler version and several libraries that will be used as testing models on the container and host system in the VM Environment.

A. Load Data Testing on MongoDB

Testing image data inserted on Hbase an MongoDB is presented on the chart below with each parameter:

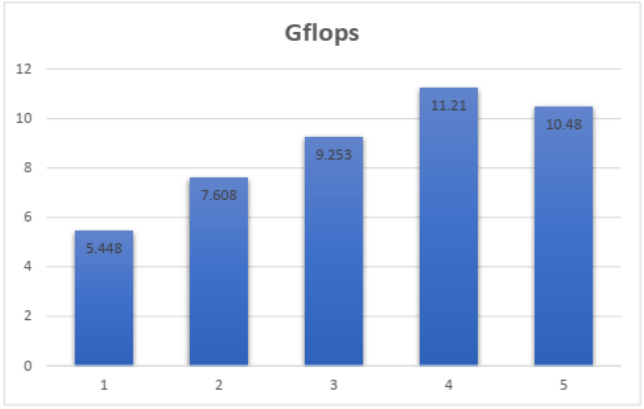

Fig. 2. Insert data (in thousand)

Figure 2 shows testing image data by using the Runtime parameter with 1,000 data results in 5.448s Gflops. Location data with Runtime parameters with 3,000 data yields 9.253s GFlops. Data with a Runtime parameter of 5,000 data generates 10.48s GFlops.

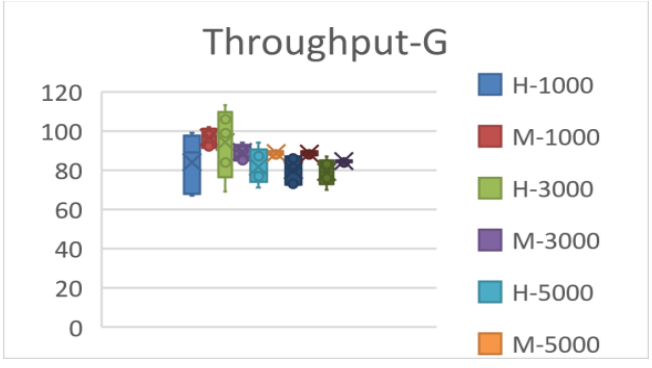

Fig. 3. Throughput-G

Figure 3 demonstrates testing the image data by using the throughput parameter with 1,000 data producing 84 ops/ s Hbase, 97 ops / s MongoDB. An image with throughput parameters with lots of data 1.500yields 94 Hbase ops / s, 89 ops / s MongoDB. An image with a multiple data throughput parameter of 1,800 yields 82 Hbase ops / s, 89 ops / s MongoDB.

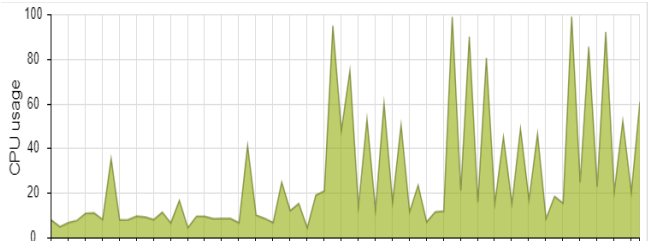

Fig. 4. CPU Usage Test

Figure 4 indicates testing image data using the CPU Usage parameter with a lot of 1,000 data results in 83% Hbase, 27% MongoDB. An image with CPU Usage parameter with lots of 1.500 data yields 93% Hbase, 52% MongoDB. An image with a CPU Usage parameter of 1,700 data returns 96%, 50% MongoDB. An image with CPU Usage parameter with a lot of 1,800 data yields 98% Hbase, 53% MongoDB.

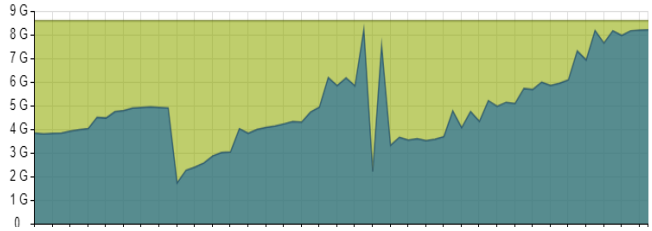

Fig. 5. Memory Mapping

Figure 5 shows testing image data using the Memory Usage parameter with a lot of 1,000 data results in 86% Hbase, 47% MongoDB. An image with a Memory Usage parameter with a lot of 3,000 data yields 93% Hbase, 87% MongoDB. Images with the Memory Usage 5,000 data parameter yield 92%, 93% MongoDB. An image with a Memory Usage parameter with a lot of data of 7,000 yields 94% Hbase, 95% MongoDB. Image with Memory Usage parameter with lots of data 10,000 yields 94% Hbase, 95% MongoDB.

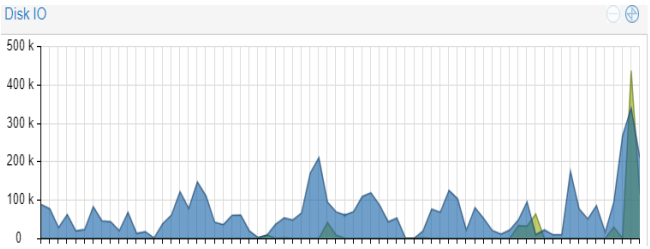

Fig. 6. Disk IO Output

Figure 6 reveals testing image data using the Disk I / O parameter with a lot of data of 1,000 results in 180kb. Hbase, 16,779kb MongoDB. An image with Disk I / O parameters with a lot of 3,000 data yields 310 kb Hbase, 87% MongoDB. An image with 5000 disk I / O data parameters yields 380 kb MongoDB. An image with Disk I / O parameters with 7,000 lots of data yields 410kb Hbase, 410Kb MongoDB. Image

with Disk I / O parameters with 10,000 lots of data yields 420 Kb Hbase, 43,761kb MongoDB

V. CONCLUSION

It can be concluded that Hbase was built using VPS and VPS has the same specifications as previous research. Hbase can store and display heterogeneous data from sensor nodes in the previous environment using the MongoDB environment.

From the results of testing Hbase and MongoDB uses two text and image tests with parameters, having runtime, throughput, cpu usage, memory usage, and disk i / o and five test data results in data comparisons, average Text data with runtime parameters of 54s for Hbase, 33s for MongoDB. The throughput parameter is 938 ops / s for Hbase, 1599 ops / s for MongoDB. CPU usage parameter is 54% for Hbase, 48% for MongoDB. The Memory Usage parameter is 81% for Hbase, 67% for MongoDB. And the Disk i / o parameter is 482kb for Hbase, 4354kb for MongoDB.

While the average runtime parameter image data is 64 Hbase, 60s MongoDB, 84 ops / s Hbase Throughput parameter, 90 ops / s MongoDB. CPU usage parameter 83% Hbase, 46% mongo. Parameter Memory usage 93% Hbase, 84% MongoDB. And Disk I / O parameters 99.003kb Hbase, 37.885kb MongoDB.

From the comparison of the MongoDB test, it is superior to Hbase in storing text and image data. However, if it is seen from storing image data with Disk I / O parameters Hbase has a good performance compared to MongoDB.

REFERENCES

[1] C. Aditama and A. Priadana, "IMPLEMENTATION AND PERFORMANCE ANALYSIS OF PRIVATE CLOUD USING OPENSTACK SWIFT DAN RCLONE," vol. III, no. Ix, pp. 317–322, 2018, doi: 10.9790/0661-1805064858.

[2] P. Xu, "Performance Evaluation of Deep Learning Tools in Docker Containers," 2017.

[3] N. Nguyen, "Distributed MPI Cluster with Docker Swarm Mode," 2017.

[4] B. Rochwerger *et al.*, *Cloud Computing: Principles and Paradigms*. 2011.

[5] Q. Zhang, L. Liu, C. Pu, Q. Dou, L. Wu, and W. Zhou, "A Comparative Study of Containers and Virtual Machines in Big Data Environment," *IEEE Int. Conf. Cloud Comput. CLOUD*, vol. 2018-July, no. December, pp. 178–185, 2018, doi: 10.1109/CLOUD.2018.00030.

[6] A. M. Potdar, D. G. Narayan, S. Kengond, and M. M. Mulla, "Performance Evaluation of Docker Container and Virtual Machine," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1419–1428, 2020, doi: 10.1016/j.procs.2020.04.152.