Wei Ye

Dr. Fan Zhang

Data Science Research Program

Sep, 15th 2019

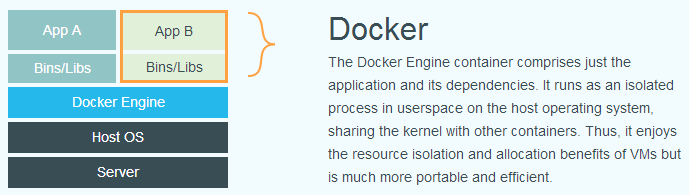
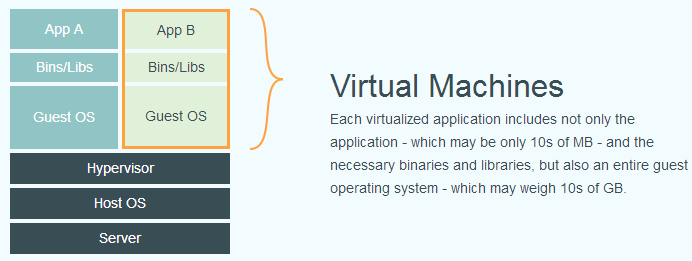
Project Report

First of all, I am honored and happy to participate in this research program. I came in with a mindset of learning and understanding. I also gained a lot from this study. As a mathematics major, I have hardly ever dabbled in such fields as container technology before. I learned a lot in the process.

Through this study, I have a deep understanding of big data. The meaning of big data is accompanied by the increasingly popular network behavior of human beings, which is collected by relevant departments and enterprises. And it contains the right intention and preference of data collectors, not traditional structure and meaning data. Extracting useful information from vast amounts of data is also a massive challenge for network architecture and data processing capacity. Big data is a kind of Internet development to the present stage appearance or characteristics. It is not necessary to myth or keeps it the heart of fear, in innovation by cloud computing technology. These were difficult to collect and use data to easy to use, through the eyes of continuous change, big data will gradually create more value for human beings. Computing and storage hardware is now very cheap, and with a lot of open sources of big data tools, people have the luxury of grabbing large amounts of data before thinking about analytic propositions. It's fair to say that cheap computing is changing the way we use data. Also, significant improvements in processing performance (such as in-memory computing) make the real-time interactive analysis more comfortable to implement, and "real-time" and "predictive" take BI into a new world -- the unknown. This is also the most significant difference between big data analysis and traditional BI. Computing and storage hardware is now very cheap, and with a lot of open sources of big data tools, people have the luxury of grabbing large amounts of data before thinking about analytic propositions. It's fair to say that cheap computing is changing the way we use data. Also, significant improvements in processing performance (such as in-memory computing) make the real-time interactive analysis more comfortable to implement, and "real-time" and "predictive" take BI into a new world -- the unknown. This is also the most significant difference between big data analysis and traditional BI. In today's society, cities are facing multiple problems, such as budget overruns and infrastructure problems. Cities are the perfect testing ground for big data.

I majored in pure mathematics as an undergraduate. I have many choices when choosing my Ph.D. major. I can continue to study mathematics, I can analyze statistics, and I can pursue my career in computer science. Under such circumstances, it is difficult for me to make a decision. Data science is one of my choices, which is also my preference. The reason why I apply for this project is to learn about this field and see if it can be used as my Ph.D. research direction. Although big data is now a relatively hot field, it is also excellent employment. However, I still need to practice to see if it is suitable for me, and cannot blindly follow the trend. After studying this project, I am always interested in this field, and I also realize that I still need to learn a lot of basic knowledge to transfer from mathematics to computer field.

In this study and practice, I mainly recognized Docker and container technology. Based on the Container, Docker further encapsulates, from the file system, network interconnection to process isolation, etc., greatly simplifying the creation and maintenance of containers. This makes Docker technology lighter and faster than virtual machine technology. Traditional virtual machine technology is to run a complete operating system on a set of virtual hardware, and then run the required application process on the network. The application process in the Container runs directly on the host kernel, which does not have its core and does not do hardware virtualization. So containers are much lighter than traditional virtual machines. The following images compare the differences between Docker and traditional virtualization.



The relationship between an Image and a Container, like classes and instances in object-oriented programming, is static, and the Container is the entity that mirrors the runtime. Containers can be created, started, stopped, deleted, paused, and so on. Containers are essential processes, but unlike processes that execute directly on the host, container processes run in their separate namespace. So the Container can have its root file system, its network configuration, its own process space, and even its own user ID space. Processes in a container run in an isolated environment and operate as if they were working on a host independent system. This feature makes container-encapsulated applications safer than running directly on the host.

Cassandra is a highly scalable, high-performance distributed database for handling large amounts of data on large commercial servers, providing high availability without a single point of failure. Cassandra is designed to handle big data workloads across multiple nodes without any single point of failure. Cassandra has a peer distributed system between nodes, and the data is distributed among all nodes in the cluster. The following figure shows how Cassandra USES data replication between nodes in the group to ensure that there is no single point of failure diagram.

In conclusion, I have learned a lot from this research, and I have also obtained the information I need to make decisions on my future research direction. This is meaningful scientific research, and I am honored to participate in it.