Big Data Project Report

Chendu Li

2019.08.18

Prof. Fan Zhang

Description:

This project can help users to recognize the numbers in their uploaded pictures through virtualization containers. The project includes the use of Docker and MNIST training model.

Content:

1. Lecture Content:

In this course, I learned to use some basic application of system and tools from the lecture. Some of them are relating and necessary to the project, and others are widely used in the area of Big Data. Applications and tools include the followings:

- **RESTful API:** RESTful Web Service uses HTTP method to execute operations and URL to access various resources including XML and JSON. An API is called application program interface, which was designed to utilize the HTTP protocol better.
- **NoSQL**: NoSQL is designed for a non-relational database, such as a graph, key-value, JSON, and Column Family. Cassandra is one of the most useful NoSQL language. In the CAP Theorem, Cassandra is mainly focused on availability and partition tolerance. Furthermore, Cassandra can boost performance in proportion with the increase in the numbers of hardware to overcome the diminishing utility problem.
- **Container Technology**: Docker is the container technology I learned and used in this project. It is an open platform to build, ship, and run distributed applications. Docker was constructed with a layered filesystem, which makes the user capable of transmitting information by pushing and pulling layers from Docker Registry efficiently. During the lecture, I learned the basic examples of the Docker CLI, and

Docker Daemon, which is one of the components of Docker. Docker can create an environment that helps users to transmit Docker Images and download other images to users' Docker Container without install of numbers of applications.

- **Flask:** Flask is a microframework written in Python. While using Docker, I also applied Flask to the project coding to help me access its extensive functions, including routing and HTTP Method.
- **Spark:** Spark improved the shuffle stage in Hadoop that increased the processing speed tremendously while supporting a wide range of languages and software. Spark can be recognized as a resilient distributed dataset. It can deal with different parallel operations in one cluster with high speed.
- **GitHub:** For this project, I created my GitHub Account to update my process and upload my works. The following is the project GitHub Link: https://github.com/lichendu/BigData-Project

2. Project:

Step 1: Save and use TensorFlow MNIST Model:

The first part of the project is to save MNIST training model to local and name it as model.ckpt.

(抱歉老师，这部分还不太会写)

Conclusion:

From this project, I obtained a brand-new understanding of Big data. This project taught me to use Linux system and Mac terminal. Throughout the process, it was the first time that I learned to use different kinds of tools, like Anaconda, GitHub, Docker, Flask, etc.

It was an unprecedented experience for me to write my code. While I was working on the project, I found the training process of MNIST was creative and interesting. I found the use of Docker makes the coding process easier and more efficiently. I also found there are more interesting tools and systems are waiting for me to explore. This project not only taught me how to use computing and coding applications but also evoked my curious in this Big data trend.

In this project, I got to know the current trends of Big Data and the relating industries. From Prof. Fan Zhang's lecture, I was introduced to the Hadoop MapReduce and its application in Google. I found that the way of MapReduce helped people to process data efficiently was interesting. The lectures gave me a new perspective on how data are utilized, stored and visualized. Applying MapReduce, Google File System, and Bigtable, Google is the most successful company in this Big data world. In other words, Data can help companies and people solve problems, and it is the system and those tools I learned from this project that helped the world move forward. Therefore, I would like to pursue further understanding in the area of Big data.

In conclusion, this project gave me extensive learning experience in all kinds of system, application, and tools. But more importantly, this project confirmed my interest and determination in pursuing this area.