Bigdata Remote Project Report

Name: Hongwei Li

School: University of Pittsburgh

Major: Computer science

1. **Introduction**:

Different from traditional data management and data processing, bigdata is a field that people extract and analyze extremely large data and this process can never be done by one or two programs and tools. Data size is no longer limited to a small data set that can be stored locally in our PC or laptop, instead the data we are going to use or analyze could be the data of the whole population and most of the data are more likely to be accessed remotely. Even though today we have extremely large hard drive disk and fast GPUs that can compute billions of multiplications of matrix in a second, more importantly, how to manage and apply them to the bigdata is what we need to concern at first. Keep in mind with three key concepts of bigdata: volume, variety and velocity, to store and extract data efficiently, we have to use systematic ways to accomplish it. This is also the main goal of this project: we are going to learn and write programs based on different tools and finally build a system running programs all together that can efficiently store, analyze and post the data we are going to use.

1. **Course content**:

**Class 1**:

-Basics of Python and practical usage of Linux.

-Why Linux:

First of all, Linux is an open source platform. Due to its flexibility, scalability and reliability, Linux is a powerful environment for data scientist to fully use the resource of computers.

-Why python:

Python has plenty of packages that are friendly to data analysis. Its support of unconventional data like image and voice makes us access data easier. There are also many mature and advanced libraries built in python which makes the process of data learning much easier. Of course, easy to learn and use is another reason that it is more friendly to users.

**Class 2**:

-Introduction of Docker and the advantages of it:

Docker containers are like virtual machines, but it is much lighter. Packaging all applications with their requirements into the images, docker is like an engine that loads images and is able to run them on all OS. Different from traditional VM, it doesn’t need a guest OS. Docker gives data scientists a Big data environment on their local computer which is a convenient way to share and develop programs.

Tasks:

-How to build and run docker container, images

-How to make sure the program installed python and required modules

Usage of Docker file, Requirements, Application

-Introduction to Restful API

Rest (representational state transfer) an architectural design for web services and web API based on HTTP protocol. (Miguel Grinberg, *Designing a RESTFUL API with python and flask* ,2013)

We learned how to build restful web service using python and the flask microframework. From clients’ console, we can send requests to URIs using methods defined by the HTTP protocol and then get the result from the Restful web server. After this, we are able to write our own HTTP method (get,post,put,delete) in python.

**Course 3**:

-Introduction of database management

We learned about features of different databases. If we want to build a database system, there are there characteristics we need to concern: volume, velocity, variety. There are also three key concepts of database: Consistency, Partition tolerance, Availability. The database has to be available to multiple regions. Therefore, partition tolerance should always be satisfied by database system. However, if we want consistency that makes sure that every database returns the same, we do need to spend a lot more computing time to check. As a result, we would lose speed of data transportation. Theoretically speaking, no database can satisfy all three conditions. Therefore, when we start to build our own database, we need to judge and weigh which feature does the data need and choose the best one for our bigdata. One of the most popular databases in bigdata, Cassandra database, excels with flexible, real-time data ingestion and analysis with no single point of failure across commodity hardware clusters. (Fan Zhang, *Course PPT*,2019) We then focus on Cassandra database and learn about how to build Cassandra database in docker container.

**Course 4**:

Introduction of data stream, spark and the native landscape of cloud. The whole cloud system is like an ecosystem that consists of infrastructure, provisioning, runtime, Orchestration & management, Application, Platform, and Observability & analysis. Every part listed there is indispensable to the big ecosystem of cloud. We generally talked about different tools from different sections and the usage of them.

1. **Final project:**

Applying most of the points from courses we learned, I build the Mnist hand-written recognition program into a docker container and by applying restful web service API, the program is able to handle the request from client console and print out the result of prediction of hand-written picture, and finally post all of the information into Cassandra database which is also built in another container in docker. The most challenging part of this project is how to connect two containers and let them communicate with each other. Redis, Flask, RESTful API, Docker, Cassandra, Mnist, all of these tools are new terms for me, so making them work together is also a challenge for me. This is also why the project takes me lots of time on researching.

Finally, thanks to Dr. Fan Zhang’s guidance and direction throughout the whole project. He does contribute a lot to our projects.

Reference

1. Grinberg, M. (May 20 2013 ). *Designing a RESTful API with Python and Flask.* miguelgrinberg.com.

2. Zhang Fan (2019) Course PPT from video 3.