Name: Qiushi Teng
CS510, Data Science, Reflection Paper

Thoughts of the Project

This project provides me a great opportunity to learn about the Apache Spark system, including some background knowledge and some hands-on experience of performing queries. I have been learning a lot by taking courses in data science since last fall, and also tried several different data processing systems, including SQL, NoSQL, and some cloud services. I think Spark is a very powerful and efficient system among all the systems that I have learned so far. One reason is that it processes data much faster than some other data processing systems (as tested in some academic publications). Another reason is that it has a set of libraries, which allows it to process both structured and unstructured data. Also, its multiple programming language APIs and the ability to run almost anywhere make it accessible for a lot of use cases. Overall, Spark is a very interesting system and technique to learn.

There are two things that I really like about this project. The first one is that we have great flexibility by the different available choices, doing research papers or implementations. In my mind, this does not only provide people more options, but also leaves people room to think about what they want to do, and how they can achieve the goals. We learn by being told about what to do. Sometimes, we could learn even more by figuring out what we are interested in and what we want to try. This allows us to dig deeper into something we already know or try something that is new. These different options allow more flexibility, and the process of making decisions from these options is also one form of learning to me.

The other thing I like about this project is the overall design along the timeline of this course. To me, different assignments/write ups during this term act like a guidance for what should be learned at different times and how things should go step by step. Take my project as the example, since I had neither theoretical experience nor hands-on experience of Spark , it could be very confusing about where and how to get started. But the assignments really help me to sort things out. The project plan is the first assignment due, which allows me to decide what topic I would like to focus on. Then by getting the research paper done, I read two ebooks and some academic publications of user cases of Spark, which helped me to build up some background knowledge about the system. Then I learned how to set up the system and perform different queries in the implementation part. I would say going along with the timeline makes studying this new system a lot more clear than I thought.

There are also two things that I think I could have done better. One thing is that to compare how traffic changes, I pick the data from the most recent four years, because I think usually people like to see something that is most up to date. However, I forgot this year’s special situation. I actually expected to see some trends in my comparison. But choosing data from 2017 to 2020 actually is not the best choice. If I redo the project, I will probably change the four year range to

2016-2019, or simply adding more years prior to 2017. I think doing so may help to see if there is any trending in the traffic change.

The second thing I would like to try is to include some large datasets. In a lot of resources and references, it is mentioned that Spark runs a lot faster than a lot of other data processing systems. However, since the dataset I am using is not large, I actually did not have the opportunity to see how fast Spark runs a task. I still remember that we had a very large dataset in the CS488/588 project, which allows us to compare how different time it took to perform a query on datasets of different sizes. Therefore, I think it will be interesting to compare performing the same query on a large dataset on different data processing systems.

Thoughts of the Class

Data science is a subject that covers a broad range of aspects in our life. I feel like almost everything can be turned into some formats of data, and also data can be used to manage a lot of things in both a positive way and ways that might not be so good. But overall, I think data science is everywhere, and has various features. Due to this property, I think it is not easy to have a course in several weeks to cover a topic that almost covers everything around us. I would not say this course covers everything, but we have so many different pieces in this course and these different pieces provide me views of different aspects. Also, having these different pieces makes me realize that these different pieces, or even non-related pieces actually are good representations of how many varieties that data science covers.

For example, part of this course is to watch the Calling B.S. videos, which discusses what we should know and how to deal with information at this age when more and more data is present. In addition, we have four guest lecturers. Although they are all about data science and share some similarities, to me they all seem to be different and focus on different aspects. Also, we have a lot of different options for doing research papers and doing projects. All of these let me know that there are tons of things that data science covers. Overall, after learning different pieces in this course, I feel like that data science is not something that is limited to some areas. Even more, anything I can think of can be turned into parts of data science. In addition to learning particular knowledge, I also learned the general concept and a view of how broad data science can be.

There is only one thing that I can think of that could be more helpful to me. In this term, we had two guest lectures in one class. Since all guest lectures cover a lot of information that freshes my mind, I think it would be great if we can split these guest lectures into different class periods so that we can have one guest lecture every class period. At some point, I felt that this lecture was very interesting, but my mind was too full to take in more information. Therefore, I think it would be nice to have only one guest lecture per class period so that I might be better at learning these interesting things. But I know that all guest lecturers we have are very busy, and they have their own jobs. In particular at this special time and situation, I really appreciate that they gave us these wonderful lectures.

Overall, I am a little surprised that I can learn so many things in eight weeks, and I think the overall design of this course helps me to accomplish everything that I have done. Thank you so much for all your help and support along the way.