Introduction

Throughout the deciphering big data module of this degree, students were given the task of documenting assignments and activities assigned to ourselves. These activities came with the purpose of giving students the stepping stones that would allow growth when manipulating large datasets, the end goal was to improve our familiarity with big datasets and databases. Throughout this reflective piece I will critically analyse my growth and development I endured during this module, as well as any challenges and limitations I may have come across. My intention is for this piece of writing to be tethered to the reflective model of Rolfe et al (2001) which is, “what? So what, Now what?”.

Learning about Big Data

Data Science and the concept of big data were relatively novel to me during the beginning of this module. A reason why I chose to study this degree was to learn about the different aspects of big data and how it can be manipulated for the benefit of the user. Due to my current occupation within finance I’m more accustomed to tabulated and integer-based big data, which is why the first three units of this module allowed me to develop my knowledge on non-relational data. As part of these units myself and the rest of the students within this module were encouraged to take part in a collaborative discussion on the internet of things (IOT) and how there may be opportunities, limitations, risks and challenges. As mentioned previously non-relational data was an almost foreign term to me and the IOT is closely associated with non-relational databases. Ceresnak & Kvet (2019) found that “non-relational databases have better performance than relational databases for data sourced from IOT”, thus showing that non-relational data and environments are more likely to be used. My initial post showed that I had conducted the required reading on IOT and the challenges faced when cleaning and wrangling data sourced from IOT, this gained knowledge was also presented in the required peer responses where I highlighted the limitations and benefits to data wrangling despite not receiving any responses to my initial post from students on the course, thus being unable to provide a final summary post for the exercise. Overall, I can confidently state that I have developed a larger understanding in relation to IOT and the challenges organisations and individuals may face when they are tasked with sourcing and tidying data for analytical purposes.

Data Cleaning Tasks

During units 4 & 5 we were assigned data cleaning tasks for the purpose of populating our e-portfolios for one of the assignments. A problem that I faced was that I was still relatively new to the programming language Python, and this task involved using Python at an intermediate level to clean a large source of data although I was a novice. Familiarity with Python had to be gained to prepare for this task and the rest of the deciphering big data module. I tackled this by using my study time to look into the required and suggested reading of Data Wrangling with Python: Tips and Tools to Make Your Life Easier. Where the book states that “data analysis skills can be taken beyond Excel and to the next level through this hands-on guide shows nonprogrammers how to process information that can be too difficult or messy to access, without needing to know a thing about Python programming language to get started” (Kazil & Jarmul, 2016). Along with this reading paired with Python based activities, enough progress was made to which I could work through the UNICEF data cleaning activity which entailed cleaning csv data that was previous incomprehensible into more understandable data. Despite the book being extremely helpful for this task the syntax it was written in was in Python 2 whereas, JupyterLab’s syntax was in Python 3 so there were some differences that had to be tackled. I did this by using my initiative and seeking out problems to errors via Stack overflow, allowing me to overcome these road blocks.

Project Report

The deadlines for unit 6 & 11 related to a project report. We first had to establish a foundation within a team and determine the approach that would be taken for the proposal of a database. Since all of my team members had a background within finance there was a similar perspective with the approach that should be taken. Although this sounds good on paper there needs to be variety within a group to allow the contribution of different ideas. We overcame this when I suggested to the team members that we should assign ourselves particular roles within the group that we would specialise at, members would then have to evaluate the proposals made by each specialist which in turn would generate solutions to problems or limitations that we would encounter with the project report. I suggested that evaluation could be done verbally through our regular scheduled calls, or non-verbally which was could be sub mitted via a cloud environment we all had access to. As a result, this helped our team receive a merit as a result of our submission and we also managed to pick up new skills and pieces of knowledge from one another. We all benefitted from the approach we took as we now had a clear vision for our individual executive summaries that we were to submit 5 weeks after. This knowledge attained from collaborating with my team mates was applied to my executive summary where a relational database was created for American Express.

Conclusion

Upon reflection of this module, I would say that as an individual knowledge and understanding of big data has improved vastly. I came into this module only having general knowledge when it came to big data, and now I will be finishing the module with the attained knowledge of how to clean data that has been sourced in an SPSS or csv format, to make this data more understandable. Furthermore, I have completed the challenge or creating a relational SQL database and populating it with multiple tables, while drawing up an executive summary complete with an entity relation diagram and an application layout to compliment the relational database. One thing I do wish to work more upon is populating a NoSQL, non-relational database which I could then present within my eportfolio. To achieve this goal, I will seek out examples or non-relational databases which I could then apply to my own work.

Reference List:

Ceresnak, R. and Kvet, M. (2019) 'Comparison of distributed data transformation and comparing query performance in relational and non-relational database', *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)* [Preprint]. doi:10.1109/iceta48886.2019.9040085.

Kazil, J. and Jarmul, K. (2016) *Data wrangling with python: Tips and tools to make your life easier*. Sebastopol, CA: O'Reilly.

Rolfe, G., Freshwater, D. and Jasper, M. (2001) *Critical reflection for nursing and the helping professions: A user's guide*. Basingstoke: Palgrave.