# Reflective Piece

The module Deciphering Big Data has expanded my knowledge both professionally and personally. From the start of the module, the data wrangling exercises have helped me understand how to efficiently prepare, clean, and explore data to solve problems. This has further expanded my understanding that data in recent times does not only come in a structured format, like from relational databases but also from unstructured and semi-structured formats (Koehler et al., 2021). In the first three units, I researched the benefits of the recent proliferation of Internet of Things (IoT) devices in the collaborative discussion. IoT devices send data that can bring several benefits to an organisation. For example, energy management systems can schedule heating and air-conditioning effectively to reduce costs and carbon emissions (Liang, et al., 2023). Furthermore, for IoT's benefits, it is vital to have a secure data transfer from IoT devices. For this, examples of blockchain-encrypted data transmission can be used. This has given me a wider knowledge of how important data transfers are protected over networks from cyberattacks or data breaches, which might be very harmful to organisations, both in a legal and a reputational manner (Janani & Ramamoorthy, 2024).

Further, during the module, web scraping was another interesting topic that broadened my knowledge of how data can be accessed from an online environment. Scraping and data cleaning methods can lead to a dataset from which insights can be gleaned for better business decisions. Boegershausen, et al., (2022) state that a methodological framework for collecting web data is an ideal thinking process that identifies the limitations of collecting an ideal dataset online. From source selection to the design of the collected data that tackles which information you need to extract to answer your business questions or needs. At the end of the thinking process framework is extracting the data efficiently. The entire framework is encompassed by the legal and ethical risks of extracting data from resources on the web. The latter is an important consideration in my professional career as, despite the availability of information on the web, it is essential to consider the ethical and legal implications to avoid potential legal consequences.

The next task, before the mid-module team project, was data cleaning. This was an interesting task that enhanced my technical knowledge as I tackled the task with another library rather than the one that was asked of us. I used Pandas as it is a library that is more commonly used and this helped me to clean data and finally visualise missing data from a survey. Even though I cleaned data manually, it was interesting to delve deeper into the topic of automating data cleaning as in an industry process, data cleaning’s efficiency comes from automation by data cleaning algorithms, from which insights will improve the organisation’s processes (Ding, et al., 2019).

In the module's second half, we explored database design and creation. These topics are the largest gap in my knowledge. In these units, we started with database design and data normalisation, which are needed to organise a database efficiently (Wong Ting Yan & Fong, 2021). In our normalisation task, we normalised the database till the third-normal form. There are other forms, but most databases are normalised till this form as it would be enough to have a functional database. In addition to normalising the given table, I also used SQL code to build a database. This task again, helped me in expanding my knowledge of programming languages.

The next task in this module was the collaborative discussion about compliance. The differences between the UK and the EU GDPR are not monumental but significant for UK organisations or EU organisations that want to operate in both countries. There are interesting differences, mainly from the UK’s side, as their GDPR was largely based on the EU’s GDPR. This research has increased my knowledge of how Data GDPR can increase your workload, even in database design. Due to the proliferation of metadata, data which must be stored separately under GDPR, the workloads of databases have increased and must be designed for (Shastri, et al., 2020). This has been quite an interesting topic, as I did not know that GDPR could have affected databases to this depth.

In the last part of this reflective piece I will focus on the evaluation of the initial project and the final project. In the team project, along with my colleagues we created a database solution for an online bookstore inventory management system. This was my first experience in my professional career, and breaking down user requirements into a database design model was very interesting. From creating an ERD diagram to understanding the relations between attributes and entities. We also gave some points on compliance, along with database type, storage design, and the data cleaning process in the design process. The discussions with my colleagues gave me experience in discussing a data-related project, which will help me further understand the process of creating a database at work.

In comparison, the task in the final executive report was to present the initial project, along with further adjustments and recommendations from my personal perspective. In this report, I enhanced my knowledge of the SQL and NoSQL database types. SQL is good for structured scalability, while NoSQL is more flexible. On the other hand, NoSQL suffers from accuracy when it comes to big data processing. Furthermore, I also researched two different architectures, Kappa and Lambda, which helped me identify how to architect a data pipeline from data sources when streaming is needed or when batch processing is enough. This knowledge opens a lot of doors for me that helps me to understand data pipeline structures which I hope to use in my profession in the near future.

The journey through this module was quite intense. I personally believe this is one of the modules in which I had no experience. I believe that this knowledge will help me further in my career because now I understand how to design and architect a data pipeline.

# References

Boegershausen, J., Datta, H., Borah, A. & Stephen, A. T. (2022) Fields of Gold: Scraping Web Data for Marketing Insights. Journal of Marketing 86(5): 1–20. <https://doi.org/10.1177/00222429221100750>

Ding, X., Wang, H., Su, J., Li, Z., Li, J. & Gao, H. (2019) Cleanits: A data cleaning system for industrial time series. Proceedings of the VLDB Endowment 12(12): 1786–1789. DOI: <https://doi.org/10.14778/3352063.3352066>

Janani, K. & Ramamoorthy, S. (2024) PIoT ‐fortifying IoT device identity and data access: A security framework empowered by blockchain. Security and Privacy. DOI: <https://doi.org/10.1002/spy2.443>

Koehler, M. et al, (2021) Incorporating Data Context to Cost-Effectively Automate End-to-End Data Wrangling. IEEE Transactions on Big Data 7(1): 169–186. DOI: <https://doi.org/10.1109/TBDATA.2019.2907588>

Liang, X., Chen, K., Chen, S., Zhu, X., Jin, X. & Du, Z. (2023) IoT-based intelligent energy management system for optimal planning of HVAC devices in net-zero emissions PV-battery building considering demand compliance. Energy Conversion and Management 292: 117369-. DOI: <https://doi.org/10.1016/j.enconman.2023.117369>

Shastri, S., Banakar, V., Wasserman, M., Kumar, A. & Chidambaram, V. (2020) Understanding and benchmarking the impact of GDPR on database systems. Proceedings of the VLDB Endowment 13(7): 1064–1077. DOI: <https://doi.org/10.14778/3384345.3384354>

Wong Ting Yan, K. & Fong, J. S. P. (2021) Data Normalization. In *Information Systems Reengineering, Integration and Normalization.* Springer International Publishing AG. DOI: <https://doi.org/10.1007/978-3-030-79584-9_8>