Deciphering Big Data – Database Planning, Cleaning & Design (Units 1 to 3)

Overview

These first units of the module introduced key foundations in managing big data environments. We covered how to navigate data tools, workflows, and strategies—alongside what's needed to keep data secure, structured, and usable in real-world contexts.

Through both technical and collaborative work, I explored how data is collected, cleaned, and prepared for analysis, while also developing Python skills for working with APIs and file formats like XML and JSON. One of the most engaging aspects was the critical group discussion on the **Internet of Things (IoT)**, where we examined its potential and risks using key readings like Huxley et al. (2020).

What I Learned

- How to clean and transform raw data to make it reliable and usable.
- Different data structures, formats, and types—and how to choose the right one based on the context.
- The importance of data quality and how poor cleaning affects downstream insights.
- How to use Python for data tasks, including web scraping, API interaction, and handling file formats like JSON and XML.
- The role of key fields in linking relational databases.
- How anomalies and outliers affect data integrity.
- Why normalisation matters when designing efficient databases.

IoT Collaborative Discussion: Key Insights

The main focus of our discussion was to critically evaluate the opportunities and challenges of IoT, guided by Huxley et al. (2020). My peers and I agreed that IoT unlocks massive potential across sectors, from healthcare and agriculture to smart cities and industry (Islam et al., 2015; Wolfert et al., 2017; Zanella et al., 2014; Atzori et al., 2010). We all highlighted that the key to making IoT successful lies in **data quality**.

I emphasised that while large volumes of IoT data can be messy, not all outliers should be dismissed—some might be meaningful anomalies (like a valid sensor alert). Cleaning must be handled carefully, and blindly removing data can cause more harm than good.

Initial Post

by Chiamaka Ndudirim - Monday, 12 May 2025, 5:59 PM

Internet of Things (IoT) allows for the massive, continuous collection of real-time data from connected devices, unlocking major potential across sectors. As Huxley (2020) points out, when this data is clean and reliable, it becomes incredibly powerful for automation, predictive insights, and smarter decision-making. In healthcare, wearable devices enable remote patient monitoring and detection of early warning signs, thereby easing pressure on hospitals and enhancing care (Islam et al., 2015). In agriculture, real-time data supports more efficient crop and soil management (Wolfert et al., 2017). Smart cities rely on IoT for better traffic flow, energy use, and emergency response (Zanella et al., 2014).

But the scale of this data also introduces serious limitations and risks. Huxley (2020) highlights how inconsistent or messy data can skew results and lead to faulty conclusions, including Type I and II errors. And beyond the technical issues, there are real concerns around privacy, ethical use, and governance (Perera et al., 2014). With so much data flowing from so many sources, making sense of it—safely and ethically—is a challenge in itself.

Ultimately, the promise of IoT hinges on how well we manage and clean the data. Without strong data standards, clear cleaning protocols, and accountability, the risks may ultimately overshadow the rewards.

References

Huxley, K. (2020) 'Data Cleaning', *SAGE Research Methods Foundations,* Available at: https://doi.org/10.4135/9781526421036842861

Islam, S.M.R. et al. (2015) 'The Internet of Things for Health Care: A Comprehensive Survey', *IEEE Access*, 3, pp.678–708. Available at: 10.1109/ACCESS.2015.2437951

Perera, C. et al. (2014) 'Context Aware Computing for The Internet of Things: A Survey', *IEEE Communications Surveys & Tutorials*, 16(1), pp.414–454. Available at: 10.1109/SURV.2013.042313.00197

Wolfert, S. et al. (2017) 'Big Data in Smart Farming – A review', *Agricultural Systems*, 153, pp.69–80. Available at: https://doi.org/10.1016/j.agsy.2017.01.023

Zanella, A. et al. (2014) 'Internet of Things for Smart Cities', *IEEE Internet of Things Journal*, 1(1), pp.22–32. Available at: 10.1109/JIOT.2014.2306328

We also discussed **privacy and governance**, especially in contexts where users aren't even aware their data is being collected (Weber, 2010). Security risks, data misuse, and lack of standard frameworks were raised by peers as serious limitations. Despite all these, we agreed: if IoT is done right—with clean, ethical, and well-managed data—it can be transformational. If done poorly, it becomes a liability.

Database Design & Management: My Approach

In these units, I began planning a full database structure by:

- **Mapping entities and attributes** to design the logical layout.
- **Choosing the right data types and formats** to match how each item would be stored and accessed.
- Proposing a **DBMS** that suits the project needs—factoring in storage capacity, secure user access, and the ease of data manipulation and retrieval.
- Building out a **data cleaning workflow** to prepare raw data—documenting issues like duplicates, missing values, and formatting inconsistencies.
- Evaluating how cleaning methods help not just with analysis, but also reduce storage load and improve performance.

Personal Reflection

These early units gave me a solid foundation in how data should be handled before any analysis begins. I've realised that the strength of any data project lies in the behind-the-scenes work: how the data is collected, cleaned, and structured. Without that, even the best tools and models won't deliver useful insights.

Core Readings

- McKinney, W. (2022) Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter. 3rd edn. Sebastopol, California: O'Reilly.
    - Chapter 1 & 7
- Sardar, T. H. and Pandey, B.K. (2024) Big Data Computing. CRC Press

    - Chapter 5, 8 – 8.1 – 8.29, 8.35

- Huxley et al. (2020) Data Cleaning. Sage Foundation.

Recommended Reading

- McKinney, W. (2022) Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter. 3rd edn. Sebastopol, California: O'Reilly.
    - Chapter 3 and 6