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

## Overview

The first units laid the groundwork for managing data in real-world settings. I learned how to work with tools, workflows, and strategies to keep data secure, clean, and usable. We covered how data is collected, cleaned, and structured, while I also improved my Python skills—especially with APIs and file formats like JSON and XML. One standout task was the IoT group discussion, where we explored both the potential and the risks of connected devices (Huxley et al., 2020).

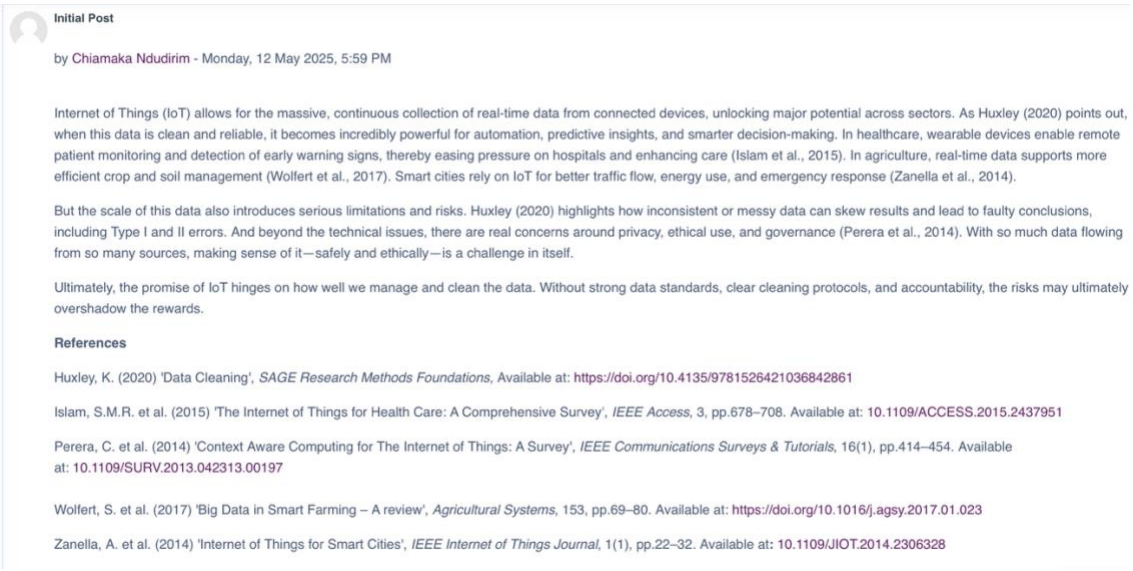
## What I Learned

- How to clean and transform raw data so it's usable.
- Choosing the right data structure or format based on the task.
- Why data quality matters and how poor cleaning affects results.
- Using Python for tasks like API calls, web scraping, and handling JSON/XML.
- Understanding primary/foreign keys in relational databases.
- The impact of outliers and anomalies on data integrity.
- The importance of normalisation when building databases.

## IoT 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

During these units, I began shaping the database project by:

- Mapping out entities, relationships, and key attributes.
- Choosing formats and data types that matched usage.
- Recommending PostgreSQL for its scalability and security.
- Documenting issues during cleaning like duplicates and formatting errors.
- Showing how clean data improves both analysis and storage efficiency.

## 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.

## References


Atzori, L., Iera, A. and Morabito, G. (2010) 'The Internet of Things: A survey', *Computer Networks*, 54(15), pp. 2787–2805. doi:10.1016/j.comnet.2010.05.010.

Huxley, J. et al. (2020) *Data Cleaning*. Sage Foundation.

Islam, S.M.R., Kwak, D., Kabir, M.H., Hossain, M. and Kwak, K.S. (2015) 'The Internet of Things for Health Care: A Comprehensive Survey', *IEEE Access*, 3, pp. 678–708. doi:10.1109/ACCESS.2015.2437951.

Weber, R.H. (2010) 'Internet of Things – New security and privacy challenges', *Computer Law & Security Review*, 26(1), pp. 23–30. doi:10.1016/j.clsr.2009.11.008.

Wolfert, S., Ge, L., Verdouw, C. and Bogaardt, M.-J. (2017) 'Big Data in Smart Farming – A review', *Agricultural Systems*, 153, pp. 69–80. doi:10.1016/j.agsy.2017.01.023.

Zanella, A., Bui, N., Castellani, A., Vangelista, L. and Zorzi, M. (2014) 'Internet of Things for Smart Cities', *IEEE Internet of Things Journal*, 1(1), pp. 22–32. doi:10.1109/JIOT.2014.2306328.