Dark Data: Software Diagnostics Case Study

Aaro Fulton

October 14, 2024

Abstract

This case study explores the concept of dark data within the realm of software diagnostics, such as software logging, to better understand the scale of dark data and the associated costs.

1 Introduction

According to Gartner (2023), dark data consists of information collected and stored during regular operations but rarely used for analysis or business value creation. This study will examine data that fits this description within the context of software logging and other diagnostic tools. While these tools are designed to be useful and provide insights into what a piece of software is doing, it can be challenging to see why some of this data might be considered dark.

2 Why is this data dark?

Logging tools are important pieces of any software implementation, as they allow for complex systems to be monitored and enable easy debugging when something goes wrong. However, this in turn creates data that may not be utilized to its fullest. For instance, if we consider a software application like a game, when it experiences errors, it will create a log and display it to the end user. This log may not be very readable or understandable for the end user, which can render this data dark, as it is not being "processed." Another way that dark data is generated by logging is through the logging of regular events. This can vary by product, but in some cases, logging can be overabundant, creating data that may not be useful but is still stored.

3 How much dark data is there?

To extrapolate the amount of data to the UK population, consider the following: The UK population is 68.26 million Office for National Statistics (2023a), with 90% of households having access to a computer Office for National Statistics (2022), and the average household size being 2.36 residents Office for National Statistics (2023b). We can calculate the number of households a s follows:

$$\text{Households} = \frac{68.26 \text{million}}{2.36} \approx 28.92 \text{millionhouseholds}.$$

Calculating 90% of these households that have computer access gives:

$$28.92 \cdot 0.9 \approx 26.03 \text{million households with computers.}$$

Assuming each household with a computer generates logging data, we can work out the volume of dark data generated.

4 What is the cost of dark data?

4.1 Extrapolated Costs

5 Conclusion

6 References

References

Gartner (2023). *Dark Data Definition.* Accessed: 2024-10-14. URL: https://www.gartner.com/en/information-technology/glossary/dark-data.

Office for National Statistics (2022). *Percentage of Homes and Individuals with Technological Equipment.* Accessed: 2024-10-14. URL: https://www.beta.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/percentageofhomesandindividualswithtechnologicalequipment.

— (2023a). *Annual Mid-Year Population Estimates: Mid-2023.* Accessed: 2024-10-14. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2023.

— (2023b). *Families and Households: 2023.* Accessed: 2024-10-14. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2023.