Dark Data: Software Diagnostics Case Study

Aaron Fulton

October 21, 2024

Abstract

This case study explores the concept of dark data for 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 that is rarely used for analysis or business value creation. This case study will examine data that fits this description within the context of software logging and other diagnostic tools. While these tools are designed to provide useful insights into software behavior much of the data remains unused as such becoming dark data.

2 Why is this Data Considered Dark?

Data is considered dark when it is not utilized for analytical purposes. When examining software logging as a source of dark data, we must account for the varying levels of expertise among users who interpret the logs.

2.1 Limited User Understanding

End-users, for example are less likely to comprehend the information parsed in logs. As (Hand 2020) notes, data is interpreted and analyzed by humans who apply their own understanding to make decisions based on that data. In the case of software logs users are often searching for specific errors or events that have caused unexpected behavior, but interpreting these logs requires domain knowledge that many lack.

2.2 The Therac-25 Case

An example of the dangers of dark data in logs can be seen in the Therac-25 incident, where operators of a CNC radiation therapy machine failed to understand an error message leading to the injury or death of six patients. The vague error message, "Malfunction 54," was documented only as "dose input 2," which affected patient safety but was not clearly communicated (Leveson and Turner 1993). In this instance, the logging data was available but became dark due to its insufficient explanation and the lack of infomration for the operator.

as such software logging can be considered dark data when it is not fully utilized by its intended audience, leading to potential harm or missed opportunities for improvement.

3 How Much Dark Data Exists?

undstanding how much dark data is created from logging can be difficult because its darkness is determined by the user's ability to understand and make use of the data presented. However, we can attempt a rough estimate by examing a small sample of data and extrapolating from there.

3.1 Data Collection and Extrapolation

Based on a small survey from four games and two applications across four individuals, the average amount of logging data generated per program was found to be around 1132 MB. While games typically generate more comprehensive logs than applications, this number gives us a rough estimate for extrapolation.

The UK population is approximately 68.26 million (Office for National Statistics 2023a), and 90% of households have access to a computer (Office for National Statistics 2022). With the average household size being 2.36 residents (Office for National Statistics 2023b), the number of households can be calculated as follows:

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

Of these, 90% have computers:

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

Assuming each household generates logging data around 1132 MB per program, and an average PC has around 72 programs installed, the total amount of dark data can be estimated:

$$\text{Dark Data} = 26.03 \cdot 72 \cdot 1132 \approx 2121.55\text{ petabytes.}$$

This rough estimate gives us a sense of the scale of dark data generated by logging, though the true figure may vary based on usage patterns and software complexity.

this data is colected from a small sample size of users from the games Ark: survival evolved, Minecraft, Space Engineers and War Thunder and the applications Chromium and Discord. these were chosen as they are commonly used applications and games by the group of individuals surveyed. this dats is in no way Conclusisve and is in fact a very small sample size. as this data was collected without care for Operating system or users time with the software that could lead to larger amounts of data being generated.

4 What is the Cost of Dark Data?

while the number we have calculated is the a very rought estimate we can use this to estimate the cost of dark data.

4.1 Extrapolated Costs

the cost of storaing data can very wildly depending on the scale and the method used (e.g ssd or hdd, m.2 or sata) but as we are looking at user data we can assume that the data will be storaged on user grade hardware using most system now run m.2 SSD that are mostly 1TB in size. the cost of a 1TB m.2 SSD is around £100. knowing this we can calculate the cost of storing the data.

$$\text{Cost} = \frac{2121.55}{1024} \cdot 100 \approx \pounds 207\,\text{million}.$$

with a forther cost coming from power these drives with the most populer of these drives the Samsung 970 Evo Plus using 6.0W of power when active and as the average user will be using their computer for 4hr 23 mins a day (Online Audience Measurement Service 2024) we can calculate the cost of powering these drives. from the cost of power in the UK being around £0.24 per KWh we can calculate the cost of powering these drives. (Ofgem 2022)

$$\text{Cost per day} = \frac{6.0\,\text{W}}{1000} \times 4.383\,\text{hours} \times 0.24\,\frac{\pounds}{\text{kWh}} \approx \pounds 0.0063\,\text{per day}.$$

$$\text{Cost per year} = 0.0063 \times 365 \approx \pounds 2.30\,\text{per year}.$$

To extrapolate this for 26.03 million PCs:

$$\text{Total cost per year} = 2.30 \times 26.03 \times 10^6 \approx \pounds 59.87\,\text{million per year}.$$

5 Solutions

there are meny Solutions to this problem that can be used to currect this issue.

5.1 Better Logging Practices

One Solutions to create beter logging practices creating logs that are more human readable and that can probide mroe inforamtion of how to solve and fix issues or that allow for data to be easily filter or searched for. as well as insuring that logs are removed after a set amount of time to reduce the amount of data or when a threshold is reached.

5.2 IT litracy

another Solutions is to create programs to teach people how to understand and read these logs to allow for them to be used more effectivly.

6 Conclusion

In conclusion, dark data generated by software logging can be a issue, but this is meny Solutions to this problem that can and has been used to help fix this issue. the cost of this data can be high but with the right tools and practices in place this cost can be reduced.

References

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

Hand, D. J. (David J.) (2020). *Dark data : why what you don't know matters / [internet resource].* eng. Princeton: Princeton University Press. ISBN: 9780691198859.

Leveson, N.G. and C.S. Turner (1993). "An investigation of the Therac-25 accidents". In: *Computer* 26.7, pp. 18–41. DOI: 10.1109/MC.1993.274940.

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.

Ofgem (2022). *New Energy Price Cap Level October-December 2024.* Accessed: 2024-10-20. URL: https://www.ofgem.gov.uk/news/new-energy-price-cap-level-october-december-2024-starts-today.

Online Audience Measurement Service, UKOM Ipsos iris (2024). *Online Audience Measurement Service, March 2024.* Accessed: 2024-10-20. URL: https://ukom.uk.net/uploads/files/news/ukom/281/UKOM_Mar_2024_OMO.pdf.