RESEARCH ON METHODS OF DATA LOSS AND DATA RECOVERY

By

Christian Tabbah and Joshua Menezes

CSC492

Professor Andreas (Andi) Bergen

University Of Toronto

August 13th, 2023

In modern society, data plays a large portion of daily aspects in one's life. With approximately 300 million terabytes of data being generated and analyzed each day, precise care and effort must be put into storing data and having mechanisms in place to recover data in case it is destroyed or lost. This paper aims to study both modern and historical examples of data loss as well as data recovery through the use of physical storage mediums as well as logical (virtual) systems. In this section, we will discuss the mechanisms of some common storage devices as well as their respective functionalities.

Before tackling the subject of data loss and data recovery, it is important to first research some of the important virtual and physical mediums of data storage. These mediums are pivotal to understanding the complexities and nuances of data management. This paper aimed to take a somewhat historical approach to this, as understanding the development of storage technologies over time provides valuable insights into the evolution of data preservation and retrieval methods. Below are the physical data storage mediums that were researched:

Physical Data Storage:

- **Hard Disk Drives (HDDs):** HDDs have been a mainstay in data storage since 1956. They rely on spinning platters and read/write heads to write and read data on the disk. The disk is composed of Cobalt Chromium Tantalum Alloy, because it has small magnetic regions, who's directions can be manipulated using external magnetic fields. These magnetic regions are manipulated in one of two directions, which correspond to either a 1 or a 0. Note that it is not the direction of the region that determines the bit, but instead it is the change of magnetic direction that determines whether the bit is a 1 or a 0. Additionally, Each disk is divided into tracks and sectors. The head stack assembly will navigate these zones to find different parts of data on the disk. Each track sector has a preamble, an address, data and error correcting codes (ECC). The preamble helps synchronize the read/write head of the HDD with the data in each track sector, the address tells the head which track it's on, the data contains the actual information being stored and retrieved, and the error correcting codes help detect and correct errors that may occur during read/write operations.

- **Solid State Drives (SSDs):** Flash memory based SSDs have rapidly gained popularity due to their speed and reliability after being created in 1989. They use NAND flash memory instead of spinning platters to store data electronically, thus eliminating mechanical components and increasing data access speed. SSDs work by storing data in billions of charge trap flash memory cells. Each of these cells store 3 bits of information, by trapping different levels of electrons. These cells are organized into pages, and pages are grouped into blocks. To modify data, the SSD uses electrical charges to modify the stare of memory cells within a page. Each cell is accessed through bitline selectors and control gate selectors, which act as row and column access to each page, through the use

of electrical charges.

- **Optical Discs:** Optical discs like CDs and DVDs have played a historical role in data distribution, with the CD emerging in 1982, and the DVD following suit in 1995. Both the CD and the DVD contain data that is stored on a single data track that spirals outwards. In the aluminum on the track, there are bumps that correspond to a 0 bit, and divots that represent a 1 bit. They are both read through a handle that moves outwards to follow the spiral, which detects the bumps and divots using lasers, to read the binary. The main difference between the two is that CDs only have a capacity of about 700MB, and are normally used for audio files, while DVDs have a capacity of about 4.7GB on each side, and store anything.

Below is an introduction to the virtual data storage mediums, also known as file systems. This paper aimed to understand how they work as well as highlight the differences between them, as understanding their differences give us insight as to how some tools can recover data from so many different file systems.

File systems are essential components that provide structure to the data stored on hard drives and SSDs. Hard drives and SSDs come without any predefined structure, thus, they are formatted with file systems during the installation of operating systems, like windows. Partitioning happens when a single drive is divided into multiple logical drives, known as partitions. Each partition acts as an independent storage unit, like the familiar C: and D: drives on your computer, each of which use their own file system. Most file systems are structured like the following:

- Metadata about files and directories is stored in structures called inodes, which help the operating system keep track of file attributes and locations.

- Inodes then point to blocks, the actual storage units where data resides.

- These blocks can be allocated or unallocated, meaning they might be filled with data or ready to receive new data. Interestingly, unallocated blocks do not necessarily mean that data has been cleared. This means that even after unallocated, a block can still have remnants of the data that used to be there.

The following table provides a succinct presentation of several noteworthy file systems. The intention here is not to dissect the finer points of each notable feature or limitation, but to emphasize the existence of the differences between them. While the features are labeled as “notable”, an explanation to what they actually mean is left to the reader’s own research, as it remains beyond the scope of this paper. Instead, the focus is on recognizing that this diverse group of file systems all have their own capabilities and limitations.

File Systems	Release Date	Used in (mainly)	Notable Features (features they are known for)	Limitations
FAT32	1996	Universal	Simple structure, Cross-platform compatibility	File size limit of 4GB, Max partition size of 8TB
exFAT	2006	Universal	Large file sizes, Flash drive compatibility, No journaling	Not as compatible as FAT32 (older devices and versions of linux do not support it)
NTFS	2001	Windows	Journaling, ACLs, Compression, Encryption, Alternate Data Streams	Lacks compatibility with some linux and mac devices
EXT4	2008	Linux	Journaling, POSIX permissions, ACLs, Extents, Online resizing	Does not have secure deletion feature, has limited max file and partition size(16TB)
APFS	2017	MacOS	Cloning, Snapshots, Encryption, Fast directory sizing, Copy-on-Write (no need for journaling)	Limited compatibility for old versions of macOS and other third party apps
ZFS	2006	Various	Data integrity checks, Copy-on-Write, Snapshots, RAID-Z	Failure to check RAM health in the case of data errors, memory intensive and degrades at higher capacities

Logical Data Loss

Logical data loss involves a loss of data caused by software. This can occur from mishandling memory or file states, data lost via compression algorithms, human error, and most common, malicious data deletion. Oftentimes, logical data loss stems from rushed updates and trying to over-optimize certain tasks. This is evident in many operating systems as certain updates sometimes introduce memory related bugs, which can cause inconsistent system states leading to data loss.

Although built to be stable, file systems, if mishandled, can become unpredictable and lose data if the operating system doesn't handle the memory properly. Oftentimes, memory related bugs are caused from leaving a file in an inconsistent state or from new system updates. For instance, on a system update, if files are being modified, power loss could leave files in an uncertain state, leading to data corruption and ultimately loss. On the other hand, an example related to newer updates can be demonstrated through the windows operating system. A common error code known as "-2144927436" was reported to happen frequently after a certain windows update. The error code that was reported indicated improper file execution policies or even damaged registry files. It was later identified that this was due to Microsoft updating non-Microsoft drivers when releasing system updates, resulting in inconsistent data management, leading to data loss.

A large portion of maximizing storage capacity involves optimizing the amount of space currently being used on disk. One way to do this is by compressing files. Compression algorithms can be classified in two ways: lossless and lossy. Lossless data compression involves no data loss during compression and decompression. These algorithms are the most common as the intention behind compression data is to be able to retrieve it all back during decompression. Lossless compression is consistently used for text, where data must be preserved, such as in banking records. On the other hand, lossy compression algorithms aim to reduce data size by eliminating redundant data. Lossy compression has many applications in modern day data saving. For instance, text-summarization algorithms reduce irrelevant data to create a more concise summary. Additionally, image compression algorithms often use lossy compression to shrink images. Oftentimes the human eye cannot detect size reduction from well designed algorithms. As a result, smaller sized files remain that look the exact same as the original large file, through the removal of pixels.

The most common source of data loss is human error. Accidental file deletion happens to most individuals and in most cases data is able to be easily recovered due to system caches. Some operating systems have a concept of a trash bin, or a soft delete where the file can be reconstructed. For example, the windows operating system implements this feature in which upon soft deletion of a file, the reference to the file is removed, but within the actual recycling bin, resides metadata that can be used to reconstruct the file if needed. Additionally, a user can

permanently delete a file. A hard delete usually involves deleting all contents of the file, as well as removing all associated references and metadata that the system may hold. On a unix based system, this can be done with the “shred” command, which overwrites a file and optionally removes it, making it harder for hardware to reconstruct it, allowing for a secure delete.

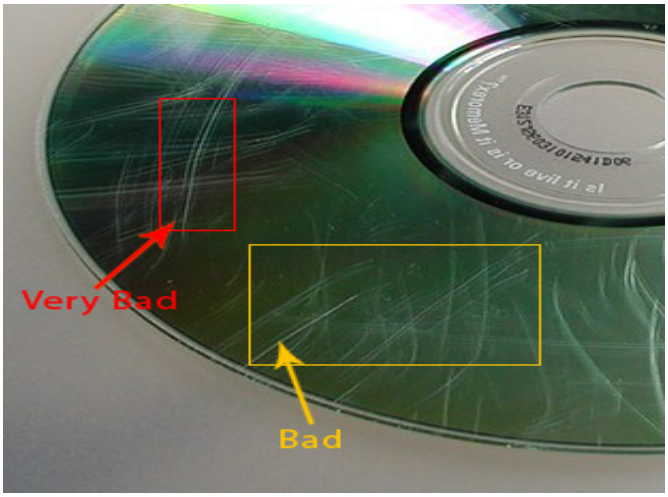
Lastly, a more modern form of targeted data loss via software involves malware. A classification of malware known as a “wiper” exists to solely delete data “beyond recovery.” Wipers will commonly target system files in order to render a system completely useless. On windows, this could be registry files and on unix based systems, shadow files. Wipers can be used in conjunction with other malware, such as ransomware, in order to extort individuals from money. Most recently, wiper software has been used in modern warfare to disrupt technologies. With the current war between Russia and Ukraine, communications for each country are in the form of extensive computer networks, and are detrimental to operations. A wiper known as AcidRain “targeted ViaSat satellite modems” destroying a considerable portion of Ukraine’s military communications. With wiper software, hard deletions are performed, making data unrecoverable after the malware has been installed. The best defense is to have up-to-date antivirus software in order to prevent the malware from intruding.

Physical Data Loss

Physical data loss is entirely hardware related. It typically involves physical obstructions, such as damaged components, resulting in data being unable to be read, power failure, or overheating.. As each hardware medium typically has its own set of faults, we'll analyze three different mediums to cover a broad range of devices: DVDs, hard drives (HDDs) and solid state drives (SSDs).

As DVDs rely on an optical drive to read grooves in the disc, it is important that the surface of the disk remains intact. Imperfections such as smudges, fingerprints, dust and scratches along this surface could cause a failure when reading the disc by the optical drive.

Figure 1: An optical disc with a scratched surface



Hard drives (HDDs) are more prone to mechanical failure. Since they consist of moving components rotating at a high rotational speed, parts can often wear down over time and fail, resulting in the entire mechanism of data retrieval failing.

Lastly, we will introduce data loss within a more modern medium, a solid state drive (SSD). SSDs store data using flash based memory, and as they are used overtime, transistors wear out and eventually lose their charging capacity. Additionally, NAND based flash memory SSDs must maintain a low charge to function consistently. When power is unavailable for an extended period of time, the device may lose data. Some research suggests that an SSD may retain data for 2-5 years with no power, before data loss occurs.

Logical Data Recovery

In the realm of data management and storage, events such as accidental deletions, hardware malfunctions or system crashes occur more often than one might think. Fortunately, there are many tools on the market to recover this lost or deleted data such as MiniTool Power Data Recovery, R-Picture, EaseUS Data Recovery and Disk Drill (and much much more). In the following discussion on logical data recovery, we will dive into a comprehensive exploration of these types of tools, by analyzing Disk Drill, which stands out to us due to its prominence and reputation. The idea of trying all of them out did come to mind, but the idea of downloading several tools that access every part of your disk seems a bit reckless if you don't fully trust them. Regardless, most of these tools are quite similar in their strategies to recover lost data, which we will examine in this section of the paper.

Disk Drill is a data recovery tool created by CleverFiles that is used to recover data from loss scenarios such as accidental deletion, emptying recycling bins, corrupted file systems, partial file corruption, ect. It supports NTFS, FAT32, exFAT, HFS+ and APFS, and can even reconstruct such corrupted file systems.

Methodology:

Before using Disk Drill, which we know will go through the entire disk, we decided to use it through a VM instead, as hopefully it would only drill through the memory allocated to the VM, and not our entire personal machines. Note that we also only ran the tool through a Windows VM as well as a macOS VM, to see if there were any differences between the two. At first, on the Windows VM, we had allocated 20 GB of memory to the VM, but using Disk Drill would result in constant crashes, thus we had to increase the capacity to 30GB for the crashes to no longer happen. We tried 5 runs on both the Windows and macOS VMs, where we downloaded a few (5) images, and promptly deleted them from the recycling bin. The hope was to recover these images through the use of Disk Drill. Additionally, we decided to use a couple different file types to see if disk drills had any trouble recovering one over the other.

Results:

After running Disk Drill on both operating systems (which is notable considering they both have different file systems to traverse), we found that we had to try several times before even being able to recover a single image. Below is a table of how the experiment went:

OSs	**1st attempt**	**2nd attempt**	**3rd attempt**	**4th attempt**	**5th attempt**
Windows	**1 picture recovered (full .png picture)**	Nothing recovered	Nothing recovered	**1 picture recovered (corrupted .jfif picture)**	Nothing recovered

macOS	Nothing recovered	Nothing recovered	**1 picture recovered (full .jfif picture)**	Nothing recovered	Nothing recovered

Through these results, we discovered 3 important facts:

1. There was no notable difference between the Windows and macOS runs. It seemed like the likeliness of a picture being recovered was completely up to chance, and was not at all impacted by any difference in file system (whether it be NTFS or APFS).

2. There was also no notable difference in the recovery chance of any particular file type.

3. We only ever got a single picture recovered at a time. This led us to believe that using a VM with only 30 GB of memory could have created a scenario where there was so little space to store the images in the first place, that maybe their clustering forced some sort of overlap in memory, only allowing a single image to be recovered.

The third fact led us to attempt to try Disk Drill on our personal machine, with around 700 GB of storage space, which allowed us to recover 3 out of the 5 images. This only reinforced our theory, although we could not figure out why this was the case.

The workings of Disk Drill:

After using Disk Drill, our goal was to explain how it worked, as well as what information it was giving us. Firstly, for the information given, it separated the files it found into three categories:

1. **Deleted or lost files:** Files that have been deleted, or files that are lost and cannot be recovered.

2. **Existing:** These files are simply all the files that you would find available on your disk.

3. **Reconstructed:** These are files that Disk Drill's algorithm was able to reconstruct, which means that these files can indeed be recovered.

Each of these categories were split into sub-categories, each of which being a partition from the drive. On our personal machine test, we saw that there was a subcategory for the C: drive, the E: drive, the EFI System Partition (ESP) as well as the System Reserved partition. This very fact let us know that Disk Drill really does scan the entire disk, and leaves no partition unread.

As for the explanation, here is a step by step explanation of how Disk Drill works. Note that we had to use our knowledge to fill in some gaps that Cleverfiles would not explain on their website.

1. Disk Drill scans the entire storage device bypassing the file system.(Deep scan)
    a. It will locate blocks specific to some files based on the file system's data allocation information. This essentially means that it will traverse the file system the same way any normal operating system would.

2. Disk Drill identifies signatures or headers. These are unique patterns of bytes at the beginning of specific file types.

3. Disk Drill looks for both the beginning and end of files by matching signatures or patterns that indicate the file's structure.

4. If a file is fragmented, Disk Drill tries to piece the segments together in the correct order (the algorithm for this is a secret).

5. Finally, we assumed that like every other recovery tool, Disk Drill verifies checksums to ensure integrity of reconstructed data (note that this is an assumption, and may not be true).

Physical Data Recovery

The recovery process of physical devices tends to be more involved than the logical side. When hardware devices fail, the most common solution is to often replace them with new ones, or replace the corresponding parts that failed. This is the case in devices such as hard drives. For other devices, corresponding fixes may be more niche to the specific device. For instance, optical disks not reading properly could be that the disc is dirty or scratched. As a result, wiping the disc with a cloth to clean any debris off or even polish the disc may allow for an optical reader to perform read operations on the disc.

Figure 2: Replacing an actuator arm on an HDD

Conclusion

To conclude, our data driven society showcases the vast importance data storage and recovery plays over various storage mechanisms. With large volumes of data being created, stored, and destroyed each day, it becomes vital to be aware of data preservation and recovery strategies. In our investigation, we have analyzed modern and historical instances of data loss and recovery efforts distributed across a multitude of storage mediums through the physical and logical lens of technology. Our findings highlight the importance of implementing dynamic mechanisms for data protection and recovery. With a continuous evolution of technology, we must reflect on past methodologies of data storage, in order to develop new strategies for future challenges in data that we may face.

References

Exploding Topics. "How Much Data Is Generated Per Day?" Exploding Topics Blog. https://explodingtopics.com/blog/data-generated-per-day#how-much

UGetFix. "How to Fix File System Error -2144927436 in Windows 10." https://ugetfix.com/ask/how-to-fix-file-system-error-2144927436-in-windows-10/#:~:text=There%20could%20be%20many%20reasons,installing%20a%20new%20Windows%20update

Wikipedia. "Lossy Compression." https://en.wikipedia.org/wiki/Lossy_compression

How-To Geek. "How Exactly Does the Windows Recycle Bin Work?" https://www.howtogeek.com/166806/how-exactly-does-the-windows-recycle-bin-work/

GeeksforGeeks. "Ways to Permanently and Securely Delete Files and Directories in Linux." https://www.geeksforgeeks.org/ways-to-permanently-and-securely-delete-files-and-directories-in-linux/

Linux Documentation. "Shred - Delete a File Permanently." https://linux.die.net/man/1/shred

CrowdStrike. "The Anatomy of Wiper Malware - Part 1." https://www.crowdstrike.com/blog/the-anatomy-of-wiper-malware-part-1/

Macgasm. "Mac Disk Repair Software: Compare the Best Disk Repair Apps." https://www.macgasm.net/data-recovery/mac-disk-repair-software/

EaseUS. "EXT2, EXT3, EXT4 File System Format and Difference." https://www.easeus.com/partition-master/ext2-ext3-ext4-file-system-format-and-difference.html

Stanford University. "How Hard Drives Work." https://cs.stanford.edu/people/nick/how-hard-drive-works/#:~:text=The%20hard%20drive%20contains%20a,the%20stored%200's%20and%201's

YouTube. "How Do SSDs Work?" https://www.youtube.com/watch?v=wtdnatmVdIg YouTube. "How Does CD-ROM Work?" https://www.youtube.com/watch?v=5Mh3o886qpg&t=667s

Canon Global. "The Structure and Manufacturing Process of CDs and DVDs." https://global.canon/en/technology/s_labo/light/003/06.html#:~:text=The%20inner%20face%20of%20CDs,on%20essentially%20the%20same%20principle

Explain that Stuff. "How Does a CD Player Work?" https://www.explainthatstuff.com/cdplayers.html

EaseUS. "File System: Types and Explained Differences." https://www.easeus.com/diskmanager/file-system.html

CleverFiles. "Top 15 Data Recovery Software of 2021: Reviewed and Ranked." https://www.cleverfiles.com/data-recovery-software.html