Langzeitarchivierung

Eike Scharf

September 24, 2015

Contents

1 Vorwort und Erklärung des Themas

Diese Ausarbeitung befasst sich mit der Langzeitarchivierung von Daten. Hier zu erst ein paar Definitionen von Dingen, die zum Verständnis hier von notwendig sein sollten oder zum weiteren Verständnis hinzugefügt sollen.

Information: Information ist ein abstraktes Konzept, das Wissen auf einer Ebene repräsentiert und ist an und für sich genommen formlos, aber ausdrückbar. Zum Beispiel der Satz "Peter trägt ein blaues Hemd." beinhaltet zwei(2) Informationen, A) Peter trägt ein Hemd, und B) es hat die Eigenschaft das die Farbe jenes Hemdes blau ist. Aber dieser Satz benötigt auch eine Information um vollständig verstanden zu werden. Nämlich die um die Identität von Peter. Ansonsten ist es nicht möglich, alle Informationen und was dahintersteht zu verstehen. Benötige Rahmeninformationen nennt man auch Kontext.

Daten: Daten sind eine Repräsentation von Informationen, welche allerdings kein abstraktes Konzept sind, sondern meist eine in der realen Welt existierende Repräsentation von Informationen. Es gibt verschiedene Möglichkeiten Informationen zu repräsentieren, beispielsweise die Zahl 27. Es ist möglich sie in Schrift zu repräsentieren. "siebenundzwanzig", in binär "0001 1011" oder in hexadezimal "1B"

Speichermedium: Die physikalische Verankerung von Daten in unseren realen Welt. Wobei die Arten wie gespeichert wird, ist der zentrale Punkt in diesem Dokument.

MTBF: kurz für Mean Time Between Failures zu deutsch: Durchschnittliche Zeit zwischen Ausfällen. Ist eine Zeitangabe bei Hardware, welche die Zeit angibt, die das datenspeichernde Objekt im Durchschnitt als Betriebszeit hat. Dies ist daher nicht auf das gesamt Gerät bezogen, sondern nur auf das Datenspeichernde. Beispielsweise das speichernde Halbleiterelement in einer SSD, hierbei sind andere Fehlerquellen wie mechanische Bauteile oder Controller ausgeschlossen. Die Zeit wird gemessen von einem Ausfall und nach Instandsetzung zum nächsten Ausfall.

2 Anforderungen an Daten Speicher

2.1 Allgemeine Anforderungen

Im allgemeinen lassen sich 2 besondere Merkmale herauspicken, welche für jede Art der Daten Speicherung relevant sind. Dies sind Haltbarkeit und Kompaktheit der Daten. Diese beiden Faktoren sind in einigen Technologien sich entgegnen über nicht unbedingt exklusiv.

Haltbarkeit als Eigenschaft ist im eigentlichen Sinn nur die Möglichkeit nach einer Zeit wieder die Daten aus dem Medium zu extrahieren. Dies ist eine der zentralsten Punkte da der Sinn hinter dem speichern ist, diese irgendwann wieder abrufen zu können. Anzumerken ist hier aber auch, dass sich je nach Anwendung der Zeitraum stark verändern kann. So kann es möglich sein, dass ein gespeichertes Dokument zum Beispiel nur für eine Woche verwahrt werden muss, was für die meisten Speicher Medien kein Problem darstellt. Es könnte aber auch möglich sein, dass ein Dokument ein Jahrzehnt aufbewahrt werden muss, was hingegen schon für einige Speichersysteme ein Problem darstellt.

Die Kompaktheit lässt sich analog zur klassischen Definition von Dichte halten $\frac{Daten}{Volumen} = Datendichte oder Kompackheit$ Diese Eigenschaft ist wichtig, da größere Datenmengen eine menge Platz einnehmen können. Grade bei wissenschaftlichen Experimenten wie beispielsweise beim LHC 30 peta byte im Jahr anfallen[1]. Dies archiviert würde einiges an Raum verbrauchen.

2.2 Besondere Anforderungen

Es gibt aber auch Situationen wo diese grundlegenden Anforderung erweitert werden. Hier im weiteren wird sich mit den zwei Fällen der physikalischen Sicherheit und der Digitalen Sicherheit.

Hierbei ist an zu merken das die physikalische Sicherheit auch den klassischen Diebstahl mit einschließt, in der Theorie aber dies hier nicht weiter beleuchtet wird, da es zu diesem Thema bereits gute Lösungen existieren und diese auch von den meisten Menschen verstanden wird. Wohin gegen die Digitalen Seiten der Datensicherheit etwas komplizierter sind und dementsprechend genauer erläutert werden.

Doch zu erst der teil zur Physikalischen Sicherung der Daten vor Manipulation; es gibt Szenarien, wo es von extremer Wichtigkeit ist, das Daten nicht manipulierbar sind da es sonst zu Konsequenzen kommen könnte, die eine Schädigung einer menge Menschen zu Folge haben könnte. So ist es zum Beispiel im Bereich der Finanzindustrie einen Transaktionslog zu führen. Aus dem einfachen Grund, dass hieraus Dinge wieder hergestellt werden können,

sollten Fehler unterlaufen sein oder Manipulationen geschehen sein. Dies ist natürlich nur solange möglich solange die Transaktionslogs auch der Realität entsprechen und selber nicht manipuliert wurden. Hier für existiert die sogenannte WORM Technologie. WORM steht für "Write Once Read Menny", also zu deutsch einmal schreiben und oft lesen. Dieses Prinzip ist der realen Welt nachempfunden, wo es möglich ist ein Blatt Papier einmalig mit einem Dokumenten echter Stift zu beschreiben und dann es beliebig oft zu lesen, außerdem ist es nicht möglich das geschriebene zu verändern ohne das Papier zu verändern, zB. ein Wort ausschneiden. WORM Speichermedien sind entweder Scheiben oder Bänder die sich nur ein einziges Mal beschreiben lassen und danach nur noch gelesen werden können. Im professionellen Bereich sind dies meist Sonderformate und Spezialdatenträger, aber auch im Heimanwenderbereich finden wir dies Konzept in der klassischen CD-R, DVD-R und der entsprechenden Blue-ray Variante. All diese Medien sind einmal beschreibbar und ohne physikalische Veränderung nicht veränderbar, sobald sie voll sind.

3 Heute Verfügbare Medien

3.1 HDD

Die klassische Festplatte im englischen auch Hard Disk Drive, abgekürzt HDD, ist einer der weit verbreiteten Wege Daten zu speichern, sowohl im Endnutzer- als auch im professionellen Bereich an zu treffen. Dieser Weg Daten zu speichern zeichnet sich durch seine Vielseitigkeit aus, welcher Datenvolumen von bis zu 4 Terabyte pro Einheit und einer Zugriffszeit welche im Bereich zwischen 10-100 ms liegt; was ausreichend für webbasierte Protokolle ist eine rechtzeitige Antwort zu liefern, bevor die Anfragezeit überschritten ist. Dennoch gibt es auch einige Argumente die gegen Festplatten als Möglichkeit sprechen, Dateien zu archivieren, im besonderen ihrer mechanische Natur welche ihnen eine 3-5 Jahre [23] Lebensdauer gibt. Des weiteren ist die MTBF des Materials von 0,8 Mio. bis 1,4 Mio. Stunden recht hoch, wobei diese immer noch unter den Silizium basierten Flash Speicher liegt. Dies macht die klassische Festplatte zu ein wichtigen Bereich der Langzeitarchivierung. Besonders im Bereich der Web-Archive spielen sie eine zentrale Rolle, da sie eine Rückmeldung in der entsprechenden Zeit liefern kann. Die Nachteile der HDD ist der hohe Energieverbrauch im Betrieb bei großen Populationen.

3.2 SSD/Flash

SSD oder auch Solid State Drive ist eine Technologie, die Daten in Halbleiterelementen speicher. Der Vorteil der Technologie ist, dass sie eine hohe Zahl an Input-Output-Operationen hat und zusätzlich ist auch eine hohe Bandbreite möglich. Diese Punkte sind für die Archivierung allerdings nicht hauptsächlich interessant. Die Haupteigenschaft die SSD interessant macht, ist das keinerlei bewegliche Teile verbaut werden. Dies ist interessant, da es mechanische Probleme, die bei Festplatten oder anderen bewegten Medien auftreten können, ausschließt. Ebenfalls die Lebensdauer von 3 bis 10 Jahren (Hersteller abhängig) ist den Festplatten ebenbürtig oder sogar überlegen. [4,5] Ebenfalls die höhere MBF von 2 Mio. - 2,5 Mio. Stunden ist der Festplatten aus mechanischen Gründen überlegen. Das Problem der SSD, was sich in der aktuellen Zeit verringert hat, das die Speicherkapazität noch geringer ist als bei der HDD, mit maximal 2 TB. Außerdem sind die Anschaffungskosten für die gleiche Speichermenge größer, was das zentrale Problem der Wirtschaftlichkeit in den Fokus rückt. Allerdings ist durch das nicht vorhanden sein von Motoren aber auch der Stromverbrauch geringer.

3.3 Optische Medien

Der Bereich der optischen Medien wird hier durch CD, DVD und Blue-ray-Disk vertreten. Diese umfassen die dominanten und dadurch wichtigsten Formate, außerdem hält es die Liste kurz und das Prinzip der restlichen Formate sollte trotzdem ausreichend abgedeckt sein. Ausgenommen sind hier die Millennium Disk und die Glas Master Disk. Diese werden in einem separaten Abschnitt behandelt, da sie noch recht junge Technologien darstellen und nicht etabliert sind. Die Speichermenge von optischen Medien hängt vom jeweiligen Datenträgerstandart ab. Von 700 MB bei einer CD-ROM bis zu 50 GB bei einer duallayer Blue-ray.
Bei den optischen Medien muss zwischen 2 verschieden Typen unterschieden werden. Die Gepressten und die Gebrannten. Diese beiden Arten unterscheiden sich in Materialien, Herstellung, Lebensdauer und Zielmarkt.
Der Markt der gebraten Medien ist der eigentliche Handel mit blanken Medien, die dann vom Benutzer selber beschrieben werden. Das heißt es ist möglich, dass auch eine einzelne Kopie an zu fertigen. Im Gegensatz zum Pressen, wo es nicht möglich ist ein einzelnes Medium einmalig zu pressen, aus ökonomischen Gründen. Dafür ist die Haltbarkeit von gebraten Medien auch geringer, und diese sind anfälliger gegenüber Umwelteinflüssen. Dies hängt mit der Eigenschaft des Brennen selber zusammen, wo ein Laser mit

einer bestimmten Wellenlänge das Material so verändern und die Daten in das Material eingraviert. Und da das Sonnenlicht auch diese Wellenlängen enthält, ist es möglich das die Sonne ebenfalls das Material verändert, bei einer Aussetzung über längeren zeit; wegen der geringen Intensität. Bei einer Lagerung in guten Konditionen können selbst gebrannte optische Medien einige Jahrzehnte überdauern; wenn das Material hochwertig ist.
Im Vergleich mit selbst gebraten optischen Medien sind gepresste optische Medien weit aus länger haltbar. Was mit anderen Werkstoffen zu tun hat. Die maximale Lebenszeit dieser Datenträger ist noch nicht ganz klar. Bekannt ist, dass einige CDs langsam verfallen[8]. Dies gilt allerdings für CDs die nicht optimale Bedingungen haben. Beispielsweise bestimmte Lacke und Lösungsmittel in den Aufdrucken haben die Eigenschaften sich zu den Daten durch fressen und diese zu zerstören. Allerdings dauert dies in den oben genanten Fehlen über 20 Jahre. Der größte Nachteil an gepressten optischen Medien ist, dass diese Datenträger nicht individuell sind.

3.4 Bandmedien

Bandmedien stellen den Klassiker in der digitalen Archivierung dar. Wobei diese hier durch den LTO Standard, der um die Millennium-wende entstand, repräsentiert wird. Davor existierten andere Formate, die physikalisch anders aufgebaut waren, andere Breite, Tapelänge, Dichte der Daten oder ähnliches. Diese machen alle aber im Grunde das gleiche, also hier zusammen gefasst von der frühen Magnetbänden zu den modernen LTO Standard.
Magnetbänder eigenen sich besonders für lange und statische Speicherung von Daten mit geringen Zugriffszahlen. Dies ergibt sich aus der Langlebigkeit von bis zu 30 Jahren und hören Speicherkapazitäten bis zu 2.5 Terabyte (ohne Kompression) mit LTO-6. Und da die Bänder ein Abspielgerät benötigen um gelesen zu werden, verbrauchen diese selber kein Strom, weder bei Lagerung noch beim Lesen.
Allerdings hat diese Speichervariante auch einige Nachteile. Der erste ist, dass es lange dauert bis eine spezifische Datei gefunden wird. Dies hängt mit der Länge der Bänder zusammen, welche mehr als 800 m ab LTO-4 beträgt; dem entsprechen dauert der mechanische Zugriff auf die Daten. Dies bedeutet auch, das Bänder die fragmentiert sind besonders ineffektiv sind, und am besten große lineare zusammenhängende Daten gespeichert werden sollten. Außerdem sind die Bänder nicht auf einen dauerhaften schreib und lese Zyklus ausgelegt, wie es beispielsweise bei einen Betriebssystem der Fall ist.

3.5 Cloud

Die Cloud ist ein recht neues Konzept, sei es Cloud-Computing oder Datenspeicher in der Cloud. Das grundlegende Konzept ist, dass der Nutzer den Aufbau des Innenrinnen der Cloud nicht kennen muss und diese auch nicht Instandhalten muss. Dies hat mehre Vorteile. Zum einen sollte dafür gesorgt sein, das Personen mit Fachkenntnissen sich um die Datenspeicher kümmern. Dies bedeutet einfach, dass der Nutzer sich keine Sorgen machen muss, dass durch fehlerhafte Dateiträger die Daten verloren gehen. Aber das ganze hat auch einige besondere Risiken, die später noch genauer behandelt werden, sich aber um den gleichen Punkt drehen, dass die Daten in den Händen anderer sind. Dies kann sich als Problem erweisen, wen der Träger der Cloud die Dienste einstellen muss.

4 Gefahren für Daten

Die Gefahr für Daten stammt nicht nur von Datenträger selbst, sondern auch von den Daten selber. Oder anders ausgedrückt: Auch wenn es so aussieht als sei alles mit den Datenträger gut, so können dennoch Daten beschädigt werden, oder auch durch andere Faktoren unlesbar werden.

4.1 Datenspeicherung

4.1.1 Lite vs Big endian

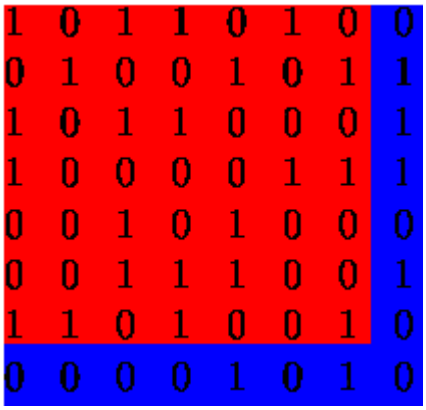
Die Frage wie Daten gespeichert und verarbeitet werden sollte, ist eine Frage, die Software und Hardware grundlegend unterliegt. Probleme hier entstehen wenn 2 verschiedene Systeme auf einander treffen. Dies kann dann dazu führen, dass die Daten nicht korrekt interpretiert werden. Hier als Beispiel werden die 2 endian Codierungen herangezogen. Diese behandeln die Handhabung des MSB (Most signifikant Bit), bei Big endian wird das höchstwertige Bit zu erst gespeichert. Im Gegensatz zu lite endian, wo das MSB als Letztes gespeichert wird. Wenn nun ein System Daten, die im anderen System Codiert sind; bekommt dann wird die Wertigkeit der Bytes nicht richtig interpretiert. Dies führt dazu, dass die Daten dann nutzlos werden. Hier als Beispiel, die Zahl 10 codiert in den beiden Systemen.

$$0000\ 1010_2 = 10_{10}\ \text{Big Endian}$$

$$0101\ 0000_2 = 10_{10}\ \text{Little Endian}$$

4.1.2 Heming code

Aber es gibt auch auf der Datenebenene diese so zu speichern, dass kein Verlust stattfindet. Ein Beispiel ist hier die Heming-Codierung. Die Idee hinter dem Hemingcode ist, dass durch die Daten generierten Paritäten die integriert gewährleistet werden. Das Prinzip ist, solange die Parität identisch aus den Daten generiert werden kann sind die Daten intakt. Durch besondere Anordnung und Algorithmen können die Paritäten auch so generiert werden, dass Fehler korrigiert werden können.
Hier ein 64 Bitblock mit einen Heming-Abstand von 2. Dieser Code kann einen Fehler korrigieren und zwei Fehler entdecken.



Rot: die Daten-Bits; Blau: die Paritäts-Bits. xxx In diesem Fall errechnen sich die Parität Bits durch das exklusive oder die Berechnung wird per Spalte und Zeile individuell gerechnet. Die daraus entstehenden Bits werden dann nochmal zusammen gerechnet und sollten das Gleiche liefern für Spalten und Zeilen. Dies wird dann in die untere rechte Ecke geschrieben.

4.2 Daten Verlust durch Ende der Unterstürzung der Formate

Eine weitere Gefahr für die Daten stellen proprietäre Dateiformate und Archive dar. Die Gefahr hier ist einfach, dass mit der Zeit keine Software mehr existiert, welche die Daten korrekt lesen kann. Zentralpunkt hier ist, das die inneren Funktionen der Archive und Dateien unbekannt sind und für einen externen Betrachter diese nicht lesbar sind. Wenn nun aus irgendwelchen Gründen der Hersteller keine Updates für die Produkte mehr anbietet und die archivierten Dateien für Jahre unbemerkt herumliegen, ist es dann in der Zukunft nicht mehr möglich diese zu lesen, was dafür sorgt das der Inhalt verschollen ist, auch wenn die Daten die Zeit schadlos überstanden haben. Hierbei ist anzumerken, dass alle Dateiformate theoretisch betroffen

sein können. Wobei es natürlich möglich ist, solange das Innere des Formates bekamt ist, neue Software zu schreiben, welche dann dazu verwendet werden kann die Daten zu retten. Aber besonders bei proprietären Dateiformaten ist dies nicht der Fall, was diese dafür besonders anfällig macht.

4.3 Clouddaten außerhalb der eigenen Infrastruktur

Der 11 September 2001 hatte grundlegende Konsequenten für die Welt, auch in der digitalen Welt schlägt sich das wieder. So ist beispielsweise der von der US Amerikanischen Regierung beschlossene "Patriot Act" ein Beispiel für die weitreichende Überwachungslegislation in der Welt. Der Patriot Act steht hier vertretend für alle anderen Gesetzte, die weltweit die Überwachung des digitalen Datenverkehres legitimieren[6]. Diese sind problematisch, da sie Zugriff auf Daten einforderbar machen und viele Cloud-Anbieter diese dann heraus geben müssen. Und dann beispielsweise sensible Daten in den Händen Unbekannter sind. Dies ist nicht nur ein Problem der Privatsphäre des gemeinen Bürgers, sondern auch ein Problem das Firmen ihre Geheimnisse, welche sie in der Cloud sichern wollen, de facto offen legen, sofern nicht vor der Speicherung verschlüsseln. Dies sorgt dafür, dass die relative Sicherheit der Cloud wesentlich gestört wird und auch dieses Speichermedium einen nicht zu vernachlässigen Nachteil mit sich trägt.

5 Zukünftige Technologien

Hier noch 2 Technologien die noch nicht ganz etabliert sind. Diese sind viel versprechend, auch wen die Lebensdauer außerhalb der Labore noch nicht bewiesen wurden.

5.1 Millennium Disc[9]

Die Millennium Disc ist eine Modifikation zu dem DVD- und Blue-ray-Standard und soll einen wesentlichen längeren Lebenszyklus haben. Da der Standard abgewandelt ist, werden spezielle Laufwerke benötigt zum Speichern und Beschreiben. Die Speichergröße ist allerdings identisch mit dem jeweiligen Standard. Die angegebene Lebensdauer soll nach Hersteller Angabe zwischen 100 und 1000 Jahren sein.

5.2 Glas Master Disk[10]

eine andere alternative ist die Glas Master disk. Diese benötigt keine speziellen Abspielgeräte und kann so mit blue-ray oder DVD Spielern wiedergeben werden. und wird in Einzel Produktionen angeboten, auch wen der preis von 160 Euro zuzüglich steuern sehr hoch ist. Die Disk besteht im Gegensatz zur normalen DVD nicht aus einem Kunststoff sondern aus Glas. Dies soll bei der Langlebigkeit helfen die ähnlich der Millenniumdisk über 100 Jahre bis zu 1000 Jahre sein soll.

6 Fazit

Im groben und ganzen lässt sich das folgende fest stellen. Um die Lieblingsurlaubsfotos zu speichern kann Otto-Normal-Verbraucher auf die gute alte DVD oder ähnliches zurückgreifen. Sollte es aber in den Businessbereich gehen, wird es komplizierter. Dort sollte dann eine entsprechende professionelle Lösung erarbeite werden. Dies ist dann eine große und eventuell komplexe Lösung, je nach Anforderungen. Eine simple Lösung existiert aber auch nicht.

7 Quellen

[1], Computing — Cern, http://home.web.cern.ch/about/computing 24.09.15

[2],Western Digiral, http://www.wdc.com/wdproducts/library/SpecSheet/ENG/2879-771386.pdf http://www.wdc.com/wdproducts/library/SpecSheet/ENG/2879-771386.pdf 24.09.15

[3], Seagate, http://www.seagate.com/www-content/product-content/hdd-fam/seagate-archive-hdd/en-us/docs/archive-hdd-ds1834-4-1412us.pdf http://www.seagate.com/www-content/product-content/hdd-fam/seagate-archive-hdd/en-us/docs/archive-hdd-ds1834-4-1412us.pdf 24.09.15

[4],Samsung Electronics,Samsung Electronics http://www.samsung.com/global/business/semiconductor/minisite/SSD/downloads/document/Samsung_SSD_850_PRO_Data_Sheet_rev_2_0.pdf 24.09.15

[5] http://www.sandisk.com/assets/docs/lightning-eco-genII-sas-ssd-datasheet.pdf

[6],http://www.gpo.gov/fdsys/pkg/BILLS-107hr3162enr/pdf/BILLS-107hr3162enr.pdf, http://www.gpo.gov/fdsys/pkg/BILLS-107hr3162enr/pdf/BILLS-107hr3162enr.pdf 24.09.15

[7], $LTO-6_LTO-5StandaloneDriveDatasheet(German)[DS00457G].pdf$, https://iq.quantum.com/exLink.asp?11903793OK68L56I59408379 24.09.15

[8], Ein Ersatzteillager für bedrohte Datenträger - Im Laufe der Jahre gehen Festplatten kaputt, http://www.3sat.de/page/?source=/nano/bstuecke/119799/index.html 24.09.15

[9] SYYLEX AG,GlassMasterDisc - Syylex, http://www.syylex.com/glassmasterdisc.html 24.09.15

Herstellerseite der M-Disc [10] https://www.mdisc.com/ 01.08.15

[11] Eduardo Pinheiro, Wolf-Dietrich Weber and Luiz Andre Barroso, Google Inc.,$http://static.googleusercontent.com/media/research.google.com/de//archive/disk_failur$ http://static.googleusercontent.com/media/research.google.com/de//archive/disk_failures.pdf 24.09.15