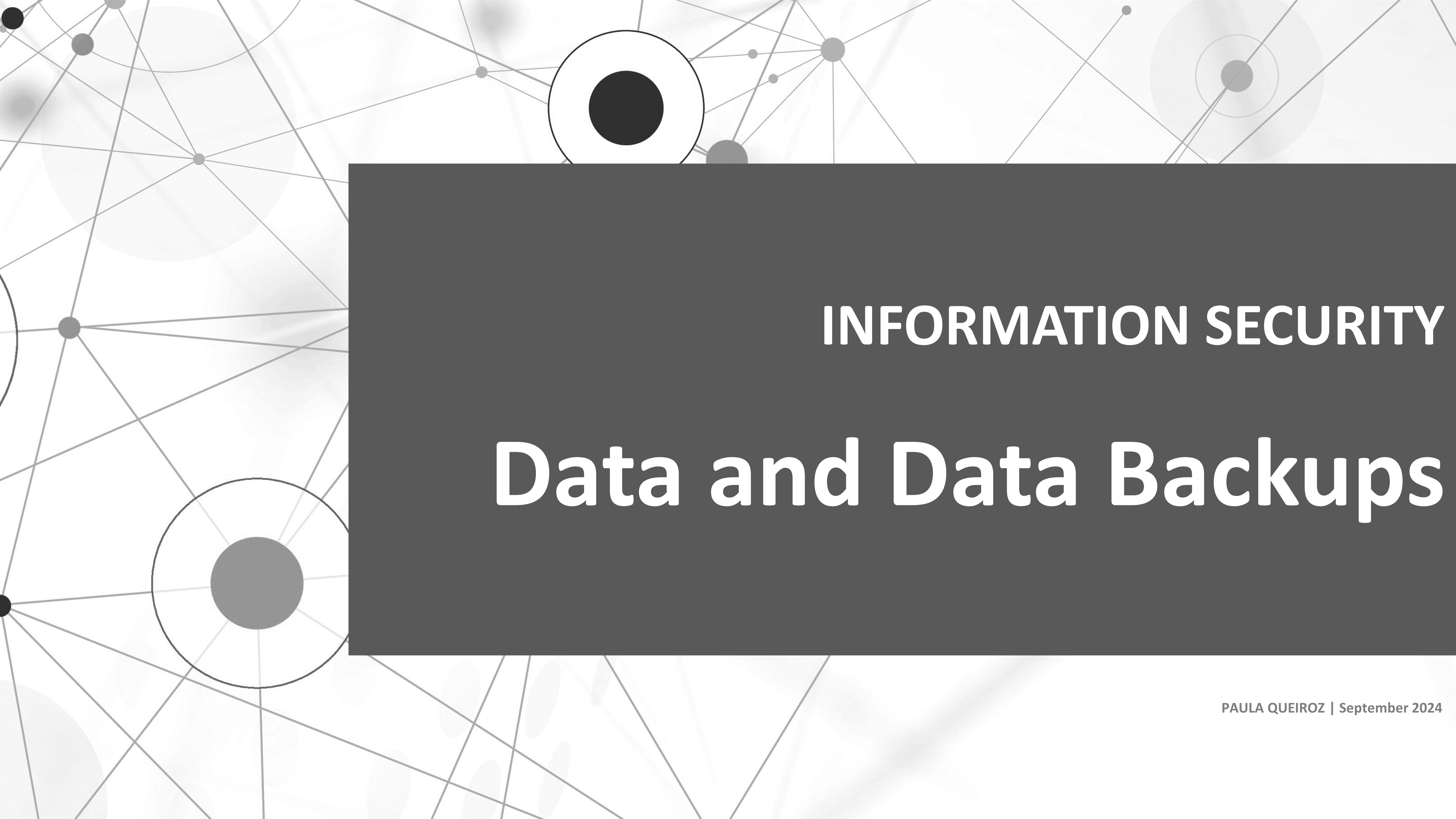


INFORMATION SECURITY

Data and Data Backups

PAULA QUEIROZ | September 2024

INTRODUCTION

Due to numerous past strange work experiences related to data backups, I believe this post is worth writing because, as plain, easy and trivial a subject as “data backups” may appear to be to some sysadmins (hopefully not the majority thereof) at first glance, there is a lot more to them than what’s perceivable from the surface.

The periodically backing-up of data to disk, tape or the cloud and the infrastructure involved in doing so is just the visible part of a complex iceberg of which the submerged (invisible) structure needs to be understood. This invisible structure is the intertwining of, amongst others, the data lifecycle, the laws and regulations we are subject to when handling data, and a lot of knowledge about information security – as all in the IT world.

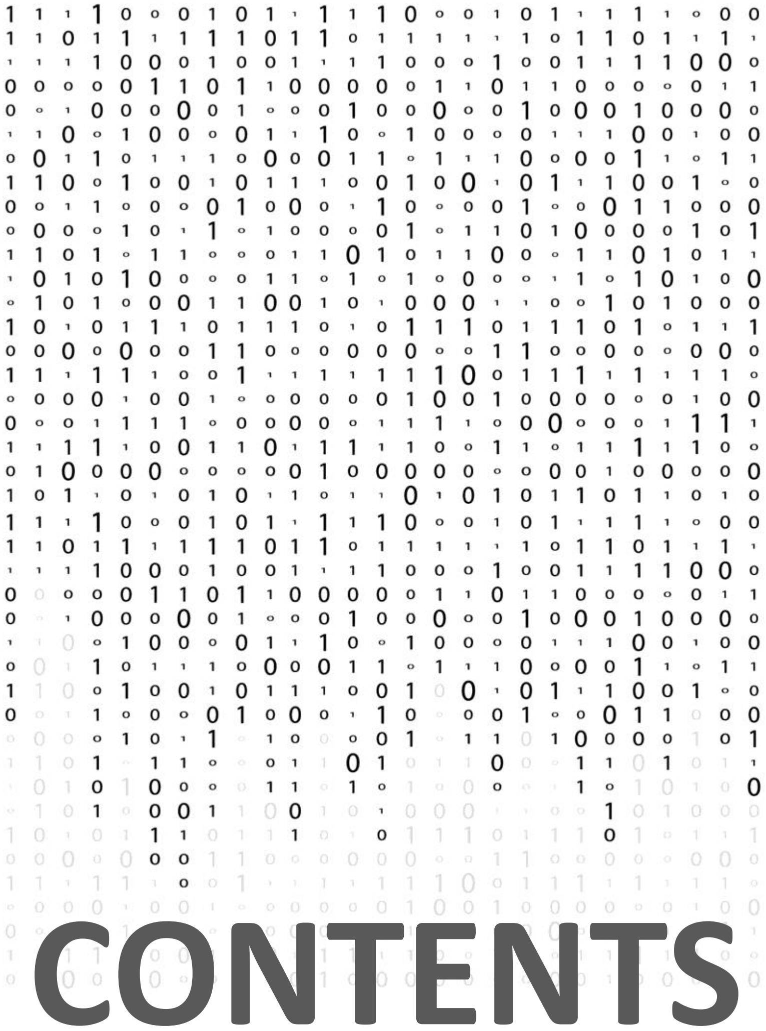
Because, in the IT world, when we do things, if we want to do them properly, we don’t do it just “because we want to”, this “because we want to” which is still very much a part of experienced systems administrators’ mindsets must be replaced with knowledge of why, in formal terms, according to industry best-practices and always guided by information security principles, we do what we do.

So, prior to talking about data backups and certainly long before talking about disaster recovery and business continuity planning (we will discuss the often misunderstood relationship between the two later on), we must take a moment to talk information in an enterprise context, and its lifecycle.

to ensure we are following a coherent, industry-recognised framework and not just making things up as we go along, we will be using the CISSP as our main source of information, best-practices and thought-framing.

GLOSSARY

DR:	Disaster Recovery
API:	Application Programming Interface
BIA:	Business Impact Analysis
RPO:	Recovery Point Objective
RTO:	Recovery Time Objective
TCO:	Total Cost of Ownership
CAPEX:	Capital Expenditure
OPEX:	Operational Expenses
LTO:	Linear Tape-Open
WORM:	Write Once, Read Many



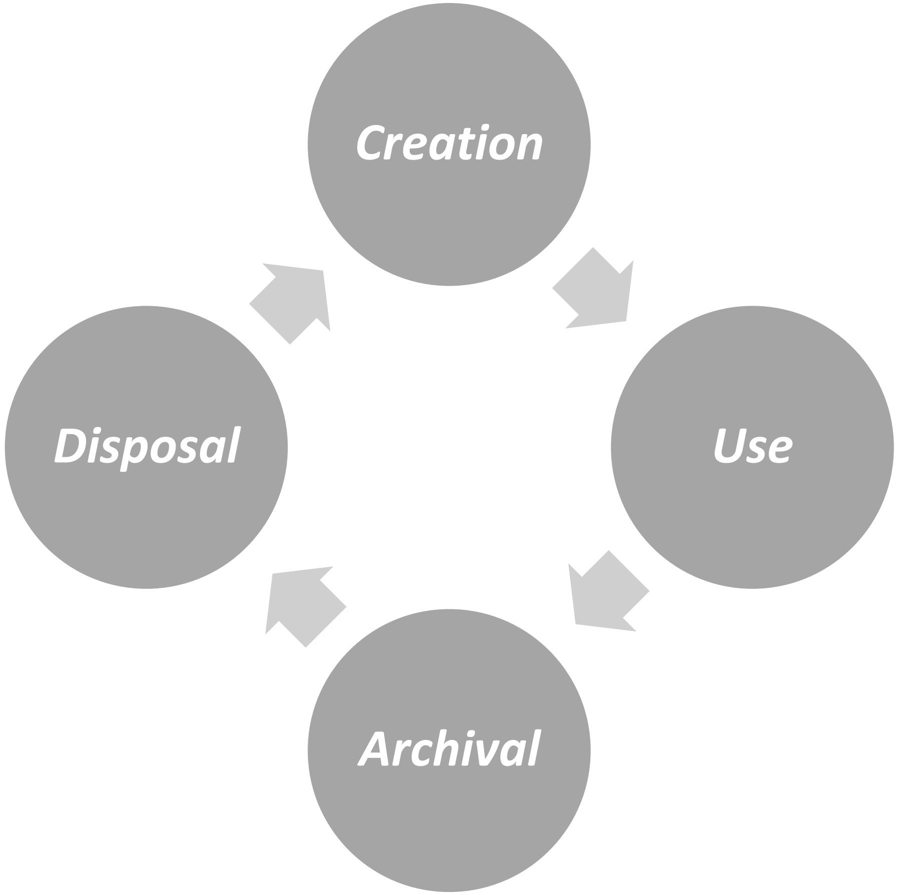
Contents:

- The information lifecycle
- Data Archival and backups
- Data retention, controls and frameworks
- Data backup strategies: 3-2-1, “the golden rule”
- Storage formats
- Data structures
- Snapshots and application-aware backups
- Backup strategies
- Cloud vs tape storage
- Focus on cryptovirology
- An end-to-end backup solution

PART 1

The Information Lifecycle, Archival and Backups

Information Lifecycle



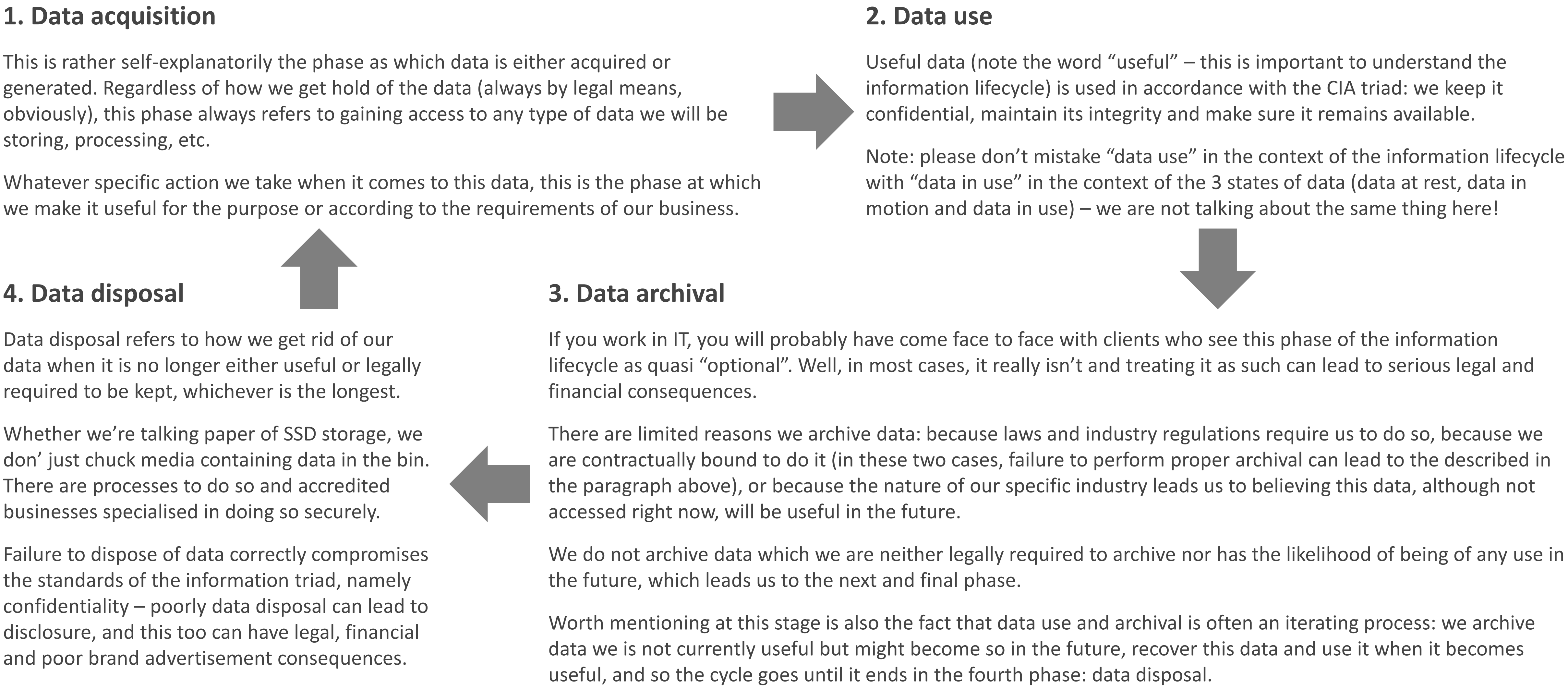
The handling of information follows a rather obvious and self-explanatory lifecycle of which, albeit necessary and self explanatory, some phases are often neglected even in corporate contexts.

There is room for tailoring how this lifecycle is managed in operational terms – from local legislation to industry specificities, many things can have a smaller or larger impact on how this is done in practical terms.

Here, for the sake of simplification, we will focus on the mandatory presence of these four phases:

- **Data creation (or acquisition)**
- **Data use**
- **Data archival**
- **Data disposal**

Information Lifecycle



1. Data acquisition

This is rather self-explanatorily the phase as which data is either acquired or generated. Regardless of how we get hold of the data (always by legal means, obviously), this phase always refers to gaining access to any type of data we will be storing, processing, etc.

Whatever specific action we take when it comes to this data, this is the phase at which we make it useful for the purpose or according to the requirements of our business.

2. Data use

Useful data (note the word “useful” – this is important to understand the information lifecycle) is used in accordance with the CIA triad: we keep it confidential, maintain its integrity and make sure it remains available.

Note: please don’t mistake “data use” in the context of the information lifecycle with “data in use” in the context of the 3 states of data (data at rest, data in motion and data in use) – we are not talking about the same thing here!

3. Data archival

If you work in IT, you will probably have come face to face with clients who see this phase of the information lifecycle as quasi “optional”. Well, in most cases, it really isn’t and treating it as such can lead to serious legal and financial consequences.

There are limited reasons we archive data: because laws and industry regulations require us to do so, because we are contractually bound to do it (in these two cases, failure to perform proper archival can lead to the described in the paragraph above), or because the nature of our specific industry leads us to believing this data, although not accessed right now, will be useful in the future.

We do not archive data which we are neither legally required to archive nor has the likelihood of being of any use in the future, which leads us to the next and final phase.

Worth mentioning at this stage is also the fact that data use and archival is often an iterating process: we archive data we is not currently useful but might become so in the future, recover this data and use it when it becomes useful, and so the cycle goes until it ends in the fourth phase: data disposal.

4. Data disposal

Data disposal refers to how we get rid of our data when it is no longer either useful or legally required to be kept, whichever is the longest.

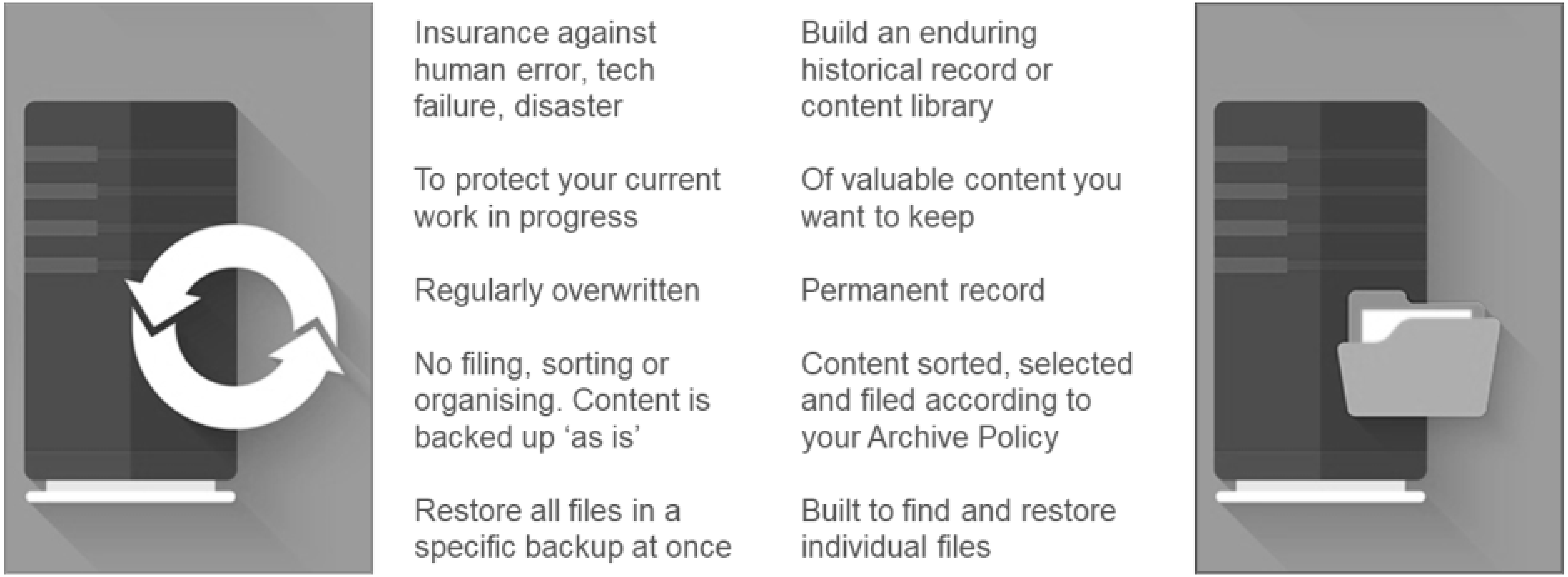
Whether we’re talking paper of SSD storage, we don’ just chuck media containing data in the bin. There are processes to do so and accredited businesses specialised in doing so securely.

Failure to dispose of data correctly compromises the standards of the information triad, namely confidentiality – poorly data disposal can lead to disclosure, and this too can have legal, financial and poor brand advertisement consequences.

DATA ARCHIVAL VS BACKUPS

There is often some confusion between these two so, before going any further, let’s clarify things:

No, data archival and data backup are not the same. We may achieve both using the same technology and store them in the same media. We may even perform them simultaneously, but there are big differences between the two, and we must know what these are.



Insurance against human error, tech failure, disaster	Build an enduring historical record or content library
To protect your current work in progress	Of valuable content you want to keep
Regularly overwritten	Permanent record
No filing, sorting or organising. Content is backed up ‘as is’	Content sorted, selected and filed according to your Archive Policy
Restore all files in a specific backup at once	Built to find and restore individual files

DATA BACKUPS

We will go deeper into this subject later on in this post but, just to get the differences right prior to going any further, we need to understand that we perform data backups to mainly respond to the antithesis of 2 of the 3 elements of the infosec triad: destruction (vs availability) and alteration (vs integrity).

Going back to the information lifecycle, data backups are within the scope of the 2nd phase thereof: data use.

We perform backups of data we are actively using in order to quickly and swiftly recover it in the response to alteration and/or destruction, which can come in many shapes and forms (intentional administrator action, hardware failure, corruption, malware, etc etc etc).

The point is that the data we backup is the data we need to use in the present, not the data we don’t need right now but may need at some point in the future, nor the data we don’t actually use but need to keep for legal and compliance reasons – for that we use archival.

NOTE ABOUT BACKUPS VS IT DRP

Backups are not to be mistaken for “disaster recovery” either. Data backups will likely be a part of the DRP but they are just a small part thereof, so “backups” and “DR” are two terms which must never be used interchangeably!

However, given the importance of backups for successful DR, here are some interesting statistics provided by Gartner and related to both:

A. Only 6% of companies affected by a major outage which didn’t have a proper DRP in place managed to survive for longer than 2 years after the outage.

B. The most common causes of data loss are:

- Hardware/system failure (31%)
- Viruses, malware and ransomware (29%)
- Human error (29%)

C. Globally, the average downtime cost for an enterprise is $5,600 per minute.

DATA ARCHIVAL

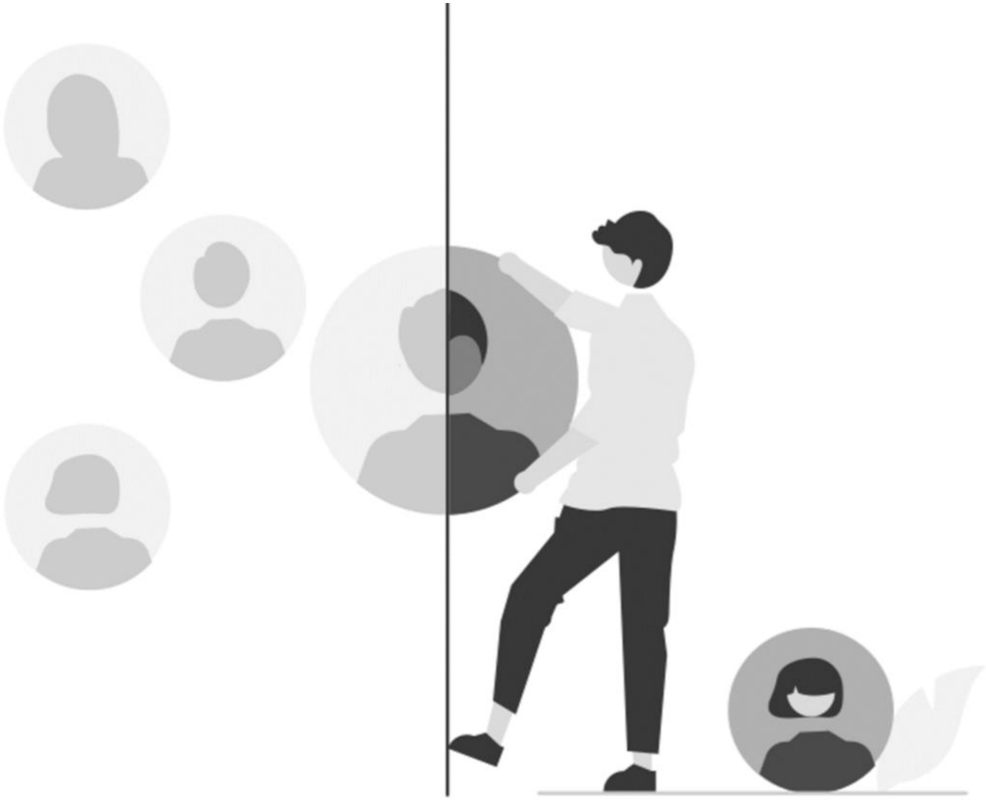
This shouldn’t need stating again given what’s already been stated above but, again, we archive data for two main reasons:

- **because we don’t need to use it now but have valid reasons to believe we may need to do so in the future, or**
- **because of laws, regulations or contractual stipulations.**

The differences between data archival and data backup, depending on the nature of the business and data volumes, will likely justify in financial terms the use of different media to host the data in each scenario.

This will be discussed later on…

DATA ARCHIVAL



This shouldn’t need stating again given what’s already been stated above but, again, we archive data for two main reasons:

- **because we don’t need to use it now but have valid reasons to believe we may need to do so in the future, or**
- **because of laws, regulations or contractual stipulations.**

The differences between data archival and data backup, depending on the nature of the business and data volumes, will likely justify in financial terms the use of different media to host the data in each scenario.

This will be discussed later on…

PART 2

Data Retention, Controls and Frameworks

DATA RETENTION

When we talk about data archival, we are implicitly talking about data retention which, again, takes us to the legal and compliance realm where “free will” will seldom apply.

We must do what we are legally obliged to do (for instance, in the health industry, when we host patient records, we must keep them ad eternum, and that’s the end of it).

If we get too creative and don’t, we must be ready to face the consequences.

As stated before, do remember that data should never be kept beyond the period of usefulness or legally required, whichever is greater. The standards and regulations stipulated by PCI-DSS, HIPAA, GDPR, etc, are meaningful in this context.

SECURITY CONTROLS AND FRAMEWORKS

As we all know, in security, the "one size fits all" does not exist. Different contexts will be subject to different controls, security is tailored to the context in question and, from industry to geography, many factors influence what's applicable to us in our specific circumstances.

There are, however, industry-recognised control patterns and which are applicable to all, providing a comprehensive and methodical way to organise ourselves in terms of how we handle security, data backups and data archival (and retention) included.

The following concepts describing a simple methodology for handling not only data but also systems are industry best-practices that help streamlining the process of controlling security.

SECURITY CONTROLS AND FRAMEWORKS

1 SCOPING

1. Scoping refers to analysing and isolating the part of a certain standard are applicable to us (what’s in scope).

Depending on who we are and what we do, not everything contained in a standard will be within scope. For instance, if we host no patient records, nothing included in any standard which relates to handling patient records will be applicable.

2 TAILORING

2. Once scoping is done and we know what parts of a standard are applicable to us, we are allowed to customise those parts of the standard in question in order to match our specific corporate requirements.

The catch is: we can go above, but we can’t stay below. For instance, if a part of a certain standard which applies to us requires certain encryption, we can use stronger encryption than recommended if we think we should. What we cannot do is use encryption that is weaker than that required for the standard to be applied.

3 CERTIFICATION

3. During the certification process, we apply to our data and systems the security measures deemed required by data and system owners as well as those required by the laws and regulations we are subject to.

Certification must not be confused with accreditation, which is described next.

4 ACCREDITATION

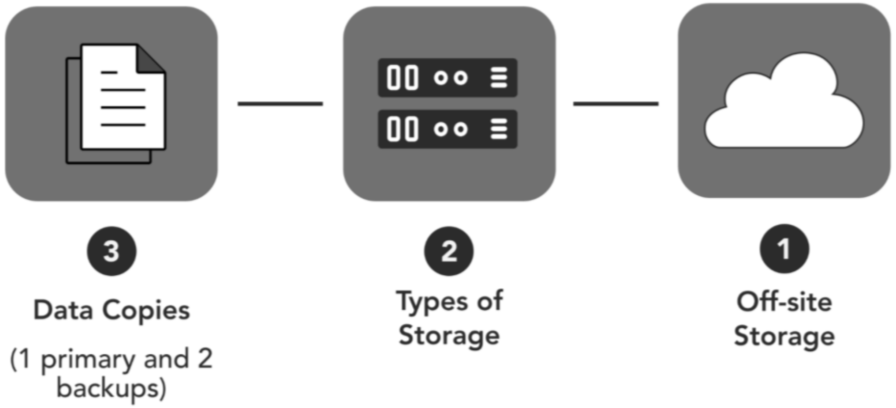
4. Successful accreditation means that the data/system owner(s) accepts our certification of implementation of the security measures and standards required and accepts the residual risk we informed them about.

If the certification is not accepted or the residual risk is not tolerated, we go back go back to work to produce a new certification to be submitted for new accreditation.

PART 3

3-2-1: The Golden Rule

3-2-1: The Golden Rule



There are many data backup strategies out there to choose from. In the context of this presentation, I will highlight the so-called "3-2-1", because it revolves around simple concepts, it's been labelled an "industry best-practice" but security experts, and it's a standard known to be applied by governments, including the American one.

As illustrated in the image above, the 3-2-1 strategy consists simply of:

3 copies of the data: the original and two copies.
2 different types of storage, least.
1 offsite copy, safely stored outsides the corporate premises.

Sound easy enough, right? However, this is just the start.

When we talk data backups, we are squarely in the realm of infosec, and in infosec nothing is as easy as it seems at the surface and many variables, some controllable, others not quite so, must always be taken into account to build the best solution in the specific context.

Remember another basic of information security – it's never a "one size fits all scenario" and the complexity theory always applies: you can't gain something without the inherent risk of losing something else.

Picking the 3-2-1 as the base for this thought experiment, let's go a bit deeper into its theoretical components.

3-2-1: The Golden Rule

Note on Data Copies

It’s not just about the number of copies, it’s also about their contents, their recoverability, frequency and retention period, just to name a few.

If you have a small business with a couple TB of data, you can most likely get away with just backing up all your data, replicating the backup to a 2nd storage media, and making sure one of the resulting 3 copies goes offsite – and there you go, here’s your 3-2-1.

However, if you’re a large enterprise who owns many PB of data, you probably won’t be able to just “back it all up and then replicate the backups” – more than costly in terms of money, this will take time and resources, which can be harder to manage than money.

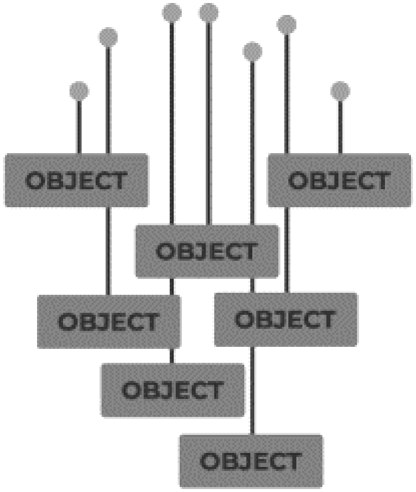
This is where our BIA becomes highly useful, as it will help us prioritise what needs backing up on a usefulness basis (remember, we don’t keep data which is neither useful nor legally required to be kept).

Besides not backing everything up (unless we’re really lazy and have either very little data or infinite financial and time resources resources), we also don’t back all information the same way. On this note, we’ll move on to a deeper dive on backup strategies.

PART 4

Storage and Data

STORAGE FORMATS



Object Storage

- Object storage (aka “object-based storage”) is a flat structure in which files are broken into discrete units called (surprise!) “objects” and spread out among hardware, in a single repository. Each repository corresponds to a storage volume working as a modular unit.

- An unique identifier will allow for specific objects to be found, and metadata, which can be extremely detailed, will describe the data in the object. To retrieve data, identifier and metadata are used by the storage OS. Load distribution boosts performance, rendering searches for specific data faster. Administrators also have easy ability to apply policies to add performance to complex searches.

- Object storage can be quite cost-efficient and is widely available in the cloud in a pay-as-you-go basis. It’s probably the first choice for static data, for its flat and agile nature allows for excellent scalability.

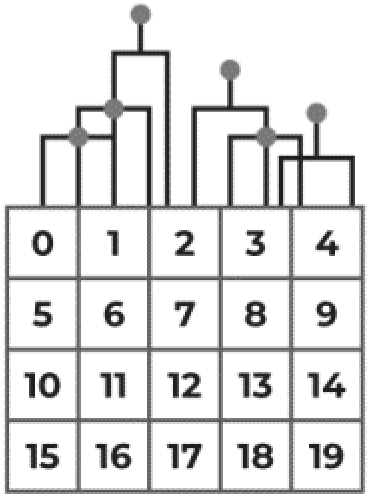
Object storage downsides

- Object storage requires clients to use an API (nothing too fancy, just HTTP). However, this may not be immediately compatible with some legacy systems, so the use of object storage may require some development, which can be costly and time consuming.

- Writing objects is, per se, a slow process, and each object is written all at once. Once written, an object cannot be modified which, in many cases, makes the process more complicated and time-consuming than using simple file storage.

- Concerning the inherent slowness of the object-writing process, it’s worth mentioning that this makes this type of storage less than ideal for use with traditional relational databases, which play pivotal roles in many enterprises’ technical ecosystems.

STORAGE FORMATS



File Storage

- File storage is precisely what it sounds like, and the format storage every standard computer user will be familiar with: storage shaped as files. These files are stored inside folders and organised in a hierarchical fashion. Files are accessed through paths, which can be long or short, depending on the location of the file within the hierarchy.

- This hierarchy is the reason why this format storage is also referred to as hierarchical storage.

- Metadata (which, in object storage, can be quite rich), is rather plain here. It works mainly like a catalog that matches files to their locations, so as to provide whatever is looking for the file with the right path where to find it.

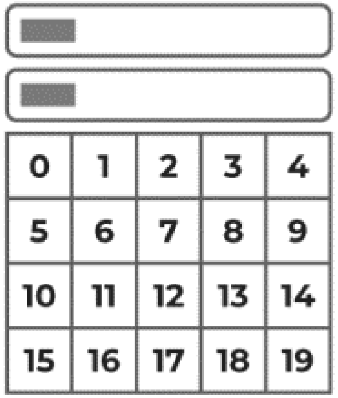
- Local and network-attached storage will traditionally use file storage. File storage is very versatile in terms of the immense file formats it can host, and it’s very user-friendly also, just about anyone can look for files in a hierarchical folder structure.

File storage downsides

- File storage has a few downsides too, one of its greatest benefits – the ability to store just about anything in just about any file format, being one of them: as the number of files and different formats increase, manageability decreases, with finding the right files becoming ever more difficult and resource-consuming.

- In terms of scalability, for the obvious reasons, scaling is done horizontally, more capacity is added through the addition of more storage, rather than vertical scaling by adding performance which, when handling vast volumes of data, can become rather costly.

STORAGE FORMATS



Block Storage

- Block storage breaks data into blocks and distributes them across different systems, wherever it’s most convenient for them to be placed. Each block has a unique identifier and blocks are transparently reassembled by the storage system’s software (normally a SAN, but the storage software can be the something other than that which comes with the box) whenever access to the data is needed.

- Unlike file storage, which relies on a single path to access each file, the distributed nature of block storage means that a file, during the process of reassembly, is retrieved through multiple paths corresponding to where the pieces of it are located. This tends to make file retrieval quicker by comparison to file storage.

- This format storage is widely used in enterprises due to its efficiency and reliability, as well as it’s manageability and compatibility with frequent, large transactions, like those normally associated with enterprise-grade databases.

- The more data is stored, the more visible the benefits of block storage becomes, at it is with heavy use that the benefits of efficient distribution of blocks across devices becomes ever more meaningful.

Block storage downsides

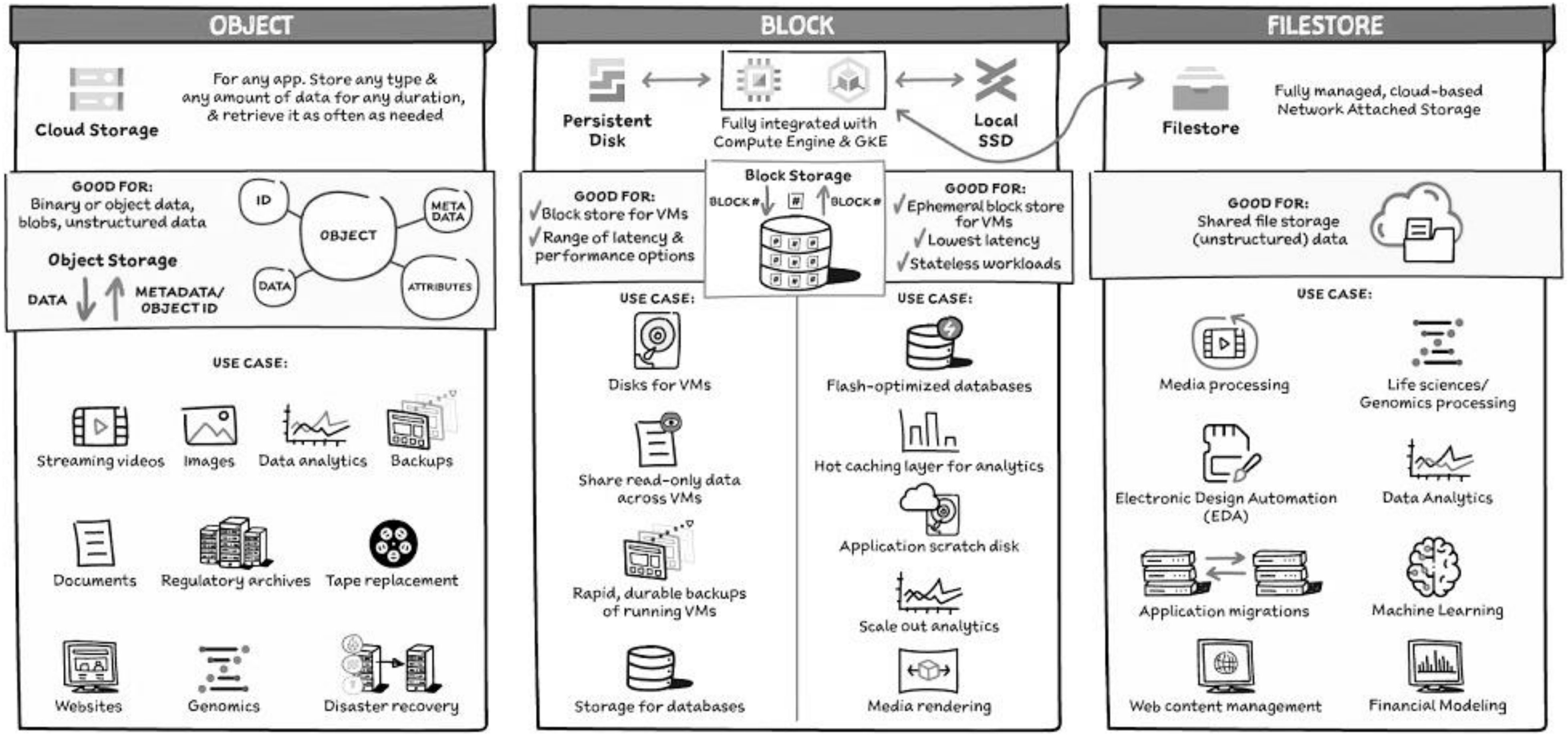
- Block storage can be frankly rather expensive. It too relies on horizontal scaling and, even when using block storage in the cloud, it doesn’t normally work on a pay-as-you-go basis: we need to pay for whatever block storage is allocated to us, whether we’re using it or not.

- Limited metadata is available, so anything beyond the basics we may wish to add needs to be handled at an application and/or database level, adding some complexity to the whole.

- In general, everything which has to do with permissions, versioning, filesystem choices, etc, will need to be handled by developers and/or sysadmins, so the need for significant human intervention is a factor in this storage format.

STORAGE FORMATS

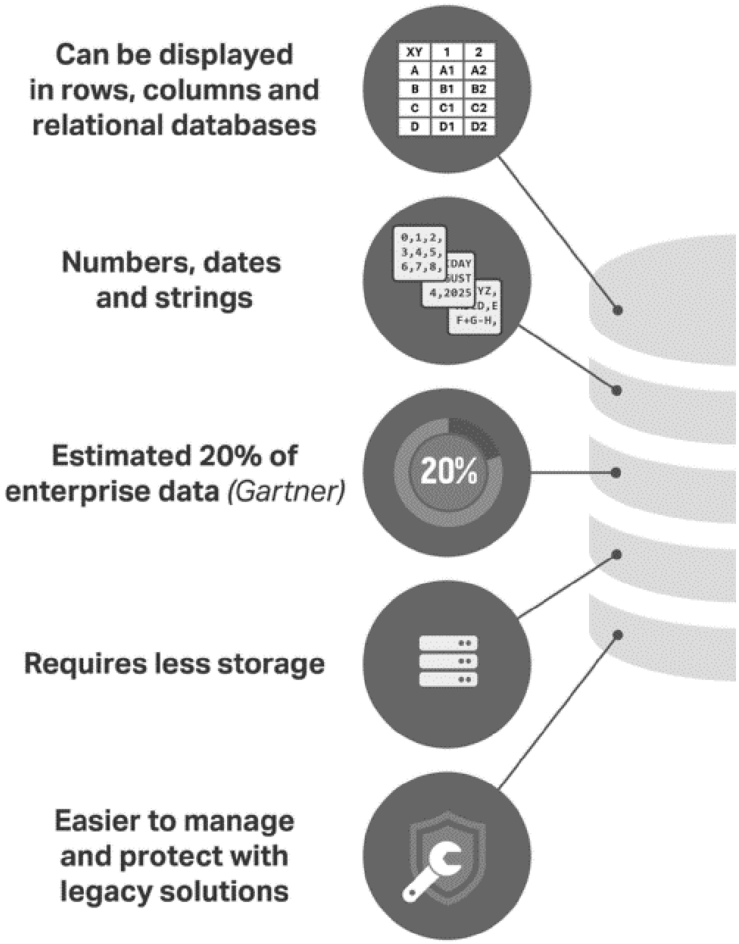
For a quick comparison of use cases, I have stolen a cute picture from Google:



PART 6

Data Structures

STRUCTURED DATA



Can be displayed in rows, columns and relational databases

Numbers, dates and strings

Estimated 20% of enterprise data *(Gartner)*

Requires less storage

Easier to manage and protect with legacy solutions

What structured data is

Structured (often called “quantitative”) data, is data formatted to fit a certain structure, which makes it easier to work with and faster to search.

A simple example of structured data is that generated by, for instance, relational databases.

Structured data includes quite a few benefits, such as:

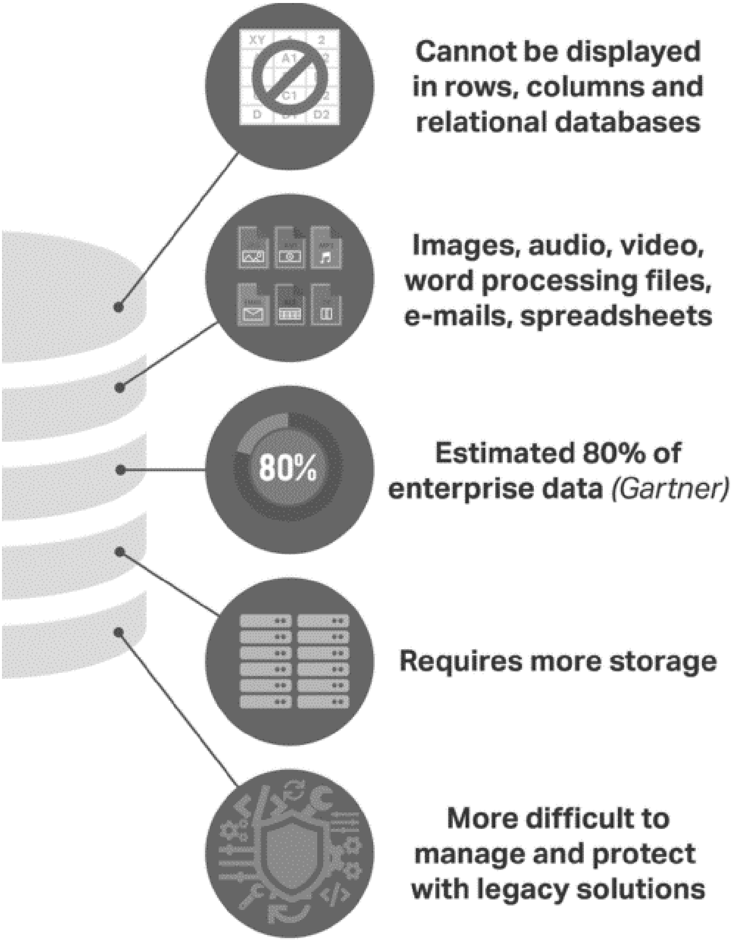
- Less processing is required when compared with unstructured data, with increased manageability.
- Querying is highly simplified through the used of algorithms built to crawl and use this data.
- It’s been around for a long time, so there are many well-known, easy to use tools available to manage it.

Structured data downsides

There are some downsides to structured data also. For instance:

- Its usability is limited due to of its pre-defined structure/format, as data with a predefined structure can only be used for that structure’s intended purpose.
- There are limited storage options. In fact, structured data is usually stored in data warehouses which are not very scalable or flexible.
- These are basically data storage systems with rigid schemas where any changes will require updating all the structured data to meet the new needs, potentially resulting in significant investment (costs can often be reduced though the use of cloud storage).

UNSTRUCTURED DATA



Cannot be displayed in rows, columns and relational databases

Images, audio, video, word processing files, e-mails, spreadsheets

Estimated 80% of enterprise data *(Gartner)*

Requires more storage

More difficult to manage and protect with legacy solutions

What unstructured data is

Unstructured data (often referred to as qualitative) is basically just data stored in whatever native format and not processed until it is used (aka “schema-on-read”).

There is a lot more unstructured than structured data, and it comes in many more formats, so some of its benefits are:

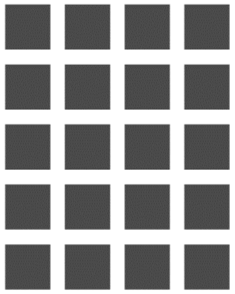
- A larger variety of native formats increases the number of possible use-cases.
- The lack of need to predefine data results in unstructured data being collected quickly and easily.
- Unstructured data is regularly stored in on-premises (or, more recently, cloud data lakes) which are easily manageable and highly scalable.
- Great volumes of unstructured data provide better insights and create opportunities to turn data into a competitive advantage.

Unstructured data downsides

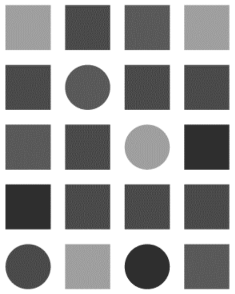
Unstructured data also comes with a couple of downsides. For instance:

- A greater number of formats also means greater difficulty in analysing and leveraging unstructured data.
- Large data volumes of undefined formats make data management challenging and requires for specialised tools.

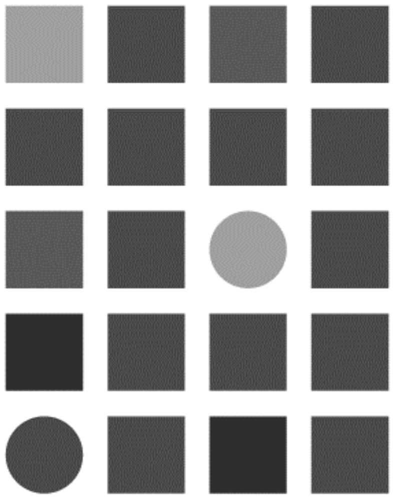
SEMISTRUCTURED DATA



Structured Data



Unstructured Data



Semi-structured Data

What semi-structured data is

Semi-structured data is a middle ground between structured and unstructured data. It doesn’t fit neatly into a traditional database schema but contains tags or markers to separate semantic elements and enforce hierarchies.

Semi-structured data is more adaptable than structured data, but easier to process than unstructured data and contains both elements of structured data (like tags in XML) and unstructured data (like the mixed content within an XML file).

Some characteristics:

- Flexible schema
- Human-readable
- Metadata
- Mix of data types
- Hierarchy
- Partial consistency
- Scalability

Some examples:

- JSON
- XML
- CSV
- YAML
- HTML
- Log files
- NoSQL databases

Semi-structured data limitations

Semi-structured data can still be challenging to integrate and analyze compared to structured data due to its varying formats and lack of a uniform structure.

PART 6

Snapshots and Application-aware Backups

SNAPSHOTS VS BACKUPS

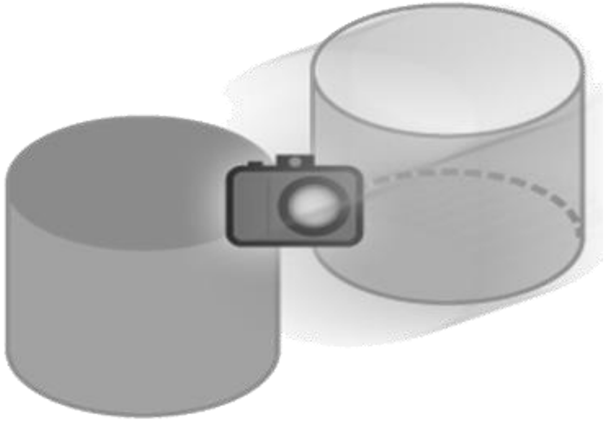
For starters, contrary to the strange beliefs of many, snapshots are not backups. They are different processes meant to address different needs.

Whilst, in some cases/when using specific products, snapshots may play a role in the backup process, the two are not one and the same.

So no, we cannot replace backups with snapshots.

So, to make this clear, let’s start by understanding snapshots.

WHAT A SNAPSHOT IS



- Briefly, a snapshot is like an instantaneous "picture" an entire system at a specific point in time. A snapshot-based rollback while affect the whole thing, not only specific parts of the system in question.
- A VM snapshot is the process of saving the data state of a VM with the possibility to revert to that point in time.
- A storage snapshot is a set of reference markers for data at a particular point in time which acts like a detailed table of contents, with PiT reversion capabilities.
- Multiple snapshots can be taken which will be organized in a parent-child type hierarchy.
- Snapshots are designed for short term storage as they will be hosted locally. When space runs out, new snapshots will typically overwrite older ones which makes them only useful when rolling the system in question back to a recent state.
- Snapshots are typically used in the context of development, testing OS and software updates and so on.
- Different products and vendors will provide significantly different snapshot technologies, with different benefits, requirements and limitations. To understand Veeam snapshots, we need to check the Veeam literature, and the same goes for VMware, AWS, etc etc, but the above is valid in all cases.

Snapshot backups

- In specific cases, administrators may choose to backup snapshots - this is what I was referring to when I mentioned that snapshots could play a role in the backup process.
- A backup of a snapshot is (surprise!) a backup. The snapshot on its own is not.

SNAPSHOT BACKUPS

- In specific cases, administrators may choose to backup snapshots - this is what I was referring to when I mentioned that snapshots could play a role in the backup process.
- A backup of a snapshot is (surprise!) a backup. The snapshot on its own is not.

APPLICATION-AWARE BACKUPS

- Application-aware backups will capture the exact state of an application (or several when, for instance, backing up an OS) at the precise moment of the backed-up.
- They rely on application-aware processing that allows for creating transactionally consistent backups.
- Data in memory and pending transactions will be included, which simplifies the process of restoring an application and starting using it immediately.
- This being said, application-aware backup tools will normally require quiescence to ensure transactional consistency, which is crucial for highly-transactional applications.
- Depending on what you’re backing up, quiescence may be achieved in different manners and these “manners” may have different names. Once again, please read the product(s) literature to ensure you are applying the configuration that fits your situation.

PART 7

Backup Strategies

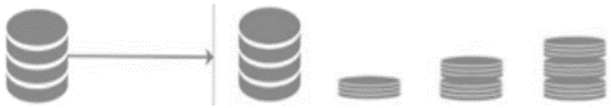
BACKUP STRATEGIES

I guess we're all familiar with this, so we won't dwell too much on it, just as a brief reminder of what full, differential and incremental backups are:

Full backups

- Full backups are rather self-explanatory: as the name indicates, these are backups of the entire system/dataset, independent of any previous backups.
- Out of the three (full, differential and incremental), these backups will obviously be the ones which take the longest. Regardless, it is imperative that they should be taken frequently.
- The frequency with which they're taken, and which systems/datasets are subject to full backups with which frequency will, in real life, obviously depend on several factors that go beyond the theory of the "ideal backup scenario" – volume of data to be backed up, storage available and performance of the backup infrastructure will all have an impact o what can be done in practical terms.
- Again, in a world of finite resources, we will rely in our BIA will give us some guidance in terms of prioritisation.
- Very importantly, since the aim of backing up is being able to restore, we must bear in mind that restoring a whole system from a full backup will be faster and require less backup media than restoring from a differential or incremental one. And even if restoring from one of the latter, the more recent the full backup the incremental of differential one relies upon, the easier and faster it will be to perform the full restore.

BACKUP STRATEGIES



Differential backups

- Differential backups will only backup any additions and/or alterations since the last full backup.
- As the differential backup distances itself in time from the last full backup the volume of data per backup will still remain significantly less than between full backups, but more than when incremental backups are performed, so they will be faster than a full backup, but likely slower than the average incremental one, as we’ll see next.
- In terms of restoring from a differential backup, we need the media for two backups: that of the differential backup we are restoring from, and that from the last full backup.

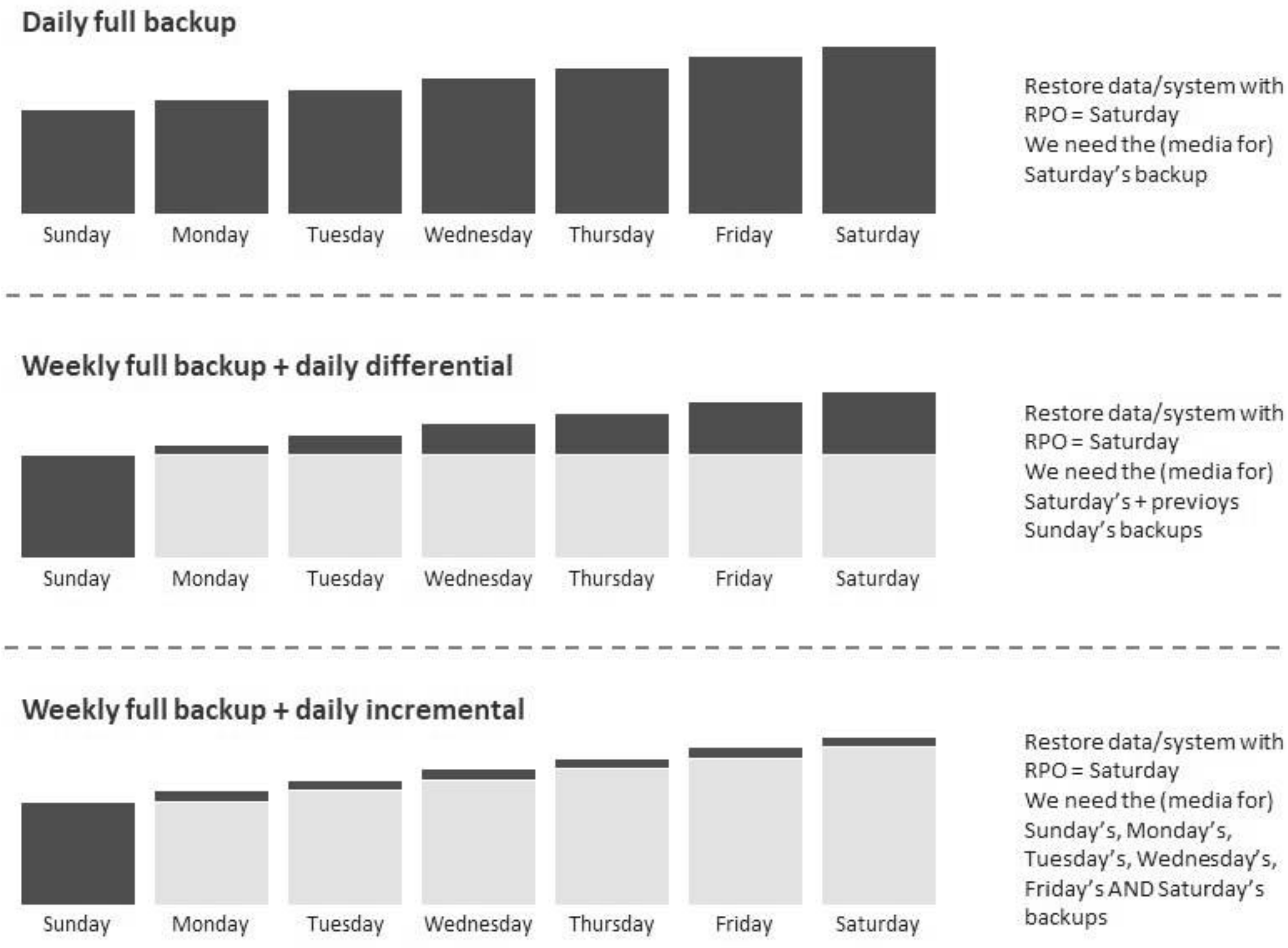
BACKUP STRATEGIES

Incremental backups

- Incremental backups are like “breadcrumb” backups, and I frankly don’t like them and avoid them at all costs whenever possible – restoring from them being the reason why.
- Take a full backup – every incremental backup will backup additions and/or alterations not since this full backup, but since the last incremental backup.
- This makes them fast to complete as there won’t normally be as much data to backup as if a differential backup strategy was in place instead.
- However, restoring from incremental backups can be a huge nightmare – we will need all the media for all incremental backups since the last full backup plus the media for that last full backup in order to perform a full restore.
- In a situation where we have proper defense in depth in place, which removable media stored offsite at a 3rd party’s location, the retrieval of the media plus the effort of restoring from heaven-knows how many different devices can be perfectly incompatible with our RTO.

BACKUP STRATEGIES

The following presents an easy way to visualise what full, differential and incremental backups are…



Daily full backup

Restore data/system with RPO = Saturday
We need the (media for) Saturday's backup

Weekly full backup + daily differential

Restore data/system with RPO = Saturday
We need the (media for) Saturday's + previoys Sunday's backups

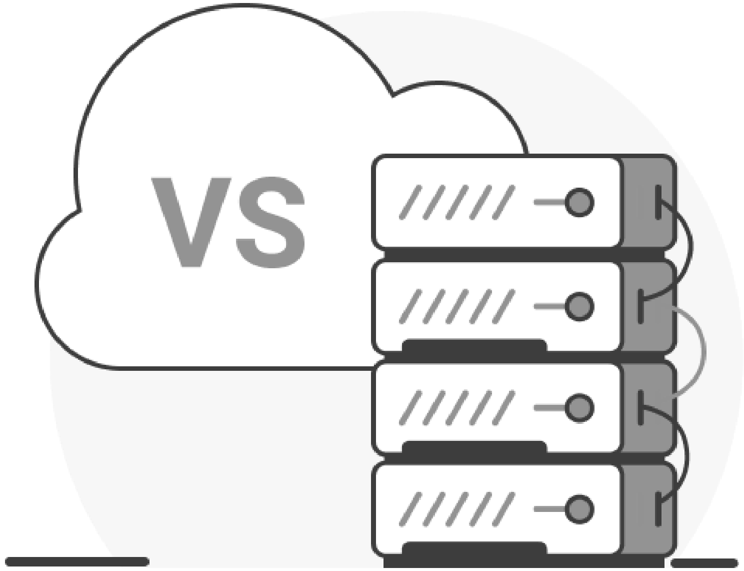
Weekly full backup + daily incremental

Restore data/system with RPO = Saturday
We need the (media for) Sunday's, Monday's, Tuesday's, Wednesday's, Friday's AND Saturday's backups

PART 8

Cloud and Tape Storage

CLOUD STORAGE VS TAPE STORAGE



Back to our 3-2-1 and the need to follow it as our “golden rule”, the “2” will normally refer to either tape or cloud storage. Meaning, to ensure at least 2 media and 1 offsite location, either backups will be stored in tapes and taken to a 3rd party secure location, or we will have these backups to the cloud, making it our de facto 3rd party site.

Both options have the good, the bad, and the ugly but, as usual, the choice will depend on the specific enterprise, on financial means, contractual constraints, and the enterprise’s technical trajectory in general.

In recent years, in different contracts, I have heard a lot of strange things presented as the “downsides” of both tape and storage, the most frequent (and frankly annoying) of which being:

“The cloud is not safe” and

“Tapes are outdated/not efficient”

Certainly, both can be true, but it’s not inherently so - it all depends on what goes in the heads of the people putting the solutions in place, and the hands touching those solutions - pretty much like everything else, really.

CLOUD STORAGE

Per se, cloud storage is pretty safe, okay? But remember one of the basis of information security: your security chain is only as strong as its weakest link, and the weakest link is always humans.

So yes, naturally, if you leave your S3 buckets open for everyone to look into the contents, AWS storage isn’t very secure but, again, this is clear-cut human error, AWS is not to blame for the bizarre things their clients do in terms of security.

And no, the American “Cloud Act” doesn’t mean the Americans can just go in and check all data in the cloud (which is a different subject which I might get into in a future post, as it bothers me how misinterpreted such clearly written law has been).

So yes, cloud storage is safe (again), it is highly scalable and highly available by nature, and it might as well be the best solution for our backups and to enforce our data retention requirements. S3, for instance, is compliant with PCI-DSS, HIPAA/HITECH, FedRAMP, GDPR, and FISMA.

We are not getting into DRaaS in this post but, we’ll certainly do so in the future, for it is an alternative to traditional DR certainly worth investigating - in any case, DRaaS relies heavily on cloud storage, the message being that cloud storage is perfectly fit for DR - if certain conditions are met.

I am a big admirer of cloud storage, namely S3, I truly am, however, there are things to bear in mind when we talk about it, especially when we’re talking about hosting vast amounts of different formats of data in the cloud, which can be needed asap at a moment’s notice:

Cloud storage can be costly. It all depends on the type and class of storage, the volume, how frequently it’s used, and how quickly you want it to be made available, but these are things which need calculating before making a rush decision to move storage - whether backups or something else - to the cloud.

CLOUD STORAGE

Like elsewhere, when it comes to cloud storage, not all data is the same, and different data beggars being treated in different manners. We need to know our data in order to be able to choose where and how to store it, which is why we've spent some time talking about data earlier in this post.

Cloud storage is not just about the storage, it's also about how you access it so, in corporate scenarios, the infrastructure linking you to the cloud (and to the data hosted there) is likely to have to be set up.

Cloud providers, especially AWS, have a lot to offer in terms of establishing a fast and reliable connection to the cloud - storage gateways, direct connect, etc - but, again, these come with a price and the need to perform studies and proceed to careful implementations (I've dealt with a specific terrible implementation of direct connect and, let me tell you, the thing was useless until it was fixed 6 months into my contract with the client).

For more information on AWS S3 (remember, it's object-based), please have a look at my previous post https://paulastechblog.blogspot.com/2019/12/aws-architecture-storage-s3-intro.html - you'll have a quick intro cross-region replication, security, storage gateways, virtual tape libraries, multipart upload APIs and other cool stuff.

In my point of view, when it comes to designing a comprehensive backup solution, cloud or no cloud storage boils down to the technical trajectory of the enterprise: if the enterprise is serious about embracing a cloud provider, then it makes sense to consider cloud storage for this purpose. Otherwise, "cloud" makes no sense from a complexity or investment perspective to be seen and invested in as simply yet another component of a backup solution.

TAPE STORAGE

We'll start this section quoting from Tech Radar, in a sentence which I think sums it all up:

Despite some outdated perceptions, tape technology is a key player in fulfilling all of these criteria when it comes to storing inactive data. Tape has proven to be the cost leader in backup and archive for years and that is not changing anytime soon. With significant capacity advancements with every new tape generation, the cost per gigabyte of tape storage continues to decrease, meaning even rapidly growing environments where data growth often outpaces budgets can harness the advantages of tape for backup and archive.

With the word "cloud" being in everyone's heads and mouths in recent years, my perception is that, whilst trying to stay "cut-edge" with what they consider to be such (ie "cloud storage"), many IT people have actually achieved the opposite, which is that their knowledge of what tape storage is and offers got stuck in the past, so their perception thereof is simply outdated, and outdated knowledge is no good when designing up-to-date solutions.

When choosing storage for whatever purpose it may be, TCO will always be in the forefront of our minds, and when it comes to TCO, tape storage has clear benefits as compared to cloud storage: firstly, it's rather predictable, secondly, it's become cheaper and cheaper over the last decades.

It is true that upfront investment (CAPEX) may scare off some people, especially those with an aversion to number-based projections and analysis, whilst smaller OPEX numbers for cloud services look a lot more appetising but, in the long term, for big data volumes with moderate-long retention periods, it is broadly agreed that TCO for cloud storage is far higher than for its tape-based counterpart.

TAPE STORAGE

LTO, as an adaptable and scalable industry standard, is perfectly aligned with ever-increasing demands of data protection. It assures secure and reliable long-term archival data storage at substantially lower cost than disk, flash or cloud.

The ninth generation LTO exceeds generation 8's capacity by 50%, at 18 TB native storage capacity, at 400 MB/s, with R/W 8 compatibility. In the future, we can expect an LTO-10 achieving up to 36 TB native; LTO-11 up to 72 TB native, and LTO-12 up to 144 TB. That's a lot of data inside a very small tape – even without compression!

Impressive as the future LTO generations may sound in terms of capacity, with the existing LTO-9 are already able to build exabyte libraries, with incredible R/W speeds thanks to multiple, highly fast drives (an LTO-9-based tape library with a maximum of 144 drives can transfer up to 207.4 TB (518 TB compressed) of data per hour).

But, really, tape storage has been around for so long and has such a good track record, that we shouldn't need to dig any further to sell it – if in doubt, "just google it".

The only reason why this section is being added to this post is to ensure readers do know that LTO, tape drives and libraries haven't become frozen in time since the cloud became mainstream – although our knowledge thereof might.

PART 9

Focus on Cryptovirology

FOCUS ON CRYPTOVIROLOGY

▪▪▪ Characteristics
Death God Level: 6
Type: Male
Likes: Apples and games
Dislikes: Boredom

▪▪▪ Death Chart
Intelligence ☠☠☠☠☠☠
Curiosity: ☠☠☠☠☠☠☠☠☠☠
Proactiveness: ☠☠☠☠☠☠☠☠☠
Sympathy: ☠☠☠☠☠
Murder count: ☠☠☠☠☠☠☠

A unique Deathgod who dreads boredom

In order to elevate himself from boredom, Ryuk dropped a Death Note into the human world. Judging from his actions, he can be said to be the culprit who indirectly created chaos in the human world. He has no special feelings for human beings or their actions. He watched Light commit many human crimes without lifting a finger to help or stop Light. But from his reactions, he was obviously pleased with Light's evil doings.

He has the calmness of a Deathgod but also displays certain human characteristics.

Backups on their own offer little to no protection against malware, ransomware being notorious in terms of the damage caused in recent years, and how much cleverer it’s become with time.

Needless to say, protection against malware should be achieved through defense in depth, there should be no single mechanism to protect our data against in, and many controls should be in place prior to the infection getting anywhere near our backups.

As usual, this defense in depth should start with user training, are users are the most likely targets to malware attack vectors.

Some ransomware examples which are actually pretty cool you may, out of curiosity, look at are WannaCry, SamSam, Petya and Ryuk (yes, Ryuk like the hilarious Death God in the Death Note manga/anime).

Although I suppose we all must have a pretty good understanding of what ransomware is, here is, however, a quick definition of ransomware, by Spectra:

> ***Ransomware is a type of malicious software from crypto-virology that breaches an individual’s or organization’s data and threatens to publish the victim’s digital content or perpetually block access to it unless a ransom is paid. Ransomware attacks cause downtime, data loss and potential intellectual property theft. One of the first cases of ransomware was reported in 1989, but its widespread use dates back to 2005, with newer versions being especially virulent in their propagation and depth of infection.***

FOCUS ON CRYPTOVIROLOGY

▪▪▪ Characteristics

Death God Level: 6
Type: Male
Likes: Apples and games
Dislikes: Boredom

▪▪▪ Death Chart

Intelligence	☠☠☠☠☠☠
Curiosity:	☠☠☠☠☠☠☠☠☠☠
Proactiveness:	☠☠☠☠☠☠☠☠☠
Sympathy:	☠☠☠☠☠
Murder count:	☠☠☠☠☠☠☠

A unique Deathgod who dreads boredom

In order to elevate himself from boredom, Ryuk dropped a Death Note into the human world. Judging from his actions, he can be said to be the culprit who indirectly created chaos in the human world. He has no special feelings for human beings or their actions. He watched Light commit many human crimes without lifting a finger to help or stop Light. But from his reactions, he was obviously pleased with Light's evil doings.

←↑ *He has the calmness of a Deathgod but also displays certain human characteristics.*

The typical steps in a ransomware attack are:

A) Infection:

After delivery to the system via email attachment, phishing email, infected application or other method, the ransomware installs itself on the endpoint and any network devices it can access.

B) Secure Key Exchange:

The ransomware contacts the command and control server operated by the cybercriminals behind the attack to generate the cryptographic keys to be used on the local system.

C) Encryption:

The ransomware starts encrypting any files it can find on local machines and the network. In some cases, it will look for and delete data backups.

D) Extortion:

With the encryption work done, the ransomware displays instructions for extortion and ransom payment, threatening destruction of data if payment is not made.

E) Unlocking:

Organizations can either pay the ransom and hope for the cybercriminals to decrypt the files, or they can attempt recovery by removing infected files and systems from the network and restoring data from backups.

However, it is worth noting that, as of a 2020 report, 42% of organizations who paid a ransom did not get their files decrypted.

RANSOMWARE AND BACKUPS

Whilst there was a time when we thought about ransomware as that type of malware that encrypts our data, ransomware has evolved and, today, it will tend to track down and eliminate backups.

There are several ways backups themselves can be protected against ransomware, all of which should be used simultaneously, for the obvious reasons:

1. IAM

- Identity and access management is one of the barriers that can help prevent malicious access to backups. We want specific, purpose-created user accounts to be able to manage our backups, not generic administrator accounts or accounts used for any other purposes but managing backups.
- Preferably, we want a dedicated authentication system to manage access to our backup infrastructure.
- The rationale being compartmentalisation: should an administrator account be breached, we want that account to have access to as little as possible, and we certainly don’t want that breached account to have access to our backups.

2. Airgapping

- Airgapping, the practise of having media containing backups offline is considered to be the golden standard in terms of assuring the immutability and safety of data backups.
- WORM technologies also offer immutability, which protects backups against malicious encryption, but airgapping, especially when combined with offsite secure media storage, expands this protection far beyond just malware, adding physical security against physical tampering, theft and natural and/or man-made disasters.

RANSOMWARE AND BACKUPS

3. Anti-malware backup protection

- The malware worming capabilities can make it quite insidious and hard to detect during the first stages of the infection.
- Fortunately, many vendors offering backup solutions (either just software or integrated storage + software) offer built-in malware detection technology, which should be thoroughly explored.

4. Full backup frequency and retention

- Restating the above, malware worming capabilities can make it quite insidious and hard to detect during the first stages of the infection.
- Even with malware detection technology in place, this won’t give us 100% assurance that we aren’t backing up infected data which hasn’t been flagged as such.
- As seen above, incremental and differential backups rely on a previous full backup for the restoration process. This means that, no matter how frequent the incremental/differential backup, if the full backup they rely upon includes any sort of infection, the information restored will be infected too.
- For this reason, it is important to perform frequent full backups of the most critical systems and data, which should be kept for long-ish periods of time (given the fact that malware isn’t normally noticeable straight away, it can take several months until we realise it’s there).
- These full backups which are kept for longer periods are those which should be airgapped, safely stored offsite.

RANSOMWARE AND BACKUPS

5. Frequent testing

- There is no way we can know for sure if we can perform a restore and, even if we perform the restore, if the restored data/system will be operational unless we test it.
- When a backup is performed, it is frequently unaware whether it's backing up encrypted data or not and, even if it knows the data is encrypted, it mostly cannot tell whether it's supposed to be so, or has been subjected to ransomware encryption, so it's important to test our backups regularly to make sure we are not backing up compromised information.

PART 10

An end-to-end backup solution

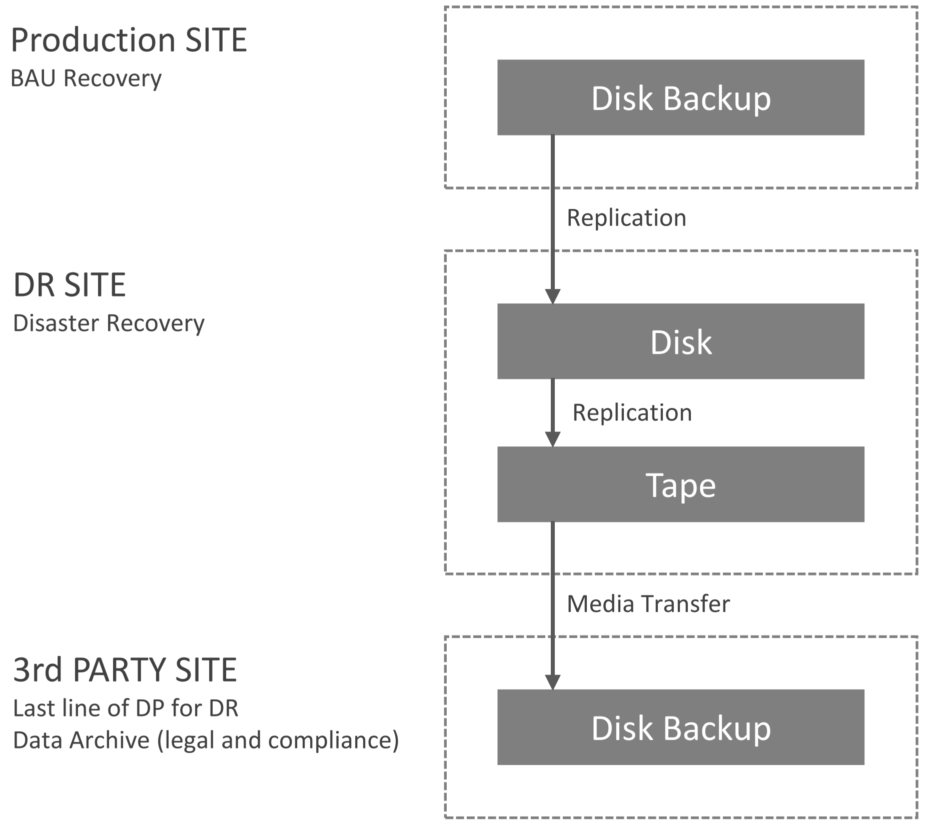
AN END-TO-END BACKUP SOLUTION

Designing a simple, yet robust backup solution

To provide a highly customisable example of a backup solution, we will take into account the following:

- Backups serve different purposes that range from being used for recovery in benign BAU situations to recovery of critical systems and services following major disasters, to data retention in compliance with laws and regulations.
- The “golden rule” (3-2-1) and infosec best-practises will be taken into account, focusing on defense in depths, as they should always be.
- Resources are finite, and the solution must cater for this finitude whilst ensuring all scenarios where data recovery may be required are covered.

AN END-TO-END BACKUP SOLUTION



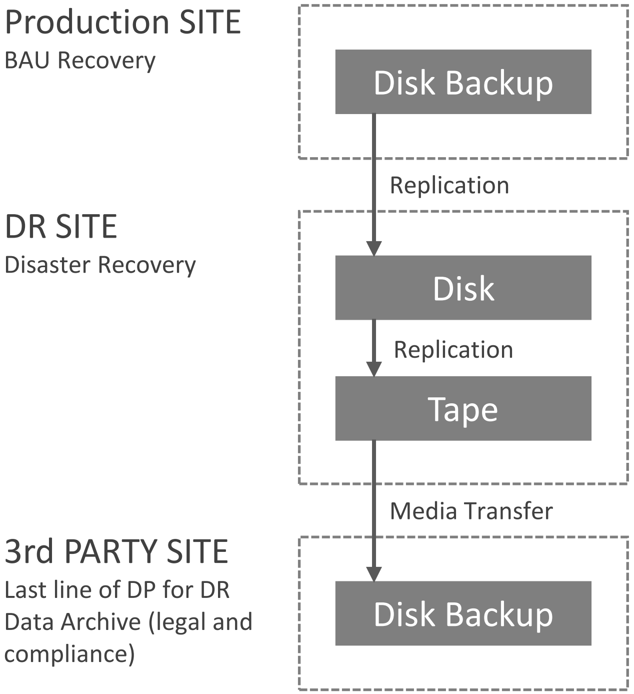
Solution Topology

The solution is spread across three sites: the production (BAU) site, the DR site (for DR, obviously) and a 3rd party site, mainly used for data retention purposes, but also as an additional layer of security including airgapping.

It is certainly technically possible to have no backups present in two datacentres and go straight from one datacenter to a 3rd party site, whether cloud storage of Iron Mountain-style facilities. This is, however, no advisable in terms of information security, as it breaks our 3-2-1 golden rule.

Since we put security above all, we won’t consider such option.

AN END-TO-END BACKUP SOLUTION



1. The production site: BAU recovery

The production site backups have the main goal of catering for quick recovery of data accidentally modified or deleted.

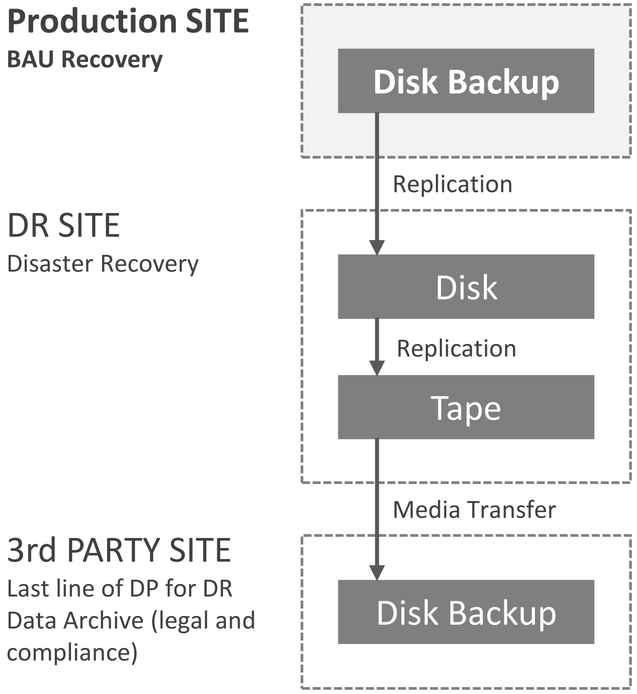
This is a BAU solution: these backups are not used for DR or data retention purposes, restores made from these backups will tackle every day, non-major incident-related needs for data recovery.

Given the purpose for which restores from this site are done, there is no requirement to keep these backups in this location for long – some 2 weeks should suffice so storage can be recycled.

This is where the backup chain starts and backups to be hosted in other sites and used for different purposes will have their origin here.

Given my deep dislike for incremental backups but understanding that daily full backups are just not quite possible, we will assume a weekly full backup of P0 and P1 systems at least, with daily differential backups. The periodicity of P2 and P3 systems' backups will depend largely on the capacity available.

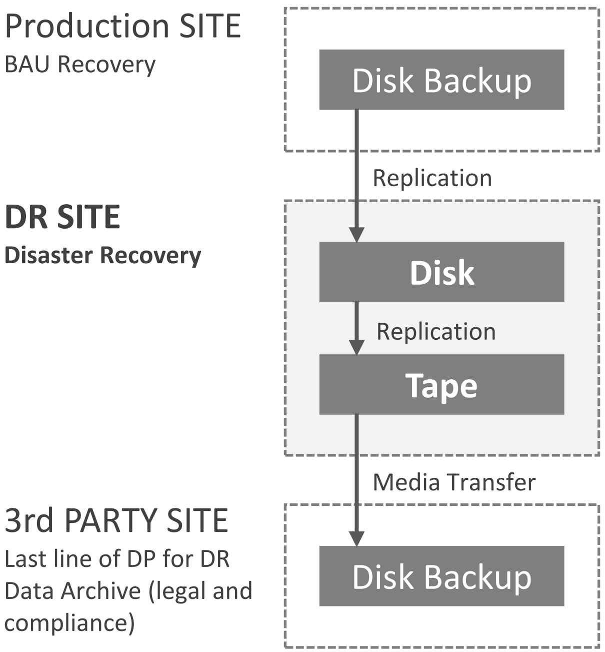
AN END-TO-END BACKUP SOLUTION



1. The production site: BAU recovery

- The production site backups have the main goal of catering for quick recovery of data accidentally modified or deleted.
- This is a BAU solution: these backups are not used for DR or data retention purposes, restores made from these backups will tackle every day, non-major incident-related needs for data recovery.
- Given the purpose for which restores from this site are done, there is no requirement to keep these backups in this location for long – some 2 weeks should suffice so storage can be recycled.
- This is where the backup chain starts and backups to be hosted in other sites and used for different purposes will have their origin here.
- Given my deep dislike for incremental backups but understanding that daily full backups are just not quite possible, we will assume a weekly full backup of P0 and P1 systems at least, with daily differential backups. The periodicity of P2 and P3 systems’ backups will depend largely on the capacity available.

AN END-TO-END BACKUP SOLUTION

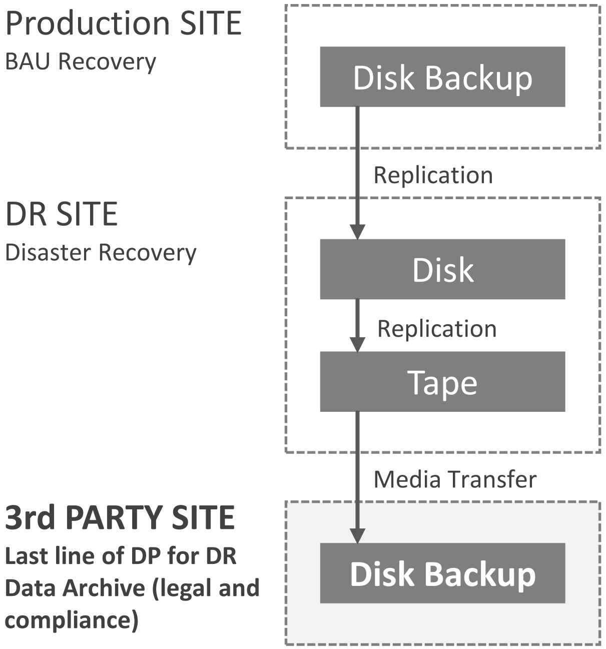


2. The DR site: disaster recovery

- Backups in the DR site are replicated from the production site and used for disaster recovery, whether minor or major, should the circumstances allow.
- The length of time the backups remain in the DR site's disk storage will be greater than that in the production site.
- Retention period policies will, therefore, differ between the disk backups in the production and DR sites.
- In previous situations where I manage to reengineer or fully commission end-to-end backup environments, I would keep differential daily backups on the DR site for a period of between 2 and 3 months (in the odd event that the forgetful user suddenly remembers they deleted some important file a few weeks before), at which point they would be erased, and weekly backups for up to a year.
- These backups are, however, not intended for data retention for legal and compliance reasons, so the time they remain onsite will be based on RPO definitions.

- Again, the main purpose why we keep backups in this site is the eventual need for to put our DRP into action, which will more than likely rely heavily on our ability to perform great volumes of data and system restores.
- For this reason, we keep the backups we may need to rely on in our fast disk-based storage solution, to ensure recovery is as quick as possible.
- We must, however, be prepared for the event of a loss or breaching of both our primary and DR sites, in which case our backup plan will be securing data offsite.
- So as to not add additional workloads and resource consumption to the production site's infrastructure, the 3rd replica of our backups, the one to be stored offsite, will be made in the DR site.

AN END-TO-END BACKUP SOLUTION



3. Third party site

- The 3rd party site will serve two purposes: it will both be our last layer of defense in terms of data protection, and it will be where data retention according to laws and regulations will be performed.
- In terms of data recovery following a major incident, this is where we literally “prepare for the worst whilst hoping for the best”. If the day comes that we have lost both our production and DR sites, we will be in a world of troubles, to which we don’t want to add “having no secure, airgapped backups in another location to restore our most critical information from”.
- In this design, if you’ve maximised the image above, you will have seen that we’re storing data in tapes. This could as well be cloud storage, in which case a couple of components such as links between the datacentre(s) and the cloud provider would be missing, but the concept remains the same – an unalterable copy of the data securely stored off-premises (remember: airgapping is not available in the cloud, but WORM is, which might be considered “good enough”).
- The main reason why I am basing the second copy of the data on tapes, given that this is a generic scenario where to specific corporate requirements are being considered is simply due to TCO.

AN END-TO-END BACKUP SOLUTION

Other considerations to highlight

There are many considerations to be had in mind when creating a comprehensive backup solution such as the high-level topology above.

Again, all our security practices and mechanisms will be customised to match the needs of our specific organisation, so it’s not possible to list all those considerations exhaustively.

I will, however, go through two crucial ones, which will hopefully help trying up together everything that’s been mentioned earlier in this post prior to talking about designing a backup infrastructure.

AN END-TO-END BACKUP SOLUTION

The principle of simplicity

Yes, I'm mentioning this again - because it's a very important principle in information security which, over and over again, I see being ignored.

A system, solution, platform, etc needs to be simple in order to be manageable. The more complex a system becomes, the less manageable it is and, most importantly, the greater the surface for errors, accidents and attacks becomes (as a side-note, I advise everyone to read about the complexity theory - it's as entertaining as it is eye-opening).

In this case, simplicity translates basically into standardisation. We centralise, standardise and rationalise in order to boost manageability through simplification.

Nowhere is it more important to centralise, standardise and rationalise than when dealing with services which are by definition transversal to many of the enterprise's other systems and processes.

By doing this, we not only make sure administration and support of, in this case, the backup platform is simplified, but the requirement for human intervention is also minimised by comparison with the need for human intervention when managing multiple backup solutions. This results in a decreased probability of human error.

Needless to say, simplification also brings financial gains. The more cohesive and standardised the ecosystems, the less the need to invest (both CAPEX and OPEX) in multiple, overlapping/redundant solutions.

AN END-TO-END BACKUP SOLUTION

Knowing our circumstances

Fun fact: it has been estimated that circa 80% of the world's corporately-held data is stored in the wrong storage tiers, which makes for an incalculable amount of money wasted storing data where it makes no sense to be stored.

Even prior to thinking about what we want to backup, we should know exactly what data we have, whether we need it (or are even legally allowed to keep it) or not and, if we do need to keep it, why and, therefore, where.

The data backup solution I've designed and commissioned which I felt the happiest about with the results was done following a massive job of data structuring and classification. When the moment came to design the backup solution, the solution was engineered to match our precise circumstances and needs, which made it extremely easy to accomplish.

Knowledge of your environment as a whole, including the data you host will allow you to see the full picture in terms of what the solution you are designing and commissioning has to cater to. The architect should have this birdseye view in order to put together a consistent and coherent solution that matches the needs of the organisation as a whole, and not those of just a few, which is always a terrible idea in corporate technical architecture terms.