Secure Data Life Cycle

SECURING YOUR MOST IMPORTANT ASSET FROM "CRADLE TO GRAVE"!

Author: Garry Kolb, CISM

TABLE OF CONTENTS

INTRODUCTION

We live in a world where "*DATA"* has become a most valuable asset to companies in every vertical of business. From financial companies to manufacturing companies and everyone in between each has data that they use to operate, enrich and grow their business in one form or another. Companies maintain a staggering quantity of data that resides in many different formats and storage media. These data include intellectual property, financial information, employee personal information, customer information and business plans just to name a few. I am sure that everyone can name other categories or nuances to the existing ones but suffice it to say that all organizations from small businesses to large enterprises or government entities have data that requires management and protection. The challenge that most organizations face is that they often have very little idea about where all of these data reside, how they are managed and protected and what happens when they are no longer needed. In this paper we will be investigating ideas, processes and techniques that will assist organizations improve their overall posture when it comes to managing data from its creation to its final deletion.

An Explosion of Data

Over the years data has been growing at an unprecedented pace. According to a recent article in Forbes Magazine [1] they quote figures from Domo, Incorporated, a business intelligence company, who reports that 2.5 quintillion bytes of data are created every day, and that 90% of all data currently on the Internet was created in the past 2 years. To put it another way:

"EMC estimates that the amount of data in the world doubles every two years, and that by 2020, we will have created and stored more than 44 trillion gigabytes' worth."[2]

This trend will only continue to increase as the number of users on the Internet increase it only stands to reason that the data they create will also increase. The problem that faces businesses is that they and their staff, are creating business critical data at an accelerated rate as well but many of the individuals in the Information Technology and Cybersecurity professions remain ignorant of much of what is being created, and worse yet where and how it is stored and protected and who can access it!

The sheer quantity of data and the ease with which it can be created and accessed poses major problems for any organization. And this problem is not limited to just the security team, but data quality is a concern as well. In the cybersecurity space they often talk about the CIA or the Confidentiality, Integrity and

[1] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#7535c72360ba
[2] https://www.cio.com/article/3224533/is-data-the-currency-of-the-future.html

Availability of data. Obviously if you don't know who is creating data, what they are creating, where it is stored and who can access it, you have no idea of its value or quality; and poor quality data can be more damaging than stolen data or no data at all.

The Lifecycle of Data

Just like anything in our world, both organic and inorganic, data has a lifecycle all its own. Dictionary.com[3] defines a lifecycle as:

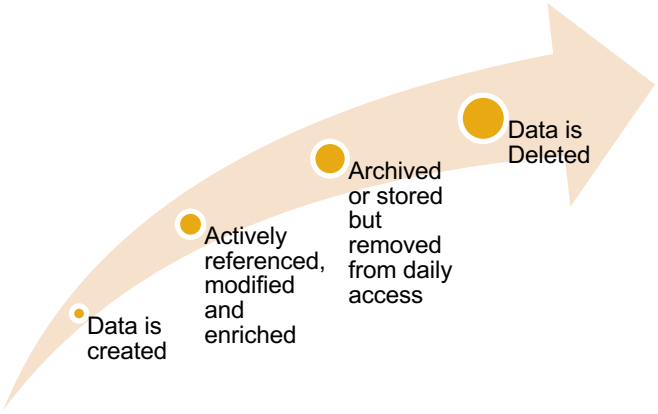
1. *Biology. the continuous sequence of changes undergone by an organism from one primary form, as a gamete, to the development of the same form again.*

2. *a series of stages, as childhood and middle age, that characterize the course of existence of an individual, group, or culture.*

3. *any similar series of stages: the life cycle of a manufactured product.*

Data is "born" the moment it is created; and as we have seen, we do an awful lot of that these days. It also goes through a "continuous sequence of changes" and eventually it grows old and dies. This is the lifecycle that we will be analyzing in this paper. At first glance that may seem like a very simple thing and one that is easily understood and followed, but as you will see that is really not the case. It is often difficult not only to identify the origin of data but also to determine its position relative to the overall lifecycle. It follows then that it can be a daunting task to ascertain the relative value of that data to the organization or the risks associated with its theft or disclosure.

So, what does this lifecycle of data look like? It is a straightforward "sequence of changes" that might seem obvious to most of us. It might be confusing to call it a "cycle" as that depicts a circle that starts and ends in the same location. However, the life of data is more of an arc than a circle; at least we hope it is not recreated every time! In short there are only four phases in the life of any data. They are; Creation, Active Usage, Long Term Archive and Deletion. Our ability to identify, track and properly protect data through each of these phases is what a Secure Data Lifecycle is all about.



There are a number of processes that are part of this Secure Data Lifecycle Arc (SDLCA) that vary depending on where on the overall continuum in the lifecycle the data happens to reside at that moment and that is what the remainder of this paper is all about.

[3] https://www.dictionary.com/browse/life-cycle?s=t

ROLES RELATING TO DATA

Before we go into each of the phases it is important to understand individuals' relationship to any data. There are five distinct roles that relate to data throughout its lifecycle. They are Creator, Owner, Custodian, Aggregator, and User. Each has their own “relationship” to the data and exactly what they can, and should, do relating to the data is inherent in the title of the role but for the sake of clarity we can discuss the details here. It should also be noted that one individual can hold more than one relationship with the data, such as Creator and Owner but that is not always the case, so they are distinct roles, nonetheless.

Data Creator

Obviously from the name this is the individual who is putting the data into the environment. This can be done by obtaining data from another source from outside the organization and manipulating it to fit the organizations needs or inputting the data onto a system in either structured or unstructured format. Either way the Creator is introducing a new set of data into the organization to be used for a specific business reason. While each of the roles mentioned is important the role of the Creator is essential in that this individual will be expected to classify the data and ensure that this is properly stored and protected.

One thing to be clear about is that once a set of data has been created and is in use by the organization new items can be entered but that does not put everyone who is entering those new items in a creator role. The Creator role is reserved for the initial introduction of the dataset into the organization. This can be an application developer who builds out the database table or the file structure that is to be used. It can be a person working with a third party and receives data from them to be introduced into the organization or an author of a document. Basically, the role of creator is reserved for anyone who is introducing a completely new set of data to the organization.

Data Owner

This role can be one of the least understood but, in our opinion, the most critical role relating to data. It is possible that the Data Owner is also the Creator but that is not always the case especially when dealing with large datasets such as a customer database. The Owner is the one who should be making decisions about the data for the remainder of the lifecycle. It is the Owner who determines what those with other roles should be permitted to do with the data. They also decide when data can be moved from the Active Phase to the Archival Phase and eventually when it is to be deleted.

Data Custodian

Not all data will have a formal custodian as this is normally reserved for structured data. A data custodian is the individual who has the responsibility for the technical oversight regarding the data. When dealing with structured data this will often be the Database Administrator who has the responsibility for the CIA of the data ensuring that proper access controls are in place, data is properly backed up and ready for use when needed. While they are responsible for the technical aspects of CIA the owner still has the responsibility for making the decisions around those areas.

Data Aggregator

Once data has been created it may often need to be combined with other data to provide the business with a specific view for reporting or processing. This is called aggregation where you are selecting portions of data from several sets of data and putting them together or aggregating them into a new set altogether. Often individuals view an aggregator as a creator, however this is not the case as they are not creating new data and they should never modify the aggregated data but rely on the original source data for changes. If an aggregator makes changes to source data that they have placed in their dataset the diminish the value of the data and destroy the integrity of it possibly rendering it useless to the rest of the organization as it "drifts" from its origin.

Data User

The final role is the Data User. This is an individual who accesses the data in the place where it is stored and performs a business function either with the data as input to the process or modifying the data as part of a business process. The Data User more commonly works with structured data but there can be users of unstructured data as well depending on the business need and process involved. For example, if there is a Word document that is being shared by several individuals they would be classified and Data Users as opposed to any of the other role.

PHASE 1 DATA CREATION

Who, What, When, Where and How of Data Creation

In every organization data is created at an amazing pace; not the pace of things like social media, but large amounts, nonetheless. The problem arises when those who have the primary responsibility for managing and securing the organizations data assets are unaware of much of what is being created. Think about it! In today's society almost everyone has enough technical acumen to open a word processor or a spreadsheet and enter information. The problem is, if the Information Technology (IT) team and the Cybersecurity team don't know that the data has been created they don't know the value of it or, what level of protection to afford it. This lack of visibility can have two very different impacts. First of all, it is possible that the larger organization will miss out on something very important simply because they have no idea that the dataset exists. Secondly, information that is important to the organization or more troubling, damaging to

the organization could find its way into the wrong hands purely because the Cybersecurity team did not know it needed to be protected.

Structured versus Unstructured Data

In order to fully understand what and where data is created is to first understand how data can be created. There are two basic categories of data organization, it is either Structured or Unstructured. Structured data is that which is stored in such a way as to group like elements of data in fixed locations such as in a database table or even an Excel workbook. Structured data is easier to track and understand through all the phases of its lifecycle and there are a few tools out there to help locate, categorize and classify these data. Unstructured data is data which does not follow a set pattern for how it is organized and stored. Some examples of unstructured data would be Word documents, Power Point presentations, simple text files and other such formats. It is this unstructured data that is often very difficult to identify at the point of creation and therefore often follows no lifecycle at all.

As stated above structured data is generally easier to track and even to determine things like the time of creation as it is often timestamped, and the event is often recorded in an application log or journal. There is also usually a preset application that facilitates the creation of the data and ensures things such as data quality and correctness. Of course, this is not always the case for structured data that can be created locally such as in a spreadsheet or local database such as Microsoft Access, but even those methods have a bit more rigor around timestamps and other such issues.

Unstructured data, as its name implies is data that can be created by anyone with a laptop or a desktop or even a connection to a server. These data are not put into any fixed pattern and can contain bits of unrelated data that is very difficult to parse through and find that which is significant. For example, this document is unstructured data that is being created over a period of time. Now even though there is a table of contents and there would be a date for when this document was first opened it would be difficult to pin down the point of creation as it was written over a period of several weeks. Additionally, since this is not a database or even a spreadsheet how would you know what parts of it are important and what is less so. Is there any proprietary data included in the document? Have I placed personally identifiable information (PII) in the document? How can that be determined? When was the sensitive data created in the document? And finally, how do we know if the location in which it is stored is appropriate for the level of data that resides in it. These are just some of the challenges when dealing with unstructured data.

Classifying Data at the Point of Creation

All data should carry with it some form of classification. Classification of data is essential to understand what protections are needed and how the data should be handled. Every major IT or InfoSec framework stresses the need to have a system by which data is classified and that classification must be applied to all data and it should be applied at the point of creation. This means that the creator of the data needs to have a good understanding of the organizations data classes and know how to apply them to their data. This puts a great deal of trust on the data creator but with proper education and procedures in place most individuals want to do the right thing when it comes to protecting the organizations data.

The importance of properly classifying data cannot be overstated. Basically, if you don't know the importance of the data how do you determine what protection it requires? Additionally, regulators have increasingly required that organizations storing regulated data (which at this point is almost everything relating to individuals) needs to have a classification scheme applied.

While there are several different frameworks that can be used to set up a Data Classification Policy it is really something that must be determined by the organization themselves. Data differs from one industry to another and then there are additional differences that are specific to each organization within that vertical. Of course, there are somethings that apply universally to specific verticals such as HIPAA for healthcare, PCI for large merchants and credit card companies and an alphabet soup of regulations in the financial industry for example so those will apply more broadly. But in the end each organization is responsible for setting their own data classification policy.

The number of classes that you will have to define and the type of data that will fit into those categories is largely dependent on your organization, the regulatory or legal environment you are in and also how you conduct business. A simple classification system that can usually be adapted to most organizations will contain four classes of data; they might be:

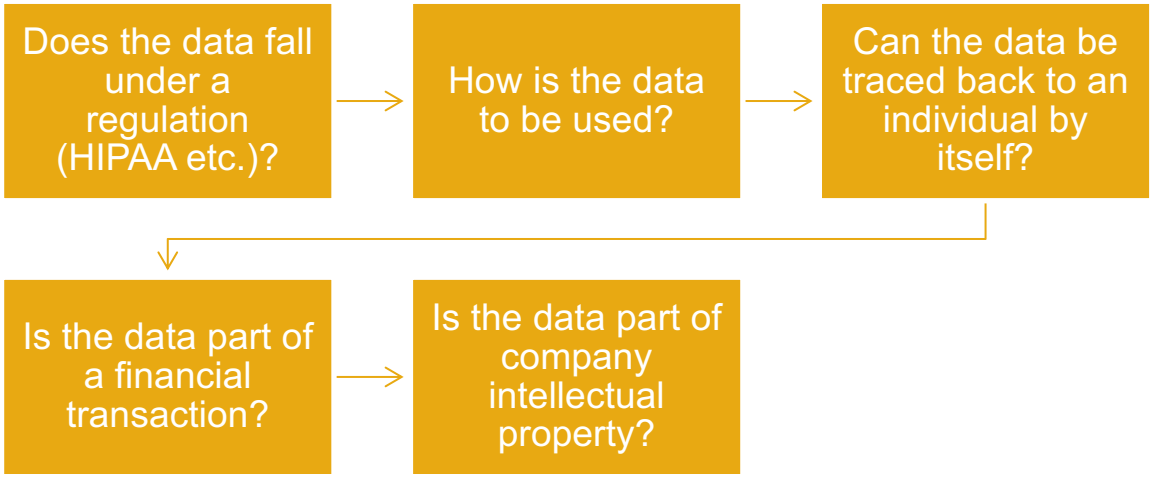
- **Highly Restricted** – this would be the highest classification and would be used for data that is covered by regulations such as HIPAA, SOX or GDPR other legal agreements such as PCI.
- **Confidential** – data that would fall into this category might include things such as customer information not covered under regulations, customer and employee personal information, future product plans and intellectual property.
- **Internal Use Only** – some data is only for those in the organization needed to do their job.
- **Public** – obviously by this name you can share this data freely outside your organization.

There may be more granularity required than what is shown above but that should give you the basic idea that not all data is created equal. For example, Confidential data might be subdivided into Customer Confidential and Company Confidential both requiring slightly different controls yet afforded the same level of protection. Additionally, one of the mistakes that organizations can make is believing that their classification taxonomy can be set at a point in time and remain there. This is a fallacy and dangerous for the organization. Data changes, regulations change, technology changes; your taxonomy must also change.

Applying the Classification

Once the organization has set up their classification taxonomy, they need to make information available to data creators, aggregators and users so that they can be assured that data is handled correctly. It is very important that everyone in the organization fully understand the classification taxonomy and how to apply it to any data they might be creating. It is best if there is a clear process that gives them a pattern or even workflow to follow to allow them to come to a decision of the classification. It is also important to teach them about aggregation and its effect on the classification of the set of data in question.

The following chart show some example questions that could be asked in classifying data. Of course, some sort of decision tree would be best as you can allow the user to answer questions and walk down paths dictated by those answers. So, in the example below you are seeing really the major headings for the various parts of the decision tree.



There are a number of examples and templates available on the Internet to help you with building out a decision tree and you can choose any of them or build it yourself the important thing is to have one. It is also essential that it is accessible and usable by everyone in your organization if you are to ensure proper classification is applied at the time of data creation.
Just to carry our example forward a bit lets discuss each of the headings that are represented above.

Does the data fall under a regulation?

This is a very important question as there are myriad regulations that dictate the way in which data must be handled, protected, processed and stored. Some of these regulations come with very heavy fines if you fail to properly protect and handle the data. Some examples are the Healthcare Insurance Portability and Accountability Act (HIPAA), Sarbanes Oxley (SOX), in the US and the EU's General Data Protection Regulation (GDPR). Other data protection dictates come in the form of voluntary or contractual agreements such as Basel 4 for banking or the Payment Card Industry (PCI) standards.

How is the data to be used?

Often the way in which data is to be used can lead a creator to an easy decision when it comes to the classification of those data. Is this data that is used as part of a financial transaction? Is this personal information on customers or employees to be used in contacting them? Is this data a plan for a new product

or project for the organization? Any number of questions come to mind once you have a good view of exactly how the data will be used in the organization.

Can the data be traced back to an individual?

There has been much focus in recent years on personal privacy and this includes data that may be maintained in your environment. One of the questions that the creator of the data should ask is if these data identify an individual or be used to gather personally identifiable information (PII). Many of the recent regulations around the world relate to this important information in an attempt to assure individuals that their PII is properly handled and not disclosed without their knowledge and consent. The thing to remember here is that this can include customers, prospects and even employees; under the regulations all PII must be protected.

Is the data part of a financial transaction?

Obviously anytime finances are involved it is important that the highest level of protection be afforded the data. In our modern world money moves though computers and networks not via paper hand to hand. This complicates the job of data creators as they need to identify those data that are part of that global financial transference and make sure that it is not exposed to breach. This includes records of financial transactions that can often be required by regulators and auditors and must be afforded the highest protection to ensure Confidentiality and Integrity.

Is the data part of the organizations Intellectual Property?

Theft of intellectual property can have devastating consequences for any organization. Intellectual property comes in different forms; it may be something as simple as a company logo to something complex and large such as a trade secret. There are a number of laws and international agreements that cover the theft of intellectual property making it a crime but even if the thieves are caught and prosecuted the damage may have already been done, especially in the case of the theft of a trade secret. So properly classifying and protecting Intellectual Property is essential to its unauthorized disclosure.

Use Technology to Help

While it is important to educate everyone in your organization about classification of data and how to do that there are some vendor products that can be deployed to help with that process. For most organizations that are very committed to classifying their data having a product that uses input from the user community combined with pattern matching and machine learning to apply classification to data both during the creation phase and even scanning existing sets of both structured and unstructured data to apply a classification.

The key to using technology to help is to remember it is there to help not to do all the work for you. That is why it is important to have solid documentation on the types of data that your organization creates and detailed guidance on what factors need to be taken into account when classifying data. The thing to remember is that the tools you use will be much more accurate based upon the input they receive. If you provide poor information, then you will have a poor result.

A Word about tools

There are several tools that claim to be data classification tools but many of them are simply a repointing of the existing Data Loss Prevention (DLP) products using the same regex-based pattern matching and applying it to the classification process. While this can be useful, the high potential for "false positives" tends to make these solutions rather unwieldy and often requiring a great deal of maintenance to ensure the data is properly identified and patterns are adjusted as needed. While all tools will have a degree of misidentification of information it is best to use tools that are multi-faceted so as not to rely heavily on just matching patterns.

Tools that enable the creator to select the classification based on a set of criteria, that can be enriched through machine learning from the broader user community, are the best tools to use. In a sense you are "crowd sourcing" the classification by utilizing this approach. As each creator determines the classification of various data elements the tool "learns" and can apply that knowledge to the next user as they create new data with similar attributes. This "learning" can then be applied to static scans of existing data greatly improving the accuracy of the classification and thereby reducing the overhead involved in maintaining the rule set.

Technology Alone Won't Work

As you can see from what we have shared thus far it is essential that the Data Creator establish the classification of the data from the start. The problem is in many organizations getting the creator to correctly understand the taxonomy and take the time to assign the proper classification is not necessarily in their DNA! Without the cooperation and diligence of the creator much of the rest of the planning for the lifecycle will fall apart.

PHASE II ACTIVE USAGE

Data is created to be used. Creating data that is not used is like creating a great work of art and then hiding it away never to be seen. Now data is rarely a work of art, but its purpose is to be used; used to further the business or the goals of the organization in some way or another. If data that is created in your organization does not have a purpose for bettering the organization then it is wasting valuable time and resources. However; in order to use data correctly it must be handled in accordance with the importance that it has to the organization. One obvious way to determine its importance is by the classification that it is given, and each organization should have policies, standards and procedures to provide the details needed for proper handling.

Handling, in this context refers to several different aspects relating to the usage of data. Those areas include; storage location, data focused controls, encryption, and monitoring of the movement and copying of data. In this section we will cover each of these areas individually.

Storing Data

Once data is envisioned and the classification has been determined it is essential that the data creator and subsequent custodians, and aggregators have clear guidance on where and how those data are to be stored. Those organizations who store regulated data or data that is covered under a contractual agreement such as PCI should have their network divided into segments with specific segments that are used to store data with higher classification. There should be additional controls on that portion of the network to ensure that it limits access and has higher levels of network monitoring including things such as a firewall, that will even be traversed from other parts of the organizations network before a connection can be made. It is also essential to have other OSI Layer[4] 3 and 4 protections in place such as Network Intrusion Detection Systems (NIDS) and Host Intrusion Detection Systems (HIDS).

Data Focused Controls

Unfortunately, many organizations implement Network based controls and think that they have adequately protected the data. Nothing could be further from the truth. Network focused controls operate under the assumption that inside of the controlled network segment data is handled correctly. That being said it is important, for many reasons, to segregate the network according to the classification of the data that should reside there. However; in addition to network controls it is important to implement other controls such a data tagging, role-based access controls (RBAC) and data encryption.

Data Tagging

Data tagging as the name implies involves having an identifier that travels with the data and provides the classification of the data. By tagging data it can be afforded protections that are appropriate regardless of

[4] https://www.lifewire.com/layers-of-the-osi-model-illustrated-818017

their location in the network. Now this is not to say that data of a higher classification should be intentionally placed in a network zone of a lesser classification, but in reality, we all know that such things happen so data tagging can provide a better way of identifying that situation and ensuring that other controls that are more data centric are applied.

Role Based Access Controls

Role Base Access Controls or RBAC means that access to data is tied to a role that is properly applied to an individual's identity. This is a means of enforcing what Cybersecurity professionals often refer to as "least privilege" which is simply ensuring that data can be accessed only by those who have a business need to do so and no one else is permitted that access. RBAC is most effectively implemented as part of an overall Identity and Access Management (IAM) strategy that should enforce periodic recertification of access roles and is most effective if it can be tied to something such as the user's cost center. I have found that utilizing the cost center is a very effective way of ensuring accuracy as managers may not always notify the IAM team when someone leaves their team but, they will always let the payroll department know to get them out of their cost center.

Encryption

One additional control that is normally reserved for the most sensitive, highest classification of data is encryption. The reason that this is only used for highly classified data is due to two simple things, cost and ease of use. Encryption, when done right, is usually expensive in that it often requires hardware as well as software and some cryptography staff to support it. That is, if you want to do it right! For regulated data or other sensitive data such as PCI it is often required that the encryption technology that is deployed meets certain minimum standards, such as AES 256 or better and that keys be managed through a Hardware Security Module (HSM). Encryption, along with proper key management, is a way to ensure that data is only accessed by authorized individuals or processes as only they can be granted access to the proper keys.

While we are here let's talk about how data gets encrypted and run through some scenarios that might provide more insight. When working with data inside your network you will probably use symmetric keys; that is a single key that is used to both encrypt and decrypt the data. In order to ensure that the data is safe the way the symmetric key is managed is essential. This is why the use of an HSM is essential. An HSM ensures that the key is properly built and managed without a human needing to know the key. HSM's are packaged with key management software that controls how the key can be retrieved and used when a process requires access to the unencrypted data.

But what about the use of Asymmetric keys; where there is a public, private key pair. You will usually see Asymmetric key pairs used when you need to share sensitive data with someone else, either in your organization or usually, outside of your organization. To make it simple the way an asymmetric key pair works is that one side shares their public key with the other but keeps their private key, well, private! The public key can only be used to encrypt data it has no ability to decrypt data that has been encrypted. The private key is then used to decrypt the data. So, for example, if Company A needs to share sensitive information with Company B and it needs to be encrypted before being sent, Company B would share their public key with Company A who would use it to encrypt the data. Once it is received at Company B

they would decrypt the data using their private key. If Company B needed to send back modified data to Company A then the process would be reversed. As you can see through these examples the security of the keys involved is paramount to the success of these approaches.

Data in Use

As we mentioned earlier in this paper data is created to be used and there are any number of ways that happens. Often data is simply accessed by approved individuals in the organization. These Data Users simply need the data to do their jobs. What a user is permitted to do with data is limited by the access permissions that they have. In general, there are four different access levels they are Create, Read, Update and Delete. These permissions are often referred to by the first letter of each or CRUD. We have already discussed the role of the creator so that takes care of the C in CRUD. The R is for those are users who simply need to view the data. They may be reviewing a report on the financials of the organization or searching for a customer record in the database, but they are simply viewing the data in its current state without modifying it.

The U stands for Update so users with that level of permission to the data will have the ability to modify it in some form or another. If the data is structured it will most likely be modified through a user interface with a process module controlling just how the data is modified and the update stored. With unstructured data things are not always so well ordered. While there may be a specific user interface (UI) that controls the update there is often no such rigor around an update to unstructured data so the integrity of the data can be compromised. This is why most organizations prefer structured data to unstructured data for the most important information resources.

Finally, there are those who can delete the data. This level of access is often the most tightly controlled and frequently is given only to data custodians who have policies and procedures to follow regarding that action. More will be said about this in a subsequent section of the document.

Role Based Access Controls (RBAC)

Having established that there are different levels of access relating to what a user can do with data, it is often necessary to establish set roles for users tied to their job function. This is commonly known as Role Based Access Controls or RBAC. This approach allows an organization to establish a set of access rules to various sets of data and allow a user to request a specific role which enforces that access without having to set individual rules for each user. Often this is combined with an approval and recertification workflow which is usually required by policy and regulations. However; in order for this to work well it must be set up in such a way as to ensure that a user's role is specific to their current position and if that changes they are taken out of the role for their old job and placed in the proper role for their new job. If this is not properly done, you can have users whose access exceeds the needs of their current position because they have moved around within the organization. This is why periodic recertification of access is important so that the manager who owns the roles associated to their staff can remove access from individuals who have transferred out of their department. Another way to do this is to establish roles that relate back to a cost center in the general ledger and specifically to payroll as a manager may forget to remove someone

from a role when someone transfers but they will never forget to remove someone from their payroll cost center!

A Word About Backups

One other thing to remember when dealing with active data is that critical data is always backed up somewhere so that it can be recovered in the event of an incident of disaster. Unfortunately, what many organizations overlook is the fact that the protections afforded to the data where it is active must also be applied to the location of the backups. In years gone by backups were often placed on tapes or some other form of removable media that was then afforded a degree of protection in its physical placement in a vault somewhere. Today however, backups are often placed on similar media to their active counterpart, but the protection does not always follow the data. The concept of having a digital vault is very important and this is usually accomplished through robust encryption protocols and strict access rules. An organization that wishes to fully protect their data must consider "over protecting" their backups.

PHASE III ARCHIVING DATA

There comes a point in the arc of data's useful life that it is no longer needed for active usage but must be maintained for a period of time for one reason or another. Usually this means that the data can be archived or sent to lower cost, lower accessed locations and media. There are two forms of archiving as well. One is short term or accessible archived data. This is when data is not actively being used but is still accessible to certain individuals in the organization. The other is long term archive which is often in an off-site location and is more difficult to recall.

What Data Should Be Archived?

Once data is no longer needed in active operation what criteria should be used to decide between archiving the data or simply deleting it? In general, that largely depends on a number of factors such as; the business you are in, the regulatory environment under which you operate and guidance from legal counsel. First of all, it is important to realize that the decision to archive data and the length of time it should be kept are really decisions that should be made by your legal department and enforced via a Data Retention Policy.

Most legal counsel will advise you to fully delete data once it is no longer active unless there are regulatory or business requirements that would cause you to retain it for a period of time. The main reason for this is that data which is kept in any form is discoverable under court order and it might cause embarrassment or worse for your organization if some data were discovered, not to mention the time or effort needed to find and package those data. However; depending on the business your organization is in you may have regulatory requirements to retain data for a specific period of time. For things like financial transaction

data that may be anywhere from three to seven years! Other industries also have retention requirements for certain data and a publicly traded company has retention requirement for internal financial data as well. It is important that you work with your legal team to draft specific retention policies and then follow them. Some organizations have run into trouble with courts for not having a retention policy and then deleting data once a court action has been initiated. If they had detailed retention policies and were just following those policies consistently, they could have avoided that problem.

Short-Term or Long-Term Archive

Data that is archived is often needed for a short period of time for things like reporting or metrics. These data must be stored in a location and format that allows for limited access for those specific purposes. The key here is to remember that this is meant to be for limited access as it is no longer active data. If your organization confuses short-term archive data with active data, you have just put yourselves into a very difficult data integrity situation. Short-term archival data needs to be clearly marked and named and have specific access rules applied to keep cross contamination with active data from happening.

Long-Term archive data is usually stored in an offline manner so as to be available if required but not readily accessible by just anyone in the organization. It is a good idea to have a data custodian who has responsibility for long-term archived data so that it is completely controlled an only accessed if need. It must also be stored in such a way as to be deleted when there is no longer a need to retain it.

Proper management of archived data helps keep your organization compliant with regulatory demands and at the same time ensures that only active data is in general use.

PHASE IV DELTEING DATA

As we mentioned in the previous section once there is no need to retain data it should be deleted. There are many forms of deletion that can be done and there should be a policy on data destruction which ties various methods to the classification of the data. The way in which data is deleted will also depend upon the format and media in which the data is stored. So, a data handling standard or a specific data destruction standard should call out all of the media types and the appropriate deletion methods depending on the classification of the data involved. For example, if you have Highly Restricted on a non-volatile storage media such as a data array, just issuing an operating system delete command may not be a sufficient method of deleting that data as that will often simply remove the pointer and mapping to the location of the data but not destroy the data itself. For data of that nature it might be necessary to perform a programmatic “shredding” of the data to ensure that it is completely gone. For some confidential data this might be needed as well. However; for lesser classifications a simple delete command might suffice.

Additionally, data destruction standards should cover how data is to be deleted if the media it is on is being taken offline for storage, destruction or to be re-purposed. I remember receiving a leased disk array

that contained data from the previous lessee. Fortunately for them we just immediately performed a total reformatting of the devices and wiped out the data before we brought them online to the rest of the organization but under other circumstances that could have gone very wrong for that organization. The main thing to remember here is to have good standards and operating procedures when it comes to data deletion or destruction.

CONCLUSION

Data, as we stated, for most organizations is a key asset that must be protected throughout its entire lifecycle from creation to deletion. However; in order to do that requires forethought and planning. Without proper written policies and standards regarding every aspect of data management, it will be nearly impossible to properly control data throughout its life.

This paper was probably somewhat obvious for those professionals in the Cybersecurity and Data Management professions, but we hope it was useful for business professionals who are so reliant on the data maintained by your organization.