# Data Retention and Versioning Considerations for Identifying Security Measures

**Anand Athavale**

Independent Researcher
Decades of Industry experience in Data Management
andyathavale@gmail.com

**Abstract**

**Data retentionadds to the complexity for identifying data security measures. Typically, data resides in its original form and in the primary residence or data locations built and meant for that type of data. However, various regulation requirements along with the need for reducing the cost of primary data locations, often data is either copied to a secondary location and is completely removed from the primary data location or copied to ensure that it is around for stipulated number of years. During these transitions, while some data types retain its original form, often data could lose its original structure, and it goes beyond the access control measure mechanisms put in place at its primary data location. Additionally, ever changing data types create additional challenges for keeping the classification up to date and intact. This article discusses the challenges created by the retention and versioning aspects of data.**

**Keywords: Information Security, Ransomware resilience, Sensitive Data, Data Retention, Immutability**

## Introduction

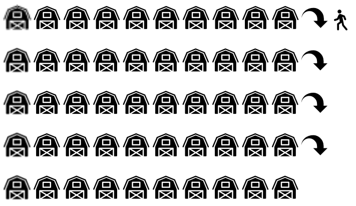
Data regulations are one of the key drivers for data retention. Typically, data regulations enforce requirements to keep data of certain nature and content for a minimum stipulated amount of time. These regulations sometimes are enforced as part of some emergency. As a recent example, California state in the United States of America, mandated employers to maintain records of written notices of potential COVID-19 exposures for at least three years [1]. Various regulations can have different time periods for minimum retentions. As technology evolved, the regulations have added more stringent languages for “ensuring the data is retained without alteration.” The common term established because of such requirements is Write Once-Read Many media for retaining data, or WORM for short. Here is where different regulations have made it tougher to follow these regulations. On one hand, there are regulations requiring minimum retentions and on the other, different regulations like GDPR call for not retaining data more than needed, especially the type of data which falls under personal data belonging to individuals vs. the organizations themselves.

**Regulation Complexities**

The “minimum retention”and“retain only until needed”requirement combination gets more complicated due to the changing nature of data.The additional complexity stemsfrom having tokeep the classification up-to date, requiring the labeling and re-labeling. Sometimes the language of the regulation or best practices is also ambiguous. As an example, the FDIC general instructions say that a bank should maintain in its files a signed and attested record of its completed Call Report, including any amended reports, and the related workpapers and supporting documentation[1] for three years after the report date. The “supporting documentation” may include, but is not limited to, overdraft reports, trust department records, and records of other material adjustments to deposits.These complications make the application and business owners defensive about deleting any data, regulated or otherwise. Such these retention habits increase the attack surface multi-fold. The information security persons must watch out for such habits even though these may appear out of their purview.

Here is an analogy for why it is important for information security team. Consider that there are two warehouse businesses. One business, is not disciplined about keeping the non-essential data to a minimum. Due to this, the essential and non-essential items are kept side by side eventually needingfifty storage sheds. These sheds are fifty meters apart and thus, overall distance for a security person to travel is two and a half Kilometers. Now, no matter how fast the person on the shift makes the rounds, the frequency between each shed visit is going to be more than twenty minutes [2]. The second business is more disciplined to retain essential items only as much as possible, hence can manage with only eighteen sheds. The sheds being fifty meters apart, the duration between two visits is only ten minutes given the distance to cover is only nine hundred meters. The risk is significantly reduced in this case with the same amount of security resources.

Business 1 (50 storage sheds, 50 meters apart)



**Keep essential and non-essential**

Business 2 (20 storage sheds, 50 meters apart)

**Keep only essential**

The same concept applies to the attack surface because various types of monitoring and leak prevention solutions would require less resource and time and could be deployed towards essential items instead of “everything.” It also gives lesser opportunity for dormant malicious code to hide. It is similar to a runaway thief hiding in the crowd of hundred compared to a group of ten. The concept of keeping only essential data is similar to data minimization at a high level. However, data minimization talks exclusively about personal data [3].

**Data retention requirements impact on information security measures**

Data retention requirements have three types of necessities to adhere to the regulations.

i. Pre-requisites

Pre-requisites typically may involve discovery and labeling. It is important to note that “not all data” is subjected to regulations. It is also important to note that one type of data may be subjected to one or more regulations, which in turn may have varying retention requirements. A good example of this is communications data [4]. General EU retention requirements asked for retaining this data for twelve months in one of the EU countries. However, trade repositories and central counterparties may be required to keep it for ten years. Now, the vast variance in retention explodes the attack surface for information security teams. What they have needed to watch for only twelve months, could become a ten year long guarding duty. Depending on the technical standards of the regulations with accessibility, information security teams may not be able to “throw the data in dungeon and throw the key in a lake”. In such scenarios, exposure control for this retained data becomes their responsibility at least for defining security measures. It is also important to keep in mind that as same data is subjected to multiple regulations, it can have more than one labels associated with it.

Historically, one of the classic media types for ensuring the WORM nature were tapes. Typically, they would copy the data to a tape and then ship the tape to a remote vault for “safe-keeping.” However, recent improvements in disk storage technologies and devices have introduced the concept of “immutable” storage, allowing to simplify the retention through disk storage and making it relatively easier for choosing the retention periods. This has been clarifiedin the SEC Recordkeeping and Reporting Requirements [5]. However, not all technologies offer same level of security to ensure that this “immutable” property of the storage, or, data retained on the storage is not easily compromised. This is additional consideration for the information security measures related to data retention.

ii. Process and Workflows

The second necessity is for having a process for the data which is labelled as matching for data type covered by one regulation or more. As an example, one organization could choose to have various storage locations marked with certain amount of retention period to match varying retention requirements. Such a process could then move or copy the data “after the use” to the location with the highest retention period setting assuming the data falls under multiple regulations requiring various retention periods.

The consideration for information security team is to ensure that the exposure remains limited even after the movement. Also, in case the data needs to be restored or recalled, it is vital that the original access control is restored along with it. Additionally, the information security team must ensure that the location of storage, whether backup or archive, adheres to the technical standards like Write Only Read Once. They must also ensure that the retention periods cannot be reduced directly through the application, or, indirect manipulations like moving ahead system clocks. The transmission to these archival or backup locations along with the storage itself must ensure encryption.

The last consideration for the process and workflow necessity is to ensure that when data is finally retired, or in other words deleted, it is still applicable for deletion. Sometimes the regulations itself may change requiring additional retention. In some cases, the data stored may not have been labelled properly and may need a final check before it is deleted. Incorrect settings may cause accidental early deletion. On the other hand, regulations like HIPAA may require more stricter erasure methods, especially when refreshing storage hardware. One of such methods simply calls for encrypting the data and then destroying only the key.

iii. Post monitoring and actions

The final necessity is monitoring changes to data. Certain complexities arisedespite of organizations classifying large portions of the data and assigning appropriate labels manually, or using one of the classification tool or technology. However, data does not remain stagnant.The constant changes to data bring three types of complexities.

First complexity is related to the new data or, items, or, objectscreated every day. Since it did not exist at the time of “initial” mass classification project, this new data is not labeled at all. So, there is a clear necessity to monitor for newly created data, classify it and then label it accordingly. This must be followed by the same process and workflows illustrated in the previous subsection. The monitoring more often must be automated, since manually keeping an eye on newly created data is nearly impossible, especially when large volumes of data is involved. Keeping sensitive and regulated data safe from theft and destruction is one of the primary jobs of information security teams. Hence that team needs to consider this complexity and necessity of keeping track of newly added data. A reminder from previous article related to data content considerations, what new data is being created is only one part of the monitoring aspect. Who is adding new data also needs to be monitored as that too has bearing on security requirements beyond the retention duration due to regulatory and other reasons.

Second complexity has two subparts to it. First subpart relates to modificationto existing data. This modification can have three types of impacts. First, newly added content to existing data would make the existing otherwise non-sensitive data which does not fall under any regulatory criteria, eligible for one or more regulations. Second impact, relatively rare, is that the data or item, already labeled, may become eligible for additional regulations, and would require additional labels. Third impact which is also rare, is opposite of the second impact, where the content making that data, or item, eligible for a regulation, is removed as part of modification of the item, and hence necessitates removal of relevant labels. This final impact is not about increase of the risk, but mainly to do with efficiency, as that data may no longer need special treatment and care from security aspect.

🗎 No Label === Data modified ==> 🗎 Label1

🗎 Label1 === Data modified ==> 🗎 Label1, Label 2

🗎 Label1 === Data modified ==> 🗎 No Label

Impacts of Data Modifications on classification

Some data types explicitly implement the concept of versions, which was covered during the data purpose article. Having versions of any data item recorded may help somewhat, if the retention technology like archiving products, are able to explicitly label different versions of the same data item accordingly. However, if the versions are not considered during labeling, the "eventual" label decision drives the fate of all the versions, even if some versions of the item may be eligible for a regulation, and some may not.This covers the first subpart of the second complexity data change brings up.

Second subpart of the second complexity is not related to data change itself but it does necessitate action. This subpart has to do with changes in existing regulations. This change could either add newer type of content eligible for any regulation, or, change the technical standards around retention and accessibility mechanisms. If there is newer type of content added to any regulation eligibility criteria, it necessitates major re-classification event, or, a project. If it is change in the technical standards, any new process and workflows acting on newly classified data must adhere to the newer technical standards, which could result in changing the storage media for the same, or, more actions like anonymization or obfuscation. Here, information security teams must collaborate and monitor the implementations by the storage teams. While storage team is responsible for enacting the changes, it is information security team responsibility to monitor that the new security standards are implemented and enforced.

The third complexity is the trickiest to handle. Many regulations like GDPR give the data subjects, in different words, original owners of the data, to act on their "right to forget."It means that specific item must be deleted, barring exceptions of continued business need. The action taken for that data item is relatively easy for the original copy. The IT team can simply remove just that piece of content from that item, or, simply delete that item altogether if it does not impact any other data subjects, or other parts of the organization. However, if the data is also present at other locations, which are meant for either backup, or regulation driven retention, or, archiving just for efficiency purposes, carrying out the same actions may not be easier. To make the situation even more complex, the location may be of the type "write once read many", preventing any change, or, removal of such data item. Here the security team consideration is to ensure that when data is restored, this "eligible for deletion" item is not restored, and if it is restored, it cannot be used by anyone. This is where labeling of data becomes extremely important for information security teams. If the organization is using storage media like tape and CD ROM, then the restore events become difficult to monitor. Last piece of the complexity to consider is that the backed up or archived data may not have been labelled at all from the regulatory classification aspect. This puts additional responsibility on information security teams to make them part of such restore events to ensure that the destination of restore is secure and the need of restore is legitimate and not an attempt by any malicious insider to let the data on the loose.

## Conclusion

Data retention is required for multiple reasons including regulatory compliance. However, the malleable nature of data works almost against the methods and techniques of the compliance and creates challenges for both, the compliance, and the security teams alike. Additionally, the lack of co-ordination among storage, data protection and information security team can create further security challenges, especially during one of events of recall of archived data, or restore of backed up data. Many times, cyber criminals target the retention locations, especially when the intent is theft instead of destruction because of the complexities and challenges listed in the article. Information security teams must be aware of these complexities and the weaknesses in security caused by these complexities.After all, security is only as strong as the weakest link.

**References**

[1] HUNTON ANDREWS KURTH, California's New COVID-19 Notice And Record-Keeping Requirements, Hunton Employment & Labor Perspectives, (November 2020), https://www.huntonak.com/hunton-employment-labor-perspectives/californias-new-covid-19-notice-and-record-keeping-requirements, (February, 2021)

[2] Alves, Fernando & Santos Cruz, Sara & Ribeiro, Anabela & Silva, Ana & Martins, João & Cunha, Inês. Walkability Index for Elderly Health: A Proposal. Sustainability. 12. 7360. 10.3390/su12187360. (2020). https://www.researchgate.net/publication/344166318_Walkability_Index_for_Elderly_Health_A_Proposal, (January 2021)

[3] REGULATION (EU) 2016/ 679 OF THE EUROPERAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016, Official Journal of the European Union, (April 2016), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679, (January 2021)

[4] Data retention across the EU, Fundamental Rights Agency – European Union (July 2017), https://fra.europa.eu/en/publication/2017/data-retention-across-eu, (January, 2021)

[5] Recordkeeping and Reporting Requirements for Security-Based Swap Dealers, Major Security-Based
Swap Participants, and Broker-Dealers, Securities and Exchange Commission, (February 2020), https://www.sec.gov/files/rules/final/2019/34-87005.pdf, (January 2021)