International Journal of Multidisciplinary Research and Growth Evaluation.

Data classification segregation and targeted approach necessity for defensive security measures

Anand Athavale
Independent Researcher, Decades of Industry Experience in Data Management, USA

* Corresponding Author: **Anand Athavale**

Article Info

ISSN (online): 2582-7138
Volume: 03
Issue: 06
November-December 2022
Received: 14-11-2022
Accepted: 13-12-2022
Page No: 703-706

Abstract
Data classification has been painfully and clearly an afterthought for security measures. Historically, data classification was a "no-one-wants-to-do" regulation requirement applicable to only those organizations and industries who were "selected" to be "regulated." As part of the regulation compliance, those "chosen" organizations had to show proof of data classification either in a direct way, or, data classification was a necessary pre-requisite for carrying out the mandated "data actions" prescribed by various regulations. However, once attackers started using dual attack techniques of threatening the access to data by encryption or locking, with the technique of exfiltrating the data to hold it ransom with a threat of a leak, security and IT teams were caught off guard. While data leak prevention systems were in place from a long ago, the scope and origins were vastly different. In the rush of addressing this issue, security solutions seemed to have missed a step of subdividing those requirements into two separate categories which needs to be clarified for the security practitioners to implement effective and rapid measures.

DOI: https://doi.org/10.54660/.IJMRGE.2022.3.6.703-706

Keywords: Data classification, credential compromise, application security, precision classification, privileged account management, one-time passwords

Introduction

Data classification has been around for more than a decade. It has been recently primary focus of data privacy regulations, which started with General Data Protection Regulation in European Union and now many countries have adopted the same. At the same time, pandemic caused remote work to explode and that in turn started rise in credential compromise. While credential compromise was and is being used for data theft, data classification is being looked at in a mingled fashion which could reduce the speed and effectiveness for security practitioners for spotting the obvious signs and sources of such a compromise. This has to do with human inability to remember passwords and secret keys and then going down the wrong path of writing those down. In combinations, sometimes the adoption of technologies for avoiding such a necessity in the first place is lower. This low adoption has many reasons including incompatibility, budget, skills and to put simply, lethargy. The lethargy portion can be correlated to the lethargy of "simply washing our hands frequently" before the pandemic.

Overall, security practitioners first need to understand data classification capabilities and the limitations of such solutions in terms of focus and scale. Then, they need to work with compliance teams to work on two separate fronts. One, is the very valid need of identifying sensitive data, but the second, to give way to rapid and focused classification solely targeted towards credential compromise. This segregation exists loosely in the data classification solutions today. But it needs to be further segregated and then deployed in different locations as per the fundamentally different purposes.

Understanding the current problem of credential storage

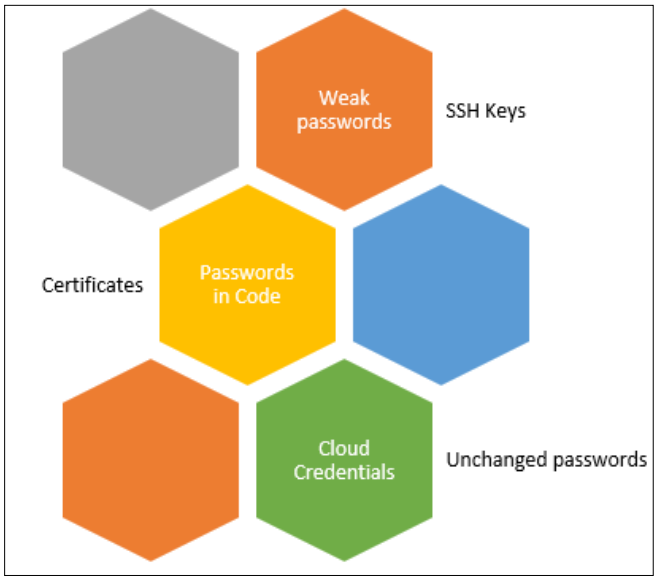
There is a hard to accept reality, especially for a large organization. Despite of best practices, mandates and security trainings, credentials, or secrets "lying around" still exist. The reality may be compared to the reality of traffic violations. There are clearly written rules for traffic and driving. But drivers cannot be solely relied upon to follow those and hence, traffic cops are needed.

So while we look at the problems and solutions for the "lying around" credentials, it must be acknowledged that those alone will not be sufficient and hence a "monitoring" aspect will be required.

By this time, post the pandemic led remote work explosion along with some "pre-existing" conditions, following is an indicative list of the sources and causes of "lying around" credentials which are one of the possible sources of credential compromise during a cyber-attack, or, sometimes a malicious insider led attack and possible solutions and practices to avoid these.

Sources of compromise prone credentials

Overall, any credential is prone to compromise. More prone are the ones which are not secured with best practices and available technologies. First and foremost are passwords. Now these have two sub-categories in them.



Sources of credential compromise

Fig 1

First, passwords stored in plain text, password files like passwd file in Unix systems, or, rare but possible hardcoded passwords in source code or scripts. The second category is not related to storage but mostly lethargy or fatigue. Either with an understanding of repercussions, or just out of ignorance, widely "known" or easy to guess passwords chosen are. "123456", "password", "111111" and "1q2w3e" are only some of the examples because the list is very long [1]. The next set of credentials prone to compromise are API keys, tokens, encryption keys and SSH keys. These are typically found because of need of some type of automation, or, avoiding a "password" prompt. Typically, these are used in operations which are not low in numbers, like twice a day or so. These could probably be combined with newer cloud technology offered cloud access credentials like access tokens. But mostly these exist because such operations may run into hundreds. The next type is typically referred as "service accounts" which are used by pre-built applications and are meant for the software applications to carry out operations without human intervention, and again in bulk.

Key rotations and password changes as mitigations

Many identity systems and applications force change of passwords on designated frequencies as recommended by general best practices [2]. Similarly, keys are supposed to be rotated manually, or automatically. Both these methods make stolen credentials less effective or usable. However, there is still risk from the first password reset, or, the key rotation till the next.

Privileged Access Management and One-time passwords

There are many privileged access management technologies which "maintain" the credentials to be used by human users and applications and mostly have lower time to live. Besides the secure storage and strong passwords, these solutions take care of stringent session management with secure detailed auditing, especially for sensitive operations carried out by privileged accounts like administrative users [3]. They also ensure password rotation because they change the password after each use, reducing the "time-to-live." This reduces the re-usability of stolen credentials. Lastly, these provide break glass emergency accounts where immediate access is needed but logging and monitoring is even more stringent for those.

However, even with the sources and reasons of credential theft and available alternatives, keys and passwords sometimes are left in the wild, either because of non-compliance to best practices or inadequate knowledge, or, the need to write quick scripts for some kind of automation.

Data classification: Existing segregation inadequateness

Data classification technologies have heavily focused on regulations in the past. With emerging ransomware threat, they have gotten smarter to cover security centric content, but it is more of the reuse of the existing classification techniques. While Artificial intelligence is employed, there is no segregation involved. Lastly, for large enterprises, the amount of content to classify is like a vast ocean. There are newer technologies claiming to be faster than the traditional ones, but the coverage of sources remains low and the accuracy questionable. Then there is question of urgency, differing objectives and control.

Regulation driven data classification for the most part is about retaining data [4]. Data Management solutions build defensible data deletion, or quarantine workflows around data classification, but eventually the purpose is data management, and not security. Here an important point needs to be made. This type of classification is important for security teams too because they want to protect the "crown jewels" data at risk from destruction and, or theft. But the basic purpose is regulations compliance.

The credentials data in the wild, needs different workflows and personas to "deal with it" when found. Some of the credentials may already be out of commission. Some may be in use. Also, there is another complication. There is no easy way to confirm it either way. Imagine if there were a central location of all passwords and key in use to "crosscheck" whether any credential found in the wild was still being used. That will be end of data security since everyone will go only after that source. It will be like finding the ultimate, or "universal" key. Thankfully, there is no such central repository of "keys and passwords in use." It means, security administrators need to carefully investigate whether the "credentials found in the wild" are being used. Also, while this is being done, that credential cannot remain where it was found. It needs to be transferred aka moved to a secure, less easily accessible location. Eventually, security team, depending on their finding would then change existing keys, or, passwords, destroy the key, notify the culprit person or the

team to "not do this again" and finally, notify legal, or, compliance teams as necessary if the key was a means of accessing any sensitive data. This entire workflow differs completely when compared to sensitive data handling.

Other than the workflows and personas, the "obvious" places to look for these two distinct types of data are also different. Analogous to "honor among thieves," the lazy or ignorant persons mostly avoid keeping these "credentials in the wild" on shared unstructured sources. If those are stored digitally vs. paper, and this is not to say that paper is secure; mostly they store it in their local drives, or devices. They may even zip and password lock these, thinking now they have secured these. As opposed to this, regulation centric data classification targets unstructured data in NAS servers, cloud SaaS applications and similar data locations. These are indeed different than the suspected locations for "credentials in the wild" locations.

The industry is trying to separate disaster recovery from cyber incident recovery [5]. Separating regulation centric classification from credentials centric classification is one of the next segregations that the data security industry needs to work on.

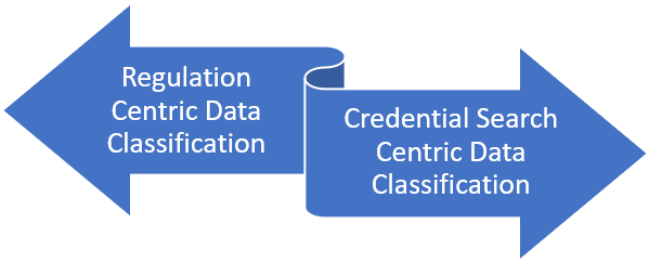


Fig 2

Changes to achieve segregation in practitioner approach and classification abilities

Two control centers with clear objectives

First and foremost, whether it is the same data management solutions, or separate ones, two separate control centers are needed for these two separate objectives of finding regulation centric sensitive data and compromise-prone credentials data. It means that the instances of the data management solution, especially the one responsible for data classification, needs to be deployed separately for these two use cases. Those deployments ideally need to be handled by different teams. For the regulation centric use case, the policy setting need to be done by the compliance teams while for the security related use case involving hunt for credentials in the wild, security operations teams need to set those up. Now, given that security teams are still involved with securing sensitive data, they also need to have access to the compliance team owned deployment. But the compliance team cannot have any access whatsoever to the security team owned deployment for data classification. This is plainly because, if there are keys or credentials out in the wild, only security operations should know about those to securely act on the containment of it. It is important to note however that this is more than just a separation of duties.

Priorities establishment

As discussed earlier, the obvious places to look for these two different content types are slightly different, but there are still some overlaps. In the overlap scenario, the high priority should be given to the security team classification project, or, runs than the regulation compliance team. This may be difficult to digest but here is an analogy. Consider a large area in which, you have scattered gold particles in many boxes which have locks and then, a very small number of boxes, you have universal keys which could open all the boxes. You do not know which boxes have keys and which boxes have gold particles. Wouldn't you first find all universal keys and secure those, instead of either looking for the gold particles first, or, in parallel? The same logic applies here.

Lastly, just like the search for credentials gets higher priority than searching for other regulation centric data, the remediation of any positive hits needs to follow the same priority. The criticality of positive hits for credentials should be much higher in a way which demands urgency for the responsible IT practitioners. The responsibility to mitigate a credential found in the wild falls jointly on security operations and other parts of IT. Security operations team based on "what was the key meant for," would lay out remediations actions which would then need to be carried out urgently by respective application and IT teams. This could involve resetting of passwords, rotation of keys, verifying audit logs for any compromise based on the key or the credential. But it could be much more comprehensive if there is any sign of lateral movement based on the credential or key.

Conclusion

Data classification has become important consideration for information security. Primary reason for that is the exfiltration type of attacks where sensitive data is stolen and threats are issued to release that, or, sell that if the ransom is not paid. Additional reason is to prioritize securing of sensitive and critical data while ensuring regulatory compliance is not breached. On the other hand, a lot of the attacks are happening due to credentials being compromised instead of "break ins." There was an initial instinct of IT practitioners to use disaster recovery for cyber incident recovery. In the same manner, apart from standard best practices and privileged access management, IT practitioners are using the regulation centric data classification workflows for finding any loose credentials or similar content to prevent credential compromise, or, at least make those difficult. As discussed in this paper, there are compelling reasons for both the use cases to be segregated for the loose credential search to be effective quicker.

References

1. Tuazon G. Most common passwords; latest 2021 statistics. Global Compliance Certification (GCC). November 2021. Available from: https://gccertification.com/most-common-passwords-latest-2021-statistics/ [Accessed 2022 Nov].
2. California Office of Information Security. Does Your Agency Implement Forced Password Changes? Access Control Information Sheet No. 7. May 2017. Available from: https://cdt.ca.gov/wp-content/uploads/2017/05/Does-Your-Agency-Implement-Forced-Password-Changes-Info-Sheet-7.pdf [Accessed 2022 Nov].
3. Miller M. What is Password Rotation and Why is It Needed? April 2018. Available from: https://www.beyondtrust.com/blog/entry/password-rotation-needed [Accessed 2022 Oct].
4. Brook C. Digital Guardian's Data Insider blog. Data

Classification as a Catalyst for Data Retention and Archiving. August 2017. Available from: https://dataclassification.fortra.com/blog/data-classification-as-a-catalyst-for-data-retention-and-archiving [Accessed 2022 Oct].
5. Platsis G. A journey in organizational cyber resilience part 3: Disaster recovery. SecurityIntelligence. September 2021. Available from: https://securityintelligence.com/articles/organizational-cyber-resilience-part-3-disaster-recovery/ [Accessed 2022 Oct].