International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Data Content Considerations for Identifying Security Measures

Anand Athavale

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

Abstract

This article covers data content considerations for identifying security measures to employ. Effective Data security measures help reduce probability and impact of a cyber or a compliance breach. In the real world, many of these concepts may be known, but those are often scattered among different IT practitioners. This article elaborates on data content considerations. If you compare data security to physical security, you can draw analogies very easily. For instance, the highest level of leaders and officials have more security. Lower-level staff and officials with less critical roles have less security. This tells us that what you are protecting itself has a lot of influence on what would be required to establish security measures. Data Content gives that "what is it that you are protecting" in essence. However, this article will also try and clearly differentiate data security from data privacy and other compliance and regulatory aspects, even though there is a good amount of overlap. In a way, this article will also help segregate the responsibilities of the data security team from the data compliance and privacy teams at the process and concepts level, while illustrating various aspects of data content.

Keywords: Data Loss Prevention, Information Security, Cyber resilience, Sensitive Data, Data Classification

Introduction

Data Content is the very thing that you are trying to secure by establishing data security measures [1]. In turn, this aspect becomes the most important aspect which cuts through other considerations of data security posture. But, before we go into this aspect deeply, we need to understand the security threats and separate those from data compliance and privacy regulation risks [2].

Data is an asset, no doubt. Everyone agrees to that. But just like any other assets, let's try and associate value to this asset and also differentiate between various data assets. Looking at more familiar tangible assets can help listing possible security risks. Consider a farming business. The business assets include a large piece of fertile land. On that fertile land, as of this day, there is a lush crop ready to cut. The farming business also has farming equipment which has a good amount of value. Then there is some storage for when the crop is cut.

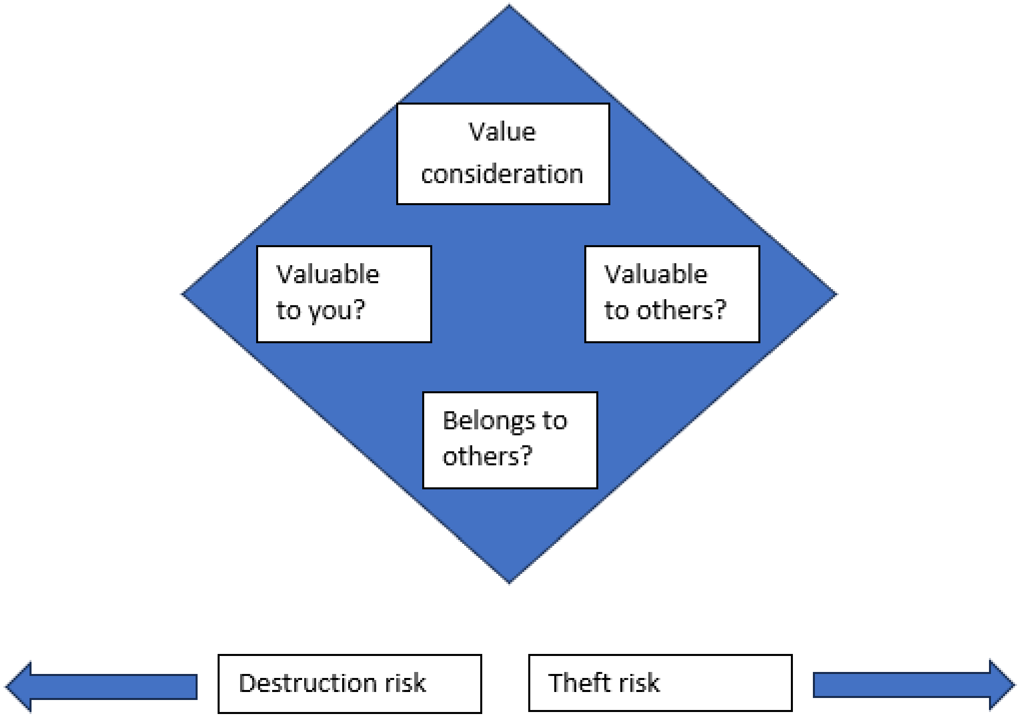
Now, what would be the risks to this business? Well, one can cut the crop partially and steal it and sell it somewhere. The other destructive actors may come and render the farming equipment inoperable. Those same actors can burn the crop once it is put in storage. They can also put something in the land which reduces or takes away the fertility of the land. They cans block the access to the land. Overall, depending on the value and the type of asset, the risks vary. Something of value to the business but not so much to the bad actors, is more prone to destruction. However, something may be of immediate value to the bad actors and that is more prone to theft.

We can take the data value discussion a bit further by considering another day-to-day life example. Consider a business that offers ornament services. That business stores jewels and price possessions for their clients, to maybe shine those, restore and many such things. Those possessions do not belong to their business. They just have those jewels for holding purposes. Now here, what bad actors would do is cause two different types of harm. One, by stealing those jewels and selling those elsewhere, and two, causing reputation damage and making the business automatically have violated the agreement between that business and their clients. This example brings us close to considering data sensitivity and privacy. For identifying security measures for data, it becomes important, not just because it is sensitive data, but the fact that it may belong to someone else and the loss of the same creates indirect business risk, unlike the immediate and tangible risk of equipment and land destruction. Now here is the separation point between data privacy and compliance and data security. Businesses must ensure having necessary processes and tools to avoid accidentally losing these jewels or damaging them. But data security teams are responsible for someone else like bad actors doing such type of harm or theft. With this background, we will need to investigate the data content types.

Data Content

Data Security measures need to consider the following aspects of data content itself. The lens we use for categorizing the data is related to the value and the type of risk. Data value considered here is not quantitative although if available, it can help in further prioritization of security measures when budgeting constraints apply. However, here data value is used to differentiate among various intentions of the bad actors and malicious insiders.

Data Value Risk



Internal operations risk

Here the data value is more significant to operating the peripheral applications for a business. The classic examples of these are the list of vendors who supply laptops, their invoices etc., internal news and communication content and similar things. If this data were to get stolen, there may be less impact, but if the data was corrupted or destroyed, then it could halt the business, or slow it down to cause productivity loss. Again, here, the security posture consideration is mainly for avoiding bad actors from doing it willfully vs. regular workers doing it accidentally. We should not assume that all violations of data protection policies are malicious. Knowledge workers may be unaware that an action adds risk, or may simply be trying to create shortcuts that make their jobs easier [2]. To consider a simple analogy, we don't let kids handle valuable objects because they don't know how to be careful with those. This type of responsibility falls under the data compliance team. On the other hand, we keep the valuables and currency safe under a lock and keycode from thieves and this is a type of responsibility for the security team. However, we don't share the keycodes with the kids either as they don't have a need for it and do not how to use the currency.

Business applications risk

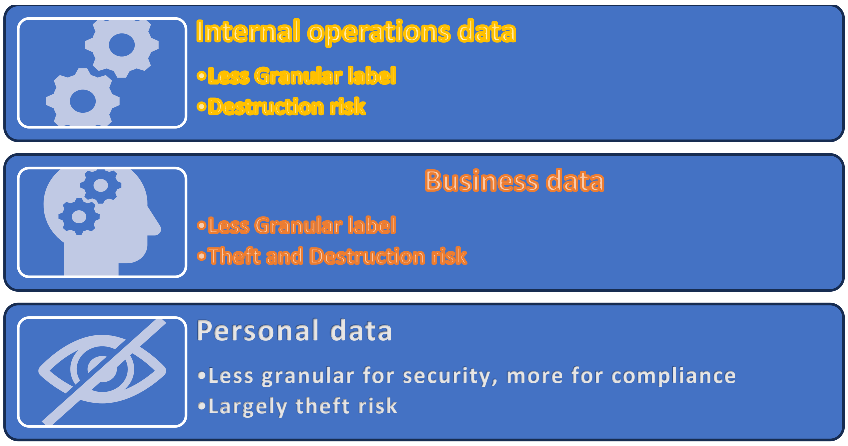
Here, data value is more significant to the core business applications which generate revenue. For a software company, their application repository could qualify as one such example. For a vendor manufacturing anything, the list of suppliers with the specifications of the supply could be another example. This is any data, which bad actors will mostly try to either destroy, or, hold ransom by encrypting it. Stealing this information also could be another way because they can threaten to hand off this type of data to their competitors, or simply threaten to make it public. In broader terms, this type of content represents intellectual property or trade secrets, which could be at risk for both, destruction/corruption or theft by the bad actors. The value of this type of sensitive data and the associated revenue loss both, could get significantly higher if destroyed or stolen [3].

It also represents data needed for the business applications to function. Typically, those would be application databases, or data stores. These too can be encrypted sometimes, but theft of those becomes more valuable to the bad actors than the destruction. What bad actors may do instead of destroying, is prevent access to that data itself and demand ransom to release the access. To bring back farming business example, this could be compared to blocking the access to the farm itself. Given that this deals more with cutting off access to critical data, it falls more under vulnerability management than data security measure identification process. However, sometimes attackers' intent could simply be to destroy the data, similar to a physical terror attack. Here, applications like banking could be targeted. Here, while banks hold money of their customers, without knowing current balance, banks can not allow further transactions like deposits and withdrawal bringing the core business itself to halt. The difference is, in such scenarios, the entire application along with the content is "obviously identified" to be critical. It may not require further qualification.

Personal data risk

To separate the personal data risk to be addressed by the data security team from the privacy risks to be addressed by the data privacy or compliance team, we can think of the following example in a crude way. For a car, some monitoring systems will make sure you have your eyes on the road so that you don't cause an accident. But another auto-sensing system will ensure that even if someone tries to jump in front of your vehicle, or pushes someone else, which can still make you liable for an accident, the car automatically stops. The former type of responsibility falls on the shoulders of the data compliance team, while the latter falls on the shoulders of the data security team. Here, the type of data at risk is mainly non-owned data of customers and clients. Examples of such data are credit card details of business customers, their name and address, or their social security or bank account numbers. Data security posture team needs to ensure that no suspicious persons or bad actors acquire access to such data and prevent stealing that and sending that from within the organization to external sources. This type of data is mainly at the risk of theft by the bad actors, rather than destruction, with the threat of exposing it if the demands are not met.

Data Content Labeling



Internal operations data
- Less Granular label
- Destruction risk

Business data
- Less Granular label
- Theft and Destruction risk

Personal data
- Less granular for security, more for compliance
- Largely theft risk

The other consideration for data content which is critical is labeling the content properly. Sometimes, it could be referred to as data classification, sensitive data discovery or just tagging. There are a few aspects within labeling which need to be looked at.

Data content labeling is of utmost importance from a data privacy perspective too. However, the granularity and type of sensitivity differs between data privacy and data security posture process. Data privacy and compliance related labeling requires it to be wider and more accurate. Data security posture is less wide and could be fuzzy from a prevention perspective. [4] On the other hand, when a data related security incident takes place, then, the data content labeling needs to be more specific.

Data Privacy related data content labeling needs to be more granular. For example, it is not enough to say that this specific share contains PII data. It is not enough because that share itself may be accessible to a lot of admin and all sorts of persons within the organization. But if the same level of access is present for the folders or the files containing the credit card data, or say, social security numbers, bank account numbers etc. it would create a violation for the regulations like GDPR, CCPA [5]. There are also healthcare regulations like HIPAA which are even stricter about who can have access to someone else' health data. Hence, data content labeling for data privacy requires more granular labeling at item level, like a file, or an object. These aspects are also important in data exposure control and monitoring.

Now, let's discuss the data security point of view for data labeling. Here, data security would need to know whether the share has sensitive data. If it does, then they need to know whether it is the data with data value risk of type internal, business or personal data risk.

They would probably consolidate the credit card, Social Security numbers, and other such data as PII data and consider those as personal data risk, more prone to theft than destruction. Hence, where possible, they would employ techniques like encryption, especially for secondary copies of that share, stored for the purpose of backup. Encrypting the actual data would be a challenge because then the real users of that data, who need access to such data to do their jobs, would have issues. Of course, if the data type itself is structured, and only the application has access to it, like databases, then they can still consider data encryption of the actual data. This would be okay because only the database application user would have meaningful access to that data. The application which interacts with that data itself would have limited access. So, the bad actors need to not only break into the network, but also get hold of application user credentials to see that type of data.

If the sensitive data is of type trade secret, intellectual property or code, the InfoSec team would consider it as a business application data value risk and guard it for both theft and destruction.

If it is internal but sensitive data, the InfoSec team would mainly guard it from destruction but they don't need granular data labeling like vendor list and invoices. But they still need to consider it at higher risk than other non-sensitive data like internal news articles.

Now, if they were handling a security incident, then they need more accurate data content labeling and also more granular to know whose data was stolen etc. in case of personal data theft. In data classification terms, this gets referred to as Named Entity Recognition [6]. This explains why data content labeling matters from both angles, one, being more accurate and specific and second, from a level or a granularity aspect.

Data content labeling becomes even more challenging for communication systems due to sheer volume and other factors [7]. Data content labeling for messages in a primary communication system is a mammoth scalability challenge. Here, in the absence of communicators themselves marking it as privileged and confidential, they need to consider the senders and recipients of the communication and try and elevate the security accordingly. For example, if it is an email from CEO to employees for a festival greeting, or wishing Happy New Year, it has low risk if leaked. But if it is a communication to his direct staff about a change in strategy, or, a new product plan, that has high business application data value risk. Also, there are certain overlaps with data compliance systems which need to be considered for communication applications. This is a very corner case scenario, however, there is a possibility for malicious actors forcing a non-compliance or disruption. Financial services for example, have to comply with trade reconstruction and surveillance requirements. However, to reduce costs and surveillance volumes, they sometimes write discard rules. Now, if a malicious actor was able to change those, many required emails with retention requirements may get discarded from archiving and eventually deleted. On the flip side, phishing emails may get archived and would have a risk of being recalled and then opened and acted upon. This is different from sensitive data labeling and almost the opposite in fact. But this too is an important consideration for security measure identification.

Data content labeling often gets ignored when considering log files, typically generated for diagnostic purposes. While the expectation is for the application developers while writing code to ensure not dumping data into the logs, it is something one hundred percent needs to be considered when identifying data security measures from data content labeling aspect.

The last but not the least aspect of data content labelling is the frequency. While most organizations struggle to classify the content even once, the security cannot rely on a one-time labelling. Here is why. Let us take an example of three cupboards in a house. One only has clothes and the other two have a lot of valuables in it. If the content is assessed only once, and accordingly only the two cupboards with valuables are always locked, then if someone accidentally puts a valuable in the cupboard with the closets, it remains non-secure. The same is possible with data repositories or items. Just because the first classification pass came back clean does not mean a few sensitive data items may not be placed there later. Data content labelling or re-labelling needs to be frequent because changed labels can trigger changes in security requirements.

Conclusion

Data Content is indeed the most critical of the aspects, not only having the most bearing on data security measures, but also touching various other security measure considerations such as data location, accessibility methods, exposure control and data types. Data content is the essence of what data security measures intend to secure. If a large number of health records get stolen, or destroyed, the most important part is that they were health records, and not as much whether they were under a folder in a share, or in a database, or somewhere else. But, for preventing it, different considerations of data type and locations do matter. This is why data content cuts through a lot of aspects of the data security measures identification process. Again, a distinction needs to be made here. For example, putting passwords into data content or keys for APIs may also come under data content labeling. But that is more closely related to vulnerability identification and that is a whole different topic. This is not to say that it is not related to security. Instead, it falls more in the realm of monitoring for deviations from best practices, which in this case would be not to store keys and passwords in the plain text. Those practices are more mature although needing constant monitoring.

References

1. Ashley Madison, 5 Steps to Prevent Sensitive Data Loss (September 2015), https://www.digitalguardian.com/blog/5-steps-prevent-sensitive-data-loss , (Jan, 2020)
2. [DataIQ with Tealium], General Data Protection Regulation 2017, Identifying its impact on marketers and the consumer’s moment of truth (2017), https://tealium.com/download/dataIQ_GDPR_report, (Jan 2020)
3. [Business Insurance Insights experts, Unknown], 4 safeguards to manage intellectual property risks in manufacturing, Liberty Mutual Insurance, (March 2019), https://business.libertymutual.com/insights/4-safeguards-to-manage-intellectual-property-risks-in-manufacturing/, (Dec, 2019)
4. Andrew Warland, Changes to security classification and records retention in Office 365 (Records About the World), (March 2018) https://andrewwarland.wordpress.com/2018/03/09/changes-to-security-classification-and-records-retention-in-office-365/, (Jan, 2020)
5. Jeewon Kim Serrato & Daniel Rosenzweig, GDPR, CCPA and beyond: Changes in data privacy laws and enforcement risks to monitor in 2019, (February 2019), https://www.dataprotectionreport.com/2019/02/gdpr-ccpa-and-beyond-changes-in-data-privacy-laws-and-enforcement-risks-to-monitor-in-2019/ , (Jan, 2020)
6. Dipanjan Sarkar, Named Entity Recognition: A Practitioner’s Guide to NLP, KDnuggets, (August 2018), https://www.kdnuggets.com/2018/08/named-entity-recognition-practitioners-guide-nlp-4.html, (Jan, 2020)
7. Vienna, THE IMPACT OF EMAIL CLASSIFICATION ON YOUR BOTTOM LINE, Critical.io (November 2019) https://www.cortical.io/static/downloads/email-classification-white-paper-2019.pdf, (December, 2019)