

How to Design a Cloud Data Protection Strategy

Evolving Data Protection for
Continuous Risk Management

Over the last decade, data has become one of the most powerful tools an organization can employ

Data is the driving force behind innovation, efficacy, and ultimately, the success of an organization. The ability to collect, analyze, and interpret data provides an organization with insights that can and should guide strategic decisions, both for internal and external matters.

At Netskope, we believe in that thesis and that data is just not only a tool, but also the key value-creation asset for organizations. Organisations service customers, and as a result, business processes and customer interactions generate the data that is needed to facilitate business operations. Without data, there is not a business process that can be executed and consequently, a service can not be provided to a customer. Data must be protected so as to maintain and ensure competitive advantage, ensure the privacy rights of customers, ensure the privacy rights of employees, and ensure the stability and accuracy of business operations.

Data is also growing at an exponential rate, which in turn is making it harder to manage and secure. As a result, a new way of thinking about data security is required.

Data protection is the process of protecting data throughout its lifecycle, from data creation, processing, modification, transmission and destruction. The old ways of thinking about data protection aren't fit for the era of digital transformation.

By 2025 IDC says worldwide data will grow 61% to

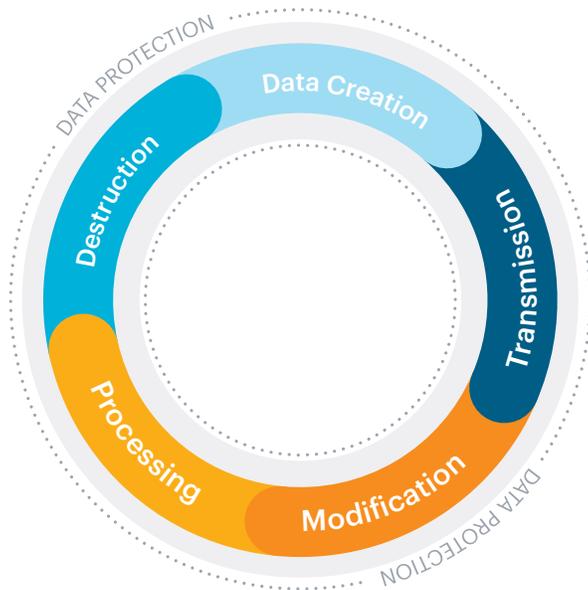


Figure 1. The Data Lifecycle

DATA PROTECTION DRIVERS

Modern data protection has five key drivers, all of which an organization must seek to understand. These drivers equally apply to Cloud and non-Cloud related data and should form the basis of any robust data protection strategy.

These five drivers are:



KNOW WHERE THE DATA IS STORED / LOCATED

Determine what information is stored locally, in the cloud or at a third party and understand jurisdictional data privacy requirements to help determine true digital risk.



KNOW THE SENSITIVITY OF THE DATA

Understand the sensitivity of the data, the importance of the data to the business, and the likely impact to the business should this data be made available to non-authorized parties (including being made public) or be modified or corrupted.



KNOW THE FLOW OF THE DATA THROUGH THE ECOSYSTEM

Understand where the data is flowing and ensure only authorized access is permitted and that data is not transferred to non-authorized or unprotected environments.



KNOW WHO HAS ACCESS TO THE DATA

Assess third party suppliers, partners and understand who has access to the data. Determine if the right identities (machine and person) do have access, and determine who should not have access to the data.



KNOW HOW WELL THE DATA IS PROTECTED

Know what controls are being used to protect the data. Are they operating as designed and are they operating effectively?

ENVIRONMENTS AND STATES

Today, data exists in two broad environments: on-premise and in the Cloud. Digital Transformation is seeing a dramatic shift with data rapidly moving from on-premise to cloud, especially public cloud. Organizations are looking to cloud environments to reduce operational costs, enhance user experience (and performance), and to make it easier to collaborate with partners and third parties. Netskope research shows that most organizations now have greater than 50% of their web gateway traffic related to Cloud services and applications—a dramatic shift from the days where data was contained within the data centre, on-premise. Furthermore, 83% of users use personal app instances on managed devices and upload an average of 20 sensitive, corporate files to personal instances each month.

This trend will continue to grow and the need for an organization to manage their own data centres will continue to decrease. As a result, an organization needs to establish, if they have not already, a **Cloud Data Protection Strategy**.



Most organizations now have greater than 50% of their web gateway traffic related to Cloud services and applications

Within these two environments, on-premise and Cloud, data is used in three states: in motion, in storage, and in memory. Each permutation of environment and state will require a different approach to adequately manage the risk, within the risk appetite of the organization, to that data asset.

WHAT IS THE DIFFERENCE BETWEEN ON-PREMISE DATA PROTECTION AND CLOUD DATA PROTECTION?

Protecting data in the cloud requires a different approach to that used when protecting on-premise data. The data's presence on the cloud increases the level of exposure as any access to it means it is sent via the public internet, whereas on-premise data could only be accessed via intranet for most internal operations.

Additionally, the universal availability and ubiquitous accessibility of this data has facilitated the shift towards increased use of Bring Your Own Device (BYOD), as well as Work-From-Home (WFH) solutions—trends that also expand the attack surface and eliminate the previously discernible “perimeter” that was the front line for controls in an older era.

95% of organizations allow personal devices in some way in the workplace.

(Source: Cisco)

Physical security, backups and disaster recovery are still an important aspect of data protection, but when considering cloud-stored data, the responsibility has (typically) been adopted by Cloud Service Providers (CSPs). This relationship brings about a shared security model which must be fully understood and agreed with their service provider so that such services are implemented appropriately for every party's needs.

Unlike on-premise data, cloud data is accessed and manipulated through API requests and JSON—the language of the Cloud. It is therefore imperative that the security tools that are being used are capable of interpreting the language of the Cloud. This will allow for a proper application of controls and for visibility that traditional on-prem security solutions simply cannot provide due to their being blind to Cloud traffic. This also means the control stack needs to be able to decode JSON and to interpret API requests. Not having this level of inspection capability will mean that context—the understanding of aspects such as data classification, user actions, inter-application transactions, anomalous user behaviour and device type—will not be able to be appropriately understood. If context isn't understood, policy or control actions will be deficient in enabling the organization appropriately to realize the value of Cloud services, inhibiting the Digital Transformation strategy of the organization.

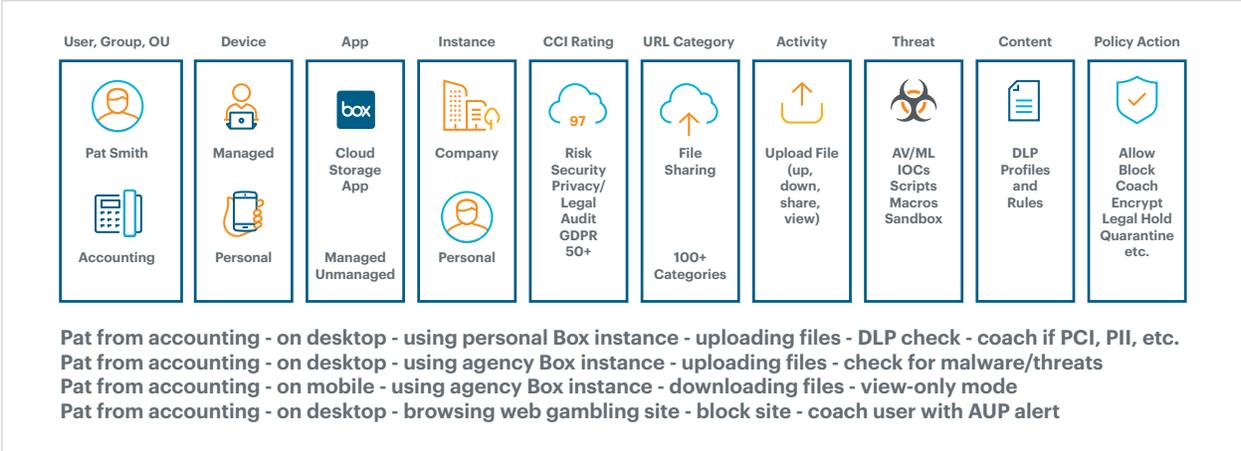


Figure 2. Cloud XD diagram

Furthermore, with cloud based systems, organizations are exposed to ephemeral security boundaries. Unlike on-prem systems with static IP addresses, and a constant number of resources, in the cloud, new resources are continually started and stopped. The resources are far more dynamic in nature and simple misconfiguration can lead to excessive exposure. In fact, Netskope Threat research has indicated that misconfiguration in IaaS and PaaS environments is currently the leading factor that is contributing to the rise in data breaches.

Having data in the cloud can increase third party risk as you are required to entrust the data with the CSP and are reliant on them having the appropriate security measures in place to not only protect your data but to also assure that you meet any regulatory compliance requirements. However, this can improve the overall technical risk of the organization as the CSP may be far more skilled and capable than the organization itself to securely manage the environment, especially with basic hygiene such as patching and maintaining technology currency. Moving to a CSP can greatly enhance these fundamentals where some organizations have typically struggled.



In 2020, errors caused 22% of the breaches and misconfigurations was the fastest growing breach cause.¹

¹Source: Verizon 2020 Data Breach Investigations Report

MULTI-CLOUD ENVIRONMENTS

The idiom “don’t put all your eggs in one basket” has always been relevant to data security. There is a tradeoff between when deciding whether or not to split data or systems across multiple CSPs. On one hand, the split increases attack surface, but on the other hand, splitting data and systems across separate networks reduces the impact of a breach or even a service delivery failure by a certain CSP.

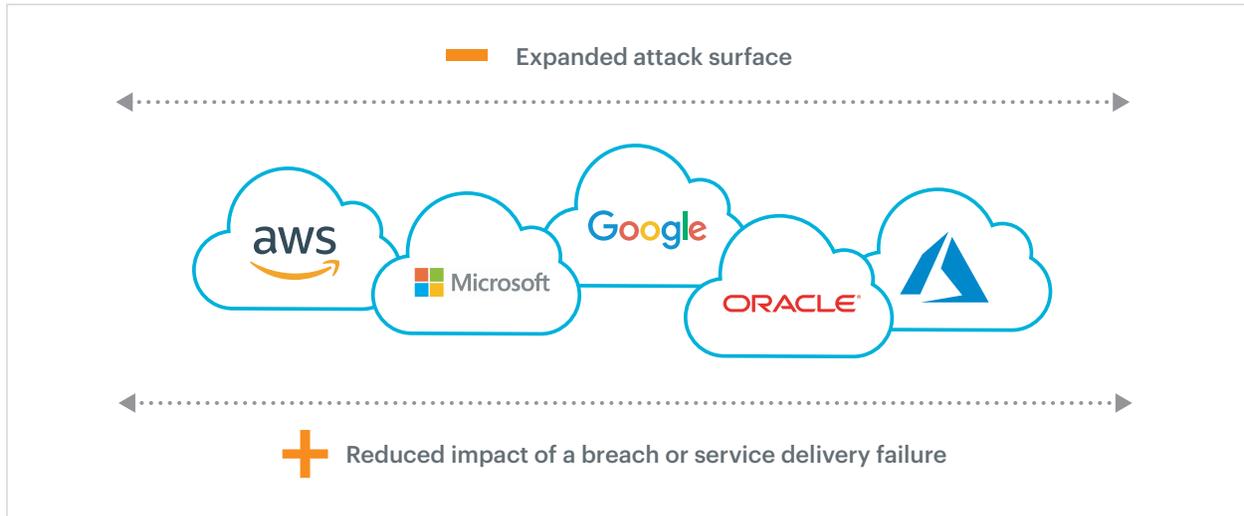


Figure 3. The tradeoff of splitting data across multiple CSPs

Having systems split across multiple CSPs yields multiple benefits such as access to a greater variety of tools, better cost optimization, improved redundancy and a lower impact in case of a breach. Existing in a multi-cloud environment does however come with an expanded attack surface and so, it is imperative to take the necessary steps to secure it appropriately. Cloud-based SaaS that is used by an organization further increases the attack surface and should also be considered in the security plan.

The shared security model increases in complexity for each cloud service that is used. Each service presents its own unique vulnerabilities that an organization’s security team must consider. Furthermore, an organization must also do its due diligence and understand the security posture of the service provider to see that it meets both the organization’s risk appetite and any regulation compliance requirements.

When running in a multi-cloud environment an organization should:

1. **Identify all the cloud services used within the organization.** Business units often adopt SaaS solutions that help them achieve their objectives without considering the security of the data that flows between the organization and the service provider. Identifying these services as early as possible is necessary in order to ensure that appropriate security controls are applied.
2. **Ensure security settings and policies are aligned across different cloud deployments.** Doing so reduces the complexity of the security environment, provides consistency for the organization’s employees.

3. **Acknowledge that the rapid deployment and turnover of SaaS solutions in the workplace make it impossible for security teams to secure each service to the desired standard.** Additionally, the constant changes increase the chance of a misconfiguration when handled manually. In order to solve for this, organizations need to implement automated security processes as they provide several indispensable benefits
 - a. They reduce/eliminate the possibility of errors/misconfigurations—increased consistency
 - b. They deploy policies and monitoring tools far more quickly—increased efficiency
 - c. They perform the tedious/repetitive tasks allowing the security teams to focus their efforts elsewhere
4. **Use effective tools that allow for effective visibility across the cloud services used by the organization.**

A NOTE ON INFORMATION SECURITY DOMAINS AND THE INFORMATION SECURITY DISCIPLINE

The following diagram illustrates all facets and/or domains of the traditional information security discipline that will be required (including but limited to) to establish a holistic approach to the data protection.



Each of these domains is a discipline in its own right. The organization will need to develop capabilities in these various domains which will then be used to enable an approach to implementing a data protection capability appropriate for the organization based on the environment, the state and driven by the five data protection drivers described above.

A PRACTICAL, STEP-BY-STEP APPROACH TO CLOUD DATA PROTECTION

Step 1: Know where the data is stored and located, aka Data Discovery

This is the process of discovering/detecting/locating all the structured and unstructured data that an organization possesses. This data may be stored on company hardware(endpoints, databases), employee BYOD, or the Cloud.

There are many tools available to assist in the discovery of data (for both in transit and in storage) and these vary between on-prem and cloud related data. This process is intended to assure that no data is left unknown and unprotected. This is the core of creating a data centric approach to data protection as an organization creates an inventory of all of its data. This inventory is critical input to a broader data governance strategy and practice.

Information assets are constantly changing and new assets are added that will make any static list out of date and ineffective almost immediately. When establishing the process for data discovery ensure to use automation. It is the only way you can keep an active view of your information assets and be able to effectively manage the risk.

Step 2: Know the sensitivity of the data, aka Data Classification

Once the data is discovered, that data needs to be classified. Data Classification is the process of analysing the contents of the data, searching for PII, PHI and other sensitive data and classifying it accordingly. A common approach is to have 3 or 4 levels of classification, typically:

3 level policy:	4 level policy:
Public	Public
Private / Internal	Private / Internal
Confidential	Confidential
	Highly Confidential / Restricted

Once a policy is created, the data itself needs to be tagged within the metadata (this is the implementation of the data classification policy). Traditionally, this has been a complex and often inaccurate process. Examples of traditional approaches have been:

- Rule based
- RegEx, Keyword Match, dictionaries
- Finger Printing and IP Protection
- Exact Data Match
- Optical Character Recognition
- Compliance coverage
- Exception management

Approaches to data classification have evolved and organizations must leverage new capabilities if they are to truly classify the large volume of data they create and own. Some examples are:

- Machine Learning (ML) based document classification and analysis, including the ability to train models and classifiers using own data sets using predefined ML classifiers (making this simple for organizations to create classifiers without the need to complex data science skills). ([See this analysis from Netskope.](#))
- Natural Language Processing (NLP)
- Context Analysis
- Image Analysis and classification
- Redaction and privacy

These approaches must have the ability to support API-based, cloud-native services for automated classification and process integration. This allows the organization to build a foundational capability to use process and technology, including models, together to classify data which then becomes a data point on additional inspection if needed. The result is to provide a real time, automated, classification capability.

Classification escalation and de-escalation is a method commonly used to classify all discovered data. For each data object that has not been classified, a default classification should be applied by injecting into the metadata the default level of classification (for example, if not classified, default to confidential or highly-confidential). Based on several tests or criteria, the object's classification can slowly be escalated or de-escalated to the appropriate level. This coincides with many principles of Zero-Trust which is fast becoming, and will be, a fundamental capability for any Data Protection Strategy. (More information on Zero Trust can be found further below [here](#) and in Netskope's document [What is Zero Trust Security?](#))

A Note on Determining ‘Crown Jewels’ and Prioritization

Data classification goes a long way in helping an organization to identify its crown jewels. For the purpose of this conversation, “crown jewels” are defined as the assets that access, store, transfer or delete, the most important data relevant to the organization. Taking a data-centric approach, it’s imperative to understand the most important data, assessing both sensitivity and criticality. This determination is not driven by data classification alone.

A practical model to determine the importance of the data is to take into account three pillars of security—Classification, Integrity, and Availability—with each assigned a weighting (1-4) aligned to related policies or standards. A total score of 12 (4+4+4) for any data object would indicate the data is highly confidential, has high integrity requirements, and needs to be highly available.

Here is an example of typical systems in use by an enterprise and typical weightings.

<p>Classification: Highly Confidential = 4 Confidential = 3 Internal = 2 Public = 1</p>	<p>Integrity: High Integrity = 4 Medium integrity = 3 Low integrity = 2 No integrity requirement = 1</p>	<p>Availability (being driven from the BCP and IT DR processes): Highly available = 4 RTO 0 – 4 hrs = 3 RTO 4 – 12 hrs = 2 RTO > 12 hrs = 1</p>
--	---	---

		Classification	Integrity	Availability	Weighted Score
Banking		3	4	3	10
Procurement		3	2	2	7
Reporting Database		3	3	1	7
HR System		3	2	2	7
Marketing Databases		2	2	1	5
General Ledger		3	3	2	8

An organization can set, based on risk appetite, a total score of 12 for any data object, which would indicate that the data is highly confidential, has high integrity requirements, and needs to be highly available. An organization can set, based on risk appetite, what score determines the crown jewel rating. In addition, this enables the organization to prioritize controls and where needed, remediation activity, in a very logical and granular way. The score can then be applied to the applications, systems, and third parties that use that data, creating a grouping of assets (applications, systems and/or third parties) that would indicate crown jewel status (or not).

Step 3: Know the flow of the data through the ecosystem—be the inspection point between the user and the data.

Data is like water—it seeks to be free. As such, an organization needs visibility and must be able to inspect all traffic flows to identify the following:

1. What data is in motion, based on criticality and sensitivity (data classification)?
2. Where is it moving from and to? Do these source and destination environments reconcile with the discovery process or have we identified unknown data repositories that need to be investigated? The latter point is one that should not be overlooked. Business processes will change and with that, data flows will change. It's imperative that an organization continuously monitors for this and takes the appropriate action where new flows are identified. Typically, these actions are:
 - a. assuring the security controls or posture of the newly identified source or destination which could be a new SaaS application (or instance of that SaaS application) meets the required security standards
 - b. assuring the security controls or posture of a new third party (and consequently the security of the third parties environment) that now has access to the data meets security and privacy standards
 - c. Confirming that this data flow is appropriate and does not indicate a compromise or identifies a broken business process or user actions that need to be rectified.
3. Can we determine any geographical and/or jurisdictional data movement that may introduce privacy or regulatory requirements?

By creating a cloud-native inspection point between the user and the data, that can interpret the language of the Cloud, the organization has now created a data discovery capability to identify all cloud related data and can then leverage those capabilities discussed in step 2 to automate and highly-accurately classify the large volumes of data and doing this in real-time when the data is in use and in motion. Furthermore, an organization needs this same automated classification capability for data at rest. This way, there is a two-pronged approach to ensure that all data is discovered and classified in an automated fashion, and naturally, the automated classification engine needs to be consistently applied over both sets of data, that being at rest and in motion.

This also enhances real time analytics and visualization, both of which are key to data protection and are fast becoming new instrumentation for Security Operations teams. These analytics are not a replacement for SIEM, but they do help redefine what is needed for effective security analysis, response, and third party risk management. This capability becomes a foundational component necessary to ensure that an organization has all the information and intelligence at hand, in real-time, to enable it to understand impact and dependencies for Cloud data, making informed decisions and taking action in a timely manner.

Step 4: Know who has access to the data—effect more visibility

Being the inspection point between the user and the data not only allows an organization to understand where the data is flowing to and from, but also gives visibility into what identities (machine or user) have access to the data.

Being able to determine this enhances the Identity and Access Management capability of an organization. This information can be used to validate existing IAM practices, such as any Role Based Access Control definitions as an example, in addition to identifying anomalies that will require investigation and potentially corrective action. This will apply to both end user and privileged access.

Having this visibility enables an organization to minimize the access to data and applications which in turn minimizes the exposure and thus risk imposed. Fine grained access control is imperative to minimize opportunities for attacks.

Step 5: Know how well the data is protected—be the policy enforcement point between the user and data

In storage: With respect to Cloud related data, it is important that an organization scans and assesses the security posture of Cloud environments, such as AWS, Azure, GCP, to verify the security configuration of these environments and assure that data is not arbitrarily left exposed. Misconfiguration of Cloud environments is a leading cause of data breaches. Security configuration compliance monitoring has been a common capability for many years for on-prem infrastructure and this naturally needs to extend to the Cloud based IaaS and PaaS services.

In motion: With respect to Cloud related data, an organization needs to establish a capability that creates a Policy Enforcement Point (PEP) between the user and the data. (This is a logical extension to the inspection point described in Step 3.)

The organization is now equipped with a number of data points allowing them to make policy decisions with context, thus enabling a true, fit-for-purpose, risk based approach to the application of controls. As an example (and recommendation), with an understanding of the criticality and sensitivity of the data (derived from classification), an organization can prioritize the protection of the highest classification level of data and work their way down to the second lowest classification. Note that the lowest level is typically classified as “Public” and should not warrant many protections, if any at all.

A Risk-Based Approach to Data Protection Policies

There are two main approaches to creating data protection policies: a content-based approach and a purpose-based approach.

A Content-Based process is one in which an organization identifies sensitive types of content (PHI, PII, etc.) and applies the appropriate policies that help compliance with internal policy or regulation. This is also the faster and broader process which will apply blanket policies based on levels of classification.

Content-based policies, when planned correctly, should be fairly strict so that once a piece of data is given a certain level of classification, any access/transfer/editing/deletion can only be done under the right circumstances. This may result in policies that block legitimate actions due to their broad nature, however it is better for sensitive data to be over-protected than under-protected.

In order to combat the rigidity of content-based policies, an organization can conduct data auditing. Data auditing is the slower, and more granular process in which an organization can identify the purpose of specific objects of data, determine what additional data protection requirements need to be granted (if any) to allow the right people to access and manipulate the data in a legitimate manner.

Conditional Authorization

Conditional Authorization leads to safer access-control rules as it regulates permissions, based not only on the digital identity trying to access certain resources but also on the environment (IP address, time-of-day, location, device, etc.). These controls can help limit malicious users from executing certain actions even if they have managed to compromise the authentication process.

Conditional authorization is almost inherent to Attribute-Based Access Control (ABAC) where the policies and rules are based on four sets of attributes; subject (digital identity), resource (the data being accessed), action (edit, read, execute, delete, etc.) and environment (IP address, Cloud service, device, etc.).

Looking at the various approaches to how policy can be defined and implemented for Cloud related data, it is fundamental that an organization creates the capability (described above) that enables an **inspection and policy enforcement point between the user and data** that provides context as to how the data is being used. This inspection and PEP needs to dig deep to provide visibility into the device, the SaaS app instance, how the user is interacting with the data within the application or environment (specifically, what commands are being issued such as delete, edit share, etc.) and overlaying this with normal and anomalous behaviours.

Controls

There is a core set of controls that need to be established that will need to be applied as a result of the policies that have been defined. These controls will equally apply to the environments and states previously defined. They are:

1. Data Encryption
2. Data Masking
3. Data Tokenization
4. User Access Rights Management including Digital Rights Management

These are mature controls in their own right today and there are market-ready solutions available. But what is important is that these controls need to be able to be applied to the end-point, web traffic, email, IaaS, PaaS, SaaS, non-cloud based applications, and messaging applications—at a minimum. Clearly, any new channel of data flow (identified through Step 3) will also need to be addressed.

A Note on Endpoint Data Leakage Protection (DLP)

The end-point (laptop, PC or server) introduces 3 exfiltration scenarios that need to be addressed so that the data will be kept within the realms of the management and controls capabilities of the organization, as described in this paper. These three scenarios are removable media (e.g. USB), printing, and Copy & Paste / clipboard.

Removable Media

Any attempt to transfer to removable media (usb thumb drive, external hard drive, etc) will need to be logged and either blocked or have encryption enforced. When enforcing encryption, the key management process should be integrated into any end-point (or enterprise) data protection capability so that keys are easily managed, shared and recoverable.

Printing

An organization will typically want to be able to control where local printing is allowed, especially when off premises, ensuring that there is at least an audit trail or log of what is being printed, by whom and what the data classification is, of the data being printed. Endpoint security will need to be able to control who can print locally, and restrict unauthorized users from printing.

Copy & Paste

Data exfiltration can also be achieved by users copy and pasting between applications via the clipboard feature. The same data protection policies should equally apply and have the capability to be implemented for this scenario. This includes, but is not limited to, the ability to block copy and paste based on device type, data classification and/or user.

Where Zero-Trust Intersects with Data Protection

Data is the value creation asset of an organization and therefore, the protection of this asset is paramount. We have discussed the need to take a data centric approach and this manifests itself through the implementation of many services and capabilities across the security domain. Data is undeniably central to this approach (see Figure 4 below).

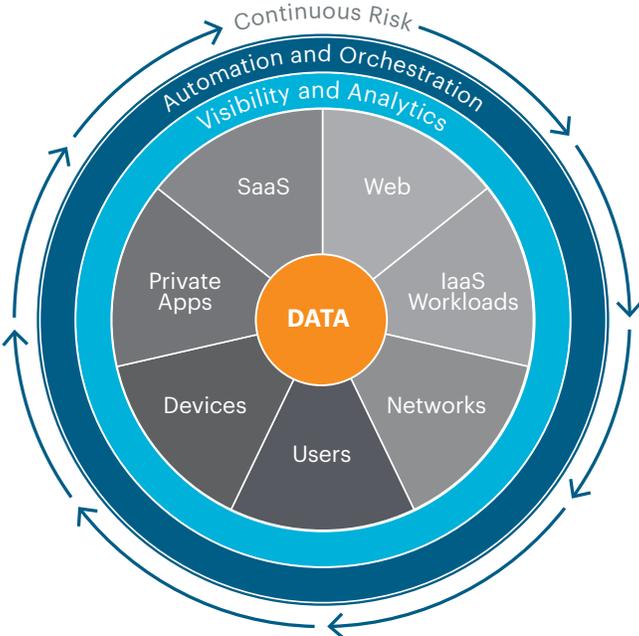


Figure 4. Continuous Risk Management

The concept of ZeroTrust is supported across the industry, but the true value of Zero Trust principles is how, when properly applied, they enable creation of an ecosystem of capabilities and controls that are continually adapted to provision the access needed at a point in time taking into account the risk posed at that point in time—in other words, real time decisioning. What is different here to how this has been approached historically, is that now we have more insights into the environment in which our users and third parties operate than ever before. We can now have deeper—and more importantly, continuous—visibility into user behaviors, data sensitivity and criticality, the end device, threats prevalent in the environment, and an understanding of the risk posed by the application in use.

Zero Trust follows a “block-by-default” scheme in which access and action are only permitted if they have been explicitly allowed. The decision to allow the action or the access is driven by a risk calculation that is derived from those many points that we now have available to use. These data points are continuously assessed and the policy is continually updated based on the calculated risk. By taking this approach, we are continuing minimizing the attack surface inherent in the data assets. We have knowledge of the interplay between user, device, app, and data, which enables teams to define and enforce conditional access controls based on data sensitivity, app risk, user behavior risk, and other factors. A net result is more effective security overall, thanks to **continuous risk management**.

The Future of Data Protection

With the anticipated exponential growth in data, the ever increasing interconnected world being serviced by higher speeds and more devices growing at unprecedented rates, the challenges for data protection will not only continue, but will also become more challenging. However, there is hope.

We will continue to see significant advances in AI/ML and Natural Language Processing (NLP) as a means to automatically classify data in near-real time. Consider PII data classification from an AI/ML perspective. The difficult part of PII detection is accurately attributing a sensitive piece of information, e.g. date of birth of an individual. This is due to the fact that most common words in the English language can be real first or last names of people. This is the challenge in identifying subjects when processing documents. Named Entity Recognition (NER), an application of NLP, is an effective way to locate and classify named entities like people names, addresses, places, organizations, dates etc. in unstructured text. In the future, we will see increased application of techniques like NER to accurately identify PII information, which is key for meeting the ever-growing regulations for protecting citizens' personal information. This approach will not be limited to PII and will be used across all data types in order to classify data.

Consent management will become more of a complex and important issue than it is today as privacy obligations continue to evolve putting more onus on the collector of the data to ensure that the consumer's consent is not only given but technically enforced and that the data collector can substantiate that at any point in time, as requested by the consumer.

This leads into API protection. As APIs become richer and richer, passing data from third parties to fourth, fifth, and -nth parties, we are going to see improvements in API security protection technologies that are identifying the flow of information between systems and mapping dependencies between services.

Lastly, with continued adoption of zero trust, data protection becomes more and more important. Looking for new technologies and processes for protection of data and devices with advanced zero trust (as described earlier) at the heart of the architecture. All leaders should stay current on advanced technologies that provide real time visibility into the daily interactions with organization data. The key to success is being able to gather and analyze telemetry such as data sensitivity, identity, application, device, source and destination location, device, and user behavior in real time and use an advanced risk engine to enforce the appropriate actions (allow, deny, restrict, redirect, etc.) is the deployment of a true zero trust architecture.

The more telemetry you are able to analyze, the better risk decisions you will make. The result is the ability to find the right balance of enabling the business, managing the risk portfolio, and protecting data—increasingly, the most important asset you have—wherever it lives and is accessed.

Netskope, the SASE leader, safely and quickly connects users directly to the internet, any application, and their infrastructure from any device, on or off the network. With CASB, SWG, and ZTNA built natively in a single platform, Netskope is fast everywhere, data-centric, and cloud-smart, all while enabling good digital citizenship and providing a lower total-cost-of-ownership.

To learn more visit, <https://www.netskope.com>.

©2022 Netskope, Inc. All rights reserved. Netskope is a registered trademark and Netskope Active, Netskope Cloud XD, Netskope Discovery, Cloud Confidence Index, and SkopeSights are trademarks of Netskope, Inc. All other trademarks are trademarks of their respective owners. 05/22 WP-450-2