NextGen DLP: Data Misuse Protection

Mark Settle and Sid Trivedi

About the Authors

Mark Settle is a seven time CIO, two time book author and three time CIO 100 honoree. He has led IT teams in public and private companies that supported global business operations, software development and online Ecommerce. His most recent book is *Truth from the Valley, A Practical Primer on IT Management for the Next Decade*.

Sid Trivedi is a Partner at Foundation Capital where helps to lead the firm's focus on cybersecurity and IT investing. Prior to Foundation, Sid was investor at Omidyar Technology Ventures and STG Partners. He started his career as an investment banker focused on technology M&As and IPOs at Barclays Capital. Outside of work, Sid serves on the advisory council for Entrepreneurship at Cornell, the California Israel Chamber of Commerce and is a board member of Cornell Venture Capital, which he co-founded.

Acknowledgements

Tom Baltis, Brian Castagna, David Hahn, Declan Morris and Jared Thorkelson provided thoughtful reviews of an earlier version of this report. Many thanks for their insights and suggestions. Special thanks is also due to Mike Rothman at Securosis and Joerg Fritsch at Gartner Research for their thought leadership on contemporary issues confronting the data security industry.

Disclaimer

References to the products and services of specific commercial vendors that appear within this report areincluded solely for illustration purposes. They do not represent personal endorsements on the part of the authors.

Table of Contents

Introduction

Data loss prevention is a fundamental security principle within the IT industry. The term DLP has been used for over two decades to refer to the framework of policies, procedures, tools and teams that enterprises employ to prevent the inadvertent loss or intentional theft of sensitive information.

DLP tools and procedures are used to protect intellectual property, prevent exposure of business-critical information and ensure compliance with financial and privacy regulations. DLP safeguards have historically been designed to combat 'insider threats' posed by individuals who may release sensitive information through negligence or malice using legitimate or hijacked authorization credentials.

Conventional DLP tools were designed to detect sensitive information crossing some type of infrastructure boundary such as a firewall, load balancer or network gateway. More recently, DLP tools have started to leverage machine learning and artificial intelligence (ML/AI) technologies to detect patterns of data usage that are potential indicators of sensitive data exfiltration.

Why is DLP so hard?

The difficulties involved in safeguarding sensitive digital information are obvious to any IT professional but will be briefly listed here to underscore the magnitude of the DLP challenge.

- The *sheer volume of data* collected and maintained by modern enterprises is growing rapidly due to the relative ease of collection and the low cost of data storage. IoT technologies are accelerating data collection in many industries such as construction, transportation, logistics and facility management.

- The *storage pricing policies* of many cloud service providers (CSPs) incent data collection by focusing charges on data export or queries, not on the acquisition or maintenance of additional data. In effect, the storage of additional data is free until it's actually used in some fashion.

- The *volume of unstructured data* (e.g. text documents, photo images, engineering diagrams, video recordings) collected by many enterprises is growing at a faster rate than structured data. The sensitivity of unstructured data is inherently more difficult to ascertain and classify.

- *SaaS proliferation* across modern enterprises has created new sources and sinks for business-related information that are rarely governed in a centralized fashion.

- The *automated transfer of data* across multiple business applications is accelerating through the use of APIs (application programming interfaces) and RPA (robotic process automation) bots.

- *Authorized data users* are no longer limited to the full time employees of a modern enterprise. Contractors, consultants, suppliers, partners, managed service providers and even customers require routine access to enterprise systems and databases to conduct daily business operations.

- *Collaboration tools* (e.g. texting, file sharing, videoconferencing, virtual whiteboards, etc.) facilitate the casual and spontaneous exchange of data among team members both inside and outside the enterprise.

- Modern CSPs enable the *rapid construction and deconstruction of computing environments* that frequently import enterprise data for development, test or production purposes. Monitoring the data employed by these transient environments is problematical.

- *Definitions of data sensitivity* can vary significantly depending upon an enterprise's business model, its reliance on proprietary knowledge or intellectual property, the expectations and concerns of its customers, and the regulations governing its operations in different geographic locations.

- *Regulatory requirements* and consumer concerns regarding privacy data management are a moving target as new regulations are introduced, existing regulations are being clarified through judicial challenges and the privacy concerns of individuals differing in age, gender, race and nationality are continually evolving.

- The *governance of data stores* across an enterprise's SaaS portfolio, cloud computing environments and legacy on-premise systems is typically shared by IT groups, security teams and business operations staff members. As a consequence, governance is frequently fragmented, inconsistent and incomplete.

- The proliferation of data types, stores and transport mechanisms has been accompanied by a *proliferation of security tools* that are used to monitor data at rest, in motion and in use. These tools collectively generate a significant and sometimes overwhelming number of false positive exfiltration alerts, thwarting the efforts of even the most sophisticated security teams to prevent the loss of sensitive information.

The factors listed above constitute a perfect storm that every enterprise confronts on a daily basis. Sadly, empirical evidence suggests that the forces of technical complexity, organizational confusion and unpredictable end user behavior frequently overwhelm the safeguards provided by conventional DLP solutions.

A brief history of DLP solutions

Conventional DLP solutions are grounded in the castle-and-moat defensive strategies that were developed decades ago to thwart inbound cyberattacks by malicious actors. Prior to the widespread use of cloud services sensitive enterprise information was housed in proprietary data centers and accessed via corporate networks. Elaborate firewall strategies were employed to protect data assets. Intrusion detection and prevention tools were deployed to block incoming exploits and malware.

As laptops and home computers were used more frequently to conduct business remotely, network-based defensive strategies were extended through the use of virtual private networks (VPNs). VPN technology provided a means of extending network-based safeguards to an expanded perimeter of remote devices. Branch office and remote worker internet traffic was commonly backhauled to central nodes on the corporate network to provide an added measure of control, inspection and protection.

As SaaS tools and consumer applications became more widely adopted Cloud Access Security Brokers (CASBs) were deployed to block employee access to questionable or suspicious Internet sites. CASB technology matured over time and became an additional infrastructure boundary controlling the flow of data to and from cloud-based data stores. Contemporary CASB tools are considerably more sophisticated and can be used to monitor data-related activities in the cloud, detect potential threats and initiate certain types of remedial actions.

The proliferation of smartphones significantly increased the complexity of an enterprise's defensive perimeter. VPN technology was not readily applicable to the management of smartphone devices and a new generation of mobile device management (MDM) tools were widely adopted. MDM solutions extended corporate control to mobile devices by creating data containers on individual devices that could be managed remotely by IT teams. In principle, MDM tools provided a means of managing data on mobile endpoints that could access

internet-based applications and services without connecting to the corporate network. In practice, the administration of MDM tools was complicated by the reluctance of many employees to allow their employers to manage functions on their personally-owned devices.

IT organizations universally adopted defense in depth (DiD) strategies to deal with the steady expansion of their security perimeters. Just as it sounds, DiD refers to the deployment of layered controls on key infrastructure elements – systems, networks and endpoints – to thwart incoming cyberattacks. The widespread use of SaaS applications and cloud computing resources compromised DiD strategies because these technologies enabled the creation and maintenance of data stores beyond the reach of conventional infrastructure boundaries.

In a world of distributed work teams, diverse endpoints and ubiquitous cloud-based services, traditional infrastructure-focused defensive schemes were supplemented with a heightened focus on managing the data access permissions and authorization privileges of end users. User authentication procedures have become increasingly sophisticated, relying upon multiple authentication factors (MFA) and contextual information regarding the timing, nature and source of an authentication request. Many security vendors claim that "people are the perimeter" and the current emphasis on authentication and authorization mechanisms to protect sensitive information – whether it resides in the corporate data center, endpoint devices or the cloud – substantiate that claim.

Conventional DLP solutions are largely based upon the DiD framework but they turn that framework upside down to prevent the ***egress*** of sensitive information instead of the ***ingress*** of cyber exploits and malware. Cyberattack defenses are designed to detect and deter incoming threats that transgress a series of end user, endpoint, network, system and resource boundaries. DLP safeguards are designed to detect and deter the egress of sensitive information across a similar set of boundaries, only in the opposite order. DLP solutions could easily be characterized as Retention in Depth (RiD) strategies because they're focused on

the same infrastructure perimeters (i.e. moats) employed by DiD cyberattack strategies.

DiD strategies continue to dominate contemporary DLP frameworks because so many production workloads and data assets continue to exist in proprietary data centers and dedicated colocation facilities. A 2020 survey of over 700 global enterprises conducted by Flexera revealed that many companies are still in the process of transitioning to cloud operations. Respondents indicated that roughly half of their production workloads and corporate data assets were still hosted on hardware resources under direct proprietary management. They anticipated that the use of public cloud resources would increase by 9% and cloud-based data storage would increase by 8% on average throughout 2020. Those projections preceded the Covid crisis.

Breach statistics presented in Verizon's 2020 Data Breach Investigations Report reflect continued reliance on privately managed infrastructure resources as well. On-premise assets were involved in 70% of the 3,950 breaches analyzed by Verizon.

The current reliance on proprietary resources highlighted in the Flexera and Verizon reports should not instill a sense of false confidence in the utility of current DLP solutions. The migration of business applications and infrastructure resources to the cloud is inexorable. CapitalOne – a $30B Fortune 100 financial services firm – achieved its transition to the cloud in November 2020, completing the transfer of all legacy operations at eight proprietary data centers to AWS. This is the future state that next generation of data security solutions needs to fully comprehend and secure. In the words of Wayne Gretzky, we need "to skate to where the puck is going and not to where it's been".

Note that the term 'data assets' is used broadly in this report to refer to data stores, data bases and file systems that may be freestanding or embedded in business applications or cloud services. It includes both structured and unstructured data types.

Why do we need to rethink DLP?

Conventional DLP solutions were never designed to deal with the data sprawl, ubiquitous cloud services, end user behaviors and regulatory environments that constitute the modern working world. They have been extended in a piecemeal and fragmented fashion to deal with an ever-expanding collection of egress boundaries. They struggle to maintain accurate inventories of sensitive enterprise information, a perennial problem that's been compounded by the accelerated acquisition of unstructured data. The rule-based policies they are designed to enforce are frequently eclipsed by dynamic changes in business operations and user behaviors.

Conventional solutions have realized that knowledge of user identity, data sensitivity and the context in which data is being accessed is not sufficient to guard against data loss. ML/AI technologies have been leveraged to create a new set of tools to identify data usage patterns that may be precursors of data exfiltration or exposure. These analytical tools have been applied to the behavior of end users (User Behavior Analytics), resources (User and Entity Behavior Analytics) and data itself (Data Behavior Analytics). Although these tools provide an additional measure of data loss prevention they are difficult to implement in practice because of the complexity of business operations and the idiosyncratic ways in which employees, contractors, suppliers and partners go about performing their jobs.

Consider the Covid crisis of 2020 as a case in point. The sourcing, manufacturing, distribution, retailing and customer support operations of many companies were transformed in a wholesale and sometimes radical way by the global pandemic. It's highly probable – almost certain – that user, entity and data behaviors during the second quarter of 2020 were highly anomalous relative to normative baselines established in prior quarters. Even the most sophisticated analytical tools struggled to identify 'new normal' patterns of behavior under these conditions that could be used to develop effective security controls.

Perhaps most importantly, DLP solutions only address a portion of the data risks and liabilities confronting every modern enterprise. DLP safeguards are necessary but not sufficient to protect enterprises from other forms of data mishandling and misuse that don't involve theft or loss. Corporate data can be internally employed in ways that are illegal, unethical or inconsistent with the terms under which it was originally acquired without ever being lost or stolen. Furthermore, it can be employed in ways that violate the expectations and sensitivities of employees, customers, suppliers or business partners, creating significant business problems and liabilities, even when it is used legally, ethically and in compliance with its terms of acquisition. Data misuse protection is a broader conceptual framework for securing data in the future. It encompasses the prevention of data loss but also includes safeguards to ensure the appropriate use and handling of sensitive data.

2020 public data breaches

Finally, there's abundant empirical evidence that conventional DLP solutions are not providing the deterrence that modern enterprises need or seek. Significant breaches have been publicly reported throughout 2020. In many instances the magnitude of such breaches has been revised upward as forensic investigations of their scope and impact proceed.

Conventional DLP platforms are difficult to implement, administer and tune. They provide working solutions for discovering and tagging certain types of

sensitive data and monitoring the movement of such data across well-defined infrastructure boundaries. Their principal utility in many instances is their ability to audit and enforce compliance with data security regulations such as GDPR, HIPAA and PCI DSS.

Data security concepts that have outlived their usefulness

Prior to discussing new perspectives on data protection it's important to reconsider and perhaps retire several concepts that are ingrained in conventional ways of thinking about DLP. The terms and concepts referenced below may not only have outlived their utility from a data protection perspective but may actually get in the way of reimaging the spectrum of safeguards that's needed to ensure proper data use and handling.

Insider threats and advanced persistent threats (APTs)

The use of cyberattack frameworks to avoid inappropriate data handling is a prime example of how the application of DiD strategies to deter data misuse is somewhat irrelevant and only marginally effective. Threat analysis frameworks are extremely useful for anticipating the tactics of cyberattack actors and putting the appropriate defenses in place to detect, deter and quarantine their efforts. Extending these frameworks to prevent the loss or misuse of sensitive data is less useful because inappropriate data handling can occur through negligence or malice by authorized users or properly credentialed imposters. Indicators of current or potential data mishandling are not uniquely correlated with the motivations or methods of different actors – whether they are good or bad, legitimate or illegitimate, transient or persistent. Threat frameworks may be a useful means of investigating and categorizing data breaches after the fact but they are no substitute for the broader framework of safeguards that's needed to anticipate and prevent data mishandling. Data safeguards need to be far more focused on 'what is a user doing with sensitive data' than 'how did he or she gain

access'. Nextgen DLP solutions need to detect and deter inappropriate data usage and data flows irrespective of the motives or means employed in committing such acts.

Protecting data in motion

Conventional DLP solutions have been constructed to safeguard sensitive data at rest, in motion and in use. As a practical matter, almost all data in motion is encrypted and data source systems require encryption certificates from data target systems to prevent malicious man-in-the-middle decryption. Whitelisting procedures can also be used to explicitly designate systems that can exchange sensitive data. Data in motion protection is a problem that has largely been solved, provided that encryption safeguards are employed in a consistent and comprehensive fashion.

Role based access controls (RBAC)

RBAC controls were conceived as a simplifying measure that would enable security administrators to assign common data access permissions to groups of individuals with common job responsibilities. Unfortunately, HCM (Human Capital Management) systems fail to capture the role specialization that's needed to assign data permissions and privileges in detail. They're primarily designed to manage employee performance and administer compensation. They're incapable of monitoring fluid changes in the roles and responsibilities of individual employees that are pervasive in a workplace that is becoming increasingly distributed, virtual and less hierarchical. In practice, RBAC controls have proven to be difficult to administer, difficult to extend to nonemployees and too coarse grained to manage the full spectrum of actions that users can perform on data. Modern DLP solutions employ role designations as a single element within more sophisticated attribute-based or risk-based control schemes that include information about user identity, location, IP address, source device, device characteristics, time of day, target application or database and other variables. RBAC controls are no longer an effective standalone access management tool.

End user inconvenience

Most data security teams live in fear of inconveniencing end users. They try to minimize alerts regarding suspicious data activities, curtail requests for additional authentication credentials and, whenever possible, avoid suspending or delaying data privileges unless required to do so on the basis of prescribed operational procedures. This phobia needs to be overcome through more effective end user education and also by ensuring that the level of such inconvenience is commensurate with the sensitivity of the data it is designed to protect. Alert tuning should be based upon business risk associated with data and actions and not on end user inconvenience. Not surprisingly, end user complaints occur far less frequently in financial services, wealth management, legal and pharmaceutical research firms where the financial consequences of data mishandling are universally understood by all staff members.

"Security perimeters are completely destroyed and they're not coming back"

Doug Merritt, CEO, Splunk

Data security perimeters

DLP solutions have inherited perimeter-based models for safeguarding sensitive data from the DiD models employed to deter cyberattacks. While the ability to monitor the transmission of data across certain types of infrastructure boundaries will continue to play a role in ensuring the retention of sensitive data, such boundaries are becoming increasingly ambiguous and permeable in a cloud-first, choose-the-handiest-device, collaboration-obsessed working world. Perimeter-based data retention models need to be replaced by frameworks that are built to secure data pipelines and data flows, irrespective of the convoluted infrastructure paths such pipelines and flows may follow.

Common sense data hygiene principles (aka table stakes)

There are some principles regarding the safe handling of sensitive data that verge on being simple common sense. The following principles are immutable and should serve as the foundation of any data security framework. If these principles have not been adopted and translated into everyday practices, there's no justification for investing time and resources in implementing the next generation principles presented later in this report. This is not to imply that the following principles constitute a foolproof or comprehensive security framework. However, failure to operationalize these principles would constitute culpable negligence on the part of any data security team. They are prerequisites for any type of data protection initiative.

Establish a clear understanding of the data that needs protecting

It's almost too obvious to state, but a clear understanding of the relative sensitivity of different forms of data is essential in establishing effective and efficient safeguards. Enterprises that fail to establish a classification framework that reflects business-critical concerns regarding intellectual property, regulatory controls and internal operations are doomed to waste inordinate time and energy overprotecting some data assets at the expense of properly protecting others.

Data safeguards inherently add friction to business processes. This friction can be minimized and made commensurate with the business risks involved in data-related operations if the sensitivity of different forms of data can be categorized on a graduated scale developed with and approved by IT's business partners. A rational and transparent scale of differential data sensitivity is essential in developing meaningful awareness training for end users who will ultimately be impacted by security safeguards.

Finally, an explicit understanding of an enterprise's sensitive data assets enables the identification of the source systems of such data. The application of appropriate controls to the management and dispersion of source system data can eliminate a wide variety of misuse scenarios in downstream operations.

A variety of tools are available to support the discovery, classification and cataloging of sensitive data. **BigID, One Trust DataDiscovery** (formerly Integris Software), **Collibra** and **Alation** are leaders in this space. Newer entrants such as **Concentric.ai** employ semantic intelligence to discover sensitive data within unstructured data stores. **Nightfall.ai** employs ML techniques to discover sensitive data within SaaS applications, API calls and data stores. **Amazon Macie** provides similar capabilities for data stores hosted within AWS.

Develop and enforce stringent data retention policies

As noted earlier, it's far too easy to collect, store, share and replicate data within the modern enterprise. One of the biggest threats that many enterprises face is the retention of data that has little or no business utility, now or in the future. Highly regulated industries tend to have the most stringent retention policies. In some cases such policies are required by regulation. In other cases they're self-imposed. However, many companies in other industries have failed to formulate or enforce retention policies that would materially reduce the business risks posed by long term retention of sensitive data. Aggressive destruction of unnecessary or underutilized data is a quick win that can considerably reduce the risk exposure of many enterprises.

Develop and enforce rigorous data backup procedures

Regular backup or continuous mirroring of sensitive data stores provides security teams with far more flexibility in responding to data misuse, contamination, exposure or exfiltration events if they occur. Although multiple copies of sensitive data represent a potential security liability, backup stores can be secured with the most draconian safeguards since they have no operational use except in an emergency.

Minimize data in the clear

There are many ways in which sensitive data at rest or in motion can be rendered useless to a casual, unauthorized or malicious actor. The most common mechanism is encryption. Other methods include hashing (involving the use of mathematical algorithms to replace data values with unique number codes) and tokenization (involving the use of lookup tables that substitute data tokens for original data values).

Encryption tools have become increasingly sophisticated.

- **Baffle** offers a cloud-based service that can apply tokenization, format preserving encryption (FPE) and AES-256 file encryption to all stages of a data pipeline, including data in memory. Baffle's service is capable of handling structured and unstructured data.

- **Fortanix** leverages Intel's Software Guard Extension (SGX) technology to manage encryption keys within protected hardware enclaves established within microprocessor memories.

- **Skyflow** employs polymorphic encryption in which a mutating algorithm changes the computation employed in encrypting and decrypting data during each encryption/decryption cycle. Skyflow provides a cloud-based vault for sensitive data that can be accessed via an API, eliminating the need to host sensitive data in any form.

- **Enveil**, **Duality** and **Inpher** are leaders in homomorphic encryption which enables data to remain encrypted through a wide variety of computational processes.

The simplest procedure for minimizing data exposure – whether it's camouflaged or readable – is to limit data replication whenever possible. Data virtualization platforms broker access to data assets by serving as a control plane between

data users and data stores. These platforms can enforce the data rights of end users and audit certain forms of data usage. When properly configured, they serve as a retail data storefront, greatly reducing the need to replicate data on a wholesale basis to satisfy the requirements of different business teams. **Actifio, AtScale, Delphix, Denodo** and **Dremio** offer leading virtualization solutions. **Informatica, TIBCO, IBM** and **Oracle** also provide virtualization capabilities as a component of broader offerings.

Assets containing PII information are routinely de-identified by simply bifurcating PII and non-PII attributes into two wholly distinct data stores that can be linked at the record level by some type of hashed key or token. This is the data-centric equivalent of the segmentation strategy that's routinely employed to manage network security. PII data can be further protected by masking, anonymization (removal of all PII attributes) or pseudonymization (removal of selected PII attributes).

Meticulous management of end user permissions and privileges

As noted earlier, end user authentication and authorization procedures are a critical control plane in safeguarding sensitive data. They are the primary control plane in safeguarding cloud-based applications and data stores. Least privilege principles need to be employed continuously to avoid granting access permissions or entitlement privileges for which there is no immediate need. Unused privileges should be automatically revoked on predetermined timetables. Privileges used on an intermittent basis should be challenged and perhaps allowed to lapse until needed sometime in the future. It's not uncommon for employees to accumulate a broad variety of permissions and privileges over time in response to changing roles, responsibilities, activities and assignments. Scrupulous diligence is required to ensure that current permissions and privileges are actually needed to support near term business operations. Distributed administration of cloud-based services procured by functional teams outside of IT makes authentication governance more challenging but also more necessary.

Authorization Privileges Are Complex

Once an end user has been given permission to access an application or data store they receive explicit authorization to perform one or more of the following data actions

View – read/inspect data

Modify – change existing data values

Delete – remove/destroy existing data

Create – add/construct new data

Replicate – make a duplicate copy of a data asset in the same environment

Export – send a copy (whole or partial) to another environment

Process/Transform – create a derivative data asset that may preserve some or none of the data values in the parent asset

Share – share any of the above privileges with another user who presumably has the same authorization rights

Data Misuse Protection – the next generation of DLP

Data loss prevention practices of the past are not sufficient to protect modern enterprises from the liabilities associated with the inappropriate use or mishandling of sensitive data. DLP frameworks need to be replaced by a new Data Misuse Protection paradigm that safeguards data from unauthorized or inappropriate use within a corporate environment *in addition to* its outright theft or inadvertent loss beyond a company's boundaries.

Conventional DLP solutions employ metadata tags characterizing the sensitivity of specific data assets and infrastructure-based boundaries to detect and deter data loss. Unfortunately, data assets are not static objects. They're continuously transformed by a variety of actions performed by a myriad of users. Conventional solutions are incapable of refining tagging schemes and expanding usage controls at a pace that can keep up with the ways in which data is used in a modern enterprise.

Next generation DMP solutions will provide data assets with more sophisticated means of protecting themselves. It's a subtle but significant distinction. *Instead of applying tags and policies to data assets, the assets themselves should 'own' a rich set of metadata characteristics and subscribe to services that protect their integrity and control their usage.*

Data Misuse Protection safeguards data from unauthorized or inappropriate use within a corporate environment ***in addition to*** its outright theft or loss beyond a company's boundaries

DMP frameworks will be based on the principles outlined below. These principles may give rise to a new generation of products and services or they may be incorporated in existing tools and platforms. They are not necessarily comprehensive or complete but they are essential elements of any future framework for safeguarding sensitive data.

The following principles are not wholly new or revolutionary. What is new is the linkage of enriched metadata and subscription services with individual data assets, giving assets a measure of self-protection that cannot be achieved with current tools or operations. These principles are far better suited to safeguarding fluid, metastasizing data pipelines in the modern enterprise than the perimeter-based retention safeguards that form the foundation of current DLP practices.

Data lineage

Critical data assets should possess a comprehensive understanding of their genetic family tree. Primary assets should contain metadata describing how, when, why and where they were originally constructed. Similar information should be generated every time they are modified or materially transformed. Business leaders need to assume responsibility for the integrity and security of

sensitive data and should be specified by name or title for every version of a critical asset. All derivative assets should inherit the lineage metadata possessed by their parents.

Lineage is doubly important in protecting PII data from misuse. Consent agreements are employed to collect many forms of PII data. These agreements impose variable constraints on the ways in which such data may be used. Metadata needs to represent these constraints in ways that can be consumed by the policy subscription services discussed below. PII-related metadata needs to translate the declarative prose found in privacy statements and consent agreements into prescriptive constraints that can be used to control the proper use of PII. These prescriptive constraints may be expressed in conventional terms such as the attributes of approved users or the characteristics of approved hosting environments. They may also be expressed in a less conventional manner in terms of business use cases (e.g. discount targeting) or usage scenarios (e.g. multichannel marketing campaigns).

Lineage is triply important in developing inference and forecasting models based upon ML/AI techniques. Small differences in the heritage of different data assets may produce significant differences in model outcomes or introduce subtle, undetected biases in model predictions. Precise knowledge of historical data collection and transformation practices is needed to ensure model accuracy and avoid unwanted and sometimes unethical side effects.

Lineage information may provide a new type of control surface to guard against widespread data dispersion. Third, fourth or fifth generation assets may be assigned 'use by' dates and automatically destroyed at the conclusion of an approved use period. Sensitive fields may be suppressed in Nth generation assets unless the exposure of such data is explicitly authorized by business leaders. Access permissions in Nth generation assets may be periodically suspended, pending approval by business management to reinstate pre-existing permissions in whole or in part. Periodic suspension of permissions provides a valuable opportunity to ensure that least privilege authorization principles are being continually and consistently enforced.

Lineage metadata should be thought of in the broadest possible terms. It should include but not be limited to information concerning data types, entity relationships and compliance requirements. It should also include historical information concerning hosting environments, access and authorization privileges, usage logs and backup procedures. Rich metadata expands the scope and sophistication of the subscription services that can be employed to guard against potential misuse.

Perhaps most importantly, lineage metadata will immeasurably improve the scope and accuracy of data asset security designations. Conventional DLP solutions take a brute force approach to designating the sensitivity of individual assets by searching for character strings (regular expressions), key words, lexicons or hashes that can be correlated with existing assets that are known to contain sensitive information. In the absence of lineage information concerning the sensitivity of parental assets this search, scan and classify procedure must be performed continually across an enterprise's entire data estate. Accurate specification of data sensitivity at the time of asset creation and the inheritance of this information in all derivative products would dramatically improve the completeness and accuracy of sensitivity designations and allow protective safeguards to be focused accordingly. In contrast, DLP solutions are primarily focused on assets whose sensitivity is defined on the basis of data types and terminology that are within the scope of regulations such as GDPR, CCPA and PCI.

Informatica, **IBM** and **Oracle** are established vendors that provide metadata versioning capabilities as features within broader data management offerings. Their solutions are primarily designed to propagate conventional forms of metadata such as data definitions, entity relationships and sensitivity flags in derivative assets. All three vendors have taken steps to extend the legacy versions of their solutions into the cloud. In contrast, **Alation** and **Collibra** are cloud-native solutions for asset discovery, data classification and metadata enrichment that have been widely used to comply with GDPR regulations.

Cyberhaven and Manta Security are emerging companies offering new lineage management capabilities.

- **Cyberhaven** provides a data tracing solution that continuously tracks file movement and ownership through multiple channels such as email, Box, Zoom, MS Teams and Slack, without employing any classification or tagging procedures. File lineage can be determined retroactively and monitored prospectively.

- **Manta Security** captures lineage information by continuously scanning software algorithms that act upon data, not the data itself. Many conventional solutions infer lineage by detecting identical data in multiple assets. Manta definitively establishes lineage relationships by monitoring the code being used to construct derivative assets. Manta cannot detect data manipulations performed by end users on endpoints.

The emergence of Policy as Code tools holds great promise for establishing and populating extended metadata schema at the time of parental asset creation. The use of these tools in constructing applications that will produce new assets will ensure a level of schema consistency and coverage that has rarely been achieved in the past. They can play a major role in implementing the highly enriched metadata schema envisioned in this report. They can also eliminate the recurring remedial rework required to extend conventional schema on an incremental basis. Leading vendors in this emerging field include **Stacklet**, **Accurics**, **Bridgecrew** and **Concourse Labs**.

Environmental awareness

Sensitive assets should possess an awareness of their current hosting environment and subscribe to services that provide continuous, real time information concerning environmental integrity. Policy services described in a later section can be configured to respond to inherent integrity risks or deteriorating integrity conditions. Evidence of potential compromise may be

used to impose restrictions on the scope and nature of the privileges users can exercise within specific environments. Evidence of current or imminent compromise may be used to trigger the preemptive destruction or redaction of an asset.

Conventional SIEM (Security Information and Event Management) vendors such as **Splunk**, **QRadar** and **LogRhythm** rely upon log data to monitor environmental conditions. Potential risks are identified and ranked on the basis of known vulnerability signatures or variations from standard operating conditions.

New entrants in this space provide continuous surveillance capabilities that enable a wider variety of near real time responses to environmental concerns.

- **Horizon3.ai** is a cloud-based service that performs continuous penetration testing of environmental assets and ranks environmental vulnerabilities in terms of relative risk exposure.

- **Kenna Security** merges data from existing security tools with global threat intelligence. It employs proprietary risk scoring algorithms to rate the potential severity of vulnerabilities associated with individual infrastructure elements within a data asset's hosting environment.

- **Traceable** monitors end user activities, API interactions, data movements and code execution. It uses ML techniques to detect anomalous departures from historical baselines. Traceable provides a deep application context for identifying anomalous data actions and movements.

Clean room designations employed in satellite construction and pharmaceutical research may provide a useful analog for characterizing the relative security of different hosting environments. Conventional clean room standards are defined on the basis of the size and density of airborne particles or the concentration of airborne gases. The relative security of different hosting environments could be classified on an analogous scale based upon vulnerabilities detected by one or

more continuous surveillance tools. To some degree this concept is practiced today. Walled garden environments commonly used by ML/AI researchers handling sensitive data represent one of the highest (if not the highest) 'secure room' designations within an enterprise. A graduated scale of such environments could be established by individual enterprises based upon their business operating model and potential security liabilities. The use of a 'secure

Retention surface

Cyber defenders have coined the term ***attack surface*** to describe the various points of entry that malicious or negligent actors employ to access, misuse, expose or appropriate data assets. Data protectors need to re-imagine this concept and define the retention surface within which sensitive data can be properly controlled and used. This differs from conventional definitions of data retention that are focused on the destruction of sensitive assets on predetermined timetables. As used here, ***retention surface*** refers to the boundary at which end users and machines interact with sensitive enterprise data.

As noted earlier, end user access permissions and authorization privileges constitute the primary retention surface for critical data assets in a cloud-based working world. Meticulous management of end user permissions and privileges was referenced earlier this report as a basic form of security hygiene. DMP frameworks of the future will place even greater emphasis on safeguarding critical assets by continuously monitoring and minimizing end user data rights.

In its simplest form, an enterprise's data retention surface is the collective set of end user permissions and privileges associated with critical assets at any point in time. The ability to exercise these rights may be restricted on a conditional basis (e.g. certain forms of data may not be exported beyond pre-specified firewalls).

Note that the use of the term 'data rights' in the discussion of DMP principles refers to the access permissions and authorization privileges of data users, not the rights of data providers.

It may be constrained on a contextual basis in which different risk factors are considered before a specific data action is permitted. It may also be time-based or time-limited (e.g. available only during specific times of day or during a specific project or subject to revalidation every 90 days).

Many financial firms have already instituted fine grained contextual controls on permissions and privileges. Some employ risk-scoring schemes to weigh different contextual attributes such as user identity, location and IP address; user device type, identity and characteristics; time of day; and historical user behavior before granting access or authorizing specific actions. Risk-scoring algorithms can easily be customized to reflect the sensitivity of the data they are protecting or differences in regional operating conditions.

Attribute-based controls have not been widely adopted outside highly regulated industries. Where they have been instituted, they've primarily been used to regulate access permissions, not authorization privileges. In the case of structured data they can be implemented at a table, column, field, cell and sub-cell level. They can also be used to constrain the movement of data via API calls.

Failure to consistently and persistently apply least privilege principles to end user data rights will inevitably result in permission creep and privilege escalation. Enlighted DMP teams will maintain a strict accounting of the number of new permissions and privileges that have been granted over a specific period of time versus the number of permissions and privileges that have been suspended or deleted. A negative balance in favor of suspension/deletion is the desired outcome. A secondary metric of security vigilance is the number of instances in which the exercise of granted rights is prohibited on a conditional, contextual or temporal basis.

Critical data assets should maintain historical records of their usage as a form of enriched metadata. Derivative assets can leverage lineage information to link usage logs to parental assets, all the way back to primary data sources. These

merged logs can provide an instantaneous depiction of how the retention surface of sensitive assets has morphed over time and whether such assets are experiencing permission creep or privilege escalation. The creation of derivative assets provides an opportunity to reset data rights using least privilege principles that should never be overlooked.

The importance of data rights management as a primary control plane in cloud-based operating environments is underscored by the high level of current entrepreneurial activity in this space.

- **Authomize** employs a proprietary analytics engine to establish a comprehensive inventory of end user data rights by correlating user identity, access histories and usage behaviors across all applications, infrastructure resources and identity providers. This inventory can be monitored on a continuous basis and used to identify lapses in usage, inconsistent provisioning and unnecessary privilege escalation.

- **Concentric.ai** employs deep learning techniques to identify semantic clusters of files within data stores containing structured and unstructured data. Peer comparison of files within individual clusters provides an effective means of detecting inconsistent and potentially inappropriate access rights.

- **Secure Circle** ensures that data files transferred to end points inherit the conditional access rules that were established for their source systems. Additional rules related to device security and user behavior can be established on end points to strengthen inherited access controls.

- **Satori** employs a cloud-based proxy service that functions as a data access controller to any type of data store, enabling the use of sophisticated attribute-based procedures to manage structured data access at the row and column level.

- **Okera** enables entitlement rights to be managed on an asset-specific basis by data stewards distributed across multiple business functions.

- **Saviynt** has constructed an identity management platform that extends the functionality of conventional IGA (Identity Governance & Administration) and PAM (Privileged Access Management) tools into multi-cloud environments.

- **CloudKnox** monitors the activity of human and machine identities within public cloud platforms and uses that information to detect permission creep and privilege escalation.

- **Ermetic** and **Sonrai Security** provide platforms for discovering and monitoring entitlement rights to cloud-based resources within AWS, Azure and GCP and enforcing least privilege entitlement principles.

And finally, a variety of new services are emerging that provide deeper insight and greater control over access to data in SaaS applications.

- **AppOmni** provides visibility into the distribution and usage of data rights within Salesforce, Slack, Zoom, Microsoft Outlook, Box and Github. AppOmni leverages the unique user management frameworks embedded within each of these applications to discover inconsistent entitlement privileges, detect anomalous administrative actions and enforce consistent security configurations.

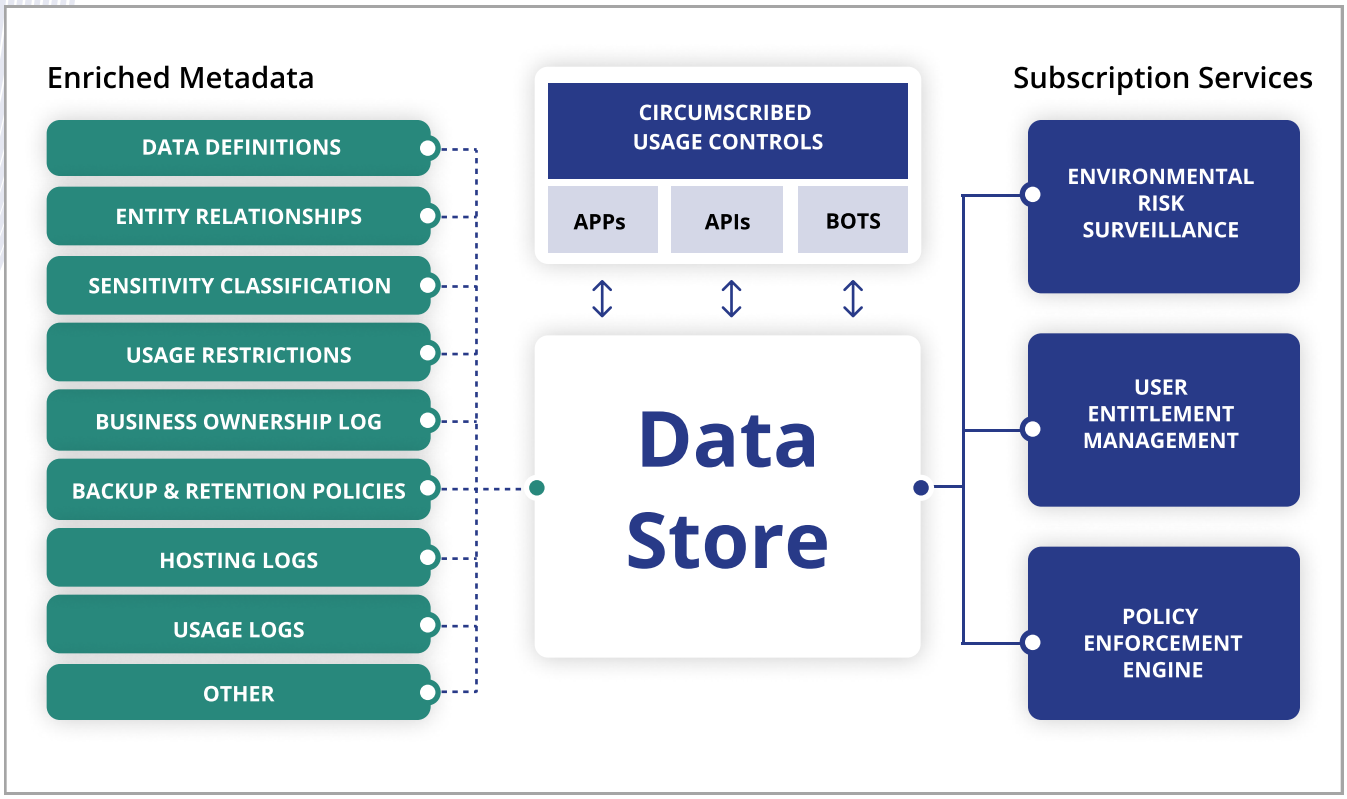
- **Obsidian Security** monitors the access and utilization of popular SaaS application to detect potential indicators of data compromise such as anomalous logins, persistent connections, unusual data movements and OAuth token abuse. Obsidian provides a means of identifying SaaS misconfigurations and stale user accounts.

- **Altitude Networks** indexes all applications downloaded by employees and monitors their access to sensitive corporate information stored in cloud applications.

Circumscribed usage controls

Many technologies in common use today circumscribe the way in which data can be used, reducing opportunities for mishandling or unnecessary exposure. The most obvious usage controls are embedded in business applications. Applications, especially those that are highly specialized to support unique business processes, are designed to implement a normative series of data flows that produce consistent business outcomes. Business applications can be configured in many different ways, but their data inputs, transformations and outputs generally adhere to a circumscribed set of normative patterns.

APIs and software bots are additional mechanisms for circumscribing data usage in a predictable and repeatable fashion with no human intermediation. APIs and bots possess data rights that were assigned by their creators or inherited from their developers. These rights can be subjected to the same conditional,



Conceptual framework for Data Misuse Protection

contextual and time-based controls discussed in the preceding section, providing further protection against mishandling or misuse. DMP teams should work in close collaboration with API and process automation engineers to proactively identify opportunities to reduce the scope and frequency of human interactions with sensitive data. The activities of APIs and bots should be subjected to the same logging and auditing practices employed to monitor human-data interactions.

APIs and bots can deter data misuse provided that their construction and deployment is properly governed. The ungoverned proliferation of these technologies can potentially become more of a security threat than a safeguard. Copy-and-paste avatars may reduce business process latency but they may also disseminate data in ways that are unintended or inappropriate.

Policy engines become subscription services

Policy sprawl is a pervasive problem within almost every security team. Actually, that's a bit of a misnomer. Corporate policies regarding data protection are generally few in number and typically stated in broad terms. Interpreting these policies and translating them into operational procedures is a far more challenging problem. The devil ***is*** in the details. Security professionals tend to use the terms policies and procedures interchangeably and that convention will be adopted here as well.

In an ideal world, individual data assets should be able to subscribe to one or more policy services that can leverage all of the forms of asset-specific metadata discussed above and be configured to optimize asset protection while minimizing the business friction created by policy enforcement. Idealized policy services of this nature are unlikely to be realized anytime soon but the concept should serve as an aspirational goal for VCs and entrepreneurs.

Abstracting different aspects of policy administration into one or more freestanding brokerage services is difficult to achieve in practice because most of the tools and systems supplying metadata and operational information to such

services have embedded policy modules of their own. These modules tend to operate as closed systems. Their business rules are rarely exposed through readable APIs which makes it difficult to discover, normalize and orchestrate pre-existing procedures.

An abstracted brokerage would have to comprehend, orchestrate and perhaps override pre-existing procedural rules distributed across multiple sources of essential input data. At the present time there are no industry-wide frameworks for abstracting security policies and orchestrating their implementation across multiple tools or systems. Business process modelers have recourse to the Business Process Definition Metamodel supported by the Object Management Group (OMG) or the XML Process Definition Language supported by the Workflow Management Coalition. There are no comparable abstraction models for security policy management.

SOAR (Security Orchestration, Automation and Response) and data virtualization technologies are the closest current approximations of the policy orchestration capabilities envisioned in this report. SOAR platforms extend the functionality of SIEM systems and, as their name implies, can be configured to enforce data protection policies in an automated fashion. In practice they're used to respond to infrastructure boundary infractions, anomalous end user access behaviors and data movements that may be indicators of compromise. They typically employ decision tree logic to invoke defensive actions and lack deep insight into data usage patterns. Significant effort is required to adapt alert and response procedures to the serial transformations that occur within most data pipelines. In short, SOAR platforms were conceived as mechanisms for detecting and responding to cyberattacks and preventing data exfiltration. They were not designed to prevent data mishandling or misuse.

Virtualization platforms function as proxy access services shielding data users from direct interaction with data assets. They can enforce a variety of controls over the ways in which data assets are exposed to end users. Access and authorization policies can be administered at the platform level and/or within

individual assets. Advanced virtualization platforms can monitor certain forms of data usage but typically don't provide the deep insight that's needed to enforce misuse policies.

Future DMP policy services will employ absolute rules and conditional logic. The movement of sensitive data beyond predetermined infrastructure boundaries will be absolutely prohibited in a manner that's equivalent to the enforcement practices of current DLP solutions. However, DMP services will also provide far more sophisticated capabilities in assessing conditional risk and invoking alerts and responses on a risk-weighted basis.

Acceptable fraud is a common operating principle in financial institutions. Financial firms realize that the costs and operational overhead involved in trying to reduce fraud losses to zero are prohibitive. Therefore, they're willing to accept fraud losses of a certain size under certain circumstances. Acceptable risk management will be an inherent capability of future DMP policy services. DMP services will assess the risks associated with different forms of data movement, transformation and usage and trigger actions to reduce such risks to levels that are deemed to be acceptable. Risk may be reduced by simply challenging an end user to present an additional authentication factor or seeking business management approval for a requested movement or transformation.

Policy infraction events and alerts reported by DMP services will be forwarded to designated systems or personnel in much the same way they're handled today. However, intervention responses triggered by a policy infraction may be executed by the service itself (e.g. suspending all access to a misconfigured S3 bucket hosting a data asset) or by the asset itself (e.g. making a call to an embedded API in the asset's metadata that redacts all PII in data views being presented to individuals outside the Marketing department). Response procedures encoded in any policy service will inevitably struggle to cope with the proliferation of data assets and the fluid ways in which data is used and exchanged. Consequently, it's imperative that critical assets become increasingly self-reliant on embedded defense mechanisms that can be triggered by future policy services.

Achieving DMP

Data can be mishandled in many different ways. The most obvious example is the criminal theft and monetization of PII data on the dark web. The second most obvious example is the theft of sensitive data by a disgruntled employee. Perhaps the third most obvious example is the theft of proprietary knowledge or intellectual property by an employee for purposes of personal gain. However, these are all extreme cases. Misuse can occur in many other ways as well.

PII data is typically collected by commercial enterprises for use in ways that are circumscribed by privacy statements or consent agreements. These statements and agreements are explicitly accepted by data providers at the time of data submission. They generally contain broad statements of intent that leave wide latitude for interpretation. However, they typically prohibit the use of submitted information in ways that are unrelated to goals of the business relationship being established between the provider and the enterprise.

For example, estimates of annual household income provided in mortgage applications are not intended to be shared with local auto dealers. Web surfing data that may provide insight into an individual's political persuasions is not intended to be shared with aspiring candidates or political parties for fund raising purposes. Cell phone geolocation information is not intended to be used to offer drive-by discounts to stores or services unless consumers specifically opt in to such offers. These examples are all forms of misuse.

Data enrichment algorithms are a particularly insidious means of subverting the stated intentions of consent agreements. Every Marketing department aspires to establish a '360 degree view of the customer' that incorporates every bit of customer information an enterprise has acquired in the past plus whatever ancillary information it can acquire from public sources, borrow from its go-to-market partners or buy. The data acquired from these various sources has invariably been collected under wide variety of usage terms and conditions. Highly enriched data records are intelligence dossiers on the behaviors,

preferences and proclivities of specific individuals. In the absence of usage controls, they can be leveraged to direct customized advertising to an individual's phone, solicit support for groups and causes or provide personal financial advice in ways that are wholly inconsistent with the intent and understanding of a data provider.

Even non-PII data can be misused in many different ways. Material non-public information (MNPI) concerning the internal operations of an enterprise may constitute a form of insider information that can be leveraged by external investors to gain an advantage over their competitors. Engineering firms may share drawings and diagrams with external design shops and manufacturing subcontractors that provide knowledge and insight they can use to solicit business from a firm's competitors. Sales representatives routinely maintain personal records of customer contacts, past sales and future prospects that provide a form of job security in the event that they seek employment elsewhere. Software engineers may maintain personal code libraries or repositories for similar purposes. Many of the data actions and movements involved in these scenarios can be performed incrementally in the course of normal everyday business making it extremely difficult to detect definitive patterns of misuse.

There will never be a collection of tools and procedures that can thwart the creativity and unpredictability of human beings who are intent on misusing sensitive corporate data or who unintentionally mishandle data through ignorance or negligence. The DMP principles discussed above can considerably reduce the risk of mishandling but they cannot provide an ironclad guarantee that all forms of misuse will be deterred in the future. Nevertheless, they are vastly more comprehensive and effective than the limited protection afforded by conventional DLP solutions that rely upon restricted definitions of data sensitivity and are designed to prevent unauthorized data egress instead of unauthorized data use.

The transition from DLP to DMP will be a journey, not an event. New tools and capabilities based upon the principles outlined above will emerge in a piecemeal

fashion. Progressive security teams will initially use these new capabilities to augment their current practices and then ultimately use them to replace legacy DLP solutions altogether. However, the lack of a reference DMP tech stack architecture at the present time should not be used as an excuse to delay a wholesale reimaging of data security as a DMP problem, not a DLP issue. Several of the concepts described in this report, such as operational metadata enrichment, continuous surveillance of critical hosting environments, attribute-based controls on the exercise of data rights and the strategic use of APIs and RPA bots to minimize human-data interactions, can be implemented today using tools that are readily available. Proactive implementation of these practices now will prepare progressive teams to obtain immediate benefits from new DMP capabilities as they become available in the future.

DMP is a business imperative, not simply a technology challenge

In the wake of customer complaints regarding false advertising and neighborhood grievances regarding customer misbehavior, Brian Chesky, the CEO of Airbnb, sent a message to his employees reminding them that “the world moves at the speed of trust”. This admonition is universally applicable to all forms of human activity, both commercial and non-commercial.

Business operations are based upon trust between suppliers and manufacturers, retailers and consumers, investors and business executives. That trust, in turn, is critically dependent upon the proper and reliable handling of information that is personal, privileged, proprietary and regulated in an array of situations and circumstances that is uniquely defined by every company’s operating model.

A new DMP framework is needed to establish and maintain the trustworthiness that all enterprises will need to survive and compete in the 2020s. This

framework will be based upon new forms of descriptive and experiential metadata that is uniquely associated with individual data assets and the ability of such assets to consume services that are customized for their protection. Through metadata critical assets will possess a historical record – some might say awareness – of their lineage, custodians, hosting environments, network utilization, end users and usage patterns. Controls will be strategically orchestrated at a pipeline level and tactically enforced at a data store, database or file store level.

Technology alone cannot prescribe, detect and control every potential instance of data misuse. Technology can be used to identify and rank conditions, circumstances or activities that pose the greatest potential or probable risks. It can be used to enforce risk reduction policies. It can even be used to initiate automated responses to certain risk conditions or scenarios. But operational policies or actions – whether they are automated or manual – can only be performed with the express guidance or permission of business leaders. Security tool administrators with incomplete and imperfect understanding of business conditions are incapable of exercising the judgement needed to prevent the misuse of sensitive data on their own.

This is not a revolutionary observation. Confidential data rooms with highly restricted access privileges are commonly established to support the divestiture of existing corporate assets, the acquisition of new lines of business and the conduct of high stakes litigation. The confidentiality concerns that motivate business executives to establish these secured environments needs to be extended across a broader range of sensitive data assets. ***Data security is ultimately an exercise in managing business risk. Therefore, business leaders need to take an active and persistent role in managing data security.*** Excessive common sense is needed (instead of paranoia) to differentiate the degrees of risk associated with different asset usage scenarios and ensure that security safeguards are sufficiently stringent to mitigate those risks to acceptable levels. This common sense approach to data security must be based on business intuition and judgement and not on the instincts or assumptions of technologists.

Suggested Reading

2020 Data Breach Investigations Report, Verizon

Cloud Threat Report – 2H 2020, Palo Alto Networks

2020 Global Threat Intelligence Report, NTT Limited

2020 Cost of Insider Threats Global Report, Ponemon Institute

2020 Insider Threat Report, Cybersecurity Insiders

CrowdStrike Services Cyber Front Lines Report 2020

Mandiant Security Effectiveness Report 2020

The Definitive Guide to Data Loss Prevention, Digital Guardian, 2019

Addressing the Top 5 Gaps in DLP, Cyberhaven, 2020

Securing the Data and Advanced Analytics Pipeline, Gartner Research, Joerg Fritsch, January 2020

Magic Quadrant for Metadata Management Solutions, Gartner Research, November 2020

Magic Quadrant for Data Quality Solutions, Gartner Research, July 2020

Market Guide for Cloud Workload Protection Platforms, Gartner Research, April 2020

Protecting What Matters: Introducing Data Guardrails and Behavioral Analytics, Securosis, July 2019

Data Security in the SaaS Age: Focus on What You Can Control, Securosis, June 2020

The Top 8 Benefits of Data Lineage, Erwin, David Loshin, August 2019

Cyber Security Mid-Year Review – 1H 2020, Momentum Cyber

The Top 10 Data Breaches of 2020, Security Magazine, Maria Henriquez, December 2020

2020 State of the Cloud Report, Flexera

Glossary

API – Application Programming Interface

APT – Advanced Persistent Threat

CASB – Cloud Access Security Broker

CCPA – California Consumer Privacy Act (2018)

CSP – Cloud Service Provider

DiD – Defense in Depth

DLP – Data Loss Prevention

DMP – Data Misuse Protection

FPE – File Preserving Encryption

GCP – Google Cloud Platform

GDPR – General Data Protection Regulation of the European Union (2018)

HCM – Human Capital Management

HIPAA – Health Insurance Portability and Accountability Act (1995)

IGA – Identity Governance and Administration

IoT – Internet of Things

IP address – Internet Protocol address

MDM – Mobile Device Management

MFA – Multi Factor Authentication

ML/AI – Machine Learning/Artificial Intelligence

MNPI – Material Non-Public Information

OMG – Object Management Group

PAM – Privileged Access Management

PCI DSS – Payment Card Industry Data Security Standard

PII – Personally Identifiable Information

RBAC – Role-Based Access Management

RiD – Retention in Depth

RPA – Robotic Process Automation

S3 – Simple Storage Service (Amazon Web Services)

SaaS – Software As A Service

SGX – Software Guard Extensions (Intel)

SIEM – Security Information and Event Management

SOAR – Security Orchestration, Automation and Response

VPN – Virtual Private Network

XML – Extensible Markup Language