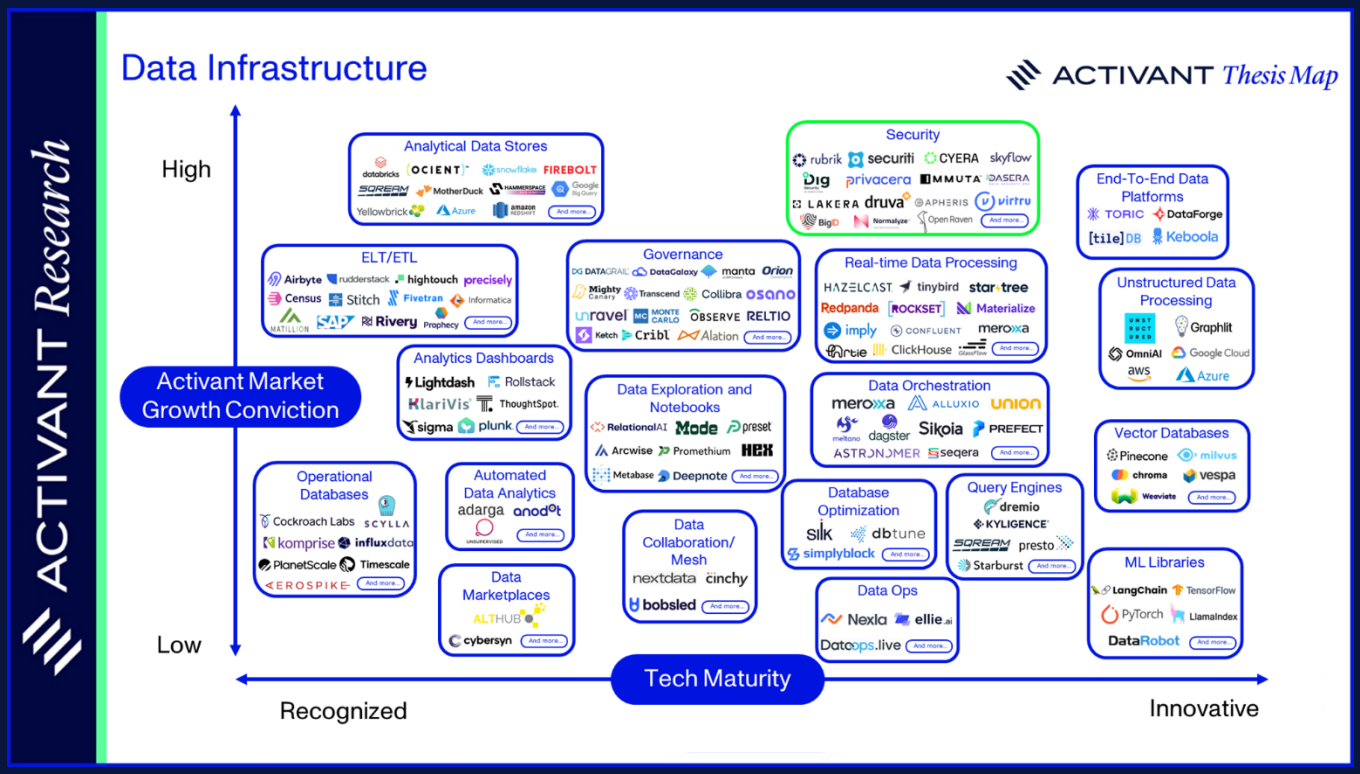
ACTIVANT RESEARCH

Data Security

Shifting *data* left: how this minor piece of the cybersecurity stack could be the most fundamental



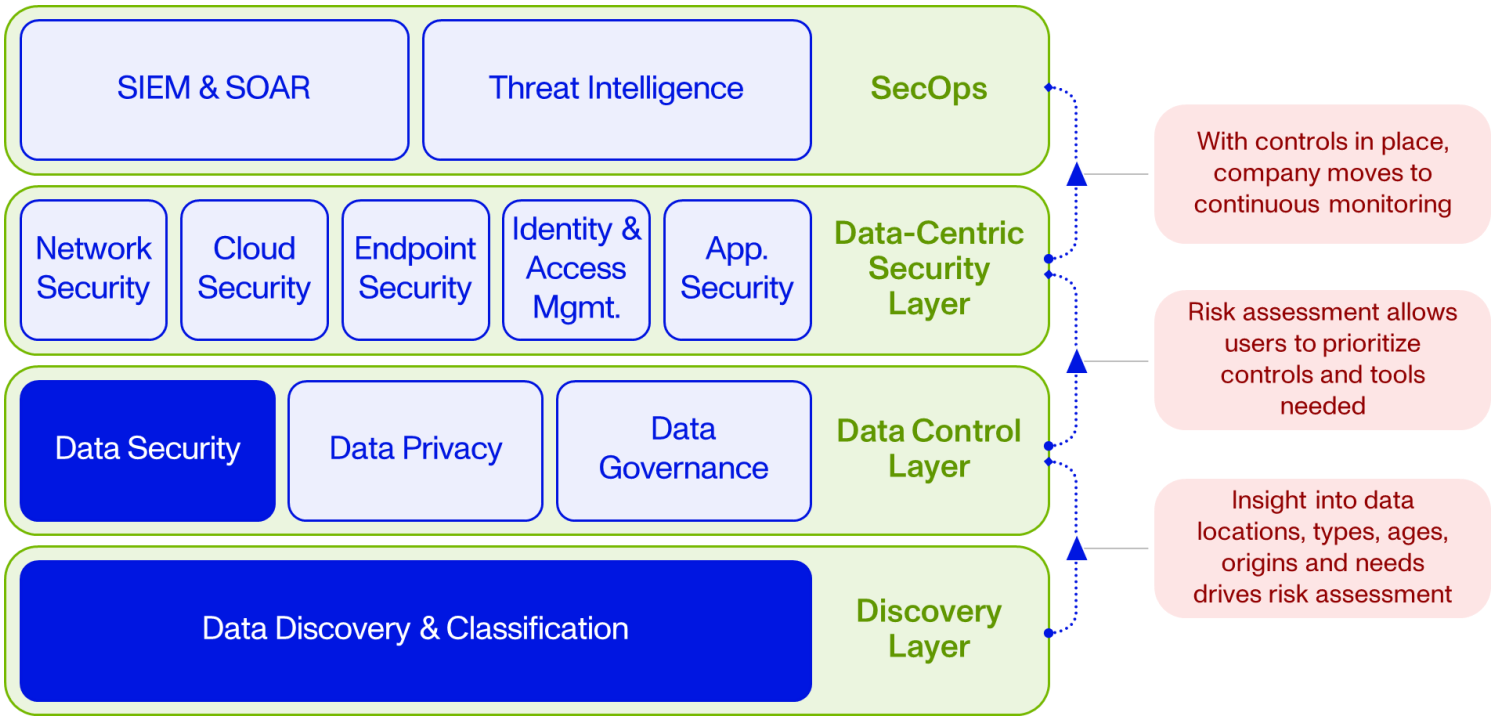
Q2 2024

Jonathan Vickery, Nina Matthews

Introduction

Safeguarding the Mona Lisa is a tough job – a prized masterpiece that could be subject to any number of elaborate heists. That's why it is protected by powerful security tools like security cameras, motion detectors and bulletproof glass. However, the Mona Lisa does not change, replicate, or move, which makes the security job a lot easier. Data Security teams, on the other hand, must protect an asset that flies around the globe through clouds and third-party apps, where hundreds of employees demand daily, unrestricted access to that data to get their jobs done. While the Mona Lisa hangs safely in the Louvre, **data breaches are occurring at a rate of five every single day**. Data Security is a hard problem.

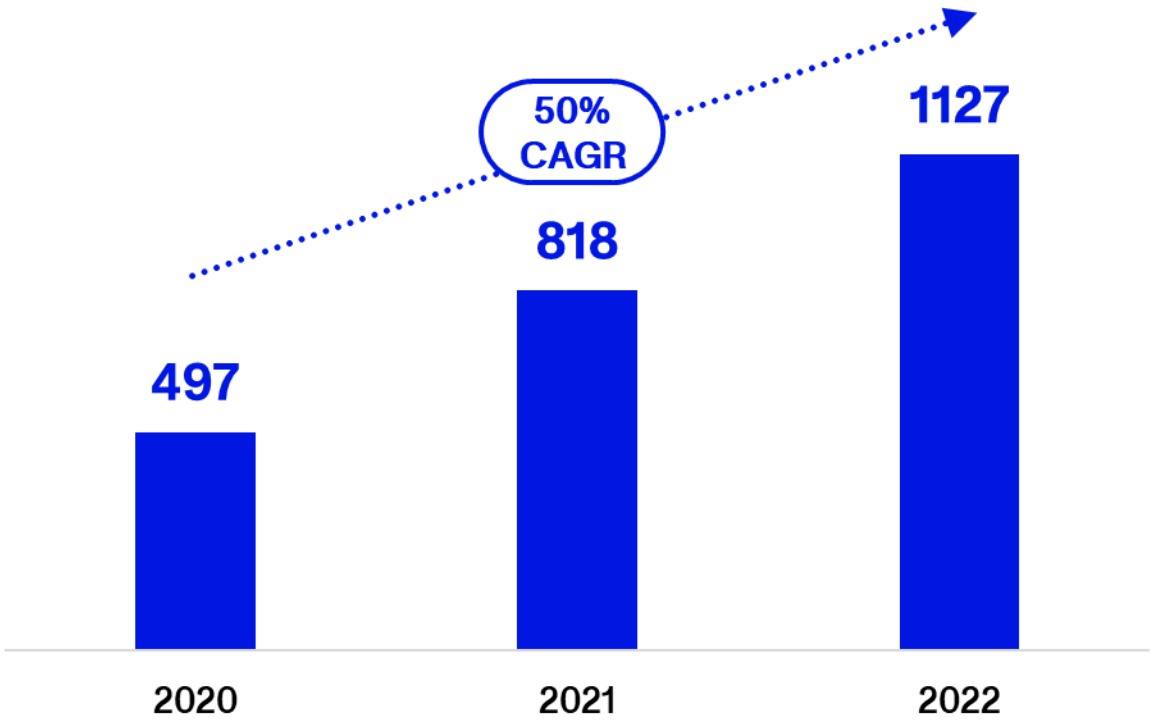
In this report, we explore what makes that problem so hard, why it's about to get even harder, and how we see the solution evolving through **the inversion of the cybersecurity stack**. Today, companies invest over $200bn on traditional security tools, but the $5bn Data Security market is too often an afterthought. To overcome existing data breach challenges, companies need to make Data Security the building block of their security strategy.



Data Security is a Major Issue

We've all been the subject of phishing emails. From personal experience, we know **the cyber-threat environment is intense**. And for large enterprises with valuable IP, customer data and multi-million-dollar transaction flows, it's an order of magnitude worse. In 2022, the average enterprise experienced over 1,000 cyber-attacks per week, or 150 every single day – more than doubling since 2020.

Average Weekly Cyber Attacks[1]



With sophisticated actors frequently attacking, data breaches occur at an alarming rate, with 1,862 breaches recorded in 2021.[2] These breaches impact a wide range of companies – 53% of organisations stated that they had experienced a **material loss of sensitive information** in the last year alone.[3] And there are many recognisable examples: Dell, American Express, Bank of America, and 23andMe all experienced breaches in the last few months.

[1] Wall Street Journal, Why even big tech companies keep getting hacked – and what they plan to do about it, 2022

[2] Ibid

[3] Rubrik, The State of Data Security, 2023. *Statistics reflect 5000+ Rubrik customers across 67 countries*

Those data breaches come with **real costs,** averaging $4.35 million each.[4] That bill could rise to $10mn+ for industries with extensive sensitive data, like healthcare.

What's driving such extreme costs?

1. **Lost revenue** due to downtime, management distraction, loss of intellectual property and reputational damage: 73% of consumers would reconsider using a company's product if their personal data was exposed.[5] Change Healthcare's systems were down for over a month, creating a backlog of $14bn in payments that were due through its claims clearinghouse.[6]
2. **Ransom fees:** Over 80% of impacted businesses pay ransom fees, which totaled $1.1bn in 2023.[7,8]
3. **Fines and regulatory penalties:** For example, Equifax's 2017 breach saw them pay £11mn to the UK Financial Conduct Authority and $575mn to the US Federal Trade Commission.[9,10]

In the information age, data is a company's most critical asset. Unlike physical assets, data cannot be replaced once it's lost. And for customers, they can't change critical identifiers like names and social security numbers. Once revealed, that information is public forever. You can't un-breach data, so the stakes are incredibly high. **Securing your data is hard**.

And the Problem Will Only Get Bigger

To understand how we arrived at this fraught situation, we must examine the key factors making data security an intensely difficult challenge.

1. **More data, duplicated across more systems, accessed more widely:** New data is growing rapidly – estimates suggest that the number of sensitive data files organisations must manage will increase ~5x by 2028.[11] Even a well architected security policy can become inundated with a growing stock of data to manage over time, **yet 66% of staff believe that their data has already outpaced their ability to**

[4] IBM, Turning data into value, 2022. *Statistics are from a survey of 3,000 Chief Data Officers across 30+ countries*

[5] Deloitte, Consumer data under attack, 2015

[6] Healthcare Dive, Change says its largest claims clearinghouses coming back online, 2024

[7] CSOonline, What is the cost of a data breach?, 2023

[8] ChainAnalysis, Ransomware Payments Exceed $1 Billion in 2023, 2024

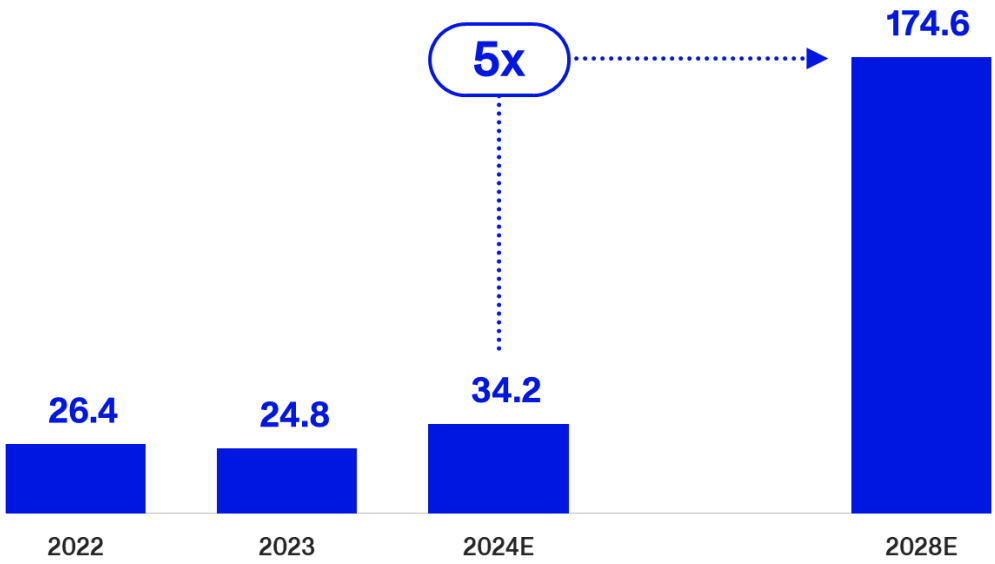
[9] FCA, Financial watchdog fines Equifax Ltd £11 million, 2023,

[10] FTC, Equifax, Inc., 2019

[11] Rubrik, The State of Data Security, 2023. *Statistics reflect 5000+ Rubrik customers across 67 countries*

secure it.[12] This is worsened by growing demands to be data-driven, which create a larger attack surface across different analytics systems and SaaS apps. Such demands also force companies to have more lax data access policies, lest the data security team becomes inundated with access requests and teams are left waiting too long for critical decision-making material.

Average Number of Sensitive Data Files under Management, in Millions[13]



2. **Disparate data sources create a lack of visibility into data:** With organisations managing an average of 187+ data sources and systems, it is incredibly difficult to appropriately classify and tag all incoming data.[14] **As a result, organizations are often unaware of the sensitive information they possess and where it resides, hindering their ability to safeguard it adequately**. Hence only 4% of companies have dedicated storage for sensitive information.[15]

3. **Inconsistent governance and security posture:** 39% of companies use more than one cloud computing vendor ("multi-cloud"), and 84% mix cloud and on-premises resources ("hybrid-cloud").[16,17] Also consider that companies may combine numerous data tools, such as using both Snowflake and Databricks. As a result, it

[12] Ibid

[13] Ibid

[14] Okta, Businesses at Work, 2022

[15] Rubrik, The State of Data Security, 2023. *Statistics reflect 5000+ Rubrik customers across 67 countries*

[16] Cloud Security Alliance, CSA Official Press Release, 2023

[17] Fortinet, Accessed via: Cloudzero, 101+ Cloud Computing Statistics, 2024

is difficult to manage a consistent and unified policy, leaving vulnerabilities, especially in the cloud. Cloud resources can become vulnerable due to poor access controls, unrestricted ports and unsecured backups – 99% of cloud IDs are "excessively privileged".[18] As a result, 80% of data breaches involve data stored in the cloud and cloud misconfigurations are **the most prevalent reason**.[19,20]

4. **A regulatory maze:** 137 countries have instituted data protection legislation, the most prominent being GDPR and the CCPA.[21] Individually, these regulations burden data teams with extensive considerations surrounding how they handle data, with equally painful costs for non-compliance (GDPR penalties can be as much as 4% of global turnover). But collectively, these disparate regulations create a maze that is near impossible for global businesses to navigate. With the recent proposal of an American Privacy Rights Act, all signs point to this situation intensifying.

5. **AI will supercharge these issues:** Text-to-SQL models will 10x the number of data users, while AI co-pilots drive more SaaS systems to demand enterprise data access. 64% of companies report being "under pressure" to adopt Generative AI, but 84% see cybersecurity as the primary roadblock.[22] Those roadblocks exist across the model lifecycle:

    a. Model training requires centralising training data which creates a honeypot for attackers.
    b. Model development often involves the use of open-source models as well as testing with numerous models, resulting in some enterprises running over 100 different models.[23] This makes it difficult to detect security vulnerabilities in all the models being used over time, leading to **more than 40% of companies suffering from privacy or security issues** in connection with AI models.[24]
    c. The inference phase exposes companies to prompt injection, where attackers can use specific prompts to drive the model to reveal sensitive

[18] IBM, 2022 IBM Security X-Force Cloud Threat Landscape Report, 2023
[19] HBR, Why Data Breaches Spiked in 2023, 2024
[20] Security Intelligence, Why are cloud misconfigurations still a major issue?, 2022
[21] UN Trade & Development, Data Protection and Privacy Legislation Worldwide, 2024
[22] IBM, What Generative AI means for your data security strategy in 2024, 2024
[23] Gartner, Top Strategic Technology Trends for 2024, 2024
[24] Ibid

information or inject malicious code into its underlying database to enable hackers to gain access to a system. For example, researchers at Google forced ChatGPT to reveal sensitive information from its training data.[25] Companies need control over what data employees provide to AI systems, which is difficult to manage at scale, particularly when employees make use of non-sanctioned AI systems, giving rise to "shadow AI."

It should be no surprise then that 40% of chief security officers feel that they are unprepared for the current threat environment. Implementing stronger data governance and security controls ranked as the number one priority for data leaders, above modernizing their data infrastructure or integrating AI into business processes.[26,27]

Companies must prioritize data security, not just for compliance reasons but to maintain business continuity, brand value and competitive positioning. Data breaches cannot be treated as routine - they pose existential threats with the potential to cripple businesses.

It's Time to Invert the Security Stack

Today, cybersecurity is a $200bn+ market and at just ~$5bn, Data Security appears to be an afterthought for CISOs. Putting data first promises a more sound strategy. Once you know where your sensitive data is, you can apply tools like access control and governance, but also have a much more targeted strategy for tools like threat intelligence, network security and cloud security. Perhaps, with efficiency gains from this approach, companies that start with Data Security could even see some overall savings on their cybersecurity budget. It's why we see data security becoming the fundamental building block of the broader security stack. That sentiment is echoed by Anshu Sharma, who's highest priority security actions start with data:

> Too many companies have too much data, in too many places, and grant too much access. Companies need to implement high value security strategies like sensitive data isolation and role-based access governance. It's all about reducing your surface area for attack because a breach is not really a breach if there is no PII.
>
> **Anshu Sharma, Co-founder & CEO, Skyflow**

[25] arXiv, Scalable Extraction of Training Data from (Production) Language Models, 2023

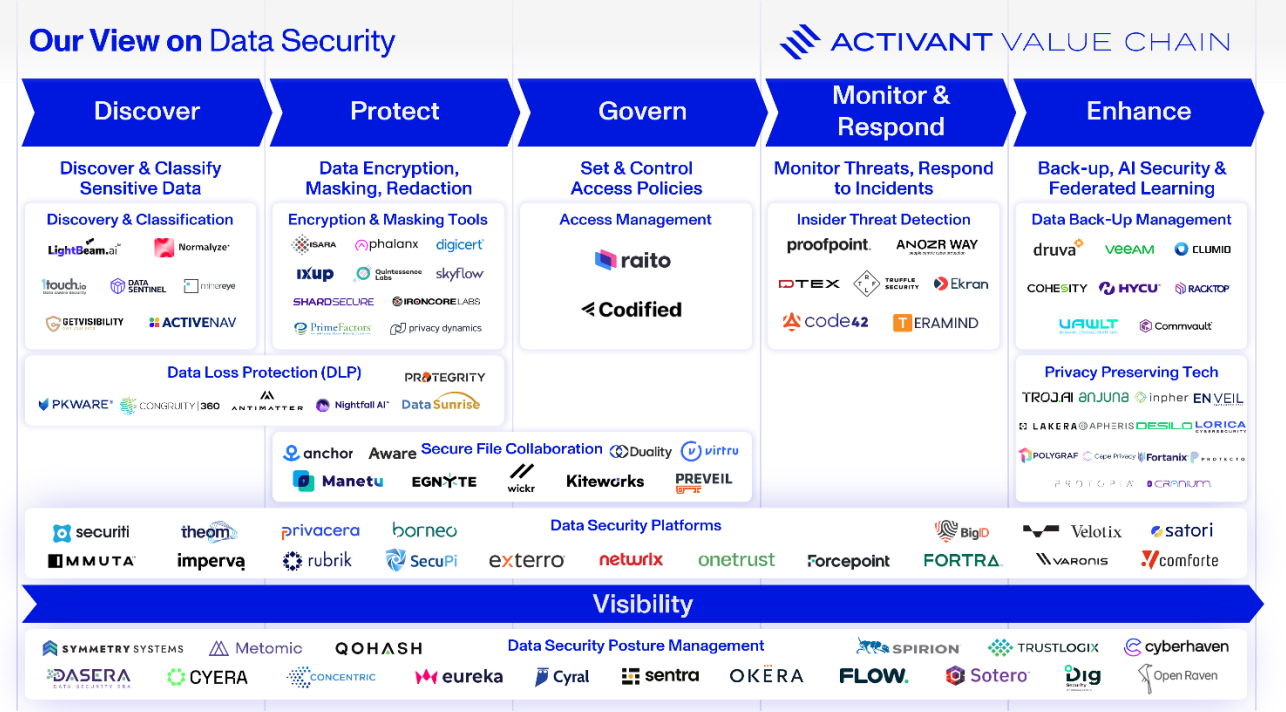
[26] ThoughtLab, Cybersecurity Solutions for a Riskier World, 2023

[27] Immuta, The State of Data Security, 2024. *Statistics reflect survey of 700+ data leaders across the US, UK Canada and Australia*

That's why we see the market evolving from its current estimates of ~$5bn to $47bn over the next decade (25% CAGR). We estimated our TAM based on industry-specific ACVs, considering the level of sensitive data and data complexity for that industry. From conversations with CISOs, companies are likely to spend anywhere from $500k - $5mn on data security, depending on their unique needs. We narrowed this range down to $500k - $1.5mn, noting that ACV's of $5mn would likely only apply to extreme outliers by size and complexity. ACV is multiplied out by the number of enterprises (firms with 1000+ FTEs) per the Bureau of Labor Statistics.

This is an emerging market segment with so much potential – data security has numerous exciting components and an ecosystem of start-ups that are leading the way forward. In the next section, we walk through how we break-down the data security market in the form of our Activant Value Chain.

Our View on Data Security



Working from left to right, we'll discuss the most critical pieces of this value chain, the jobs to be done for those software products, and highlight a few of the companies leading the way in each segment.

1. Discover

Data discovery and classification

Lack of visibility remains one of the core pain points when trying to secure data. Companies often don't know what data they have, making it impossible to adequately secure it. **Discovery and classification serves as the crucial first step.** These tools automatically locate and catalog all data, from sources like databases and SaaS apps. Classification models can tag data based on its type (PII, health, etc.) and risk level. Once companies understand their risk, the right **remediations** can be applied.

Leaders in this segment like Normalyze and Lightbeam scan enterprise data in any data store without installing agents on your devices, while working with both structured and unstructured data, where 1touch.io provide 100% accuracy.

2. Protect

Data encryption, masking, and erasure

A comprehensive encryption strategy can help companies comply with regulations and serve as a last line of defense if data is breached. Sensitive data should always be subject to some form of encryption, but traditional encryption can make data ***hard to use***. Companies might consider technologies like tokenization (a key tool used by credit card networks), format preserving encryption, and homomorphic encryption (which could power machine learning on encrypted data).

Skyflow is building a new piece of this architecture – the data privacy vault, a centralized point of control for sensitive data– to ensure that data doesn't proliferate across a company's systems. Shardsecure's Microshard technology protects data from ransomware by shredding it into 4-byte pieces and distributing it across cloud storage locations.

3. Govern

Data access governance

An access control policy dictates ***who*** may access a given dataset, and under ***what conditions***. No data security strategy can be successful if permissive access policies nullify other protection measures. But, if access control is too tight, teams can't get access to the data they need to do their work on time. Managing this tension, together with data access privileges that can vary by role, project, geographic region and time of day, makes these policies extremely complex. Luckily, activity-based access control

(ABAC) can consider contextual information together with attributes like users or files to grant fine-grained access controls with minimal administrative burden.

Tools like Raito provide a centralized point of control to write and enforce access policies, resulting in a 73% reduction in time to employee data access and an 86% reduction in time spent on access policies. Immuta provides tools to author policies in plain-language and mask data dynamically, powered by k-anonymization.

While the finer details of Single Sign-On (SSO) and Identity Providers (IdPs) are beyond the scope of this report, it's worth noting that access controls only work if apps support standards like SAML. Not all applications do, but Cerby have built a suite of automations that bring these non-standard applications under the umbrella of proper access control.

> The explosion of AI and remote work necessitates a revolution in access control – from outdated perimeter-based security to a user-centric model. Employees must be empowered to act independently across apps, devices, and locations, with technology stepping in to automatically fix any misconfigurations or inadvertent mistakes.
>
> **Belsasar Lepe, Co-founder & CEO, Cerby**

4. Monitor and Respond

Insider threat detection

While data breaches may provoke the thought of shadowy super-coders, many breaches occur due to company employees, and 75% of those insider breaches are actually non-malicious.[28] Insider Threat Detection seeks to make use of contextualized user analytics to detect these risks, and stop data leaks before they occur. DTEX looks to do just that, with a deep focus on behavioral science and the use of LLMs to automate the investigation of insider risk.

Data Security Posture Management (DSPM)

Like Insider Threat Detection, DSPM is a monitoring and analytics tool, but is focused on the company's total data security posture. DSPM vendors are generally uniform in their ability to perform data discovery and classification, provide visibility into data access and alert risky behaviors, as well as identify whether data has been appropriately encrypted and access controlled. The risk monitoring capabilities of DSPM have attracted significant investment in the space, including Cyera's $300mn

[28] Ponemon Institute, Cost of Insider Risks, 2023

Series C, and acquisitions of Dig Security and Flow by Palo Alto Networks ($95bn Market Cap) and Crowdstrike ($76bn Market Cap) respectively.

5. Enhance

Backup and recovery

When data breaches occur, companies need a way to resume operations. One way is to keep backups that are air-gapped from your operations. Backups can't protect companies from much of the significant costs of a data breach, and resuming operations may not be seamless, but the ability to return to operations in a doomsday scenario is paramount. Companies like Rubrik, Clumio and Veeeam provide the ability to automate backups to an append-only, air-gapped database, while also providing valuable attack detection analytics.

Privacy-Preserving Technologies (PPTs)

PPTs refer to a suite of technologies that ensure data privacy and security in AI/ML systems. For example, confidential computing creates a hardware-isolated enclave which protects the security of infrastructure and data being executed on. Lakera is protecting LLM users from prompt injection and sensitive data leaks, while offering an easy-to-integrate approach that has detected 100mn+ real-world vulnerabilities. Apheris's federated compute gateway allows users to collaborate on machine learning projects from different regions in a secure and compliant way, while Polygraf deploys any LLM interface into an enterprise's on-premises environment, so companies can leverage all of their existing security infrastructure with AI.

6. Data Security Platforms

Best-of-breed or a platform? It's an age-old question and as always, there are arguments for each. In an area with significant complexity like data security, there is a lot of room for best-of-breed players to build features that platform players don't have the focus to match. However, our view is that to fight issues like data siloes, lack of visibility and inconsistent policies, consolidating on a platform makes the most sense. Companies like Satori, Privacera and BigID are building out suites that cover the entire data security value chain. Securiti's Data Command Center operates across security, privacy, governance, and compliance. With a unified data discovery and classification engine, teams across all four domains can be powered by a common data model, boasting best-in-class accuracy rates and value-add tools like LLM security.

Conclusion

For too long, Data Security has been some combination of a budget afterthought or too difficult to get right. Now, companies are in crisis with a data breach cadence that measures ***more than daily.*** As teams race to implement AI, these issues just get worse. It's time for security teams to invert the stack, making data security the first thought and the fundamental building block of their cybersecurity stack.