UNIVERSITY
OF TURKU

Enhancing cyber assets visibility for effective attack surface management

Cyber Asset Attack Surface Management based on Knowledge Graph

Cyber Security
Master's Degree Programme in Information and Communication Technology
Department of Computing, Faculty of Technology
Master of Science in Technology Thesis

Author:
Marco Carmine Rossi

Supervisors:
Petri Sainio
Antti Hakkala

October 2023

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Cyber Security
Programme: Master’s Degree Programme in Information and Communication Technology
Author: Marco Carmine Rossi
Title: Enhancing cyber assets visibility for effective attack surface management
Number of pages: 63, 10 appendix pages
Date: October 2023

The contemporary digital landscape is filled with challenges, chief among them being the management and security of cyber assets, including the ever-growing shadow IT. The evolving nature of the technology landscape has resulted in an expansive system of solutions, making it challenging to select and deploy compatible solutions in a structured manner. This thesis explores the critical role of Cyber Asset Attack Surface Management (CAASM) technologies in managing cyber attack surfaces, focusing on the open-source CAASM tool, Starbase, by JupiterOne. It starts by underlining the importance of comprehending the cyber assets that need defending. It acknowledges the Cyber Defense Matrix as a methodical and flexible approach to understanding and addressing cyber security challenges. A comprehensive analysis of market trends and business needs validated the necessity of asset security management tools as fundamental components in firms' security journeys. CAASM has been selected as a promising solution among various tools due to its capabilities, ease of use, and seamless integration with cloud environments using APIs, addressing shadow IT challenges. A practical use case involving the integration of Starbase with GitHub was developed to demonstrate the CAASM's usability and flexibility in managing cyber assets in organizations of varying sizes. The use case enhanced the knowledge graph's aesthetics and usability using Neo4j Desktop and Neo4j Bloom, making it accessible and insightful even for non-technical users. The thesis concludes with practical guidelines in the appendices and on GitHub for reproducing the use case.

Keywords: Cyber Asset Attack Surface Management, Shadow IT, Knowledge Graph, Cyber Security

Table of Contents

1 Introduction

The rapid proliferation of cloud computing has revolutionised the way businesses operate, offering unparalleled advantages in terms of cost, scalability, and accessibility. However, this shift to the cloud has also exponentially increased the complexity of managing cyber assets and shadow IT, rendering traditional security measures insufficient and exposing organizations to new vectors of cyber attacks. Recent high-profile breaches and the evolving threat landscape underscore the urgent need for a comprehensive and strategic approach to cyber security.

Central to this discourse is examining Cyber Asset Attack Surface Management (CAASM) as a method to address emergent cyber security challenges. This thesis focuses on one crucial aspect of CAASM, which is "visibility" in cyber security. In exploring the relationship between cyber security and visibility, this thesis is guided by the following key research questions:

- **Visibility's role**: how is the concept of visibility within cyber security received and valued by organizations? Is there a tangible emphasis on its indispensability?
- **Market perception**: how effective are visibility-centric tools in the vast marketplace of cyber security solutions? What is the market size of visibility tools and its expected growth?
- **Potential vulnerabilities and applicability**: what risks and threats might organisations face due to a lack of visibility prioritization? How feasible is it to implement additional visibility mechanisms within organizational frameworks?

In order to answer our research questions, we examined surveys and interviews from multiple studies to investigate the importance of visibility in cyber security. We gained insights into how technical and managerial figures perceive its value. To assess the effectiveness of CAASMs, we conducted extensive market analyses, including vendor assessments and Gartner reports. Through the review of studies on shadow IT and cloud security, we demonstrated some of the potential threats, and our case study highlighted the role of visibility.

The thesis is divided into five chapters, each contributing to a comprehensive response to the research questions. Chapter 2 provides an overview of cyber assets and shadow IT, discussing their impact on managing the cyber attack surface. It introduces the Cyber Defense Matrix as a framework for understanding the cyber security landscape and examines the effects of cloud computing on cyber assets, the business needs driving cloud adoption, and the challenges posed by shadow IT. It also analyses the economic losses and cyber security risks associated with shadow IT. Chapter 3 focuses on the importance of visibility in Cyber Asset Attack Surface Management. It analyses market trends, referencing the Gartner security report to understand the evolving landscape of visibility tools.

It also briefly discusses various Attack Surface Management tools, including Cloud Access Security Brokers (CASB), Endpoint Attack Surface Management (EASM), and CAASM, highlighting their features, benefits, and limitations. Chapter 4 focuses on Starbase, a Neo4j-powered tool that consolidates digital assets and relationships for a comprehensive view of an organization's digital ecosystem. It outlines Starbase's three pillars, deployment methods, and the importance of API integrations in continuous monitoring. The chapter also discusses the role of graph databases, knowledge graphs, and search algorithms in managing cyber asset connections. Additionally, it explores the challenges in graph drawing, precisely the force-directed model and aesthetic considerations. Chapter 5 presents a case study on how Starbase, integrated with GitHub, enhances situational awareness, cyber security, and management of directories, employee accounts, and permissions in a GitHub organization project. It demonstrates the tool's capabilities in streamlining processes, ensuring policy compliance, and optimizing visibility in digital ecosystems. Key processes involved in the integration include deploying a GitHub application, simulating user interactions via APIs, and leveraging Neo4j for visual data management. The chapter emphasizes the necessity of multi-factor authentication (MFA) for admin users and the role of Starbase and Neo4j in simplifying policy compliance administration. It also highlights the aesthetic and functional advantages of Neo4j Bloom in graph visualization and interaction, which we integrated since it is not adopted by default when deploying Starbase. Finally, the chapter discusses the importance of knowledge graphs in cyber security, the challenges faced in their application, and proposes avenues for future research.

The research was conducted through a comprehensive review of existing literature, an analysis of market trends, and a practical case study involving the deployment and testing of the Starbase tool in a simulated environment. While the findings provide valuable insights into the effectiveness of CAASM technologies and their potential to enhance cyber security, it is essential to acknowledge the limitations of this study. The research focuses on a specific CAASM tool, Starbase, and its integration with GitHub, which may not represent all CAASM tools or cloud environments. Additionally, the study is conducted in a limited environment, which may only partially replicate the complexities and challenges faced in a real-world scenario.

The thesis concludes with a synthesis of the essential insights, strategic advice for entities considering the adoption of CAASM technologies, and proposals for forthcoming research and applications. Additionally, the appendices section includes detailed information for reproducing the same use case and performing a test run, with all specific data and instructions made accessible on GitHub.

2 Cyber assets and shadow IT: the role of visibility in managing cyber attack surface

The following paragraph has been intended to analyse the role of visibility in typical business environments. The main scope is to understand why visibility is a critical requirement when considering the cyber security posture of an infrastructure.

Besides, before introducing the current solutions available in the market and the selected one, it is essential first to understand cyber security problems and business requirements.

Nowadays, the technology landscape is getting broader and broader, resulting in a complex system of solutions that negatively affect the already significant cyberspace dimension. This topic is well discussed in [1], where every tool is reported in the so-called *Cyber Defense Matrix*. This investigation underscores how the cyber ecosystem is grappling with internal intricacies, thereby complicating the task of selecting and implementing inter-compatible solutions in an organized manner.

2.1 Cyber asset

As we want to examine, in a clear way, possible security requirements, we should clarify what we have to secure. Companies can be seen as a complex system of connected physical assets, hardware-based assets (devices), physical tokens, digital assets (software), licenses and data.

Cyber assets come in various forms. As a first macro segmentation, we can dive assets into two classes: **tangible** and **intangible**. Furthermore, anything with an intrinsic value for the company's system must be considered part of one of these two *hard-clustered*[1] sets. This approach of classifying cyber assets reflects the definition of the NIST [2]: *"the data, personnel, devices, systems, and facilities that enable the organisation to achieve business purposes."*

Recognizing and conceptualizing all potential assets can be challenging, as anything can be utilized to generate value for an organization. Employees are a crucial component of a company's assets, not only due to their expertise but also as cyber assets that could be one of the weak points of the chain (it is relevant to mention the high use of social engineering attacks[2]).

[1] Hard clustering algorithms defines data partitions in a fixed and limited number of sets mutually exclusive.

[2] A social engineering attack refers to the manipulation of individuals through psychological tactics to deceive them into performing actions compromising the CIA.

It is imperative to acknowledge the significance of employees as assets and take necessary precautions to safeguard against potential risks. Besides, the digitalisation of the main processes and infrastructures has introduced an exponential source of assets; one of the most common is cloud computing. According to the last report of Gartner: *"New ways of visualising and prioritising management of an organisation's attack surface are required as enterprise IT becomes more dispersed, owing to the expansion of public-facing digital assets and increased use of cloud infrastructure and applications." [3]*

The attack surface includes all the vulnerabilities an unauthorised entity can exploit to harm at least one of the CIA properties. Confidentiality, integrity, and availability, often called the CIA triad, are the fundamental pillars of cyber security. Confidentiality ensures that sensitive information is accessible only to authorised individuals or entities. Integrity ensures the accuracy, consistency, and trustworthiness of data and systems. Availability ensures that resources and services are accessible and operational when needed. [4]

Organisations must prioritise safeguarding cloud services to minimise their attack surface. Digital attack surfaces include websites, servers, databases, and laptops. Weak passwords, misconfigurations, shadow IT, shared databases, and outdated devices increase cyber attack vulnerability. Additionally, addressing the physical attack surface is crucial, involving securing physical assets and mitigating risks such as theft, insider threats, and baiting. Social engineering attacks rely on psychological manipulation to exploit human ignorance, often through phishing attempts and tricking individuals into sharing sensitive information or downloading malicious content. Organisations can strengthen their security by understanding and addressing these attack surfaces. [5]

There are numerous cyber assets that organizations must safeguard against cyber attacks. The number and variety of these assets make it challenging to identify vulnerable points and define an attack surface. As a result, a more adaptable, systematic approach is often necessary. Practices like Bring Your Own Device (BYOD), the rise of the Internet of Things (IoT) and Software-as-a-Service (SaaS) products only add to the complexity of the task. Recognizing that cyber assets encompass a wide range of valuable digital resources is vital.

Furthermore, as technology evolves, so do the nature of cyber assets and threats. Therefore, organizations must adopt a dynamic and flexible approach to cyber security that can adapt to new challenges as they arise. Given the vast and evolving landscape of cyber assets and threats, a one-size-fits-all approach is unlikely to be effective.

Instead, organizations should adopt a layered, defence-in-depth strategy that includes measures such as regular risk assessments, employee training, and robust security policies and procedures. By doing so, organizations can better safeguard their critical cyber assets and protect themselves against severe threats and consequences.

2.1.1 Understanding the Landscape: Cyber Defense Matrix

Listing all the possible cyber assets and related threats demands complex work, which could hardly cover all the possible scenarios. Indeed, Sounil Yu, the author of [1], has defined a unified tool that allows representing and distinguishing the cyber security landscape in a clear and mutually exclusive way. This framework uses a *MECE* approach, which stands for *mutually exclusive and collectively exhaustive*. The result is that a large problem is divided into smaller ones.

The Cyber Defense Matrix splints along two dimensions. The first dimension presents five distinct functions based on the NIST framework: **Identity**, **Protect**, **Detect**, **Respond**, and **Recovery**. Additionally, assets are divided into five classes: **Devices**, **Networks**, **Applications**, **Data**, and **Users**.

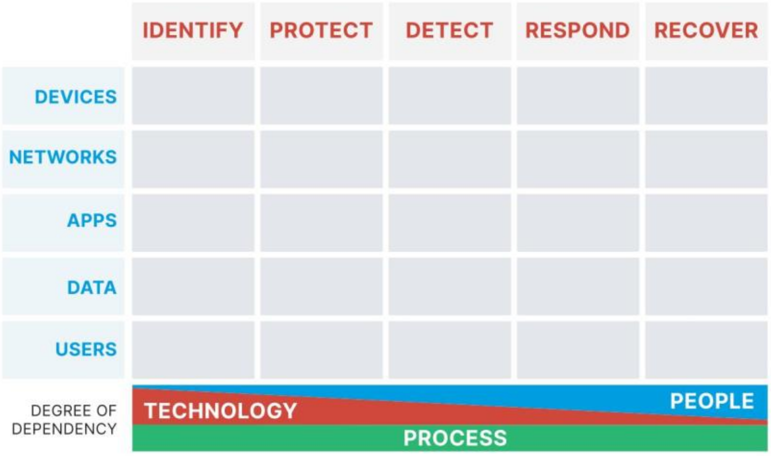


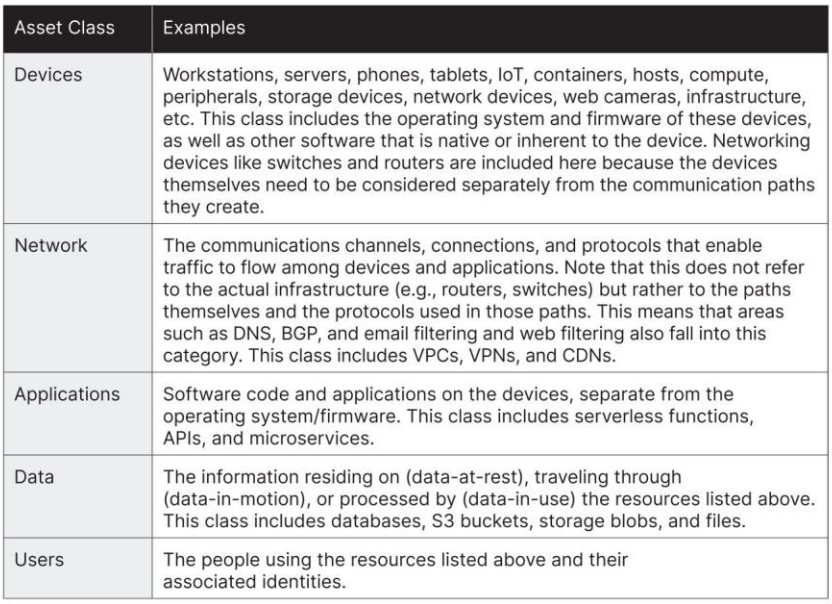
Figure 1: The Cyber Defense Matrix, which offers a simple grid schema to categorise cyber assets and cyber security measures. The Cyber Defense Matrix offers a structured approach to navigate the landscape of security solutions better. It acts as a guide for professionals to identify which products address specific security concerns swiftly and clarifies the primary functions of these products. [1, p. 7]

Technology plays a crucial role in the initial stages of cybersecurity, such as identification and protection. However, human involvement becomes increasingly important as one moves towards detection, response, and recovery. Processes remain vital across all five stages. This aspect of the matrix highlights the significance of balancing our investment and trust between technology, human resources, and established procedures to address cybersecurity issues effectively.

The value of this framework is not only the self-explanatory format, which permits identifying assets that are usually easy to forget, but it also helps to define the best technical solutions in market language and business oriented way. Core functions are immediately determined, and problems can be discerned to select the right products. [1]

As mentioned above, we use the term cyber assets to group a large set of heterogeneous assets. Each of them may have a different priority in some enterprise context and should be treated in a specific manner. The term *cyber* does not help to clarify what these assets are, but instead, it represents that both tangible and intangible are considered in the cyber security landscape. Furthermore, a cyber asset can be considered responsible for introducing liabilities. The second meaning of the term *cyber* should now be more precise, as liabilities typically result in adding new attack surfaces and require security measures.

An excellent example of assets divided into the five classes is shown in the following table.



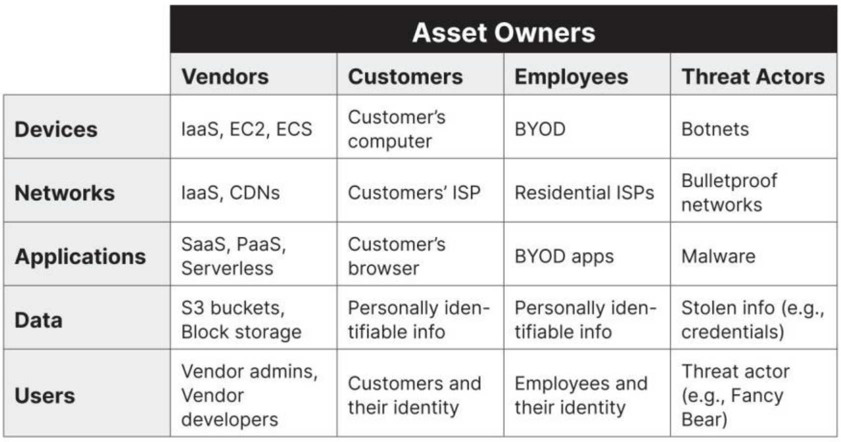
Asset Class	Examples
Devices	Workstations, servers, phones, tablets, IoT, containers, hosts, compute, peripherals, storage devices, network devices, web cameras, infrastructure, etc. This class includes the operating system and firmware of these devices, as well as other software that is native or inherent to the device. Networking devices like switches and routers are included here because the devices themselves need to be considered separately from the communication paths they create.
Network	The communications channels, connections, and protocols that enable traffic to flow among devices and applications. Note that this does not refer to the actual infrastructure (e.g., routers, switches) but rather to the paths themselves and the protocols used in those paths. This means that areas such as DNS, BGP, and email filtering and web filtering also fall into this category. This class includes VPCs, VPNs, and CDNs.
Applications	Software code and applications on the devices, separate from the operating system/firmware. This class includes serverless functions, APIs, and microservices.
Data	The information residing on (data-at-rest), traveling through (data-in-motion), or processed by (data-in-use) the resources listed above. This class includes databases, S3 buckets, storage blobs, and files.
Users	The people using the resources listed above and their associated identities.

Figure 2: By adopting the Cyber Defense Matrix, JupiterOne has identified a list of possible (and main) cyber assets in standard organizations' environments.[1, p. 10]

In the real world, businesses face limited resources, making prioritizing cyber assets according to their specific requirements and context an absolute must. While it is crucial to carefully analyse the organization's cyber assets to determine the most critical ones to its operations, allocating resources efficiently to maximize cyber security posture is equally important. However, prioritizing some assets should not mean neglecting others, as even assets with lower priority values can still pose potential entry points for cyber attacks or lead to significant losses if compromised.

We must understand that, as we describe in the following sections, visibility is crucial to define a clear and solid perimeter. To best manage and defend the infrastructure, a perimeter or, more specifically, limiting its attack surface is crucial. Every cyber asset has a crucial effect on the attack surface and could represent a door into the protected environment. Ignoring a small point of the attack surface may lead to catastrophic consequences.

The last aspect related to cyber assets should be seen as an additional label, a characteristic or better as an additional variable, which is asset ownership. Additionally, some organisations focus their strategies only on owned cyber assets. However, it is good to consider that in an era characterised by cloud computing and third-party software, accounting for not directly owned cyber assets is essential.



	Asset Owners			
	Vendors	**Customers**	**Employees**	**Threat Actors**
Devices	IaaS, EC2, ECS	Customer's computer	BYOD	Botnets
Networks	IaaS, CDNs	Customers' ISP	Residential ISPs	Bulletproof networks
Applications	SaaS, PaaS, Serverless	Customer's browser	BYOD apps	Malware
Data	S3 buckets, Block storage	Personally identifiable info	Personally identifiable info	Stolen info (e.g., credentials)
Users	Vendor admins, Vendor developers	Customers and their identity	Employees and their identity	Threat actor (e.g., Fancy Bear)

Figure 3: The Cyber Defense Matrix can be applied in different contexts. The figure shows a use case where the tool is used to identify the entity responsible for managing and controlling the liabilities of cyber assets. [1, p. 11]

In the context of business-oriented cyber security, certain assets require prioritised attention. These assets, known as critical assets, according to the definition provided by CISA (Cyber Security and Infrastructure Security Agency), are fundamental to maintaining operational continuity and achieving the organisation's mission. Breaches affecting these critical assets' confidentiality, integrity, or availability can lead to substantial business consequences. [6] Before taking any action, it is crucial to identify the cyber assets involved. Once identified, it is necessary to assess the risk associated with each asset to determine which ones are the most critical. This aligns with the definition of critical assets and ensures that the most important assets are given priority attention.

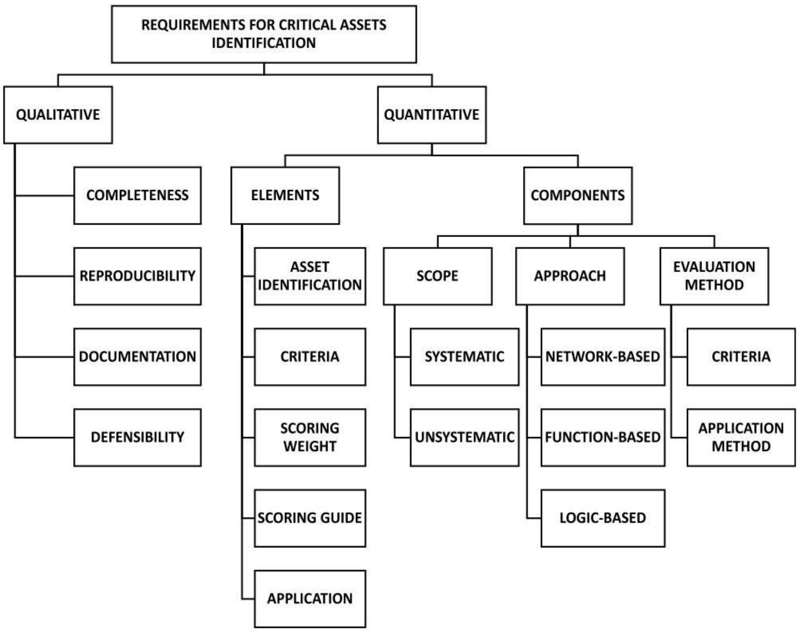


Figure 4: Requirements for critical assets identification. The first macro-segmentation in categories is characterized by qualitative and quantitative classification. Every path identifies the main points to be considered when we want to identify cyber assets. [7]

As shown in Figure 4, the paper [7] analyses critical assets in Cyber-Physical Systems (CPSs). Protecting essential assets within these systems is crucial, as failures can lead to severe consequences like loss of equipment, injuries, and disruptions in essential services.

The proliferation of Internet of Things devices has increased cyber threats, making situational awareness crucial for protecting CPSs. The article focuses on monitoring critical assets in CPSs to improve situational awareness. It analyses various techniques and tools to gather detailed information from these systems. Identifying and comprehending assets in CPSs is essential for efficient recovery, security maintenance, and patch management.

However, CPSs have IT and Operational Technology (OT) resources, making traditional asset discovery methods less efficient for appraising OT assets with unique features and time-critical considerations. In the past, OT equipment, such as industrial control systems (ICS) and industrial Internet of Things (IIoT) devices, were considered secure because they were isolated from other networks. However, these devices are often connected to corporate networks and the internet in today's interconnected world, making them vulnerable to cyber attacks even without direct connections. As a result, any breach of the OT infrastructure can lead to severe consequences, including compromise of confidential organizational data, production disruptions, or product quality. [8]

In [7], the authors examined various asset discovery approaches, including passive and active scanning techniques. Passive scanning is the safest way to discover ICS/OT resources, as it listens without intercepting or requesting responses from assets. Active scanning may not be ideal for sensitive assets. However, a hybrid approach that combines active and passive scanning emerges as a potential solution to gather data from network traffic without disrupting ICS assets. [7] Moreover, the article emphasizes the significance of dependency modelling as a critical aspect of asset discovery in CPSs. Dependency modelling entails understanding the correlations between different elements within the system. By mapping out relationships like cause/effect dependency, location dependency, resource dependency, and input/output dependency, it becomes possible to recognize the direct and indirect elements more susceptible to compromise. This insight allows for prioritizing key components and understanding potential vulnerabilities in the system. [7]

The dependency approach confirms that linking data sources and context information is fundamental to improving overall visibility. Moreover, the focus on knowledge graph-based techniques in the last chapter further strengthens the significance of this approach in enhancing system visibility and understanding interconnections within the system.

In conclusion, the paper [7] shows that by employing appropriate asset discovery tools, such as passive and active scanning, and integrating dependency modelling, CPSs can enhance their situational awareness and fortify their cyber security defences. Understanding and safeguarding critical assets are essential steps in mitigating potential cyber threats and ensuring the reliable operation of CPSs.

2.2 Cloud security and shadow IT

As discussed in the previous section, today’s infrastructures are characterised by the massive use of cloud computing services. If the management of owned cyber assets poses IT departments with challenging tasks, the additional surfaces of cloud services reduce the infrastructures’ controllability.

In recent years, cloud-native technologies have seen an explosion in the number of companies adopting these technologies. Cloud computing has gained popularity thanks to its ability to enable business agility, maximising flexibility and reducing operational costs. While cloud service providers are responsible for enhancing their security measures to earn the complete trust of their customers, it is crucial to educate users about the risks associated with adopting these technologies. The study [5] reveals that 86% of the companies are active or have recently started using cloud-native services. Of these companies, 58% state that growing cloud API volume is causing security problems.

Shadow IT *"represents all hardware, software, or any other solutions used by employees inside of the organisational ecosystem which have not received any formal IT department approval."* [9] The proliferation of shadow IT is an alarming issue for companies, given its rapid expansion and prevalence. Industry analysts have identified that unsanctioned applications and services pose a significant risk to Chief Information Security Officers (CISOs) in managing and securing digital ecosystems.

According to Gartner, shadow IT could be responsible for approximately 30% of breaches in 2020. Frost & Sullivan's report also highlights this problem, with more than 80% of respondents admitting to using non-approved applications in their daily work. The study shows that shadow IT is growing at a rate of 5% annually and now accounts for 30% of IT budgets within companies. In larger organizations, shadow IT is impact is even more pronounced, with studies by Gartner indicating it comprises 30 to 40 per cent of IT spending. At the same time, Everest Group suggests it can make up 50% or more of IT spending in some organizations. CISOs face multifaceted challenges with shadow IT, including centralizing control, maintaining oversight over assets and applications, and promptly identifying unknown IT risks. [10]

The main forces that move employees to use services within the organisation's perimeter are the lack of awareness and the willingness to improve the quality of their work by using these services. The introduced risk should be classified as internal and could lead to severe threats. The strong correlation between shadow IT and cloud security is due to today's trends seeing cloud computing as the primary resource of services. Even if companies try to limit the allowed applications that can be installed on a device, most services are accessible in the SaaS model. Employees can access services using a browser and their private accounts or installing browser extensions.

To maintain a robust security posture, CISOs must confront the complexities of shadow IT head-on, implementing a comprehensive and unified security approach that mitigates risks and fortifies their organizations against potential threats. Understanding and addressing the challenges posed by shadow IT is critical for safeguarding digital assets as companies navigate the constantly evolving cyber security landscape. [10]

2.2.1 The effects of cloud computing on cyber assets

A valuable study that analyses the latest trends in cyber security and cyber assets is [6] conducted by JupiterOne. In their annual report, *"The State of Cyber Assets Report"* (SCAR), they follow the Cyber Defense Matrix approach, which we already discussed [1], and clarify the evolution of the cyber assets landscape.

Jasmine Henry, Senior Director of Data Security and Privacy at JupiterOne, states that the inevitable attack surface growth has made complexity reduction impossible for cyber security teams. [6] The proposed approach is, instead, to understand cyber assets, liabilities, attack surfaces and their relationships in modern enterprises.

"By redefining the cyber security control plane, we can better adapt to our environments' growing complexity." [6]
Consistent awareness across cyber assets and environments is crucial to managing the complexity. This requires a methodical shift toward unified cyber insights and interoperable cyber security tools. [6]

The main research questions, and the most related to this thesis, are the following:

- What is the composition of cyber asset inventories?
    - How many assets and accounts are in an AWS, GCP, or Azure environment?
- Are security practitioners inventorying all types of cyber assets?
    - What is the value of a cyber asset?
- Do security teams have comprehensive, data-driven visibility across the attack surface?
    - How many security data sources are being correlated and aggregated?
    - How do security practitioners interact with security data?
- Which assets tend to have more liabilities (or vulnerabilities)?
    - Which assets are the most critically vulnerable?
    - What tools are security practitioners using to identify and detect vulnerabilities?
    - What is the ratio of vulnerabilities to assets, and which types of assets have the most vulnerabilities?
- How do security practitioners navigate their attack surface and data sources?

In [6], 89.7 million cyber assets have been analysed. Data has become the most significant category, with 34.9 million assets, while Devices are the second, with 21.1 million. The results show that security teams are understaffed and responsible for 393,419 assets on average. Organisations have increased their cyber assets by 132.86%, resulting in a mean value per cyber asset of $17,711. This value is calculated considering the intrinsic value of an asset and its created or future business value within the company context. [6]

Furthermore, the complexity is measured by the number of security data sources teams must manage. The average has resulted in 8.67; mid-sized organisations present the highest value with 10.11 sources. The assets with the lowest data sources are Networks and Data, 62.72% and 49.29%, respectively. [6]

Cloud hosts, part of the Devices class, have represented 1/3 of security findings and 96.1% of critical-tagged findings. What should be noticed is that Data has the lowest visibility in terms of data sources, 49.29%, but resulted in the most vulnerable class of assets with more than half of security findings. As expected, today's enterprises are characterised by cloud-based apps, services and assets. Considering only the assets originated from a cloud service provider [3]and not the whole number of SaaS and cloud applications, the number of cloud-based assets is 60%. On average, 225 accounts and subscriptions are managed by security teams; the highest value is for mid-sized organisations with 559 accounts and subscriptions. [6]

Employees' access to multiple services poses a significant risk to confidentiality, particularly with the increasing adoption of cloud technologies. As a result, companies must adopt new measures to mitigate potential risks and safeguard valuable internal information. This requires a comprehensive approach incorporating innovative strategies and control measures to ensure the security of sensitive information. By leveraging cloud technologies and implementing strict control measures, businesses can enhance their security posture and reduce the risk of data breaches and confidentiality loss.

[3] A Cloud Services Provider (CSP) offers digital services via the Internet, ranging from data storage to software applications. These services eliminate the need for businesses to own physical infrastructure. Users simply access resources on-demand, streamlining operations and reducing costs.

2.2.2 Business needs

Considering the previous results and analysis, we should notice how security teams are overloaded daily with work. Their proactivity and capability to identify and react are limited by the quality of data they have. However, more data does not imply better performance.

The number of data sources is only a part of the visibility capabilities and cannot be considered an indicator of maturity. Other factors must be considered too, for example, the number of security technologies, the efficacy in extracting context and meaning from the collected information, the prioritisation capabilities and more.

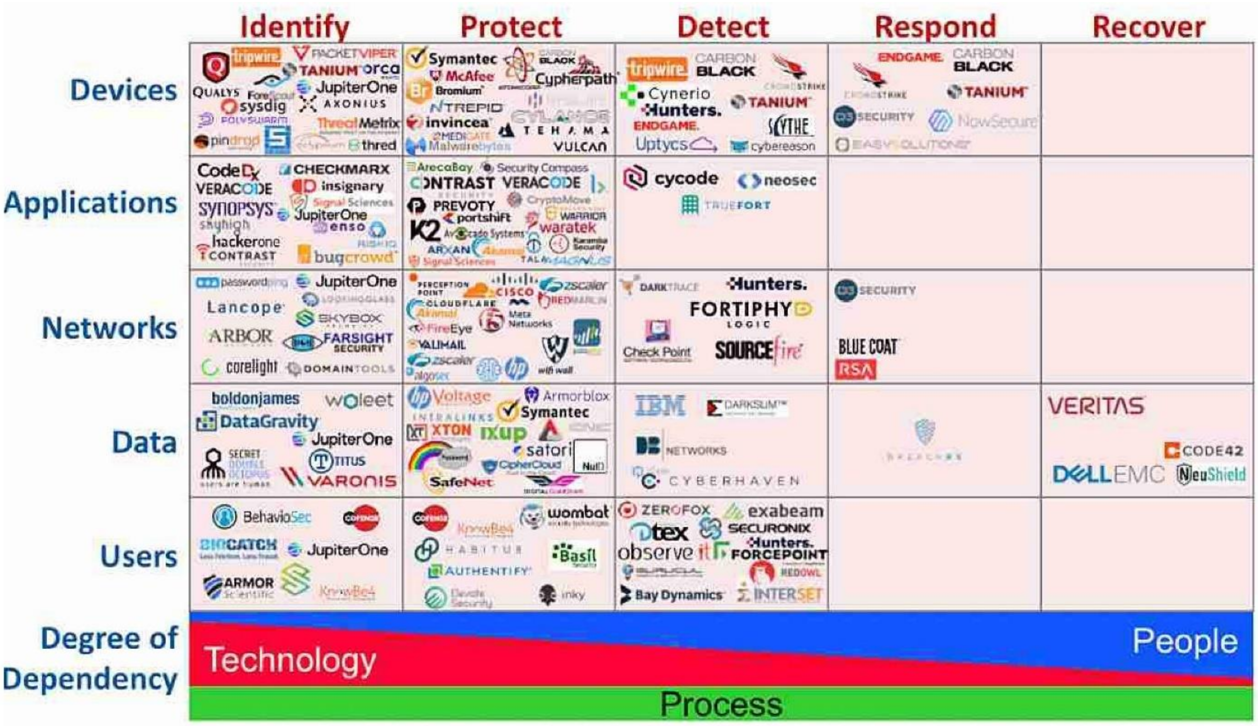
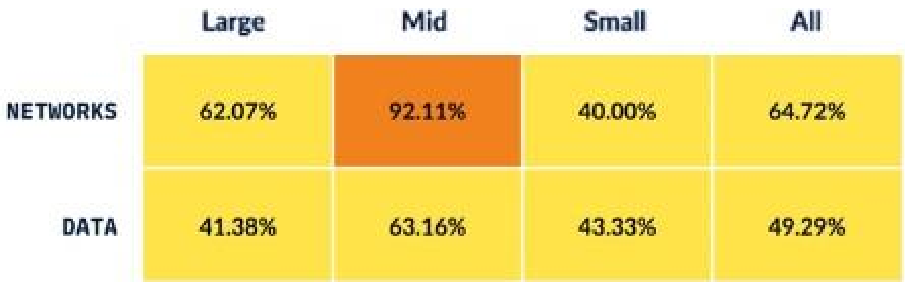


Figure 5: Security technologies by asset classes and operational functions. This figure wants to accentuate how big and complex the cyber security landscape is. Many tools are available to handle specific areas, making contextualization and cooperation more complex. The constant process of defending IT/OT systems involves technology first, to escalate to hands-on procedures in worst-case scenarios. [6]

Security teams usually have to control multiple types of devices, licences, and software and, most importantly, be familiar with multiple security technologies. As shown in Figure 5, plenty of solutions are available on the market. However, processing information and linked evidence to get the complete picture is not always trivial. The analysis showed that the categories of Data and Networks are the ones requiring the introduction of new data sources.



	Large	Mid	Small	All
NETWORKS	62.07%	92.11%	40.00%	64.72%
DATA	41.38%	63.16%	43.33%	49.29%

Figure 6: The figure shows the asset super classes with the lowest level of visibility and data source adoption by considering large, medium and small enterprises. The percentages, obtained by a deep analysis of JupiterOne, identify the small capability of teams. [6, p. 29]

Considering network assets management, Cisco offers comprehensive and holistic solutions, which are well known for their performance and quality. These technologies, namely Cisco Application Centric Infrastructure (ACI) for data centre environments, Cisco Digital Network Architecture (DNA) Center for campus networks, and Cisco Network Services Orchestrator (NSO) for enterprise and service provider administration, are presented as compelling choices for effective network asset management.

Regarding data security, [6] highlights the challenges faced by security teams, who often struggle to consolidate and integrate vast amounts of data from multiple data sources. A unified dashboard collating data from diverse sources is convenient and essential. It enables these teams to move from reactive to proactive strategies, utilizing data more efficiently to identify intricate dependencies, anomalous patterns, and potential threats.

Understanding the relationships between different digital assets is of utmost importance. Graphs are a powerful tool that visually represents these relationships, capturing the meta-data associated with these connections and providing a more holistic view of potential vulnerabilities.

Context is crucial in interpreting security data. As [6] has pointed out, the data class is particularly susceptible to breaches and vulnerabilities. It constitutes 59.51% of security findings, translating to an alarming 77.64 million potential points of failure. Host scanners contribute 30.16% of these vulnerabilities, while Cloud Security Posture Management (CSPM) is responsible for 21.34%. The empirical data from [6] validates the assumptions held by many experts about typical business ecosystems and solidifies the importance of the tools and methodologies elaborated upon in the succeeding chapters. An integrative tool that harmonizes cyber assets and data sources is required in today's digital age. Such a tool, backed by the representational power of knowledge graph models, can significantly enhance the ability of organizations to derive contextual and actionable insights from a maze of complex, interwoven relationships.

Furthermore, the intrinsic value of cyber assets in today's business landscape is undeniable. Their proliferation, both in terms of value and sheer number, is evident. This escalation is intrinsically linked to the ubiquity of cloud services in modern businesses. As illustrated in [6], an overwhelming 96.1% of vulnerabilities and security findings have ties to cloud-based operations. The subsequent sections will offer a more profound exploration of shadow IT, shedding light on the principal catalysts steering this growing phenomenon.

2.2.3 Shadow IT

"Gartner Predicts 30% Of Breaches Would be Due to Shadow IT." [10]

In [6], the percentage of assets originating from cloud service providers (CSP) was 60%. The adoption of CSP for firms with one thousand or more employees was, on average, around 94%. The main CSPs are Amazon Web Service (AWS), Google Cloud Platform (GCP) and Microsoft Azure.

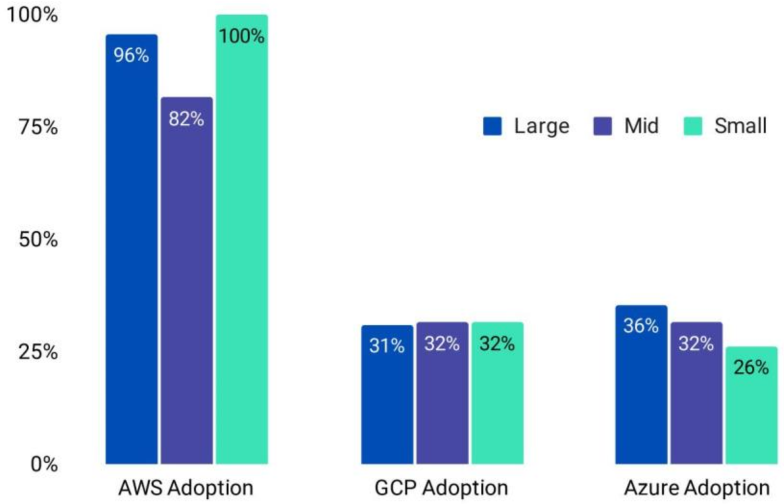


Figure 7: Average adoption of the main three CSPs among small, medium, and large organisations.

The importance of contextualising data by interconnecting findings could be noticed, considering that companies are very likely to adopt multiple cloud solutions. [6] shows that 90% of vulnerabilities have been estimated to come from cross-account security weaknesses. This means that one of the principal challenges they face is that cross-account vulnerabilities are slightly reported in centralised repositories and usually come from misconfiguration of permission rights and accesses given to those connected accounts. Relationships are then essential to identify possible permissions weaknesses and prevent information leakage.

Shadow IT *"represents all hardware, software, or any other solutions used by employees inside of the organisational ecosystem which have not received any formal IT department approval."* [9, 11, 12]
The main research we decided to use for the dissertation on shadow IT are [9], [13], and [14]. Both [9] and [14] use a methodological approach to analyse the problem of shadow IT from different perspectives, particularly considering business requirements. [9] is older than [14], however, it perfectly covers what we consider the most critical shadow in related research questions. Additionally, we decided to compare results with [13] and [14] since they analyse the Dutch market and focus more on the enterprise requirements with more extensive market research.

The previous shadow IT definition reveals the nature of this threat as an internal one. It is then important to understand the relationship between business and shadow IT. The main research questions are:

- What types of shadow IT software are used by employees?
- What are the security IT risks when using shadow IT software?
    - Does open-source shadow IT software increase the risks?
- What are the motivations for using shadow IT?

In [9], the author interviewed nine different organisations, including a public governative organisation of small-medium size. The questionnaire included open questions and was meant for Chief Information Officiers (CIOs) or IT managers. [9] It also includes a deep analysis of shadow IT in a Fortune 500 company with over 10,000 employees. The analysis was conducted with BigFix[4], a software analyser tool.

The endpoints scansion identified, from 10,001 devices, 19,633 different versions of software applications, which were installed 527,403 times among the different devices. [9] After collecting this information, they grouped software available in different versions; the non-authorised applications were 2,965. The main types of software were PDF and file tools (e.g., PDF modifier, viewer, ZIP compressor, and more) and communication applications (e.g., social media like Facebook applicant, Google Meet, Microsoft Teams, Skype, and more).

[4] BigFix is owned by IBM and it is used to manage any endpoint device, providing a lightweight management platform.

Other studies cited in [9] show that employees' primary concern at the time was the disclosure of sensitive information for 43% of them. Nevertheless, employees continued with their bad security behaviours for lack of IT knowledge or were unaware of the company's policies (e.g., file sharing policies).

From questionnaires [9] noticed that the main factors of shadow IT among employees were the new concept of Bring Your Own Device (BYOD), which originated from the always-connect capabilities of mobile devices, and the trade-off of a faster and easier work when using third parties' software. From early 2014, the current trend moved to the constant use of cloud-based applications that no longer require executables on the host system. Four of the nine organisations declared that they had no IT policy to encompass shadow IT. In contrast, the others used Windows policy internal settings or traffic analysers to manage this problem.
However, they noticed that adopting portable apps or cloud services bypassed measures like requesting an admin password to install applications. [9] Most of the employees did not see the possible risks they were exposing themselves and the company in terms of cyber risks and possible sanctions. Illegal acts should be sanctioned, but they considered applications as minor violations.

In [9], another trend identified was that 58.97% of employees used greynet applications. G*reynet* or *greyware* refers to using evasive techniques to download from network applications or files. A basic example is to use a VPN to hide the traffic, peer-to-peer (P2P) channels or decentralized file-sharing tools. The final results in [9] show that IT security departments see shadow ITs as a significant threat to the company's security posture. The measures they tended to use had no considerable effect since blocking strategies are not suited for big companies, and network or software-based restrictions can be easily bypassed. The real problem is that even though some employees are aware of the risks linked to shadow IT, they tend to see more advantages in terms of productivity.

The author of [14] focused on framework management to handle cloud-based shadow IT. The main research questions were, in this case, related to creating a framework to help organisations and measuring its effectiveness against shadow IT. The interviewed experts were all figures with an IT background and from different Dutch organisations. The main roles were Information Security Officers and CISOs.According to research findings, one of the significant reasons for shadow IT is the misalignment between the business and IT departments. When these two units fail to work together, there is a lack of collaboration and understanding, leading employees to look for alternative solutions outside the approved IT infrastructure. [14]

Another factor contributing to shadow IT is the absence of official solutions or the insufficient quality of existing ones. This situation arises when employees find the approved technologies or software inadequate or lacking in functionality, prompting them to resort to unauthorised alternatives that better suit their needs. Additionally, a lack of communication also plays a role, as some studies have shown that employees are often unaware of who is responsible or where to seek more information. [14]

In addition, studies have identified accessibility as a crucial factor in the prevalence of shadow IT. If official solutions are not adequately promoted or require complex processes for deployment, employees may opt for unauthorised tools that are easier to obtain and use. The perception that official solutions are more expensive can also drive employees towards shadow IT as they seek more affordable options, notably open-source tools. [15, 16] Employee attitudes and behaviours are also significant contributors to the spread of shadow IT. Some employees may view organisational policies as overly strict or restrictive, leading them to explore unauthorised options. Additionally, underestimating the risks associated with shadow IT can further encourage its adoption. The increasing availability and accessibility of technology have also made it easier for employees to set up IT solutions, lowering the barrier to the spread of shadow IT. [14]
Lastly, the changing workforce landscape and the rise of personal devices in the workplace allow employees to introduce unauthorised technologies. Driven by their familiarity and preference for consumer-grade tools, employees can easily bypass the official IT infrastructure, leading to the proliferation of shadow IT. [14]

It is essential to take proactive measures to address the threat of shadow IT. Three key ways include mapping the digital footprint to identify publicly exposed applications and services, continuously monitoring the digital attack surface to stay aware of any changes, and mitigating risks through immediate action when red flags or alerts are raised. By implementing these measures, organizations can better safeguard their digital assets and strengthen their security posture. To maintain a robust security posture, CISOs must confront the complexities of shadow IT head-on, implementing a comprehensive and unified security approach that mitigates risks and fortifies their organizations against potential threats. Understanding and addressing the challenges shadow IT poses is critical for safeguarding digital assets as companies navigate the constantly evolving cyber security landscape. [10]

2.3 Business economic losses and cyber security risks

"You cannot secure what you cannot see".

From previous sections, we saw how visibility is not only based on the number of collected data; visibility is more. Linking insights, understanding context and designing a picture of the attack surfaces are other steps to gain visibility.

The complexity and size of attack surfaces have put organisations under dangerous risks and prompted them to shift to automated and proactive strategies. The report [17] was conducted in late 2022 and gathered feedback from 400 highly qualified security experts around the USA and Europe. The focus was on continuous security validation technologies (CSV).

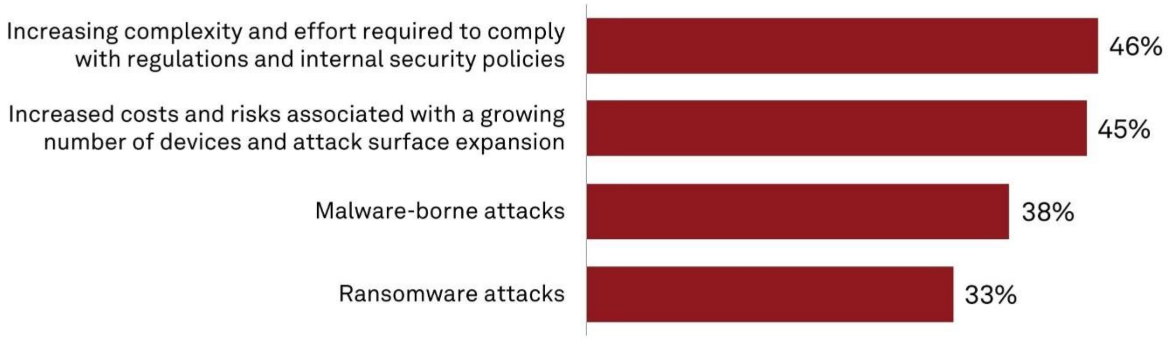


Figure 8: Cyber security risks among 96 organisations with over $750 million in revenue; S&P Global Market Intelligence (December 2022). [17]

Figure 8 represents the main organisations' cyber security risks from the last year (2022). Similar to ransomware attacks, hybrid cloud security is a concern for 33%.
The common factor among the participants was the complexity of compliance and risk mitigation increasing due to the *"explosion in the quantity of internet-connected devices and cloud deployments that result in continual attack surface expansion"* [17].

Of the risks represented in the figure, the ones with the most potential adverse business impact are malware (63%), ransomware (52%) and hybrid cloud management (48%). The most common cyber security attack was phishing in 2022 (43%); security awareness is one of the main concerns. Of the 400 organisations, 47% indicated they experienced a ransomware attack, and 56% paid the ransom. However, only 39% recovered the data after paying. On average, 60% of companies affected by ransomware or data breaches go out of business within six months. [18]

According to the FBI's report, healthcare has been the sector facing the highest number of ransomware attacks in 2021. [19] The count of losses was $6.9 billion. Over 550 organisations have been affected by data breaches, compromising the sensitive health information of 40 million individuals. [19] All healthcare organisations have to manage complex information systems and networks with vast numbers of connected devices. [19]

"Security analysts are overwhelmed by the quantity of security tooling available, making it difficult to perform their jobs effectively. In practice, many security tools are deployed tactically, often to satisfy a specific use case or compliance requirement, with little thought given to how (or whether) analysts will use them in their day-to-day jobs." [17]

Besides the number of tools, it is crucial, from a business standpoint, to assess the actual value derived by measuring the return on investment (ROI). This is a critical metric as it provides an objective measure of the utility and efficacy of any technology or tool employed by the organization. In this context, attack surface management technologies have been identified as highly valuable tools, ranking third in perceived value with a staggering 96%. This indicates that businesses perceive these technologies as almost indispensable in safeguarding their digital assets and infrastructure. Such a high level of perceived value underscores the pivotal role these tools play in an organization's overall cyber security strategy.

"Today's enterprises are looking to increase visibility across the entire IT ecosystem, gaining insights into their security posture as a basis for constructing more resilient cyber security programs." [17]

The importance of visibility in organizations' cyber security efforts cannot be overstated. With the rise of sophisticated and frequent threats, relying on infrequent and manual security assessments is no longer enough. Organizations recognise the need for increased visibility across their IT ecosystems to build strong and resilient cyber security programs. By adopting continuous monitoring and automation, businesses can gain real-time insights into their security posture, allowing them to identify vulnerabilities, promote transparency, and respond promptly to emerging threats.

Investing in solutions that provide automation, visibility, and productivity is crucial to ensure a robust security posture in today's ever-evolving threat landscape. For [20], *"the majority of security tools available today are focused on threat detection and mitigation, not device discovery and recognition".* Since organisations do not have an accurate and complete tool to protect the attack surface, 43% say it is *"spiralling out of control"* [20].

Organisations faced a rising tide of social engineering scams and ransomware attacks, and statistically, 93% of companies had vulnerabilities allowing cybercriminals to breach network parameters [20]. It is pretty concerning that just 45% of organizations possess a straightforward, well-defined approach to assessing risk exposure. In order to ensure that a network is completely safe and secure, it is essential to have a comprehensive understanding of all the devices and software currently in use. Unfortunately, this can sometimes be a rather tricky task, especially given the fact that there are so many physical devices to keep track of, along with virtual assets, operational technology, and various Internet of Things (IoT) devices. This unauthorized and often unprotected hardware and software can account for a significant portion of IT spending, estimated to be anywhere between 30-40%. [20]

2.3.1 Problem validation

The validation of the visibility problems from the literature perspective has been analysed in [9], [13] and [14]. All of these studies have focused on the Dutch market. They interviewed multiple organisations based in the Netherlands and used a similar approach, business-oriented. The multiple interviews indicate that shadow IT is a genuine concern for most companies. However, there is no one-size-fits-all solution, resulting in a spectrum of opinions on the issue.

While some organizations perceive shadow IT as a relatively minor problem compared to the security of their internal systems, others see it as a significant management challenge, among many others. A primary concern is the potential for data loss. A CISO of a municipality highlighted that when employees use their personal accounts, it disrupts the data flow. This risks data leakage and means employees might retain access to this data even after leaving the organization.

Data security and compliance experts have underscored the imperative to gain context and stack visibility as a foundational step in managing the burgeoning trend of shadow IT. A consensus exists that monitoring, starting with traffic analysis, is the most appropriate initial strategy for this approach. Although there is consistency in the initial monitoring strategy across expert opinions, subsequent recommendations vary. Some experts favour a whitelist strategy, which permits only pre-approved applications and services, while others advocate for a blacklist approach, which prohibits specific applications and services. Despite these divergent views, there is a general acknowledgement that it is impracticable to address all potential sources of shadow IT given the extensive expanse of the digital landscape.

In today's digital age, it has become increasingly difficult for organizations to monitor and control their employees' cloud usage. The rise of a mobile and connected workforce has resulted in the proliferation of personal devices in the workplace, creating a "shadow cloud" that operates outside the company's monitored network. This presents significant challenges for CISOs responsible for ensuring the security and compliance of their organization's data.

With the increase in personal devices, it has become harder for organizations to monitor their employees' cloud usage. In the past, organizations relied on monitoring tools and network infrastructure to track data flow. However, with the rise of personal devices, employees are accessing personal cloud storage and collaboration platforms that do not connect to the company's monitored network. This creates a blind spot for the organization, making monitoring and controlling employees' cloud activities difficult. Without proper oversight, employees may accidentally expose sensitive data or engage in non-compliant activities, which can expose the organization's security and regulatory compliance.

Organizations with a decentralized structure often find the whitelist/blacklist approach ineffective due to the extensive attack surface that needs protection. In many cases, they have not identified a single, centralized solution capable of providing insights and ensuring visibility. Instead, a segmented approach is more suitable when many branches operate more autonomously than the headquarters. Therefore, balancing centralized oversight and local autonomy is crucial to effectively managing the risks associated with shadow IT. Moreover, the rise of shadow IT has been run by the increased availability and accessibility of user-friendly applications and platforms. Employees often use these applications for ease of use and convenience, even if the organization does not approve them. This complicates the task of CISOs, as they need to balance employee productivity with maintaining security and compliance. A comprehensive approach that includes employee education, robust security policies, and advanced monitoring and detection tools is necessary to address the challenges posed by shadow IT effectively.

3 Cyber asset attack surface management: a new strategy toward visibility

"I genuinely believed that graph models could solve some of the security's biggest problems", Jasmine Henry, Field Security Director at JupiterOne. [21]

In this chapter, we want to introduce cyber security tools "focused" on enhancing visibility. We emphasized the importance of cyber security visibility for business continuity and growth. In this regard, we would like to clarify the distinctions among the available solutions.

Every tool presents pros and cons. Furthermore, the final scope is to highlight Cyber Asset Attack Surface Management (CAASM) since we believe this technology, with the right approach, could solve some of the current cyber security problems. In particular, it could benefit small and medium-sized organizations thanks to the knowledge graphs-based solution and straightforward approach.

Managing security has become increasingly difficult for teams today, as highlighted in the report [21]. On average, they handle over 120,000 findings, with 46% of them coming from cloud security tools. The high volume of findings coupled with an overwhelming attack surface makes it even more challenging.

3.1 Market trends: Gartner security report

This section will explore the latest endpoint security trends and how they shape the landscape. The well-known Gartner Security Reports [22], [23] are the primary references for gaining comprehensive insights into market dynamics and identifying key developments. It provides valuable data and strategic guidance for organizations seeking to strengthen their security postures.

Innovators in endpoint security have been focused on enhancing the prevention, detection, and remediation of threats through increased automation. This has led to the emergence of extended detection and response (XDR) solutions, which leverage data points and telemetry from various sources to correlate and combat evolving threats effectively. The shift from remote work to hybrid work environments has made secure remote access a top priority for organizations. The prevalence of non-company-owned devices has led to adopting desktop as a service and secure enterprise browser solutions. These initiatives aim to bolster control and strengthen security posture in diverse endpoint environments. Organizations are increasingly opting for security service edge (SSE) and secure access service edge (SASE) architectures to facilitate application access from any device across any network while ensuring minimal disruptions to the user experience.

This move towards a more flexible and secure application access paradigm is pivotal in meeting the evolving needs of modern businesses. [22]
The *Endpoint Security and Security Operations Hype Cycles* aim to monitor the advancements that assist security leaders in safeguarding their businesses against cyber attacks and breaches. As techniques and technologies evolve, two noticeable trends have emerged: an escalation in the number and complexity of endpoint attacks and the continued rise of remote work as a mainstream practice. [22]

The trend towards remote work, mobile devices, and cloud services continued to grow, and with it, organisations needed to monitor and manage risk across a broader range of digital assets. As digital business functions and third-party managed assets become more prevalent, security and risk management leaders must reassess their security strategies and tools to protect critical environments.

Effective security operations are not just about having the right department, team, or set of technologies. They require well-executed processes and skilled personnel who can maintain high resiliency. This means having access to modern security technologies that can quickly detect and mitigate threats while reducing exposure. However, finding the right skill sets and solutions can be challenging, especially with an expanding attack surface and increasing tool consolidation efforts that organizations must evaluate and support. To improve their tooling and processes, many organizations are turning to managed security services (MSS) and cloud-delivered security technologies, as well as outsourcing and “as a service” offerings. These solutions can integrate quickly with an organization’s operations, providing capabilities like ASM. Indeed, there is a growing demand for solutions like extended detection and response among smaller organizations that are not yet mature in their security operations capabilities and do not want to invest heavily in building those capabilities. These organizations prefer technology solutions that require minimal specialist skills or infrastructure to be retained internally. These solutions are best suited for new infrastructure projects rather than organizations with established security investments. [22]

The increasing number of “Innovation Trigger” profiles indicates that new market offerings address the shift in attack surfaces. These offerings include EASM, CAASM, penetration testing as a service (PTaaS), and IT threat detection and response (ITDR). Some of these offerings may merge to create more exposure-management-driven solutions. XDR, breach and attack simulation (BAS), and digital forensics and incident response (DFIR) are currently the most popular capabilities, promising to improve security visibility, response, and analysis of the root cause and solutions to threat exposures that concern many organizations. [22]

3.1.1 Trends in the market of visibility tools

In the context of visibility, the main trends and technologies identified by Gartner are: EASM, CAASM, Cyber Security Mesh Architecture (CSMA), Digital Risk Protection Services DRPS, Breach and Attack Simulation (BAS), and CASB.

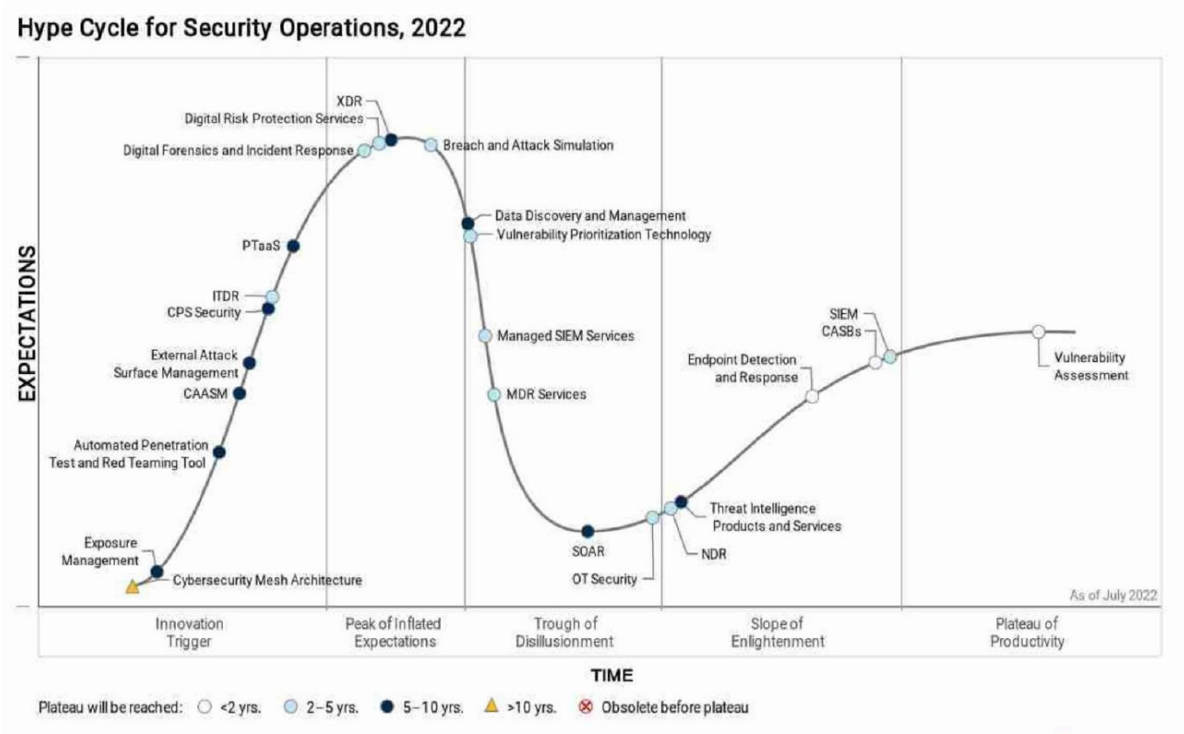


Figure 9: Gartner Hype Cycle for Security Operations, 2022. This graph shows the classic Gartner's hype cycle curve, and we clearly notice the difference between the positions of CAASM and CASB technology, representing their different levels of maturity and adoption. [22]

The cyber security mesh architecture is a game-changing concept that has the potential to revolutionize the cyber security field. However, despite its immense potential, it is underutilized, with less than 1% of its intended audience taking advantage of it. While it is still in its early stages of development and deployment, it is essential to note that it has already shown immense promise, even in its embryonic stage of maturity. [22]

The concept of a security mesh introduces a paradigm shift in how we protect digital assets and sensitive information. Unlike traditional, perimeter-based security models, the mesh approach extends security measures across a dynamic and distributed network. It weaves security into every facet of an organization's digital environment, creating a web of interconnected safeguards. This enhances protection and facilitates a more agile and responsive security posture.

Furthermore, organizations seeking to enhance their security measures increasingly turn to CAASM, which is rapidly gaining popularity. Although only a small percentage of the target audience, ranging from 1% to 5%, has adopted this technology, CAASM's maturity level is considered emerging, indicating that it is becoming more prominent and visible in the cyber security landscape. [22] Moreover, CASB tools have reached the "Plateau of Productivity", becoming an established and widely used technology. Therefore, they were not included in Gartner's latest report, released in July 2023. [23]

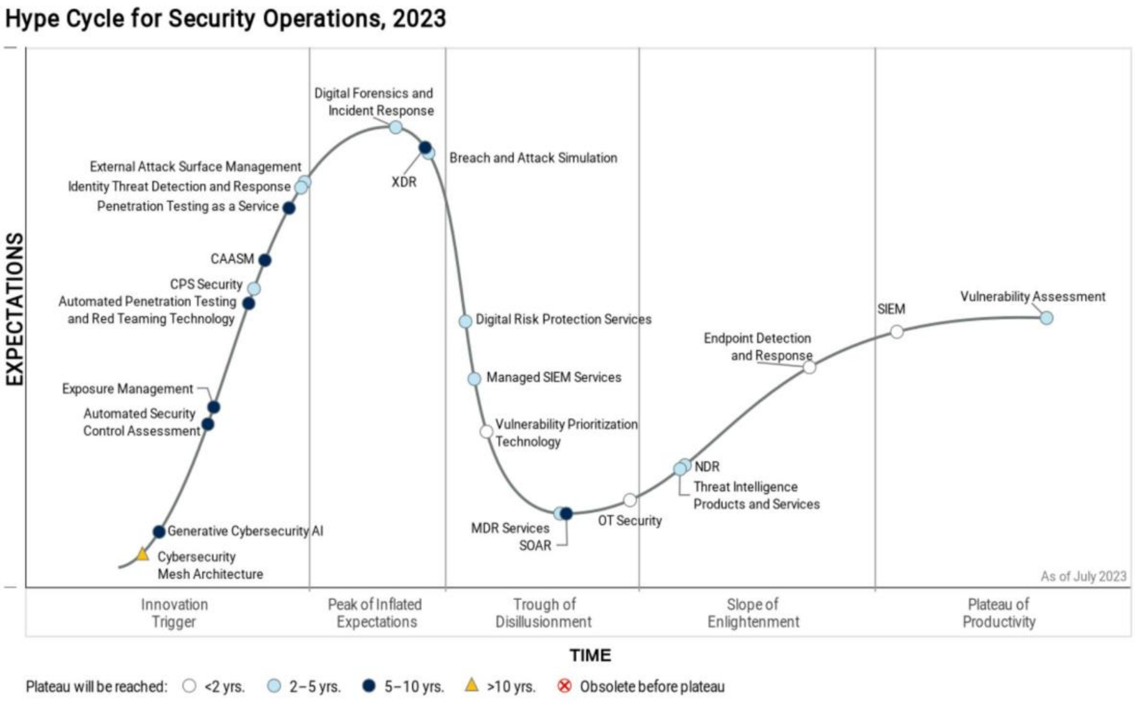


Figure 10: Gartner - Jonathan Nunez, Andrew Davies, Hype Cycle for Security Operations 2023. This last version points out that CASB technology has reached the plateau of productivity, while CAASM is getting closer to the peak of utilization. It reflects the higher number of CAASM tools available in the market when writing this thesis. [23]

External Attack Surface Management (EASM) is gaining recognition within the cyber security market. With a current market penetration ranging from 1% to 5% among security-conscious organizations, it is evident that it is gradually being adopted and gaining momentum.

EASM maturity level is categorized as early mainstream, indicating that it has moved beyond the initial stages of development and is on track to become an essential component of effective security strategies. [22] [23] Digital risk protection services are an effective solution for organizations looking to safeguard their digital assets. The fact that 5% to 20% of their target audience has already adopted these services signals a significant level of market penetration.

The “Early Mainstream Maturity” achieved by DRPS further underscores the growing acceptance and usage of such services within the cyber security industry. [23] Breach and attack simulation is a highly beneficial technology enabling companies to assess their security level proactively. With positive acceptance within the cyber security market ranging from 5% to 20% of its target audience, this technology is rapidly growing and expanding. It is a mature technology continually evolving to meet the industry’s ever-changing demands. [22] [23]

BAS tools have proven valuable in multiple domains, with a 99% positive ROI reported by survey participants [17]. BAS capabilities effectively reduce business risks, especially identifying unpublished, signatureless, and zero-day vulnerabilities. BAS tools provide increased visibility into security control performance, leading to more resilient security infrastructure. They also minimize the number of successful attacks and breaches, resulting in quantifiable risk reduction. BAS solutions can augment the capabilities of SecOps teams and reduce staffing costs while simplifying the management process and increasing operational efficiency. [17] [22] [23]

CASBs represent a robust technology that can significantly enhance the security of organizations grappling with cloud security challenges. More than 50% of the target audience has already adopted CASBs, indicating they are a mainstream solution. This reinforces that they are now widely recognized as indispensable tools for cyber security and have reached an advanced stage of implementation. [22] [23]

ASM TOOLS	BENEFIT RATING	MARKET PENETRATION	MATURITY LEVEL
CSMA	Transformational	< 1%	Embryonic
CAASM	Moderate	1% up to 5%	Emerging
EASM	Moderate	1% up to 5%	Early mainstream
DRPS	Moderate (High in 2022)	5% up to 20%	Early mainstream
BAS	High	5% up to 20%	Early mainstream
CASB (EXCLUDED IN THE 2023 REPORT)	Transformational	> 50%	Mature mainstream

Table 1: Comparison of the main ASM tools considering the last update of Gartner. Report released in July 2023. [23]

3.2 Increasing visibility with attack surface management tools

As technology advances and systems become more interconnected, organizations must implement security measures that can identify and mitigate potential risks. Attack Surface Management (ASM) is a set of practices and tools that can help achieve this goal.

As businesses undergo digital transformation, adopt API economy, migrate to the cloud, and connect more devices to the internet, identifying and mitigating potential risks becomes increasingly essential. ASM tools are designed to provide comprehensive and proactive risk assessment by mapping potential exposures to exploitability, extending the concepts of Vulnerability Management (VM). This involves referencing established frameworks such as MITRE ATT&CK[5] and utilizing both internal and external scanning techniques. [17]

ASM takes a comprehensive approach to cyber security, aiming to identify all potential weak points within an organization's IT environment. The primary objective of ASM is to increase visibility, enabling security teams to understand their entire digital landscape, including both visible and hidden assets. [17]

ASM tools cover a wider range of targets than vulnerability management and penetration testing and include various subdomains such as EASM, CAASM, and Digital Risk Protection Services (DRPS). EASM focuses on internet-facing assets, web applications, and potential attack vectors accessible to malicious actors outside the organization. In contrast, Internal Attack Surface Management (IASM) dives into the internal network, analysing endpoints, databases, and other components that might be vulnerable to insider threats or lateral movement by external attackers.

Other tools are Cloud Access Security Brokers (CASB), Vulnerability Management (VM) solutions, and Web Application Security Scanners (WASS), each with its specific focus and functionality. Organizations can tailor their ASM strategies to safeguard their unique digital ecosystems by understanding these tools and their areas of expertise.

[5] The MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) is a framework and database created by the MITRE Corporation. It aims to outline the actions and behaviours of cyber attackers throughout their attack stages.

3.2.1 CASB

Cloud Access Security Brokers (CASBs) increase the ability to gain insight and visibility into the access and exchange of data stored on the cloud.

Gartner defined CASBs as *“on-premises, or cloud-based security policy enforcement points, placed between cloud service consumers and cloud service providers to combine and interject enterprise security policies as the cloud-based resources are accessed. CASBs consolidate multiple types of security policy enforcement. Example security policies include authentication, single sign-on, authorization, credential mapping, device profiling, encryption, tokenization, logging, alerting, malware detection/prevention and so on.”* [24]

Frost and Sullivan provide an alternative definition: *“a security platform which resides between cloud service users and cloud apps. It allows an organization to consolidate multiple security policy enforcement across multiple cloud apps and extend the reach of these security policies beyond its infrastructure”* [25]

The fundamental building blocks to achieve this are:

- **Forward proxy**, used to protect client devices, which are part of the company’s premise perimeter. A forward proxy is deployed, in general, using endpoint agents, Proxy Auto-Config (PAC) or Secure Web Gateway (SWG). [13]
- **Endpoint agent**, software running on endpoints to collect and forward traffic to the CASB. It can act as a transparent proxy, which does not modify the traffic, or as an explicit proxy to have better control.
- **Reverse proxy** is generally a transit between the internet and cloud platforms. This protection affects the server side, protecting internet-facing applications. [13]

Another solution is using APIs to execute operations and control data at higher levels (from layer Session, layer 5 of the ISO/OSI stack). In this way, it is not required to act as a proxy decrypting and encrypting the traffic as data is visible at higher levels. This reduces computational complexity and latency. However, the chosen CASB must be compatible with APIs made available by cloud service providers. The CASB, in this case, acts as an out-of-bound device or remote instance that needs to communicate directly with cloud service providers. The faster analysis has as a drawback an increase in bandwidth usage. [13]

Besides, CASB's main pillars can be grouped into four key categories: **visibility**, **threat protection**, **access control** and **compliance**. Visibility involves providing a comprehensive view of all cloud activity and usage. By utilizing CASBs, organizations can obtain information about which cloud-based applications are being used, who is accessing them, and how they are being accessed. This data can be analysed to monitor user behaviour, manage policies, and improve security practices. Furthermore, CASBs provide shadow IT discovery, which offers a brief overview of the company's cloud services.

Threat protection, which includes identifying and mitigating potential risks and attacks. CASBs increase threat protection capabilities by monitoring and analysing user activity and device behaviour. In large organizations, the amount of data collected by the CASB system is significant and can be utilized by incident response, forensics, and risk management teams. In general, the integration with Security Information and Event Management (SIEM) allows for alerts generated by the CASB to be viewed together for improved analysis and easy access to IR and SOC teams access.

Access control involves regulating and monitoring user access to cloud resources and applications. This additional layer can help organizations manage and better enforce authentication, authorization and accounting controls. Moreover, CASBs can assist and improve compliance, which is conformity with primary regulation, by enforcing encryption on files and implementing sophisticated rules related to data protection. An example is DLP (Data Loss Prevention) rules for data stored on-premises. [26]

Companies such as Microsoft, McAfee, and CloudFlare have their definition of CASB. However, from their interview conducted by [13], the standard CASB market definition has been *"CASB's role is to extend the reach of an organization's security. A CASB is often found within 'X as a Service' models and is placed between the employees and the organization's perimeter."* [13]

Therefore, CASB helps to extend the security reach into cloud environments instead of only on-premises. Gartner's security report (2020) analysed the market growth of CASB and pointed out that the main factor of the growth was the high demand for remote working during the pandemic, which forced many companies to adopt cloud solutions. [24]

However, the new trend described by Gartner sees many new technologies forming a "package" for Enterrise's security: *"Software-defined Wide Area Network (SD-WAN), Secure Web Gateway (SWG), CASB, Zero Trust Network Access (ZTNA) and FireWall as a Service (FWaaS) as core abilities, with the ability to identify sensitive data or malware and the ability to decrypt content at line speed, with continuous monitoring of sessions for risk and trust levels."* [27]

The lack of visibility and shadow IT are the main concerns companies want to address with using CASBs, and gaining insights into employees' behaviours is considered crucial during remote working.

As discussed in the previous chapter, the analysis of cloud-based shadow IT [14] shows how old solutions cannot cover all the perimeter. With the advent of cloud computing, software scanning for locally installed applications lost its efficacy. Today, it is more common for employees to use SaaS applications, which, as the locally installed ones, pose a significant threat to information security and privacy, increasing data leakage risk. To prevent and respond to these adverse effects, the author of [14] proposes a well-managed framework that involves policies, security awareness, monitoring, and filtering. While the author does not explicitly define the relationship between CASBs and shadow IT, it does suggest that incorporating CASBs for monitoring, blacklisting, and whitelisting can effectively mitigate these risks. To reduce the risk of shadow IT, employees must be aware of its potential risks, and appropriate technology should be implemented. [14]

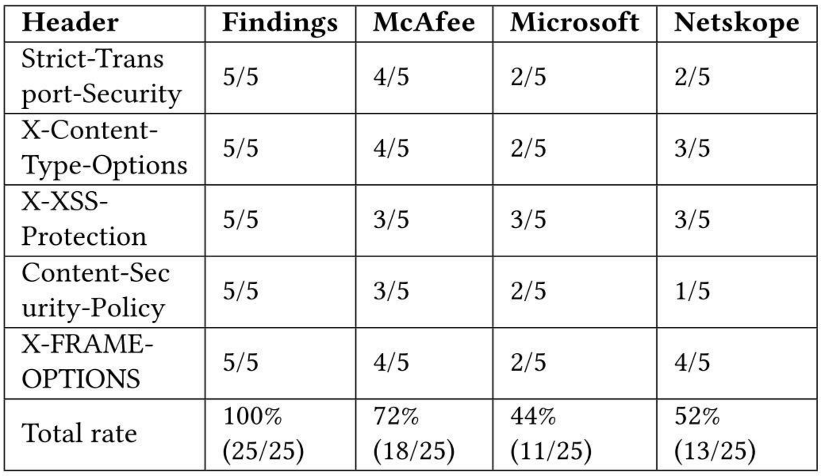
According to research conducted in [13], it has been determined that visibility and control are crucial factors in successfully implementing CASB. Organizations show a keen interest in evaluating and managing the risk of cloud applications used within their organizations. The first step towards implementing CASB involves gaining visibility through the discovery phase and determining whether the cloud solutions' risk appetite aligns with the organization's risk profile. Once a policy has been formulated, the control aspect of CASB can be enforced with confidence. [13]

The discovery phase of CASB exhibits uniformity across various CASB vendors, lacking a distinguishing characteristic. The differentiation between CASB vendors lies in their capacity to furnish contextual insights about the multitude of cloud applications.

Organizations find it impossible to examine the risks associated with each cloud application. As a result, CASB providers have introduced a well-structured framework to classify the risk level of cloud applications, commonly known as the "cloud index". The efficacy of this framework for each CASB vendor has been assessed through a use case formulated for its validation in [13].

Several methodologies can be employed to calculate the cloud index or similar indexes. However, a recent study conducted by a team of researchers has found that the security headers (HTTP/HTTPS) implemented on an application's login page can serve as a reliable factor in calculating the cloud index value.

The result of the study is shown in the following table.



Header	Findings	McAfee	Microsoft	Netskope
Strict-Transport-Security	5/5	4/5	2/5	2/5
X-Content-Type-Options	5/5	4/5	2/5	3/5
X-XSS-Protection	5/5	3/5	3/5	3/5
Content-Security-Policy	5/5	3/5	2/5	1/5
X-FRAME-OPTIONS	5/5	4/5	2/5	4/5
Total rate	100% (25/25)	72% (18/25)	44% (11/25)	52% (13/25)

Figure 11: Cloud index efficacy considering the security headers approach. By analysing the main fields of headers and network security best practices, the analysis conducted over the main CASB tools confirmed the dominance of McAfee over other providers. [13]

McAfee resulted in being the most reliable among the tested CASBs. In particular, it has been the only one raising an alert when analysing ProtonMail. Proton is a Swiss service provider characterized by high privacy; however, their ProtonMail as a Service was not ISO 27001 certified, but only their servers and data centre providers were.

The cloud index is necessary to narrow the focus of SIEM towards the most vulnerable applications. It is important to note that the cloud index scores should be regarded as a general guideline rather than an entirely precise measure.
The second differentiating feature among CASBs is the enforcement of policy control. Evaluating the extent and implementation of these policy control functionalities requires the establishment of a dual-use case. This use case ascertains how these policies are practically executed and enforced within the CASB framework. [13]

The study [11] presented a second use case involving a Dutch Hospital to explore a scenario concerning more sensitive and personally identifiable information (PII). This use case provided insights into data loss prevention practices. It has been noticed that vendors provide DLP features to recognize personal information but with different levels of easily accessible templates. Furthermore, all vendors provide choices to evaluate whether an organization complies with GDPR.

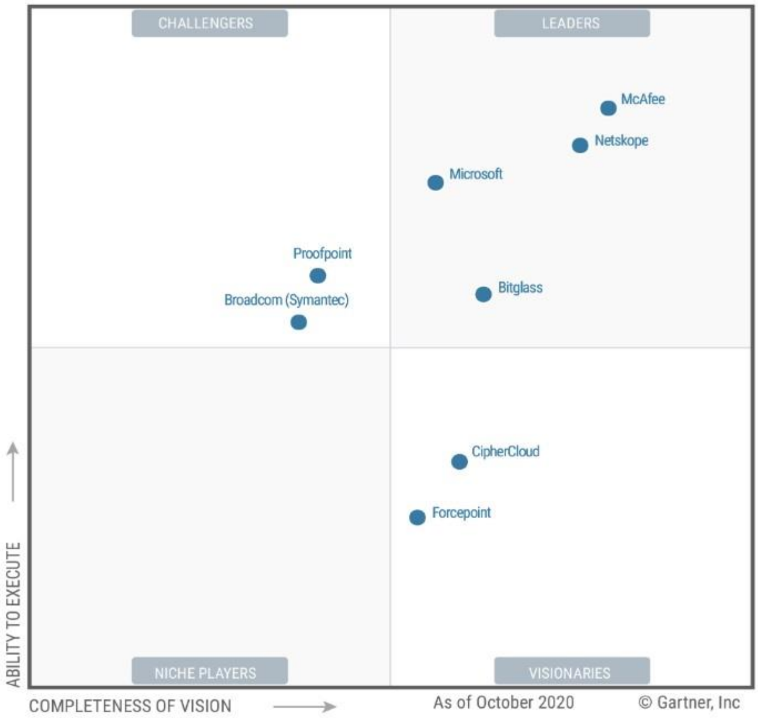


Figure 12: Gartner Magic Quadrant for CASB 2020. The analysis that Gartner conducted, which was older than the previous cited research, confirmed the the flexibility and elasticity of McAfee's product. Its various features allow it to cover a more extensive security range than others, resulting in a more accurate analysis. [26] [24]

To validate the conclusions drawn in [13], we decided to compare the results with the insights derived from the report published by Gartner [24]. Microsoft is confirmed to be the most comprehensive offering of all the vendors as it allows for using all attributes defined in their cloud index to filter traffic efficiently.

However, both McAfee and Netskope have limitations in this aspect. Microsoft's filtering capabilities, which are rich in features, make it a strong contender for handling shadow IT concerns. Most corporate candidates interviewed said that implementing CASB has administrative overhead problems. Only service providers did not acknowledge this issue. Administrative overhead can cause alert fatigue in the SOC, making CASB implementation unnecessary or just for audits. Integration with SIEM platforms can also help reduce overhead. Organizations should select a vendor that best suits their needs and onboarding trajectory and enforce policies based on organizational structure.

3.2.2 EASM

"Do not fear the unknown; manage the risks" [28].

External Attack Surface Management (EASM) has become a critical component of cyber security for companies dealing with various types of digital assets.

Gartner has defined EASM as *"the processes, technology and managed services deployed to discover internet-facing enterprise assets and systems and associated vulnerabilities. Examples include exposed servers, credentials, public cloud service misconfigurations, deep dark web disclosures and third-party partner software code vulnerabilities that could be exploited by adversaries."* [22]

EASM tools offer organizations a dynamic and comprehensive approach to managing and discovering digital assets. EASM is more than just about asset discovery. It is designed to empower teams responsible for security, governance, risk management, and compliance (GRC) within the organization by providing advanced insights into the evolving security landscape. This allows organizations to monitor changes in their risk profile, remediation efforts, and outstanding risks, which is essential for compliance with regulatory requirements. [28]

Moreover, EASM helps identify patterns of weakness and areas that need further educational efforts. Understanding contributing factors to the organization's risk profile allows for effective vulnerability management. Cloud technologies, remote work practices, and the convergence of IT, OT, and IoT are the main drivers of this risk. To help identify known and unknown exposed assets and prioritize vulnerabilities and risks, EASM has become a critical solution.

EASM has significant business implications for Security and Risk Management (SRM) leaders, providing valuable risk context and actionable insights. The five primary aspects of EASM include continuous monitoring of cloud services, IP addresses, domains, certificates, and IoT devices, asset discovery of external-facing assets and systems, risk and vulnerability analysis, efficient prioritization, and effective remediation, mitigation, and incident response capabilities. It also provides seamless integrations with ticketing systems and Security Orchestration, Automation, and Response (SOAR) tools. The adoption of EASM is driven by the increase in digital business initiatives, cloud adoption, remote working practices, and the convergence of IT/OT/IoT domains. Organizations are interested in understanding their external exposure from a potential attacker's perspective, leading to the widespread adoption of EASM capabilities.

The implementation of EASM may face obstacles due to mergers and acquisitions in the near and midterms, as well as uncertainty surrounding adjacent markets. To overcome these challenges, organizations should evaluate EASM capabilities provided by various vendors, considering the scope of coverage, accuracy, and level of automation offered. They should also prioritize use-case requirements and anticipate potential expansion into DRPS and security testing/validation scenarios. Moreover, the security team's skills, resources, and maturity readiness are essential for fully utilizing the advantages of EASM capabilities.

FortiRecon's EASM solution includes in-depth asset (and shadow IT) discovery capabilities, such as domains, IP addresses, Autonomous System Numbers (ASNs), sub-domains, and certificates. The solution also conducts vulnerability assessments, considering exposed services, historical records, and recent asset changes. This provides actionable recommendations for mitigation and risk reduction. FortiRecon's solution offers comprehensive and dynamic capabilities to help organizations manage their digital assets effectively. [29]

Microsoft has recently released a new tool called Microsoft Defender External Attack Surface Management (Defender EASM), its version of EASM. Microsoft defines its EASM as a tool to *"continuously discover and map your digital attack surface to provide an external view of your online infrastructure. This visibility enables security and IT teams to identify unknowns, prioritize risk, eliminate threats, and extend vulnerability and exposure control beyond the firewall."*

The Microsoft Defender EASM is equipped with a specialized discovery technology that conducts thorough searches for infrastructure connected to known legitimate assets. It uncovers previously unknown and unmonitored properties by connecting to these discovery "seeds", recursively revealing additional connections to create a comprehensive attack surface.

The data handled by Microsoft Defender EASM includes both global data and customer-specific data. Microsoft's EASM recognizes the concept of "dependency", which refers to infrastructure owned by third parties but considered part of the attack surface because it directly supports the operation of assets. For example, an IT provider hosting web content would be categorized as a "dependency" while the domain, hostname, and pages fall under the "approved inventory." Moreover, the EASM framework includes the monitor-only classification, which denotes assets relevant to the attack surface but neither directly controlled nor serving as technical dependencies. [30] Other EASM providers with similar capabilities are Secura [31] and CrowdStrike with Falcon Surface. [32]

3.2.3 CAASM

"Cyber asset attack surface management (CAASM) is an emerging technology area that enables security teams to overcome asset visibility and exposure challenges. It enables organizations to see all assets (internal and external), primarily through API integrations with existing tools, query consolidated data, identify the scope of vulnerabilities and gaps in security controls and remediate issues", Gartner [22].

CAASM represents a critical solution for organizations looking to enhance their security posture and overall security hygiene. CAASM consolidates assets from various products, including endpoints, servers, devices, and applications, allowing users to identify potential gaps in security tool coverage, such as Vulnerability Assessment and Endpoint Detection and Response (EDR) tools. This aggregation of internal and external cyber assets streamlines querying and analysing asset information, replacing manual and time-consuming collection efforts.

Businesses can significantly benefit from CAASM as it provides security teams with a thorough understanding of security controls, asset exposure, and security posture. Organizations can use this tool to reduce their reliance on homegrown systems and manual data collection processes, leading to more efficient remediation of identified gaps. The tool also visualizes security tool coverage and supports ASM processes, further contributing to enhanced security measures. CAASM even facilitates the correction of systems of record containing stale or missing data. [33]

The drivers behind CAASM are driven by the need for complete visibility into all assets, including IT, IoT, and OT domains. With this comprehensive visibility, businesses can better comprehend the attack surface area, identify existing security control gaps, and support broader ASM initiatives. In addition, CAASM expedites audit compliance reporting by generating accurate, up-to-date, and comprehensive asset and security control reports. By consolidating asset and exposure information from existing products, CAASM reduces manual efforts and dependencies on in-house applications.

However, implementing CAASM may present certain obstacles. Resistance to adopting an additional tool can arise, mainly if adjacent products already offer asset visibility functionalities. Not all vendors possess the capabilities to identify and integrate IoT/OT assets for visibility and vulnerability information, which may limit the tool's effectiveness. Licensing costs may become prohibitive for large organizations managing millions of assets, and scalability concerns may arise in exceptionally vast environments with substantial data requirements. Integration challenges with existing tools and reconciliation processes can also present hurdles.

To maximize the benefits of CAASM, user recommendations include taking advantage of proof-of-concept opportunities and free product versions for assessment before full deployment.
Defining primary use cases and favouring vendors with versatile asset visibility capabilities are essential considerations. Additionally, organizations should prioritize vendors with IT and IoT/OT systems expertise. Careful inventorying of available APIs and ensuring the availability of appropriate user accounts for integration are crucial preparation steps.

Widening usage to multiple teams, including compliance, threat hunting, vulnerability management, and system administration, can amplify CAASM's impact. Lastly, existing security vendors should be consulted regarding their current asset visibility capabilities and potential plans for CAASM functionality in the future. [34]

CAASM offers robust features and a structured workflow to enhance an organization's security posture and proactive threat response. The CAASM process is as follows:

- **Asset discovery**: identifying any unusual activity that may indicate potential attacks is essential for comprehensively discovering all assets in the organization's environment. This includes thorough identification and cataloguing of all assets, establishing an accurate and up-to-date asset inventory, and implementing a robust system for adding new assets to the inventory and removing or isolating assets that are non-compliant or pose security risks.
- **Vulnerability assessment**: identifying and assessing vulnerabilities within the organization's assets involves conducting scans to identify known vulnerabilities in systems, applications, and databases and assessing the potential impact of each vulnerability in the event of an attack. This phase is also essential for critical asset identification.
- **Threat analysis:** collecting and interpreting data on potential threats to understand the dangers faced by the organization. This entails using internal sources like logs and network traffic and external sources like threat intelligence feeds. Some studies focus on this feature, identifying new space for improvement through advanced methodologies such as AI and machine learning algorithms. [35]
- **Continuous monitoring**: transitioning into continuously monitoring the organization's assets to detect and respond to potential threats promptly. This real-time monitoring involves observing logs, network traffic, and other relevant data sources to identify potential threats swiftly. Real-time visibility reduces the mean time to detect (MTTD) and respond to threats, shrinking it from several months to hours.

- **Risk management**: empowering CISOs to assess and manage the risks associated with their assets effectively. This process includes quantifying the impact of each vulnerability and evaluating the likelihood of a successful attack. Subsequently, prioritized remediation efforts are employed, and informed decisions are made regarding the assets that require immediate protection. A comprehensive risk management process empowers CISOs to minimize the risk of successful attacks and preserve their assets' confidentiality, integrity, and availability.
- **Remediation**: implementing appropriate security controls to mitigate risks associated with the assets involves applying software patches, configuring firewalls, and implementing security policies and procedures. Additionally, assets with known security risks may be isolated to prevent further exposure.
- **Continuous improvement**: this iterative and ongoing process ensures that CISOs continually monitor and reassess their organization's assets to maintain a robust security posture. It involves updating the asset inventory, reassessing vulnerabilities, and implementing new security controls. The CAASM system enables organizations to ensure the safety of their assets confidently. [34]

Numerous CAASM providers and solutions are available in the current market, including Noetic Cyber, Sevco Security, Panaseer, Brinqa, Armis, Encore, Northstar.io, Axonius, JupiterOne, and Ordr. In this section, we have chosen not to discuss the implementation of CAASM. However, in the next chapter, we will focus on the deployment and implementation of CAASM. We will focus on "Starbase" by JupiterOne, the only open-source CAASM solution we have found. The objective is to provide essential information on the practical application of CAASM and to facilitate understanding of the necessary steps involved in its deployment and implementation. The study aims to offer valuable insights into the effective utilization of CAASM in real-world scenarios.

4 Starbase: open-source CAASM by JupiterOne

In the domain of cyber security, guarding against potential vulnerabilities is a substantial challenge, particularly for those at the inception of their security journey. The ongoing digital transformation of businesses, propelled by the necessity for efficiency and scalability, leads to a significant increase in cyber assets. These assets include cloud infrastructure, conventional and Software as a Service (SaaS) applications, code and data repositories, networks, user identities, and access privileges. Remarkably, the total volume of these assets surpasses human capacity by an astounding 500 to 1 ratio. [34]

This increasing complexity makes it even more difficult for organizations and security practitioners to manage their digital assets efficiently. A lack of understanding and visibility into the network can lead to potential unchecked spots, which can become opportunities for cyber attackers to exploit. Moreover, traditional security measures often fall short of comprehensively protecting the digital landscape, necessitating a more innovative approach.

A fundamental shift in approach is essential to address these challenges, beginning with a thorough and detailed understanding of the assets and their interrelationships. A graph-based methodology is a practical and innovative solution for managing and analysing cyber assets. This approach enables organizations to fully comprehend the complexity of their cyber infrastructure and digital operations. This allows them to identify deficiencies, uncover hidden risks, accelerate incident response, gain insights into cloud security, prioritize vulnerability remediation, maintain continuous control monitoring, and automate compliance. As enterprises advance in their digital transformation journey, the spectrum of potential vulnerabilities broadens, making security measures increasingly complex. Cyber assets, which extend beyond IP addresses and devices, contribute to this complexity. [34]

Navigating this landscape requires a profound understanding of assets and their interconnections. A graph-based methodology empowers organizations to thoroughly understand the architecture of their cyber infrastructure and digital operations. This knowledge is crucial for effectively addressing security concerns and establishing robust security initiatives from the ground up. [36]

4.1 Starbase

In today's ever-changing world of cyber security, effective asset and relationship management is essential. JupiterOne has introduced Starbase, a revolutionary CAASM tool that transforms how organizations manage their cyber assets and extract the intrinsic relationships between them.

Starbase is powered by the cutting-edge Neo4j database, making it a powerful solution that consolidates assets and relationships from various sources, including cloud infrastructure, Software as a Service (SaaS) applications, and security controls. By presenting a comprehensive aggregation of data in an intuitive and comprehensive graph view, organizations can gain a unique perspective on the complex interplay of their digital ecosystem. At its core, Starbase believes that security is a fundamental right and aims to democratize graph-based security analysis.

To secure any system or service, Starbase identifies three essential pillars:

- **Knowledge of Assets** (KA): an in-depth understanding of the assets within an organization's digital landscape is crucial for adequate security. Starbase meticulously captures and categorizes these assets, ushering in a new era of asset awareness.
- **Knowledge of Relationships** (KR): understanding the relationships between assets is pivotal in the intricate web of digital operations. Starbase delves deep into the connections between entities, revealing insights that traditional security measures often miss.
- **Inquisitive Insight** (II): organizations need both a comprehensive asset repository and the ability to ask insightful questions to gain actionable intelligence from their digital ecosystem. Starbase provides this empowering feature.

The value of Starbase is underscored by its three essential features. First and foremost, it offers comprehensive visibility. This is achieved by providing an unmatched, sweeping view across a multitude of external services and systems. The vast repository of Starbase, which includes thousands of entities and relationships, empowers organizations to develop a holistic understanding of their digital environment. This, in turn, is critical for making informed decisions and strategizing effectively. Secondly, Starbase uses a standardized data structure. The data collected by Starbase is organized with great precision, making it possible to create universal queries. This methodical way of organizing data simplifies the task of extracting meaningful insights from the graph, thus eliminating unnecessary complexity. It streamlines the process, making it much more efficient and user-friendly.

Lastly, Starbase offers flexibility and customization. It recognizes that different organizations have varied operational needs. Therefore, it is designed to be effortlessly extensible. Organizations can easily create custom graph integrations specifically tailored to meet their needs. This means that each organization can customize Starbase to suit its unique operational requirements best.

Starbase supports more than 115 open-source graph integrations, covering various platforms such as Azure, Bitbucket, GitHub, Google Cloud, Google Workspace, and Jira. This extensive support enables organizations to seamlessly integrate Starbase within their existing digital ecosystem, unlocking its full range of capabilities without requiring a complete overhaul of their current workflows or systems. Furthermore, the JupiterOne SDK streamlines the development of personalized integrations, making integration even more straightforward. Organizations can customize the integrations to meet their needs and operational requirements, maximizing the tool's efficiency and usefulness. The project's open-source nature also ensures an active and thriving community of developers who continuously contribute to developing and improving the integrations. This ensures that the integrations are continuously optimized and up-to-date while providing organizations with a vast resource of knowledge and expertise that they can use to improve their digital strategies and integrations. Starbase provides organizations with exceptional visibility of their digital environment, enabling them to keep up with the constantly evolving cyber security landscape.

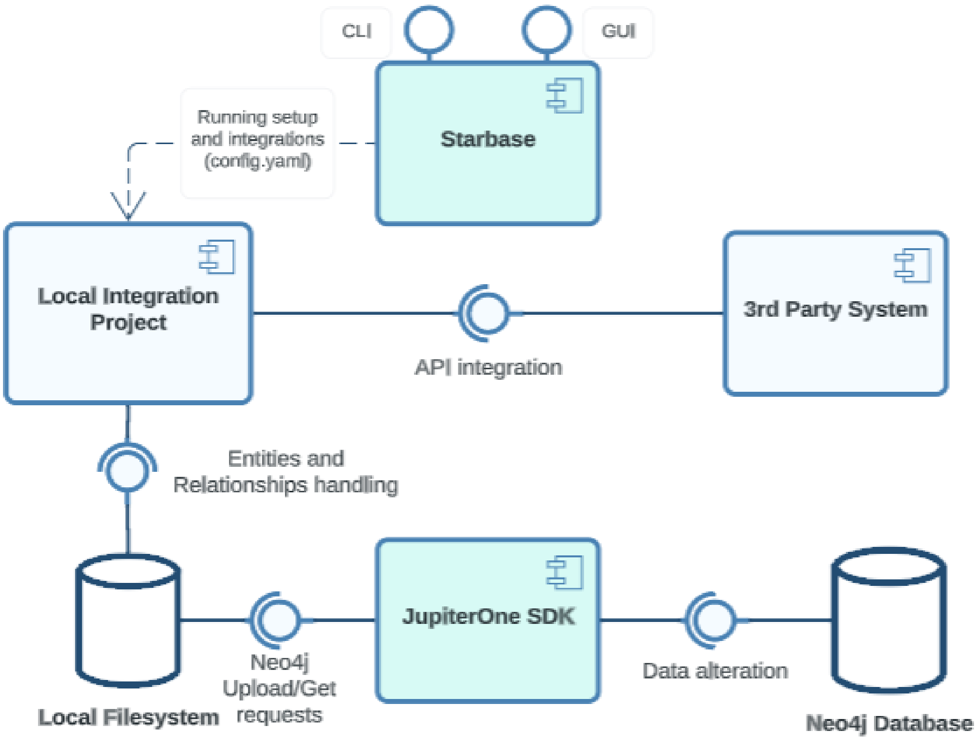

CLI
GUI
Running setup and integrations (config.yaml)
Starbase
Local Integration Project
API integration
3rd Party System
Entities and Relationships handling
Local Filesystem
Neo4j Upload/Get requests
JupiterOne SDK
Data alteration
Neo4j Database

Figure 13:Starbase framework and UML component diagram. The diagram shows the connections between the main components, emphasising their scope and interfaces.

In Figure 13, we highlight the main Starbase components and their interactions. Starbase can be seamlessly deployed by using Node.js or a container by using a Docker image. It is easy to access Starbase by interacting with the exposed Command-Line Interface (CLI) or a GUI interface.

The second method is the most suited, considering that one of the main components of Starbase is Neo4j. The integration is possible thanks to JupiterOne's SDK, which offers a well-defined interface to manage and translate requests from integrated external systems. The result is a uniform environment that enables developers to create personal integrations and manage everything in a standard way.

4.1.1 Deployment

Deploying Starbase is a straightforward process that can be done through two easy-to-use methods: the standard way or via Docker. Both options guarantee a smooth and hassle-free setup of Starbase.

Before proceeding with the standard deployment, it is essential to ensure that Node.js and yarn are already installed and that the API credentials from the systems that Starbase will interact with are available. The configuration of Starbase is simplified with a ".yaml" configuration file, *config.yaml*, which directs Starbase to operate correctly. Simply duplicate and rename the example file to *config.yaml* and input the necessary configuration values for each intended integration. The command-line interface (CLI) facilitates the usage of Starbase. To begin the setup, execute the *yarn Starbase setup command*, which will ready all integrations and install the necessary dependencies. Run the *yarn Starbase run command* to collect data from the chosen integrations.

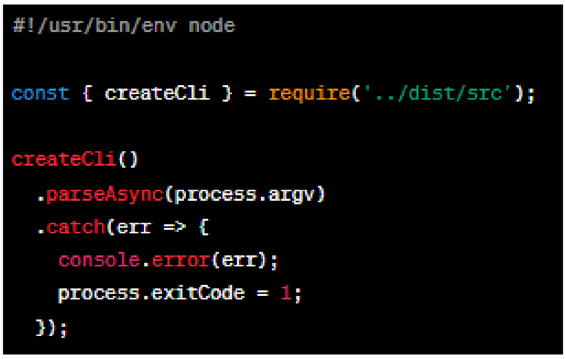
Containerization has become an essential part of modern software development, and Docker is one of the most popular platforms for containerization. It offers a range of features that simplify the deployment of applications by creating isolated environments, known as containers. Docker provides a user-friendly and efficient solution for deploying Starbase by automating processes, managing dependencies, and creating self-contained images. These Docker images encapsulate the Starbase application and all the necessary components and libraries, providing a seamless deployment experience.

The portable nature of Docker images enables them to be deployed on various host systems, offering consistency and reproducibility, which is crucial in modern software development. To create a Docker image, use the command *docker build --no-cache -t Starbase:latest*. After creating the Docker image, use *docker-compose run Starbase setup* to prepare integrations and install dependencies. To collect and store data, execute *docker-compose run Starbase run*.

4.2 CAASM and APIs integrations: GitHub use case

According to Gartner's definition [37], CAASM tools facilitate integration with an organization's digital environment by acting as intermediaries and connecting seamlessly with their current data sources through API integrations. These integrations are critical as they allow CAASM tools to create comprehensive visibility within various systems, platforms, and databases. Once these integrations are in place, CAASM tools take on the role of continuous monitoring to monitor and analyse the ever-changing landscape of potential vulnerabilities that could infiltrate an organization's attack surface.

The CLI and structured APIs simplify the integration of Starbase with external services. The first command, *yarn install* will install all the required packages. The core of Starbase is JupiterOne's SDK package, including the CLI.



```
#!/usr/bin/env node

const { createCli } = require('../dist/src');

createCli()
  .parseAsync(process.argv)
  .catch(err => {
    console.error(err);
    process.exitCode = 1;
  });
```

Figure 14: Starbase CLI integration; first code executed when launching Starbase. As we can see from the first line, this starting point is a bash script executing the CLI through Node.js (in our case, version 18.x).

Starbase interacts with GitHub using the Octokit REST client, GraphQL client, and direct curl requests. Each approach plays a distinct role in achieving comprehensive integration workflows while overcoming specific challenges.

REST API is a standardised methodology where resources are represented as URLs, and operations are performed using HTTP methods. REST APIs return data in a pre-determined structure, which may result in over-fetching. Each endpoint corresponds to a specific resource, and multiple requests may be needed to retrieve related data. GraphQL is a query language for APIs that allows developers to request the necessary data precisely. GraphQL APIs return only the requested data, reducing over-fetching and minimizing data transfer. A single GraphQL query can retrieve data from multiple resources, streamlining data retrieval. GraphQL APIs boast a strongly typed schema, ensuring clear data definitions and improving understanding. [38]

Octokit REST clients can effortlessly interact with GitHub resources such as repositories, pull requests, and issues. This method is perfect for integration scenarios that require well-defined REST endpoints. By following GitHub's RESTful principles, Starbase can seamlessly retrieve crucial data through standardized HTTP requests, such as GitHub applications and tokens. Octokit REST clients provide a dependable and organized interface. Although Starbase strives to shift to GraphQL as much as possible, REST calls are sometimes necessary for the integration landscape due to limited resource accessibility or insufficient hierarchy clarity within GraphQL. [39–41]

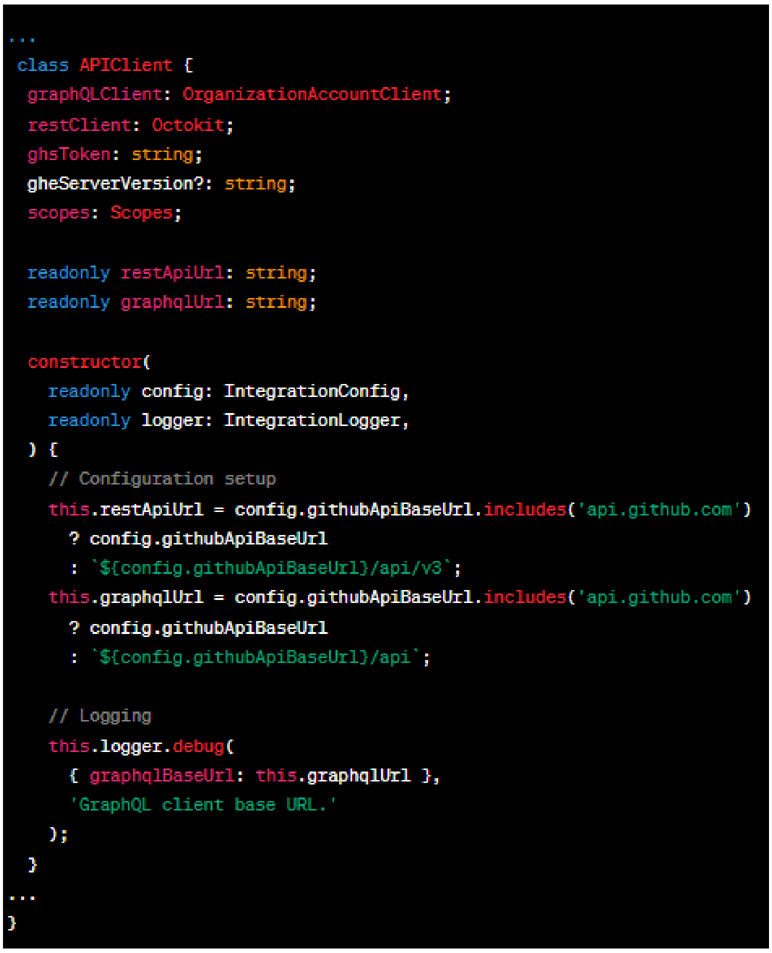
GraphQL clients allow for streamlined data retrieval by requesting only the necessary properties, resulting in increased efficiency compared to broader REST calls. The hierarchical structure of GraphQL also enables tailored data access, making it the preferred option for well-structured queries. The motivation behind the adoption of GraphQL is its efficiency and hierarchical data access, which align with Starbase's goal of optimizing integration workflows. Using GraphQL, Starbase can fine-tune data requests and reduce redundant data transfers.

However, in cases where the Octokit REST or GraphQL clients do not provide an immediate solution, Starbase exhibits adaptability by resorting to direct curl requests. This method enables direct interaction with GitHub's API endpoints. This approach selectively bridges gaps where higher-level clients cannot provide a solution. [42]

GitHub was selected as an example for integration because it is a widely recognized platform utilized by numerous organizations, especially those adhering to the DevOps methodology for version control. In particular, a specific class responsible for managing external APIs is delineated in the file and instantiated using the singleton design pattern. This design pattern guarantees that a single instance represents a class and establishes a universal access point to it. It is especially useful in scenarios like this, where managing external APIs requires consistency and centralized control. Moreover, GitHub enables the creation of personal applications within projects through a feature known as GitHub Apps. GitHub Apps are first-class actors within GitHub, which means they can perform actions on their behalf and not on behalf of a user. This is particularly advantageous as it allows more granular control over permissions and actions that an application can perform.

In the case of Starbase, leveraging the capabilities of GitHub Apps could mean setting up a GitHub App with read-only access to the repository's code and metadata while perhaps having write access to issues for automated updates or reporting. This ensures that the Starbase integration can operate efficiently and securely, fetching necessary data without risking unintended modifications while still being able to perform necessary actions.

The utilization of the GitHub App's read-only permission for APIs ensures that the integration can access necessary data without the risk of altering any repository content. At the same time, the webhook event triggered upon installation can be utilized to initiate specific workflows within the integration. This combination of features ensures that the integration operates securely and efficiently, maximizing its utility and value for the organization.



```
...
 class APIClient {
  graphQLClient: OrganizationAccountClient;
  restClient: Octokit;
  ghsToken: string;
  gheServerVersion?: string;
  scopes: Scopes;

  readonly restApiUrl: string;
  readonly graphqlUrl: string;

  constructor(
    readonly config: IntegrationConfig,
    readonly logger: IntegrationLogger,
  ) {
    // Configuration setup
    this.restApiUrl = config.githubApiBaseUrl.includes('api.github.com')
      ? config.githubApiBaseUrl
      : `${config.githubApiBaseUrl}/api/v3`;
    this.graphqlUrl = config.githubApiBaseUrl.includes('api.github.com')
      ? config.githubApiBaseUrl
      : `${config.githubApiBaseUrl}/api`;

    // Logging
    this.logger.debug(
      { graphqlBaseUrl: this.graphqlUrl },
      'GraphQL client base URL.'
    );
  }
...
}
```

Figure 15: APIClient, a class defined to encapsulate and offer a single interface (singleton design pattern) to manage third parties' APIs.

The integration strategies employed by Starbase manifest a high degree of adaptability. As the landscape of integration ecosystems is subject to evolution, the agile methodology of Starbase, underpinned by a multifaceted toolkit, equips it to address a broad spectrum of integration prerequisites. This adaptability facilitates seamless interactions with established platforms and accommodates emergent technologies, enabling organizations to adapt to technological innovations while swiftly minimising operational disruptions.

4.3 Neo4j: knowledge graphs

"When you want a cohesive picture of your big data, including the connections between elements, you need a graph database." [43]

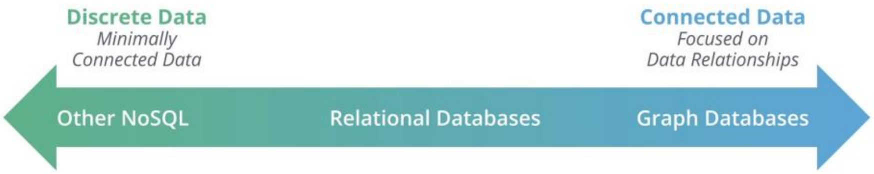


Figure 16: From discrete towards connected data. Neo4j's vision toward a better-contextualised data visualization. [43]

Graph databases are a superior solution for handling complex relationships between data points. As the data ecosystem expands, traditional databases face a noticeable decline in performance due to the complexities of relationship queries. On the other hand, graph databases maintain a steady pace even as data grows. Graph databases possess inherent flexibility that seamlessly harmonizes with evolving industry trends and changing solutions. They allow for organic expansion within the existing framework while preserving existing functionality. Creating solutions with graph databases is an ideal match for today's agile development methodologies. The ability to shape graph-database-driven applications in accordance with changing business requirements enables fluid adaptation. [43]

Knowledge graphs (KGs) have become essential models for effectively representing, storing, and querying diverse data elements with inherent relational connections, especially with the rapid expansion of the web, e-commerce, and social media during the 2000s. These data elements carry real-world meanings intricately tied to the specific domain for which the knowledge graph has been crafted. To formally define such domains, the Semantic Web (SW) community has favoured the adoption of an ontology, as explained in [44] and [45].

A knowledge graph is conventionally described as a directed graph, which is defined as *"a structured representation of facts, consisting of entities, relationships, and semantic descriptions"* [46]. Nodes represent entities ranging from standard to specialized ones like individuals, locations, cars, files, or entities within domain-specific contexts like biology (e.g., viruses, chemical components). Edges, also known as properties or predicates, indicate relationships between entities or an attribute of an entity expressed as types and properties with a well-defined meaning. Moreover, edges and nodes can also depict an attribute of an entity and the corresponding attribute value, respectively, as discussed by the paper's authors [44].

Formally, a KG is defined by using $G = (N, R, T)$, with G representing the graph characterized by labels. The set of nodes and relationships is defined respectively as $N = \{n1, n2, ... n|N|\}$ and $R = \{r1, r2, ... r|R|\}$. T is the set of triplets representing the smallest piece of information non-trivial as the intrinsic information contained by nodes (or their labels) and is formally represented as $T = \{(n, r, n) \mid n, n \in N, r \in R\}$. [47]

The concept of a knowledge graph has become increasingly prevalent and standardized in the field, particularly with the successful launch of Google's Knowledge Graph [48].

While graphs have played a crucial role in AI since its inception, the Google Knowledge Graph truly brought the term into the mainstream consciousness. As a result, research on knowledge graphs has experienced remarkable growth and interest [48].

In the realm of graph database technologies, it is essential to comprehend two fundamental aspects:

- **Graph storage**: there are different approaches to graph database strategies, but the most efficient is using “native” graph storage. This storage category is designed to handle complex graphs, leading to superior speed and efficiency. Other systems may use relational or object-oriented databases for storage, but these options can cause potential performance lags.
- **Graph processing engine**: the processing engine is at the core of graph manipulation. Native graph processing, or “index-free adjacency”, is the most efficient way to handle graph operations. This method leverages direct physical connections between nodes in the database. Non-native graph processing engines may use alternate mechanisms to handle operations, but there may be variations in efficiency.

There are various search algorithms to consider when dealing with graph-based data queries. These approaches range from basic breadth-first and depth-first searches to more complex uninformed and informed searches like Dijkstra’s and the A* algorithm. Each algorithm has unique strengths and limitations, and no single variant is inherently superior to another. In the quest to improve traversal techniques, the A* algorithm represents a significant advancement from the foundational Dijkstra algorithm, also known as “A-star”. This algorithm combines features from Dijkstra’s approach with aspects of a best-first search. At its core, the A* algorithm recognizes that specific searches involve informed decision-making, allowing the selection of optimal paths within the graph.

Similar to Dijkstra’s methodology, A* can traverse large areas of the graph, but it also incorporates the essence of a best-first search, using a heuristic to guide its exploration. Unlike Dijkstra’s algorithm, which focuses on nodes near the starting point, a best-first search prioritizes nodes closer to the destination. [43]

Neo4j is a leader in the field of graph database technologies. Neo4j is a remarkable graph database that is labelled and property-based. Engineered with a native engine designed for operational and hybrid operational/analytic use cases, it maintains ACID compliance and immediate consistency, making it an ideal solution for real-time data applications. The platform’s seamless interaction is made possible by using Cypher or OpenCypher, which are popular among its many users.

In data modelling, query languages are based on reflecting the data models they represent. SQL focuses on tables and JOINs, while Cypher centres around the complex relationships between entities, using interconnected circles and arrows to create a language that is easy to understand. Graph database models empower users to ask more precise questions and offer a wider variety of query options, with Cypher being a popular choice due to its simplicity and accessibility. [43, 49]

Neo4j is available in two editions, Community and Enterprise, and offers a range of cloud integration options. It is generally regarded as the leader in the graph database landscape and has earned a reputation for providing a user-friendly experience with consistent innovation. Multi-clustering capabilities have effectively refuted previous scalability criticisms, making it a comprehensive and versatile solution comparable to relational database giants like Oracle or SQL Server.

One of the major advantages of Neo4j is its exceptional scalability, capable of accommodating billions of nodes on a single machine. It also boasts a user-friendly API that makes interactions easy. Additionally, it facilitates efficient graph traversal to optimize data exploration. Other benefits include whiteboard compatibility to ensure uniformity of the model across design, implementation, storage, and visualization within any domain, thus empowering all business stakeholders to take part in the development process. Neo4j’s features enable swift development, which results in highly scalable applications. Furthermore, it provides data security through ACID transactions, ensuring data integrity even in cases of power interruptions or system crashes. [43]

4.3.1 Improving graph drawing aesthetic

The critical role of knowledge graphs comes from the increased number of correlations and dependencies between cyber assets. Therefore, we decided to consider graph-based solutions as the basis for managing connections, which is fundamental to acquiring the complete context picture. In this section, we want to present a method used to improve the quality of graphs.

In the paper [50], this method is called a *heterogeneous graph*. Every cyber asset is part of either a tangible (or hard) class or an intangible (or logical) class. Relationships are represented as edges, while nodes are the actual entities. By using accurate data, the authors of [50] have noticed two main structures: clusters and bridges. These are visible by using specific graph layout methods. A cluster is a structure that aggregates multiple nodes (cyber assets), while a bridge is the point of contact between clusters realizing the graph structure. This is why bridges are considered critical cyber assets.

Graph drawing is a complex field involving various algorithms to accurately represent the relationships between nodes and edges in a graph. One of the most widely used algorithms is the force-directed model, which mimics a physical system by applying attractive and repulsive forces to nodes to determine the optimal layout of a graph. Many advancements have been made to the force-directed model, resulting in the development of various graph drawing algorithms. Eades' spring-embedded approach is one of the most influential models, which depicts nodes as steel rings and edges as springs. This approach seeks to achieve equilibrium by balancing the total force on each node, resulting in an optimal graph layout. [51]

The force-directed model is highly versatile and can be applied to different scenarios. Graph drawing aesthetics is a method to evaluate them and optimise the graph layout's readability for users. The pursuit of various aesthetic considerations aims to create a graph that adheres to graph-theoretic principles while resonating with users' subjective perception of visual appeal. These aesthetic considerations include minimising edge crossings, emphasising symmetry, maintaining uniform edge lengths, ensuring uniform node distribution, and segregating non-adjacent nodes. Each consideration serves a unique purpose and contributes to the overall enhancement of graph comprehension. However, it is essential to recognise that applying these principles can be competitive and require trade-offs between them. As a result, achieving one aesthetic goal may inadvertently lead to compromising another, making deriving an optimal graph layout a nuanced challenge. [51]

The selected methods in [50] for graph drawing encompass Kamada-Kawai (KK), ForceAtlas2, SE-Barnes-Hut, and Parameter-adjusted SE-Barnes-Hut, each offering distinct advantages and applications.

- **Kamada-Kawai** (KK) is well-known for its ability to generate detailed and context-rich graph layouts. However, this comes at the cost of speed, especially for graphs with more than 30 nodes. The primary focus of KK is to achieve optimal layouts that prioritise graph drawing principles. It has been found to perform exceptionally well on small, sparse graphs with up to 30 nodes, demonstrating its proficiency in adhering to graph drawing aesthetics. One of the critical requirements of KK is that Euclidean distance should correspond closely to graph theoretic distance. This ensures that the visual representation effectively reflects the underlying graph structure. [51]
- **ForceAtlas2**: ForceAtlas2 is a highly versatile algorithm that can adjust its speed and precision according to specific requirements. Generating a graph layout requires fewer iterations than Fruchterman-Reingold (FR). ForceAtlas2 significantly emphasises short edges to create a clear and understandable layout. It has been proven to perform well on large graphs, including those with over 20,000 nodes. [51]
- **SE-Barnes-Hut** and **Parameter-adjusted SE-Barnes-Hut**: the SE-Barnes-Hut algorithm is a highly advanced method for accelerating the brute force n-body algorithm. It utilises a unique approach of grouping nearby bodies, approximating them as a single entity. This is achieved through a quad-tree, which recursively divides the body set into regions, with each node representing a specific space region. The Barnes-Hut algorithm is particularly effective when dealing with bodies near each other. It calculates net forces by traversing the tree nodes and offering simulation accuracy adjustments via the parameter θ, which balances speed and precision. [52]

SE-Barnes-Hut resulted in a better node distribution of the four layout methods, avoiding overpopulated areas with undistinguishable clusters and bridges. Furthermore, the fourth method has been created by adjusting different parameters, such as collision force strength, edge distance, and electrical charge force strength, to improve the visualization quality.

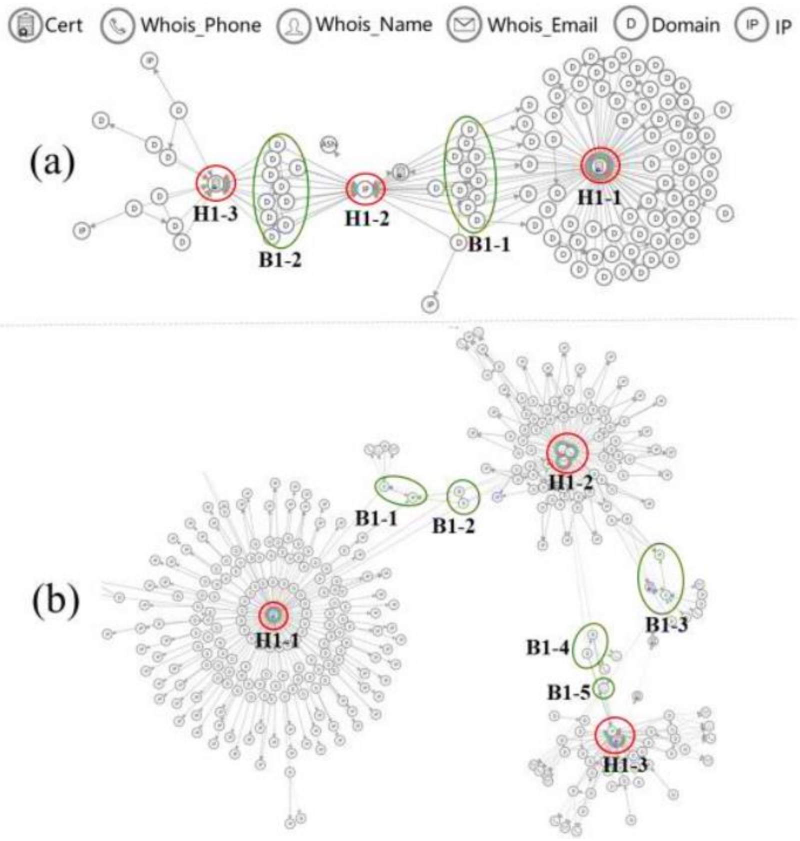


Figure 17: Heterogeneous graph modelling enhanced by using SE-Barnes-Hut algorithm. In the first case. (a), three clusters are represented by hub nodes H1-1, H1-2, and H1-3, along with a bridge structure involving B1-1 and B1-2. Case (b) displays a medium-sized cyber asset dataset with clusters highlighted in red and bridge structures in green, including single and double-hop bridges connecting various hubs. The use of clusters and bridges improves the aesthetic of the graph, making it easier to identify devices by their class. [50, p. 744]

In Figure 17, the application of the optimized SE-Barnes-Hut method, as explored by [50], is evident. This optimization offers substantial enhancements in terms of visual perception. As a result of comparing different layout methods, this chosen approach distinctly aids users in comprehending and pinpointing pivotal cyber assets. Red circles denote highlighted clusters within the visualisation, whereas other circles represent bridges or bridge hubs. The study conclusively underscores the efficacy of our method in discerning and identifying crucial cyber assets.

5 Case study: Starbase to manage GitHub as a data source

With this case study, we intend to showcase how Starbase enhances situational awareness and cyber security in managing a GitHub organization project, specifically in the management of directories, employee accounts and permissions. We aim to demonstrate how Starbase streamlines these processes, providing insights and control over critical assets while maintaining high security.

The first step to successfully integrate Starbase is deploying and integrating a *GitHub App* into the organization's profile, granting read-only permissions to collect and consolidate essential information. This sets the foundation for Starbase's operations, enabling the tool to leverage the data-rich environment of the GitHub organization. We simulated large volumes of users and interactions using GitHub's APIs, providing a robust testing ground for evaluating Starbase's functionalities within a dynamic organizational setting.

Through this case study, we want to demonstrate how Starbase contributes to the management, analysis, and enhanced visibility of a GitHub organization project. By showcasing the practical implementation of the tool's features in a real-world context, we underline its potential to bolster cyber security efforts and optimize visibility within dynamic digital ecosystems.

5.1 Integration process

The seamless integration of Starbase with GitHub, using the command-line interface (CLI) and yarn, marked a significant leap forward in cyber security management. The simulation started with using the bash shell and yarn to install all necessary packages and manage dependencies, setting the stage for a smooth integration process. Once all the packages have been installed or updated to the proper version, we have to configure Starbase by launching *yarn Starbase setup* command.
When everything was ready, the Starbase instance autonomously retrieved all data from the created GitHub organization, designed to mirror a real-world development environment. This included the utilization of the GitHub API to create cyber assets. All the simulation code, including the scripts for creating the GitHub “environment”, are available in the same GitHub repository and in the appendix of this thesis.

The integration process highlighted the tremendous capability of Starbase in making the daily responsibilities of cyber security specialists more manageable. Starbase, thanks to Neo4j Browser, offers a user-friendly interface that includes a menu meticulously organized to display all the cyber assets available at a given time.

This facilitates the quick selection and allows for a thorough examination of each asset by the users. Such a feature is of paramount importance for professionals entrusted with the management and oversight of a considerable number of assets. It effectively minimizes the time spent navigating through the system, thus enhancing overall productivity and allowing specialists to allocate more time to other critical areas of their work. Furthermore, the intuitive design of the menu aids in reducing the cognitive load of the users, which is particularly beneficial in high-pressure situations where swift decision-making is essential. Moreover, Starbase has been meticulously designed to include a set of predefined queries, a crucial feature that greatly aids in identifying irregularities within the GitHub repository. This additional functionality augments the tool's capability to monitor digital assets and fortifies the security measures in place. This proactive approach towards identifying anomalies is fundamental in preventing potential cyber threats before they can manifest and cause significant damage. Furthermore, the inclusion of these predefined queries eliminates the need for cyber security professionals to construct complex queries from scratch. This process can be both time-consuming and error-prone. Instead, professionals can leverage the existing queries as a starting point and customize them according to the organisation's specific needs. This accelerates the anomaly detection process and ensures higher accuracy in the results obtained.

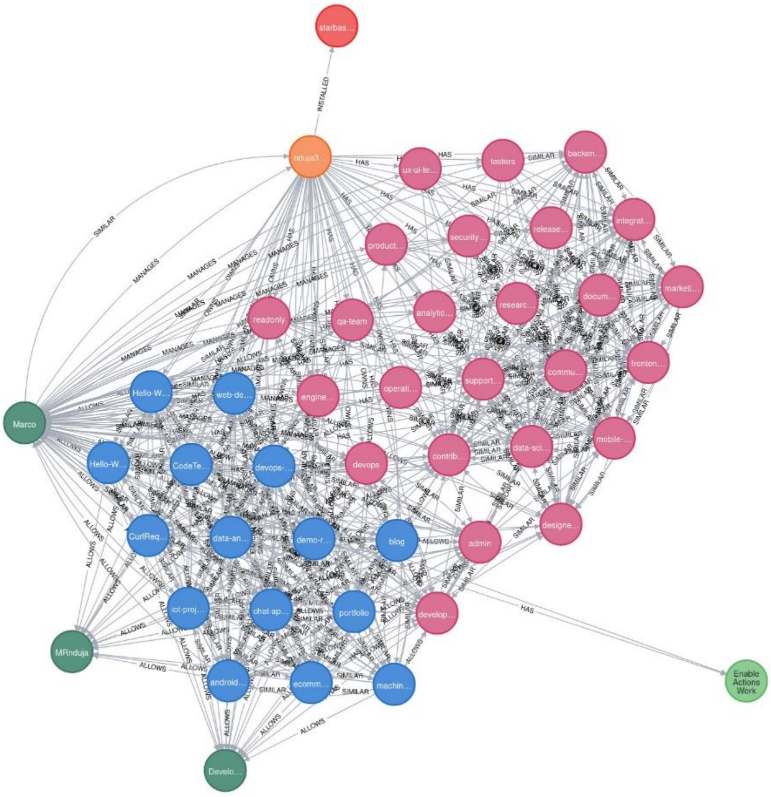


Figure 18: Figure of the return graph of the entire organization when using Neo4j Browser. This picture aims to demonstrate that when the visualization algorithm is not the proper one, it implies a drastic reduction of the quality of the aesthetic.

While insightful, the representation of an organization's structure as a graph presents significant challenges in data interpretation due to the inherent complexity and volume of the information involved. This issue was clearly illustrated in Figure 18, where an attempt was made to display the entire organization without employing any specific graph algorithms and by simply using the Neo4j Browser version. The resulting visualization was extremely cluttered, making it incredibly difficult to distinguish between the various relationships and nodes, as everything appeared tightly packed and overlapping.

While Neo4j is an excellent tool that complements Starbase by enhancing data management and transforming it into a visual format, it is evident from Figure 18 that it requires some fine-tuning to optimize its output. The default settings of the Neo4j Browser version do not suffice for clear and effective visualization of complex organizational structures. This is because the browser version employs general graph algorithms to optimize the display phase, leading to a dense and cluttered representation that hinders rather than aids understanding.

This experience underscores the necessity of utilizing advanced features of Neo4j, such as graph algorithms, to optimize the visualization of complex data structures. Specifically, the use of a force-based algorithm, as available in Neo4j Bloom, can significantly enhance the graph's quality and readability. This algorithm arranges the nodes and relationships in the graph to minimise overlaps and optimally utilise the available space. As a result, the nodes and relationships are more evenly distributed, making it easier to identify and analyse the various components of the graph.

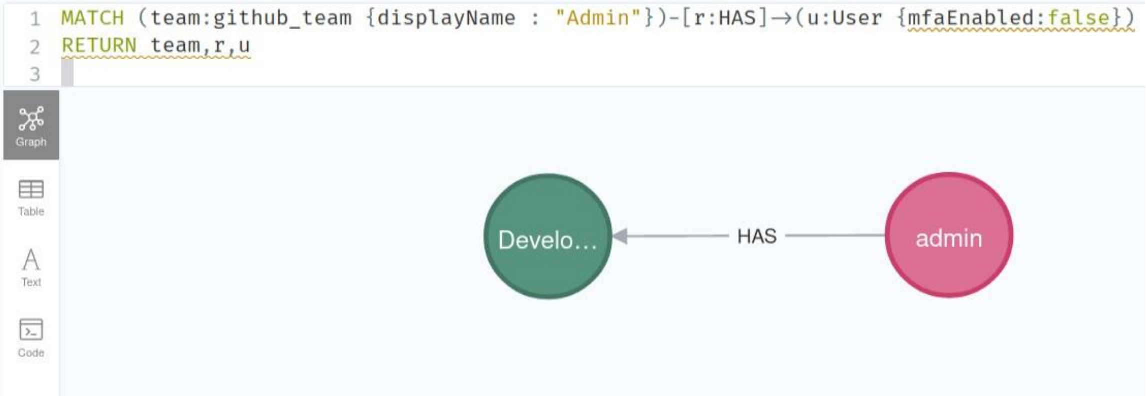
5.2 Security use case: managing account permissions

Throughout our experiment, we intentionally utilized GitHub's team feature to distribute permissions among team members. Although this tool is valuable for collaboration, it can present significant challenges if not implemented correctly. Specifically, we granted admin permissions to selected users by adding them to the Admin team. However, according to our organization's policy, all Admin team members and those with admin rights must have multi-factor authentication (MFA) activated.

Assigning and managing permissions requires careful attention and a thorough understanding of the organization's policies and tools. This task's importance is highlighted by its direct impact on the security of the organization's digital assets and the efficiency of its operations. Assigning permissions improperly can result in unauthorized access to confidential information, while overly restrictive permissions can hinder collaboration and efficiency.

MFA is an essential security feature designed to prevent unauthorized access to the organization's GitHub account. To increase security, host users must provide multiple forms of identification before logging into their accounts. This usually includes something that host users know, such as a password, and something they possess, like a mobile device capable of receiving a verification code or a token.

Enforcing an organization's policies is crucial, but it can be overwhelming. Typically, a security engineer must meticulously search copious amounts of data using the GitHub dashboard or API. This involves selecting users, confirming their active status, identifying all teams with administrative privileges, verifying user memberships, and producing a response. One particularly stressful component of this process is meticulously verifying administrative privileges, as it is essential to ensure that only authorized personnel have access. Unfortunately, this manual method is both time-intensive and prone to errors due to the numerous steps and vast amounts of data.



```
MATCH (team:github_team {displayName : "Admin"})-[r:HAS]->(u:User {mfaEnabled:false})
RETURN team,r,u
```

Figure 19: Query to get all the users who are not compliant with the organization's policy and result. The pink node is the team with a relationship with the Developer, a GitHub organisation user with admin privileges but without the multi-factor authentication enabled.

However, the integration of Starbase and Neo4j has significantly simplified this process. As shown in Figure 19, by utilizing the semantics and query language of Neo4j, we only need to specify the nodes and relationships of interest. The figure illustrates a query used to identify non-compliant users and its result. We could identify all users not adhering to the organization's policies with just a single command. Cypher, the query language of choice for Neo4j, embodies a clear and expressive syntax tailored to the nuances of graph databases. This syntax allows for precise and intuitive interactions with graph data structures, fostering a seamless bridge between the database and the user. The "MATCH" clause serves as the starting point for Cypher queries, allowing the specification of patterns to search for within the graph. Nodes, relationships, and properties are represented in ASCII art-like patterns, enabling visual alignment with the graph's structure.

Nodes are enclosed in parentheses (e.g., “(u:User)”). Relationships are represented with square brackets and an optional relationship type (e.g., “-[r:HAS]->”). Both nodes and relationships can have properties (e.g., `n.name`). In the figure, we can notice that some filters have been applied. Filters act on nodes’ properties and are represented by curly brackets (e.g., “{mfaEnabled:false}”); in the example, “mfaEnabled” is a property of “User”. Another option could be to use the "WHERE" clause to filter nodes and relationships based on specific conditions or property values.

The "RETURN" clause specifies what data to retrieve from the matched patterns. Other functions are aggregation functions like "COUNT," "SUM," and "AVG" that can be applied to calculate values, plus ordering and limiting like "ORDER BY" clause and "LIMIT" to control result set size. Cypher's syntax is unparalleled in its ability to express graph queries with elegance and ease. Its graph-centric design empowers users to confidently navigate, analyse, and extract valuable insights from graph data structures with unwavering precision and finesse.

This notable reduction in complexity and the time required to achieve the same result highlights the efficacy and potency of Starbase and Neo4j in administrating policy compliance. The effectiveness of combining Starbase and Neo4j to streamline the management of permissions and ensure policy compliance in a GitHub organization is clear and impactful. This combination simplifies the daunting task of navigating through vast amounts of data and speeds up identifying non-compliant users. It is a significant leap from the traditional manual method, which is time-consuming, prone to errors and requires much effort. In contrast, the combined use of Starbase and Neo4j enables quick identification of non-compliant users with just a single command, showcasing a significant advancement in cyber security management.

Moving on to the next section, we will improve graph quality using the force-based algorithm by integrating Neo4j Bloom with Starbase instances. This algorithm helps to optimize the visualization of complex data structures.

5.3 Neo4j Bloom: enhancing graph aesthetic

Integrating Starbase with Neo4j Desktop and Neo4j Bloom offers many essential advantages in analysing an organization's digital environment. One significant benefit is the capacity for visual exploration, which empowers individuals without a technical background to grasp the data's structure and connections without the need for complex queries. This expands the team's capability to extract valuable insights from the data. By default, Starbase uses the “browser” version of Neo4j. This limits the possibility of integrating existing tools or using advanced functionalities.

Neo4j Bloom enables users to customize the graph's visual aspects, including colours, dimensions, and symbols, to identify nodes and connections quickly. This capability is especially useful in complex networks with many nodes and relationships. The search feature in Neo4j Bloom also allows users to locate specific nodes or relationships promptly, which is crucial when handling vast datasets as it enables users to focus on the most relevant segments of the network. The force-based direct graph algorithm integrated into Neo4j Bloom enhances graph quality by refining its arrangement, making it more legible and comprehensible.

Graphs can become intricate and challenging to decipher, mainly when encompassing many nodes and relationships. This complexity can impede the ability of cyber security specialists and other interested parties to interpret the graph promptly and accurately, which is critical for making decisions quickly.

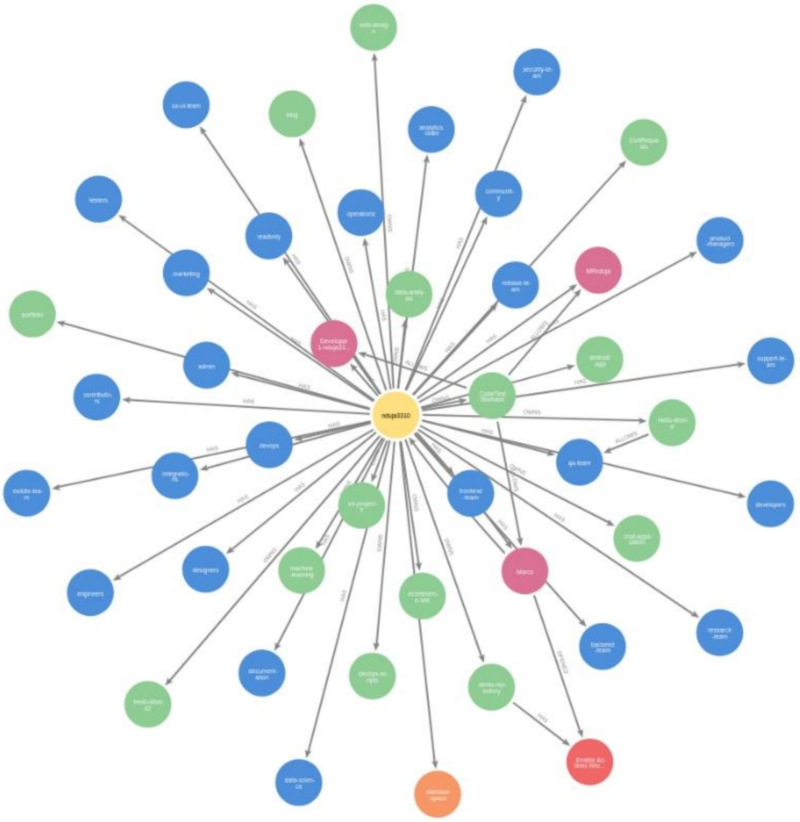


Figure 20: Graph of the entire organization by using the force graph algorithm in Neo4j Bloom. In yellow, we see a single node representing the organization, which one specific user manages. Users are in pink colour. We can notice that the organization is linked to the Starbase GitHub application to give read-only permission to Starbase. The blue nodes are the teams, while the green nodes are the repositories.

Integrating Starbase with Neo4j Desktop, notably Neo4j Bloom, has brought several key advantages that significantly enhance the efficiency and depth of analysis within an organisation's digital ecosystem. One remarkable advantage is the ability to visually explore the data, which allows even non-technical users to understand the structure and relationships within the data without writing any queries. This inclusivity broadens the spectrum of team members who can extract insights from the data. Additionally, Neo4j Bloom facilitates real-time interaction with the graph, a crucial feature for cyber security professionals who need to respond to threats as they unfold. The capacity to visualise changes in the network as they happen can lead to quicker identification of anomalies and faster response times.

Neo4j Bloom enables users to customise the graph's appearance, including the colours, sizes, and icons representing nodes and relationships through its GUI. This facilitates the identification of specific types of assets or relationships at a glance, especially in complex networks with many nodes and relationships. Moreover, the search functionality of Neo4j Bloom enables users to locate specific nodes or relationships without having to return and print a new graph instance every time. By using the main graph, Neo4j Bloom highlights the area of interest, reducing the computational time of every query. This feature is critical when dealing with extensive datasets as it allows users to focus on the most relevant parts of the network and optimize query results.

The aesthetic aspect of the organisation graph is significantly improved by the force-based algorithm, which optimises the layout, making it more readable and understandable. The algorithm operates by considering the graph as a physical system, wherein nodes are treated as charged particles that repel each other and the edges between them as springs that attract connected nodes. The algorithm successively adjusts the nodes' positions until equilibrium is achieved, resulting in a layout with evenly distributed nodes and uniform edge lengths. This minimises visual clutter, reduces the overlap of nodes and edges, and makes the graph more aesthetically pleasing. This improved layout aids in the analysis and interpretation of the graph by making it easier to identify important nodes and relationships, reducing the cognitive load on the user, and helping identify patterns and anomalies. Furthermore, the interactive nature of Neo4j Bloom allows users to customise the graph by dragging nodes and recalculating the layout in real-time. This feature enables users to tailor the visualisation to their preferences or highlight specific parts of the graph, ultimately leading to more informed decisions and insights.

Moreover, the force-based algorithm enhances graph quality by reducing node overlap and edge crossings, optimising node distribution and edge lengths, and clustering related nodes. This results in a more organised and visually appealing graph, facilitating better communication and collaboration among team members. Users can also define different perspectives in Neo4j Bloom, allowing them to view the data from various angles. It is particularly useful for cyber security professionals who need to understand the network from different viewpoints, such as from the perspective of an attacker or a defender. Furthermore, integrating with Neo4j Desktop facilitates using other plugins and tools available in the Neo4j ecosystem, including tools for data import/export, analytics, and machine learning. This contributes to a more comprehensive and robust cyber security analysis.

The integration of Starbase with Neo4j Desktop and Neo4j Bloom offers a powerful and flexible solution for managing and analysing cyber assets. The combination of visual exploration, real-time interaction, customisation, search functionality, perspective definition, and integration with other tools makes it invaluable for cyber security professionals seeking to optimise visibility, enhance situational awareness, and bolster cyber security efforts within dynamic digital ecosystems. This leads to a more aesthetically pleasing and organised visual representation and contributes to more informed decisions, better communication, and, ultimately, a more robust cyber security posture.

5.4 Knowledge graphs in cyber security use and areas of improvement

In cyber security, knowledge graphs have emerged as vital tools for navigating the complex and constantly changing landscape of cyber environments inside and outside organizations. As cyber attack threats grow, knowledge graphs offer a powerful means of gathering, visualizing, managing, and interpreting information.

Despite significant efforts to create comprehensive knowledge graphs, a noticeable gap remains in understanding how to apply them effectively to real-world challenges faced by industries in their efforts to defend against cyber attacks. The paper [47] critically examines the foundational concepts, schemas, and construction methodologies for cyber security knowledge graphs. It also presents a comparative analysis of the recent advancements in the application scenarios of cyber security knowledge graphs and proposes a thorough classification framework to categorize connected works into nine primary categories and eighteen subcategories. Moreover, it pinpoints existing research deficiencies and proposes promising avenues for future research. KGs have the potential to help security administrators gain insight into security intelligence, network conditions, and entity relationships intuitively.

This facilitates the identification of security entity attributes, which serves as a foundation for comprehending cyber security knowledge, scrutinizing cyber security data, and discerning attack patterns and abnormal traits associated with cyber attacks. [47]

Furthermore, another trend identified in [47] is the increasing importance of including event knowledge and dynamic knowledge, such as temporal information, conditional relationships, causal information, and event subordination relationships, in knowledge graphs as research on KGs progresses, and the demand for field applications grows. This will require significant efforts to represent cyber security event knowledge and facilitate relevant logical reasoning by constructing a temporal knowledge graph. It has also been noticed that there is a lack of universally accepted unified open-source knowledge graphs. Despite their undeniable value and practicality, KGs often grapple with incompleteness, redundancy, and ambiguity, culminating in uninformative query outcomes. Owing to divergent application requirements across various scenarios, researchers frequently find themselves compelled to construct a new knowledge graph from scratch.

Currently, the functionalities of KGs are predominantly confined to the query and display functions offered by tools such as Neo4j, thereby failing to fully harness the capabilities of KGs for automated reasoning. There remains a pressing need for clarity on applying KGs to address practical challenges in the cyber security domain. The semantic disconnect between KGs and logs severely hampers the application of KGs in attack path investigation. This void can be bridged by incorporating pertinent knowledge and establishing a semantic link between KGs and logs. [47] The absence of established evaluation standards for KGs is also a notable concern.

Another peculiar study [53] aimed to bolster the problem-solving prowess of researchers, educators, and professionals in cyber security through a Problem-Based Learning (PBL) lab system augmented with KG guidance. The findings revealed that KGs facilitate the exploration and acquisition of knowledge, the organization of correlated skills for tasks, the support of learning and problem-solving processes, and the enhancement of trainee confidence levels. Participants expressed a keen interest in utilizing KG guidance for future training. This demonstrates that KGs are a robust methodology to share and analyse information, which is fundamental to reducing possible human errors and helping cyber security experts with asset management.

In conclusion, knowledge graphs (KGs) positively impact cyber security by providing a comprehensive and interconnected view of an organization's network. KGs act as a digital representation of network data, allowing cyber security analysts and data scientists to identify malicious activities, vulnerabilities, and potential attack trajectories in new ways.

By consolidating network infrastructure, software information, threat intelligence, and real-time event data into a graph database, organizations can efficiently assess vulnerabilities, evaluate modifications, identify anomalous patterns, and respond effectively to cyber attacks. The ability to ingest and correlate data from various sources like vulnerability databases, intrusion detection systems, and configuration management tools empowers organizations to proactively counteract threats, monitor network activities, and identify compromised systems.

Tools such as Neo4j, GraphKer, Cartography, and Bloodhound facilitate the creation, analysis, and visualization of KGs, enabling organizations to manage and transform their cyber security landscapes adeptly. Furthermore, the availability of open-source tools and publicly accessible data from entities such as MITRE and NIST enhances the ability to analyse and respond to evolving cyber security threats. In conclusion, KGs can play a significant role in cyber security by offering an efficient means to model, manage, and transform the ever-changing landscape of cyber security. [54]

Conclusions

The evolution of the cyber landscape, characterised by the expansion of cyber assets and the growth of the attack surface, necessitates a novel, systematic approach to cyber security. This thesis started by elucidating the importance of understanding cyber assets as the foundational step in cyber security. With the growing attack surface and the proliferation of tools, finding the optimal combination of tools for a robust security posture has become an intricate task. The Cyber Defense Matrix, developed by Sounil Yu, emerged as a systematic and flexible approach that not only aids in identifying often overlooked assets but also in selecting the most suitable technical solutions in a business-oriented manner.

After validating our proposition through market trend analysis and real business needs examination, it was demonstrated that Attack Surface Management (ASM) tools are poised to become a fundamental component in many firms' security journeys. Of the various tools evaluated, Cyber Asset Attack Surface Management (CAASM) was identified as the most promising due to its capabilities, user-friendliness, and seamless integration with cloud environments via APIs, thereby addressing shadow IT. The potential market of CAASM tools has been shown in the last report of Gartner, which validated them as an emerging technology.

A scalable and reproducible use case scenario was devised to confirm the utility of CAASM and demonstrate its practical deployment, encompassing, thanks to GitHub features (i.e., organisation and enterprise accounts), small to large enterprises with multiple branches. Specifically, the open-source CAASM tool, Starbase by JupiterOne, was employed to simulate API integration and data collection from GitHub and manage cyber assets. The simulation leveraged Starbase, which uses Neo4j for its graphic user interface (GUI). While Neo4j provides an intuitive interface, the initial exploration revealed certain limitations, particularly in the aesthetics and usability of the knowledge graph. Additional tools, Neo4j Desktop and Neo4j Bloom were integrated into the setup to address these challenges and optimise the knowledge graph. Neo4j Bloom enhanced the visualisation of the knowledge graph. This enhancement facilitated the navigation and interpretation of the graph. The detailed exploration of Neo4j and its associated tools highlighted the importance of optimising graph visualisations for effective cyber security analysis. The integration of Neo4j Bloom played a pivotal role in enhancing the aesthetics and usability of the knowledge graph.

The simulation revealed that the knowledge graph facilitates easier comprehension, even for non-technical individuals. Moreover, the database query language expedited the identification of policy discrepancies, potential weaknesses, and vulnerabilities in cyber security.

This simulation underscores the efficacy of knowledge graphs in cyber security and the potential of CAASM, especially open-source projects like Starbase. Starbase not only offer the flexibility to define integrations via the SDK but also a plethora of pre-integrations managed by its open community, free of charge.

In conclusion, this thesis underscores the imperative to simplify and enhance the interoperability of cyber security tools. By leveraging methodical approaches like the Cyber Defense Matrix and innovative tools like CAASM, organisations can navigate the complexity of the cyber landscape more effectively, thereby fortifying their security posture. The demonstration with Starbase exemplifies how organisations can harness the power of knowledge graphs and the open-source community to make cyber security more accessible and robust. As cyber threats continue to evolve, it is incumbent upon organisations to adopt a proactive, systematic, and innovative approach to safeguard their cyber assets and, by extension, their business operations.

References

[1] Sounil Yu (Author), Dan Geer (Foreword), Wendy Nather (Foreword), *Cyber Defense Matrix: The Essential Guide to Navigating the Cybersecurity Landscape.* [Online]. Available: https://www.amazon.com/Cyber-Defense-Matrix-Navigating-Cybersecurity/dp/B09QP2GSGZ (accessed: May 20 2023)

[2] NIST - National Institute of Standards and Technology, *Glossary | CSRC.* [Online]. Available: https://csrc.nist.gov/glossary/term/asset (accessed: Jun. 25 2023)

[3] Gartner, Inc., *Cyber Asset Attack Surface Management Reviews 2023 | Gartner Peer Insights.* [Online]. Available: https://www.gartner.com/reviews/market/cyber-asset-attack-surface-management (accessed: Jun. 25 2023)

[4] NIST - National Institute of Standards and Technology, *CIA triad definition - NIST.* [Online]. Available: https://csrc.nist.gov/glossary/term/security (accessed: Jun. 25 2023)

[5] Scrut Automation, *Key Attack Surface Challenges Cloud-Native Companies Are Facing Today.* [Online]. Available: https://www.scrut.io/ebooks/key-attack-surface-challenges-cloud-native-companies-are-facing-today (accessed: Jun. 25 2023)

[6] Jasmine Henry, "The State of Cyber Assets Report: Jupiteone research," vol. 2023, pp. 1–71, 2023. [Online]. Available: https://info.jupiterone.com/hubfs/SCAR%202023/jupiterone_2023-state-of-cyber-assets-report_scar.pdf (accessed: May 20 2023)

[7] Y. Alrowaili, N. Saxena, A. Srivastava, M. Conti, and P. Burnap, "A review: Monitoring situational awareness of smart grid cyber-physical systems and critical asset identification," *IET Cyber-Phy Sys Theory & Ap*, vol. 8, no. 3, pp. 160–185, 2023, doi: 10.1049/cps2.12059.

[8] Pat Toth, *Cybersecurity – A Critical Component of Industry 4.0 Implementation | NIST.* [Online]. Available: https://www.nist.gov/blogs/manufacturing-innovation-blog/cybersecurity-critical-component-industry-40-implementation (accessed: Jun. 29 2023)

[9] M. Silic and A. Back, "Shadow IT – A view from behind the curtain," *Computers & Security*, vol. 45, pp. 274–283, 2014, doi: 10.1016/j.cose.2014.06.007.

[10] P. Mukherjee, *Short-Guide-On-Attack-Surface-Reduction-v1-2.* [Online]. Available: https://www.firecompass.com/wp-content/uploads/2020/12/Short-Guide-On-Attack-Surface-Reduction-v1-2.pdf (accessed: Jul. 1 2023)

[11] S. Behrens, "Shadow systems," *Commun. ACM*, vol. 52, no. 2, pp. 124–129, 2009, doi: 10.1145/1461928.1461960.

[12] A. A. B. Györy, A. Cleven, F. Uebernickel, and W. Brenner, "Exploring The Shadows: IT Governance Approaches To User-Driven Innovation," *ECIS 2012 Proceedings*, 2012. [Online].

Available: https://www.alexandria.unisg.ch/entities/publication/96a8ecae-7f07-406c-b253-68de008b7711/details (accessed: Jul. 7 2023)

[13] Marius Brouwer and Anand Groenewegen, “Cloud Access Security Brokers (CASBs) Characterization of the CASB market and its alignment with corporate expectations Commissioned by KPMG Netherlands,” pp. 1–12, 2021. [Online]. Available: https://rp.os3.nl/2020-2021/p33/report.pdf (accessed: Jun. 5 2023)

[14] © Marc Hulsebosch, University of Twente, Faculty of Electrical Engineering, and Mathematics and Computer Science, “Cloud Strife: An analysis of Cloud-based Shadow IT and a framework for managing its risks and opportunities,” Master of Science- Business Information Technology Faculty of Electrical Engineering, Mathematics and Computer Science University of Twente, University of Twente, Enschede, 2016 – version 0.6. [Online]. Available: https://essay.utwente.nl/69236/1/Hulsebosch_MA_EEMCS.pdf (accessed: Jun. 4 2023)

[15] Sandy Behrens and Wasana Sedera, “Why Do Shadow Systems Exist after an ERP Implementation? Lessons from a Case Study,” 2004. [Online]. Available: https://www.semanticscholar.org/paper/Why-Do-Shadow-Systems-Exist-after-an-ERP-Lessons-a-Behrens-Sedera/ef287c50bd64212e2019ea099a3d9acfef0b1ded#cited-papers (accessed: Jul. 8 2023)

[16] G. L. Mallmann, A. C. G. Maçada, and M. Oliveira, “The influence of shadow IT usage on knowledge sharing,” *Business Information Review*, vol. 35, no. 1, pp. 17–28, 2018, doi: 10.1177/0266382118760143.

[17] SafeBreach, *The impact of continuous security validation discovery report.* [Online]. Available: https://www.safebreach.com/wp-content/uploads/2023/04/impact_of_continuous_security_validation_discovery_report.pdf (accessed: Jun. 5 2023)

[18] Di Freeze, *60 Percent of Small Companies Close Within 6 Months of Being Hacked.* [Online]. Available: https://cybersecurityventures.com/60-percent-of-small-companies-close-within-6-months-of-being-hacked/ (accessed: Jul. 16 2023)

[19] Brad LaPorte, Gartner Veteran and Ordr Strategic Advisor, “IMPLEMENTING CONNECTED DEVICE SECURITY FOR HEALTHCARE ORGANIZATIONS,” *Ordr Startegic Advisor & Gartner*, pp. 1–27, 2023. [Online]. Available: https://ordr.net/healthcare/ (accessed: Jul. 20 2023)

[20] Lansweeper, “CAASM - Lansweeper for Cyber Asset Attack Surface Management,” pp. 1–14, 2023. [Online]. Available: www.lansweeper.com (accessed: Jul. 20 2023)

[21] Jasmine Henry, *Reinventing Cybersecurity: Latha Maripuri, Aubrey Stearn, Carla Sun, Lonye Ford, Dr. Meg Layton, Tracy Bannon, Breanne Boland, Alison Gianotto, Lisa Hall, Rin Oliver,*

Joyous Huggins, Yvie Djieya, Angela Marafino, Coleen Shane, and Rachel Harpley, for sharing your stories. Morrisville, NC: JupiterOne, Inc, 2022.

[22] Andrew Davies - Gartner, “Hype Cycle for Security Operations,: Hype Cycle,” 2022. [Online]. Available: https://www.gartner.com/en/doc/759058-security-operations-primer-for-2022 (accessed: Jul. 10 2023)

[23] Gartner - Jonathan Nunez, Andrew Davies, “Hype Cycle for Security Operations-2023,” pp. 1–84, 2023. [Online]. Available: https://www.firecompass.com/wp-content/uploads/2023/07/Hype-Cycle-for-Security-Operations-2023.pdf?utm_source=newsletter&utm_medium=email&utm_term=2023-07-31&utm_campaign=Download+Gartner+Report+Hype+Cycle+For+Security+Operations+2023 (accessed: Jul. 31 2023)

[24] Gartner, *Magic Quadrant for Cloud Access Security Brokers.* [Online]. Available: https://www.gartner.com/en/documents/3992205 (accessed: Jul. 27 2023)

[25] Store.Frost.com, *Analysis of the Global Cloud Access Security Broker Market (CASB), Forecast 2021.* [Online]. Available: https://store.frost.com/analysis-of-the-global-cloud-access-security-broker-market-casb-forecast-2021.html (accessed: Jul. 27 2023)

[26] A. Choudhary, A. Tripathi, A. Sharma, and R. Singh, “Evolution and comparative analysis of different Cloud Access Security Brokers in current era,” in *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)*, Uttarakhand, India, 2022, pp. 37–43. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10059455 (accessed: Jun. 25 2023)

[27] Gartner, *Say Hello to SASE (Secure Access Service Edge) - Andrew Lerner.* [Online]. Available: https://blogs.gartner.com/andrew-lerner/2019/12/23/say-hello-sase-secure-access-service-edge/ (accessed: Jul. 27 2023)

[28] J. W. Jake Williams, “Maximizing Security Value Through External Attack Surface Management | SANS Institute,” Jul. 28 2023. [Online]. Available: https://www.sans.org/webcasts/maximizing-security-value-through-external-attack-surface-management/ (accessed: Jul. 25 2023)

[29] Fortinet, “Swiftly Find and Remediate Security Issues in the External Attack Surface With FortiRecon EASM,” 2023. [Online]. Available: https://www.fortinet.com/content/dam/fortinet/assets/solution-guides/sb-fortirecon-easm.pdf (accessed: Jul. 29 2023)

[30] Danielle Dennis, *Defender EASM Overview: Microsoft Defender External Attack Surface Management (Defender EASM).* [Online]. Available: https://learn.microsoft.com/en-us/azure/external-attack-surface-management/ (accessed: Jul. 29 2023)

[31] Secura, “Minimizing Your Digital Footprint: The Importance of External Attack Surface Management (EASA): Secura’s External Attack Surface Assessment (EASA) service,” [Online]. Available: https://www.secura.com/services/information-technology/attack-surface-assessment (accessed: Jul. 29 2023)

[32] CrowdStrike, *FALCON SURFACE: EXTERNAL ATTACK SURFACE MANAGEMENT (EASM): The industry’s most complete EASM technology stops breaches by minimizing risk from exposed assets.* [Online]. Available: https://www.crowdstrike.com/wp-content/uploads/2022/12/CrowdStrike-Falcon-Surface-datasheet.pdf (accessed: Jul. 29 2023)

[33] Lucidium, *When it Comes to Cyber Asset Attack Surface Management Planning.* [Online]. Available: https://lucidium.io/ebooks/3-mistakes-when-it-comes-to-cyber-asset-attack-surface-management-planning/ (accessed: Jun. 5 2023)

[34] Scrut Automation, “CAASM - A Must For A CISO’s Tech Stack,” 2022. [Online]. Available: https://www.scrut.io/wp-content/uploads/2023/05/CAASM-a-must-for-a-CISOs-tech-stack_V1-2.pdf (accessed: Jun. 5 2023)

[35] V. L, M. B. Gowda, G. G. Sindhu, and K. V, “Cyber Attack Surface Management System,” *IJARSCT*, pp. 1–9, 2023, doi: 10.48175/IJARSCT-9533.

[36] E. Zheng, *Launching Starbase: A New Open-Source Contribution from JupiterOne.* [Online]. Available: https://www.jupiterone.com/blog/jupiterone-contributes-starbase-to-open-source-community (accessed: Aug. 19 2023)

[37] Gartner, Inc., *Best Cyber Asset Attack Surface Management Tools Reviews 2023 | Gartner Peer Insights.* [Online]. Available: https://www.gartner.com/reviews/market/cyber-asset-attack-surface-management (accessed: Aug. 23 2023)

[38] GitHub, *About GitHub's APIs - GitHub Docs.* [Online]. Available: https://docs.github.com/en/rest/overview/about-githubs-apis?apiVersion=2022-11-28 (accessed: Aug. 22 2023)

[39] GitHub, *Migrating from REST to GraphQL - GitHub Docs.* [Online]. Available: https://docs.github.com/en/graphql/guides/migrating-from-rest-to-graphql (accessed: Aug. 22 2023)

[40] Octokit Team, *API documentation, octokit/rest.js interacting with Github.* [Online]. Available: https://octokit.github.io/rest.js/v20#throttling (accessed: Aug. 22 2023)

[41] GitHub, *Scripting with the REST API and JavaScript - GitHub Docs.* [Online]. Available: https://docs.github.com/en/rest/guides/scripting-with-the-rest-api-and-javascript?apiVersion=2022-11-28 (accessed: Aug. 22 2023)

[42] GitHub, *About the GraphQL API - GitHub Docs.* [Online]. Available: https://docs.github.com/en/graphql/overview/about-the-graphql-api (accessed: Aug. 22 2023)

[43] Neo4J, *Graph Databases For Beginners.* [Online]. Available: https://neo4j.com/wp-content/themes/neo4jweb/assets/images/Graph_Databases_for_Beginners.pdf (accessed: Aug. 5 2023)

[44] M. Kejriwal, "Knowledge Graphs: A Practical Review of the Research Landscape," *Information*, vol. 13, no. 4, p. 161, 2022, doi: 10.3390/info13040161.

[45] P. Hitzler, "A review of the semantic web field," *Commun. ACM*, vol. 64, no. 2, pp. 76–83, 2021, doi: 10.1145/3397512.

[46] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2022, doi: 10.1109/TNNLS.2021.3070843.

[47] K. Liu, F. Wang, Z. Ding, S. Liang, Z. Yu, and Y. Zhou, "Recent Progress of Using Knowledge Graph for Cybersecurity," *Electronics*, vol. 11, no. 15, p. 2287, 2022, doi: 10.3390/electronics11152287.

[48] A. Singhal, "Introducing the Knowledge Graph: things, not strings," *Google*, 5/16/2012, 5/16/2012. https://blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed: Aug. 21 2023)

[49] D. H. Philip Howard, "Neo4j InBrief," 2020. [Online]. Available: https://neo4j.com/whitepapers/analyst-bloor-research-neo4j-inbrief/

[50] S. Fu *et al.,* "Heterogeneous Graph Modeling and Visualization for Cyber Asset Management," in *2021 8th International Conference on Dependable Systems and Their Applications (DSA)*, Yinchuan, China, 2021, pp. 743–744.

[51] H. Gibson, J. Faith, and P. Vickers, "A survey of two-dimensional graph layout techniques for information visualisation," *Information Visualization*, vol. 12, 3-4, pp. 324–357, 2013, doi: 10.1177/1473871612455749.

[52] C. Swinehart, *The Barnes-Hut Algorithm.* [Online]. Available: http://arborjs.org/docs/barnes-hut (accessed: Aug. 23 2023)

[53] Y. Deng, Z. Zeng, K. Jha, and D. Huang, "Problem- Based Cybersecurity Lab with Knowledge Graph as Guidance," *JAIT*, 2021, doi: 10.37965/jait.2022.0066.

[54] Neo4J, "Graphs for Cybersecurity: Knowledge Graph as Digital Twin," *Neo4j*, 7/26/2022, 7/26/2022. https://neo4j.com/blog/graphs-cybersecurity-knowledge-graph-digital-twin/ (accessed: Aug. 30 2023).

Appendices

The appendices section of this thesis provides a comprehensive compilation of technical materials and scripts utilized at various stages of this research. Providing this detailed information aims to facilitate any researcher or practitioner who wishes to reproduce a similar use case, verify the results, or build upon the work presented in this thesis. This section is organized into two primary subsections, each serving a distinct purpose and contributing to the overall objective of the research.

Appendix 1 GitHub Interaction via API

The first subsection encompasses the Python and cURL scripts used to interact with the GitHub API, a pivotal component of the research as it enabled the automation of data retrieval and execution of various tasks. The GitHub API serves as a programmatic interface to GitHub, allowing automated interaction with the platform and retrieving vast amounts of data that would be impractical to collect manually.

The Python scripts included here comprise functions for authenticating with the API, retrieving data, and executing other essential tasks. Python was chosen as the programming language due to its readability, extensive libraries, and wide acceptance in the scientific and engineering communities.

Creating Teams in GitHub

```
import requests

file = open("config.txt","r").read().splitlines()
token = file[0]
orgname = file[1]

team_names = [
"Designers", "QA Team",
"DevOps", "Product Managers",
"Security Team", "Support Team",
"Contributors", "Documentation",
"Marketing", "Frontend Team",
"Backend Team", "UX/UI Team",
"Release Team", "Operations",
"Data Science", "Research Team",
"Community", "Analytics Team",
"Mobile Team", "Integrations"
]

url = "https://api.github.com/orgs/%s/teams" % orgname
headers = {
"Accept": "application/vnd.github+json",
"Authorization": "Bearer %s" % token
}

for team in team_names:
data = {
"name": team,
"description": "Team for %s" % team,
"permission": "push",
"notification_setting": "notifications_enabled",
"privacy": "closed"
}

response = requests.post(url, headers=headers, json=data)
if response.status_code == 201:
print("Team created successfully!")
print("Team ID:", response.json()["id"])

else:
print("Failed to create team. Status code:", response.status_code)
print("Response:", response.text)
```

Inviting new members to join the organization

```
import requests
import requests
import random

file = open("config.txt","r").read().splitlines()
token = file[0]
orgname = file[2]

teams = {8479128: 'Admin', 8481099: 'Analytics Team', 8481092: 'Backend Team', 8481098:
'Community', 8481088: 'Contributors', 8481096: 'Data Science', 8481082: 'Designers', 8479131:
'Developers', 8481084: 'DevOps', 8481089: 'Documentation', 8479115: 'Engineers', 8481091:
'Frontend Team', 8481101: 'Integrations', 8481090: 'Marketing', 8481100: 'Mobile Team',
8481095: 'Operations', 8481085: 'Product Managers', 8481083: 'QA Team', 7937258: 'ReadOnly',
8481094: 'Release Team', 8481097: 'Research Team', 8481086: 'Security Team', 8481087:
'Support Team', 8479125: 'Testers', 8481093: 'UX/UI Team'}

teams_arr = [8479128, 8481099, 8481092, 8481098, 8481088, 8481096, 8481082, 8479131,
8481084, 8481089, 8479115, 8481091, 8481101, 8481090, 8481100, 8481095, 8481085,
8481083, 7937258, 8481094, 8481097, 8481086, 8481087, 8479125, 8481093]

url = f"https://api.github.com/orgs/{orgname}/invitations"
headers = {
    "Accept": "application/vnd.github+json",
    "Authorization": f"Bearer {token}"
}

while True:
    email = input("Enter exit or the email address: ")
    if email=="exit": break
    if "@" not in email: continue
    num_teams = random.randint(1, 3)
    max_team_id = len(teams_arr)-1
    team_ids = [random.randint(0, max_team_id) for _ in range(num_teams)]

    data = {
        "email": email,
        "role": "direct_member",
        "team_ids": team_ids
    }
    response = requests.post(url, headers=headers, json=data)
    if response.status_code == 201:
        print("Invitation sent successfully!")
    else:
        print("Failed to send invitation. Status code:", response.status_code)
        print("Response:", response.text)
```

Changing Teams’ permission

```
import requests
import random
import json

file = open("config.txt","r").read().splitlines()
token = file[0]
orgname = file[2]

with open("TeamSlug.txt", "r") as file:
    team_dict = json.load(file)

repos_data = [
    ("blog", "Personal tech blog about programming and technology."),
    ("ecommerce-site", "Full-stack e-commerce website project."),
    ("data-analysis", "Data analysis projects using Python and pandas."),
    ("portfolio", "My professional portfolio showcasing my work."),
    ("android-app", "Mobile app development using Kotlin."),
    ("machine-learning", "Repository for machine learning algorithms."),
    ("web-design", "Web design concepts and resources."),
    ("devops-scripts", "Scripts for automating DevOps tasks."),
    ("iot-projects", "Internet of Things projects using Arduino."),
    ("chat-application", "Real-time chat application using sockets.")
]

teams_for_directories = [
    # For "blog" directory
    ["8479128", "8481088", "8481082", "8481089"],
    # For "ecommerce-site" directory
    ["8479128", "8481088", "8479131", "8481082", "8481083"],
    # For "data-analysis" directory
    ["8481096", "8481088", "8479131", "8481083"],
    # For "portfolio" directory
    ["8479128", "8481088", "8481082", "8481085"],
    # For "android-app" directory
    ["8479128", "8481088", "8479131", "8481082", "8481083"],
    # For "machine-learning" directory
    ["8481096", "8481088", "8479131"],
    # For "web-design" directory
    ["8479128", "8481088", "8481082", "8481083"],
    # For "devops-scripts" directory
    ["8479128", "8481088", "8481084", "8481095"],
    # For "iot-projects" directory
    ["8479128", "8481088", "8479131", "8481083"],
    # For "chat-application" directory
    ["8479128", "8481088", "8479131", "8481083", "8481087"]
]

burl = f"https://api.github.com/orgs/{orgname}/"
```

```
headers = {
    "Accept": "application/vnd.github+json",
    "Authorization": f"Bearer {token}",
    "X-GitHub-Api-Version": "2022-11-28"
}

data = {
    "permission": "push"
}

i = 0
for tup in repos_data:
    repo_name = tup[0]
    print(repo_name)
    for teamid in teams_for_directories[i]:
        team_slug = team_dict[teamid]
        print(team_slug)
        url = f"teams/{team_slug}/repos/{orgname}/{repo_name}"
        url = burl + url
        print(url)
        response = requests.put(url, headers=headers, json=data)
        if response.status_code == 204:
            print("Repository updated successfully!")
    i = i + 1
```

Getting Teams’ slug

```
import requests
import json
file = open("config.txt","r").read().splitlines()
token = file[0]
orgname = file[2]

url = f"https://api.github.com/orgs/{orgname}/teams"
headers = {
    "Accept": "application/vnd.github+json",
    "Authorization": f"Bearer {token}"
}

response = requests.get(url, headers=headers)

if response.status_code == 200:
    teams_data = response.json()

    # Create a dictionary with team IDs as keys and slugs as values
    team_dict = {team["id"]: team["slug"] for team in teams_data}

    # Write the dictionary to TeamSlug.txt
    with open("TeamSlug.txt", "w") as output_file:
        output_file.write(json.dumps(team_dict))
    print("Team slugs written to TeamSlug.txt")
else:
    print("Failed to fetch teams. Status code:", response.status_code)
    print("Response:", response.text)
```

Creating GitHub repositories

```
import requests
import random

file = open("config.txt","r").read().splitlines()
token = file[0]
orgname = file[1]

repos_data = [
    ("blog", "Personal tech blog about programming and technology."),
    ("ecommerce-site", "Full-stack e-commerce website project."),
    ("data-analysis", "Data analysis projects using Python and pandas."),
    ("portfolio", "My professional portfolio showcasing my work."),
    ("android-app", "Mobile app development using Kotlin."),
    ("machine-learning", "Repository for machine learning algorithms."),
    ("web-design", "Web design concepts and resources."),
    ("devops-scripts", "Scripts for automating DevOps tasks."),
    ("iot-projects", "Internet of Things projects using Arduino."),
    ("chat-application", "Real-time chat application using sockets.")
]

teams_for_directories = [
    # For "blog" directory
    [8479128, 8481088, 8481082, 8481089],
    # For "ecommerce-site" directory
    [8479128, 8481088, 8479131, 8481082, 8481083],
    # For "data-analysis" directory
    [8481096, 8481088, 8479131, 8481083],
    # For "portfolio" directory
    [8479128, 8481088, 8481082, 8481085],
    # For "android-app" directory
    [8479128, 8481088, 8479131, 8481082, 8481083],
    # For "machine-learning" directory
    [8481096, 8481088, 8479131],
    # For "web-design" directory
    [8479128, 8481088, 8481082, 8481083],
    # For "devops-scripts" directory
    [8479128, 8481088, 8481084, 8481095],
    # For "iot-projects" directory
    [8479128, 8481088, 8479131, 8481083],
    # For "chat-application" directory
    [8479128, 8481088, 8479131, 8481083, 8481087]]

url = f"https://api.github.com/orgs/{orgname}/repos"
headers = {
    "Accept": "application/vnd.github+json",
    "Authorization": f"Bearer {token}"
}
i = 0
```

```
for tup in repos_data:
  for team in teams_for_directories[i]:
   data = {
       "name": tup[0],
       "description": tup[1],
       "homepage": "https://github.com",
       "private": False,
       "has_issues": True,
       "has_projects": True,
       "has_wiki": True,
       "team_id": team,
       "auto_init": True,
       "license_template": "gpl-3.0"
   }
i = i + 1
   response = requests.post(url, headers=headers, json=data)
   if response.status_code == 201:
       print("Repository created successfully!")
       print("Repository URL:", response.json()["html_url"])
   else:
       print("Failed to create repository. Status code:", response.status_code)
       print("Response:", response.text)
```

These scripts involve the use of cURL, a command-line tool employed for sending and receiving data via various network protocols. The cURL scripts contain the necessary commands and options required to interact with the GitHub API, including sending HTTP requests and managing the received data.

List of teams

```
curl -L \
-H "Accept: application/vnd.github+json" \
-H "Authorization: Bearer $(head -n 1 config.txt)" \
https://api.github.com/orgs/$(head -n 3 config.txt | tail -1)/teams -o output.json
```

Creating projects

```
curl -L \
-X POST \
-H "Accept: application/vnd.github+json" \
-H "Authorization: Bearer $(head -n 1 config.txt)" \
https://api.github.com/orgs/$(head -n 3 config.txt | tail -1)/projects \
-d '{"name":"Organization Roadmap","body":"High-level roadmap for this year."}'
```

List of repositories

```
curl -L \
-H "Accept: application/vnd.github+json" \
-H "Authorization: Bearer $(head -n 1 config.txt)" \
-H "X-GitHub-Api-Version: 2022-11-28" \
https://api.github.com/orgs/$(head -n 3 config.txt | tail -1)/repos
```

Creating repository

```
curl -L \
-X POST \
-H "Accept: application/vnd.github+json" \
-H "Authorization: Bearer $(head -n 1 config.txt)" \
https://api.github.com/orgs/$(head -n 3 config.txt | tail -1)/repos \
-d '{"name":"Hello-World","description":"This is your first
repository","homepage":"https://github.com","private":false,"has_issues":true,"has_projects":true
,"has_wiki":true,"team_id":8481083,"auto_init":true,"license_template":"gnu"}'
```

Appendix 2 Starbase Configuration

The second subsection consists of the configuration files and settings necessary to operate Starbase and configure Neo4j, two critical tools utilized in this research.

The *config.yaml* file of Starbase contains the settings and parameters necessary to deploy Starbase and establish a connection with the integrations. Besides, to modify the admin user's password of Neo4j database, one useful command is *ALTER USER neo4j SET PASSWORD 'mynewpassword'*.

Configuration file of Starbase

```
integrations:
  -
    name: graph-github
    instanceId: my-starbase-github-integration
    directory: ./.starbase/.integrations/graph-github
    gitRemoteUrl: https://github.com/JupiterOne/graph-github.git
    config:
      GITHUB_APP_ID: '336545'
      GITHUB_APP_LOCAL_PRIVATE_KEY_PATH: ./home/starbase-space.2023-05-21.private-key.pem
      INSTALLATION_ID: '37760834'
storage:
  -
    engine: neo4j
    config:
      username: neo4j
      password: devpass
      uri: bolt://localhost:7687
      database: neo4j
```