National College of Ireland

Understanding the Efficacy of Cloud Native Sensitive Data Protection Capabilities

MSc Research Project
Cyber Security

Cormac Frawley
Student ID: X21166277

School of Computing
National College of Ireland

Supervisor: Raza Ul Mustafa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Cormac Frawley

Student ID: X21166277

Programme: MSc Cyber Security – Part Time **Year:** Jan 2022

Module: MSc Research Project

Supervisor: Raza Ul Mustafa

Submission Due Date: 15th December 2023

Project Title: Understanding the Efficacy of Cloud Native Sensitive Data Protection Capabilities

Word Count: 12,571 **Page Count** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Understanding the Efficacy of Cloud Native Sensitive Data Protection Capabilities

Cormac Frawley
X21166277

Abstract

Given commercial and regulatory imperatives the area of Data Protection & Privacy (DPP) is, or should be, of central importance to any organisation storing data considered to be sensitive. Regulatory requirements requiring organisations to ensure that data is correctly identified, classified and protected are now standard in many parts of the world. In order to protect data it is necessary to know where it is located, have a means to classify it and finally have the ability to enforce data protection policies.

Technical controls used to support data identification, classification and protection have generally been classed as Data Loss Prevention (DLP) solutions. While DLP solutions are applicable, and included, in this research the term itself can be restrictive as definitions can differ on what constitutes a DLP solution. A number of definitions include an in-line preventative capability while other definitions are more broad and define DLP as any solution that provides data protection capabilities. For this reason this research paper will use the term Sensitive Data Protection which covers the broader interpretation.

The advent and subsequent acceleration to public cloud platforms has meant that data protection in the cloud has become a critical risk for large numbers of organisations. Locating and classifying sensitive information data within cloud infrastructure solutions is a key security control for organisations and as a result public cloud vendors have responded by providing native Sensitive Data Protection capabilities across their storage offerings.

Given large numbers of organisations will seek to reduce data protection risk through these capabilities it becomes vital to understand the efficacy of the proposed control. This research is intended to compare the functionality and efficacy of these Sensitive Data Protection solutions across the three main public cloud providers. A standardised testing approach will be used across the cloud platforms utilising automated configuration capabilities and allowing meaningful comparison of the different vendor capabilities.

Keywords: Data Protection, Data Loss Prevention, Public Cloud

1 Introduction

1.1 Research Question

This research aims to answer the following question - *How accurate are the cloud native Sensitive Data Protection capabilities offered by AWS, GCP and Microsoft Azure in finding SIT (Sensitive Information Types) when configured as automated infrastructure as code and how can they be compared?*

This research is intended to compare the efficacy of the Sensitive Data Protection solutions provided by AWS, GCP and Microsoft Azure public cloud environments. Key metrics that will be assessed will include the effectiveness of pattern matching (including proximity key words where applicable) across a range of sensitive information types, custom information types using regex, and accuracy across standard file types. Where possible all configuration will be coded to replicate real life cloud scenarios using infrastructure as code which replicates CI/CD pipelines where infrastructure is built and code deployed as part of a continuous cycle.

The intention is to utilise a standard testing approach across the platforms to provide meaningful analysis of the different capabilities. It is believed that this assessment of comparative abilities should allow a greater understanding of the different protections afforded by native cloud Sensitive Data Protection and how these controls can be implemented as part of an organisational risk mitigation strategy. The research will highlight

gaps, should they exist, in coverage and accuracy which would impact control effectiveness and thus the ability to ensure regulatory compliance.

1.2 Research Objectives

As stated the research project intends to understand the effectiveness of Sensitive Data Protection functionality across public cloud vendors. This functionality will be tested through structured test cases that can be applied across all three cloud vendors. The configuration will be coded through Terraform which facilitates Infrastructure as Code (IaC) to replicate real life CI/CD (continuous integration/continuous deployment) pipelines. The results will be analysed with suggestions for further research. The structure is detailed below:

- Suggestion for a comparative framework for cross platform analysis
- Analysis of out of the box Sensitive Information Types (SITs)
- Coding infrastructure & Sensitive Data Protection function to simulate use in CI/CD pipelines
- Testing of a range of provided SITs
- Testing of customised SIT (Regex)
- Analysis of results across cloud vendors
- Results with comments and recommendations

The below sections outline why data protection constitutes a risk for organisations, the difference between DLP and Sensitive Data Protection and how this relates to organisational risk management.

1.3 Data Protection & Privacy Risk Overview

As outlined by Frawley the European Union distinguishes between data protection and privacy as two distinct yet interconnected rights (Frawley, 2023). Privacy is characterized as inherently tied to human dignity, asserting individuals' entitlement to a private life where they retain control over their personal information (EDPS, 2023). Consequently, the EU views privacy as a fundamental human right, as affirmed in the Universal Declaration of Human Rights (Article 12), the European Convention on Human Rights (Article 8), and the European Charter of Fundamental Rights (Article 7). The concept of data protection stems from the right to privacy and pertains to safeguarding data concerning identifiable individuals. This right to data protection is likewise enshrined in the Charter of Fundamental Rights (EDPS, 2023). With data protection and privacy recognized as fundamental rights, it's unsurprising that the EU has enacted legislation to enforce data protection standards for personal data handled by organizations. The GDPR (General Data Protection Regulation), implemented on May 25th, 2018, imposes significant obligations on organisations processing personal data, including the authority to impose fines of up to €20 million or 4% of global turnover, whichever is higher.

Again Frawley has outlined the global trend towards safeguarding data protection and privacy through legislative measures. Following its departure from the EU, the UK introduced its own version known as the UK GDPR, aligning with similar goals and standards (ICO, 2023). In the US, although the ADPPA (American Data Privacy Protection Act) is still awaiting implementation, several states have already enacted laws to safeguard individuals' data. Notably, the CCPA (California Consumer Privacy Act) took effect in 2018, providing Californian residents with rights similar to those outlined in the GDPR (CPRA, 2023). States such as Utah, Colorado, Virginia, and Rhode Island have followed suit, with others poised to do the same (IAPP, 2023). In Brazil, the LGPD (General Data Protection Law), enacted in 2020 consolidated forty existing laws and introduced ten legal bases for data processing surpassing the GDPR's framework and enabling fines of up to 2% of revenue. Additionally, countries including India (PDP), Canada (PIPEDA), Bahrain (DPL), and South Africa (POPIA) have enacted their own data protection legislation (Thales, 2023). UNCTAD reports that 137 out of 194 countries, representing approximately 71% of the global population, have implemented data protection and privacy laws. Gartner predicts that by the end of 2024, 75% of the world's population will be covered by modern privacy regulations (Gartner, 2022). Given the widespread adoption of data protection laws, organizations are compelled to implement rigorous data protection and privacy measures to ensure compliance (Frawley, 2023).

Organisations face further challenges, argues Frawley, when the data they own is located on infrastructure not owned by the organisation itself but is located on public cloud infrastructure (Frawley, 2023). Public cloud computing has become virtually ubiquitous and can be defined simply as ‘computing services offered by third-party providers over the public Internet’ (Azure, 2023). Frawley highlights how in a recent survey of 800 business and IT leaders globally, Accenture stated that 86% reported an increase in volume and/or scope of their cloud initiatives (Frawley, 2023, Accenture, 2023). Gartner forecast worldwide spending to reach $597B in 2023 up from $491B in 2022 (Gartner, 2023). They further predict that by 2026 75% of organisations will adopt a digital transformation approach predicated on cloud. Forbes reports that the three main public cloud

providers saw 2022 third quarter annual growth of 28% to 42% dispelling any myth that the trend towards cloud adoption was slowing down (Forbes, 2023).

Public cloud services are typically offered in three primary models. Firstly, Infrastructure as a Service (IaaS) provides users with compute, storage, and networking capabilities, while leaving the operating system and application stack under the user's control. Platform as a Service (PaaS) extends this support to include the operating system and other tools, enabling users to concentrate on their applications. Lastly, Software as a Service (SaaS) offers fully managed applications to end-users which can be utilised for various purposes. Although each model presents efficiencies and flexibility, it also means that the provider manages some or all of the technology stack. This division of control and responsibility can significantly complicate data protection efforts. Hence, it is imperative for organizations to consistently understand the data they possess in the cloud, categorize it, and then enforce policies based on that classification.

1.4 Data Loss Prevention versus Sensitive Data Protection

Data Loss Prevention is one of the main mechanisms organisations can use to achieve these requirements and offset organisational risk related to data protection and privacy. As defined by NIST (National Institute of Standards and technology) Data Loss Prevention (DLP) is the 'ability to identify, monitor, and protect data in use, data in motion, and data at rest through deep packet content inspection, contextual security analysis of transaction, within a centralized management framework' (NIST, 2023). Essentially data is identified, an assessment made on whether the data meets certain criteria, then whether any policy has or is about to be breached and finally what remediating actions should be taken if needed. There are a number of ways data can be identified including regular expression, structured data fingerprinting, partial data matching, lexicon matches, statistical analysis and categorisation. Traditional DLP solutions have operated across three channels – endpoint, e-mail and cloud. For each of these channels the DLP solution can block or quarantine any data that triggers a defined DLP policy.

For public cloud the situation is more nuanced in that the offered capability will scan cloud data stores, identify applicable sensitive information types and then generate alerts. There is, as yet, no in-line capability to automatically quarantine the data in the cloud storage or to implement additional access controls on the storage itself. This is understandable given the unknown impact to business applications using the infrastructure and is materially different to stopping an email from being sent or a file being copied to a cloud SaaS instance. Alerts allow for remediation to be executed as part of an incident response workflow so it could be argued the function does provide a remediation mechanism. It remains an open question if the data protection solutions offered by public cloud vendors qualify as a DLP solution or might be better defined as a Data Loss Detection (DLD) solutions. None of the three main Cloud vendors call their solution DLP. GCP did call its offering Cloud DLP but has since changed this to Sensitive Data Protection. AWS Macie is defined as a data security service and Microsoft Defender for Cloud includes a sensitive data threat detection module which scans and identifies sensitive information. For the purposes of this research the functionality will be termed Sensitive Data Protection which can be considered to include sensitive data detection and classification.

1.5 Sensitive Data Protection & Enterprise Risk Management

Enterprise risk management is the process organisations use to manage the many and varied risks they face. Risks are identified, an assessment is made on the likelihood of the event happening and the impact if it does. The company then makes a decision on how to treat the risk and has a number of options. The risk can be avoided through stopping the activity generating the risk, the risk can be reduced through implementing mitigating controls, the risk can be transferred or shared with another entity through say an insurance policy or finally the company can accept the risk and take no action. The decision on whether to reduce the risk is predicated on being able to estimate the effectiveness of the mitigating control. If the effectiveness is not known then no estimation can be made on how or if the risk is being reduced. In this paradigm Sensitive Data Protection is a mitigating control and organisations need to know how effective it is before they can accurately estimate their risk exposure. A failure to fully understand risk exposure undermines risk management processes and exposes the company to potential financial and reputational damage. On the other hand 'by focusing attention on risk and committing the necessary resources to control and mitigate risk, a business will protect itself from uncertainty, reduce costs and increase the likelihood of business continuity and success' (IBM, 2023).

This research aims to understand how effective the native cloud Sensitive Data Protection controls are which can in turn can inform how organisations use these capabilities as a detective/preventative control.

2 Literature Review

2.1 Literature Research Approach

A number of sources were searched for papers relating to DLP in general and Cloud DLP/Sensitive Data Protection in particular. These included Google Scholar, ACM Digital Library, IEEE.org, ScienceDirect, ResearchGate and Scoupus. The searches looked for any paper referencing Data Loss Protection, Data Loss Protection in the Cloud, Sensitive Data Protection and Data Protection & Privacy. DLP itself is relatively well researched both in terms of being a critical organisational control and also in relation to various methods and frameworks that can be deployed to implement the required pattern match and behavioural heuristics. However, there is little to no research analysing the effectiveness of native (or for that matter commercial) Sensitive Data Protection controls in the cloud (or through other DLP channels such as endpoint, mail and SaaS) in the research databases and tools. Also there is very little research on how DLP/Sensitive Data Protection could or should be tested – for instance what constitutes an agreed range of sensitive information types and across which file types.

2.2 State of the Art Overview

As mentioned above a range of research papers exist and can broadly be said to follow into three categories listed below:

- **General DLP/Sensitive Data Protection**
  Papers in this section outline how information security controls are necessary. DLP/Sensitive Data Protection is the main control discussed but the analysis goes no further and does not outline which capability or how.
- **Non Native Cloud DLP/Sensitive Data Protection**
  Papers in this category outline a range of different technical controls that either could be developed or for which proto-types have been developed. None of these solutions are commercially available, they cannot be tested and do not involve the use of native cloud capabilities.
- **Native Cloud DLP & DLP Evaluation Framework/Sensitive Data Protection**
  The two papers referenced in this section address some key issues but do not answer all of the required elements outlined in the research question. A full review of GCP capabilities was undertaken by the first paper but it made no attempt to use other vendors or to show how relevant comparisons could be made. Further the testing did not involve a CI/CD approach. The second paper discusses the necessity of a DLP framework to assess DLP capabilities focusing entirely on endpoint DLP. It overlooks other DLP channels such as public cloud and fails to specify how comparisons could be conducted or how to implement a comparison framework.

2.3 General Cloud DLP/Sensitive Data Protection

A through review of the available literature was carried out by Frawley in his 2023 paper for NCI (Frawley, 2023). In the paper, he discusses how Achar, in a 2022 paper titled "Cloud Computing Security for Multi-Cloud Service Providers: Controls and Techniques in our Modern Threat Landscape," is typical of the type of research that mentions cloud DLP in a broad sense (Achar, 2022). Regarding DLP, the paper outlines what the technology accomplishes but lacks specific details on its deployment or the effectiveness of native offerings. The overall impression is that organisations require a cloud security strategy incorporating DLP, yet it does not elaborate on how this should be done.

Frawley argues that in their conference paper from 2021, Alsuwaie, Gladyshev, and Habibnia underscore the significance of DLP for organisations, albeit in a general organisational context rather than specifically within the realm of cloud computing (Alsuwaie, Gladyshev & Habibnia, 2021). The paper provides a comprehensive overview of DLP itself and its integration with data governance, data ownership, and data classification, which are essential prerequisites for implementing a DLP solution. It covers various DLP channels, including endpoint, network, and mail, and emphasises the importance of training and awareness. The paper does present a maturity assessment that organisations can utilise to evaluate their DLP maturity and consequently their level of protection. While this framework is valuable for organisational data protection overall, it lacks specifics on DLP functionality and assessment, particularly within the area of cloud computing (Frawley. 2023).

Gupta and Singh (2022) extend the analysis of DLP further argues Frawley (Frawley, 2023). Initially outlining the rationale behind DLP necessity, including statistics and the costs associated with data breaches, they outline the implementation issues facing DLP technologies (Gupta, Singh, 2022). Various technical challenges are addressed, such as the human element, encryption and steganography, the multitude of channels and data

classification. The research becomes more interesting as they analyse different types of DLP deployment schemas and analytical techniques. The strengths and weaknesses of these methods are explored, including social and behavioural analysis, which is noted for its high false positive rate. The study examines a variety of technical papers outlining diverse approaches to DLP, along with their advantages and limitations. However, it should be noted that these models are primarily academic, lacking evidence of rigorous testing in a real-world production environment, and there's no analysis of cloud DLP (Frawley, 2023).

Finally in a conference paper for the 2022 International Conference on Cyber Warfare and Security (ICCWS) Paracha et al. does provide an overview of DLP that encompasses a number of commercial solutions (Paracha, 2022). The paper starts with an examination of DLP and where it is used – at rest, in transit and in use. The authors then articulate a DLP system that incorporates a SIEM (security incident & event management) system. They review DLP functionalities from Symantec, McAfee, Forcepoint and Digital Guardian all of which are prominent DLPS vendors. There is no detail on how the functional capabilities have been verified and the assumption would be that these have been taken form the vendor literature. Also the paper focuses on endpoint DLP with a particular use case of USB drive access. This is a standard endpoint DLP feature and not relevant to the research topic. While the review of commercial DLP solutions is commendable it would appear to be superficial and no testing strategy or framework is mentioned. The papers addressed in this section, while of some interest in providing background knowledge do not address the efficacy of native cloud DLP/Sensitive Data Protection functionalities or how this can be independently ascertained.

2.4 Non Native Cloud DLP/Sensitive Data Protection

Han et al. developed a cloud-based DLP solution using JavaScript injection techniques and deep learning methods which they assert can enforce DLP policies for organisational traffic to cloud storage (Han et al., 2020). This solution, argues Frawley, serves as an internet gateway that operates inline for users accessing cloud storage SaaS instances (Frawley, 2023). The study examines conventional DLP methods utilising predefined keywords and pattern matching, highlighting their tendency to yield false positives and an inability to detect data stored in images. The research also scrutinises CASB (Cloud Access Security Broker) technology, criticising its reliance on reverse engineering network protocols for cloud traffic, which limits scalability. However, it's important to note that many commercial DLP solutions offer OCR (optical character recognition) for image scanning in addition to pattern matching, and CASB typically utilise APIs to access SaaS platforms and enforce policies, thus not requiring inline deployment. Nevertheless, the paper presents a thorough architecture for a customised application, detailing how DLP functionalities would be implemented, including statistical analysis of the application's efficacy in identifying matching information types (Frawley, 2023).

The final two papers, highlighted by Frawley, are noteworthy for acknowledging the challenges within public IaaS & PaaS cloud platforms (Frawley, 2023). They come close to recognising the necessity of DLP/Sensitive Data Protection in public cloud services, yet neither focuses on the native capabilities offered by vendors. Sharma, Dhote, and Potey propose a managed DLP security service that is available on-demand and operates on a pay-per-use basis (Sharma, Dhote, Potey, 2016). The paper delves into specific security concerns related to data in public cloud environments and offers potential solutions. The proposed model has the service running on cloud servers that can be provisioned as needed to meet demand. While the application details are not explicitly outlined, the paper focuses on outlining the framework for the service and its potential commercialisation. While the model is intriguing, there is a missed opportunity to explore the incorporation of native cloud functionality within the paper argues Frawley (Frawley, 2023).

The last paper in this section does provide a fresh viewpoint on Cloud DLP, as highlighted by Frawley, since the authors have developed a DLP/Sensitive Data Protection application designed to operate within a CI/CD pipeline in the public cloud, specifically for scanning data at rest (Frawley, 2023). "At rest" here refers to data stored in public cloud storage, such as SQL and noSQL storage, involving both structured and unstructured data. The application also emphasizes the importance of scalability for data discovery and ensuring data security at runtime as part of CI/CD pipelines (Grunewald and Schurbert, 2022). The authors themselves assert that the application "jointly leverages Infrastructure as Code definitions, multi-paradigmatic data stores, and CI/CD pipelines in cloud-native systems to inventory personal data at rest" (Grunewald and Schurbert, 2022). This aligns closely with the stated objectives of the research question in this paper, albeit with the distinction that the authors have developed their own application, while the research question aims to leverage native capabilities. The authors present a thorough argument for data protection requirements, particularly in light of GDPR, and contend that cloud platforms are optimised for accommodating millions of concurrent users, with a shift towards microservices. They argue that the native capabilities of AWS and Google are limited and lack "algorithmic transparency." Their solution utilises vendor APIs and, from this standpoint, aligns with commercial DLP solutions that complement or replace cloud vendor capabilities with their own. The paper provides detailed

insights into the development and implementation of the framework and application, although these aspects will not be analysed here. The critical point is that this solution and the others in this section are bespoke software implementations which are not commercially available, have no software support and are not tested in a production environment (Frawley, 2023).

2.5 Native Cloud DLP Solutions/Sensitive Data Protection

Two papers are covered in this section the first of which (Diaz, 2022), although not cited, is included because it was the sole paper to analyse and evaluate native DLP functionality on public cloud platforms. Frawley argues in his 2023 paper that Diaz explores various cloud paradigms and data security requirements before delving into the assessment of Google Cloud DLP functionality (Frawley, 2023, Diaz, 2022). The paper outlines the prerequisites for effective native DLP (which Google has since rebranded as Google Sensitive Data Protection), briefly mentioning similar capabilities in other cloud platforms like Amazon Macie and Azure AIP. However, the examination of native DLP/Sensitive Data Protection from other vendors is somewhat flawed. While Google Cloud DLP and AWS Macie can be compared, Azure AIP's scope is different, focusing on data classification and security within Microsoft 365 rather than DLP/Sensitive Data Protection for structured and unstructured data in public cloud infrastructure. Additionally, the comparison is limited to a basic table listing the functionalities of each solution. Furthermore, there is no identified testing framework for comparing native DLP solutions. Despite these shortcomings, Diaz conducts tests on Google Cloud DLP functionality across various scenarios and finds it to be effective based on the outlined tests (Frawley, 2023).

The second paper, (López, Richardson, Carvajal 2015), outlines the dearth of academic research covering actual DLP evaluation and how it should be carried out. The paper reviews DLP in general but focuses on endpoint and perimeter channels only excluding cloud – this may be explainable from the age of the paper which pre-dates the escalation in cloud adoption. The paper references a number of articles by Blakely, Rabe, and Duffy who review commercial DLP solutions for the online publication NetworkWorld (Blakely, Rabe, Duffy, 2010). The reviews cover installation and functionality but do not elaborate on testing mechanisms or any rigorous comparison methodology – they are primarily high level buyers guides and also only cover endpoint and perimeter solutions. However, there is a review of DLP capabilities which was not found in any of the academic papers researched. The authors go on to articulate the need for an evaluation framework which they outline in the appendix – the framework is a serious of tests and capabilities that should be examined for Endpoint DLP. The authors do not say how the framework should be used to compare DLP functionality or how any metric for assessing the effectiveness of the DLP capability overall is generated. Also as mentioned Cloud DLP/Sensitive Data Protection is not considered. The paper makes useful references to the issue of how to assess DLP technology and the need for a framework while leaving substantial gaps particularly in reference to cloud Sensitive Data Protection.

This research aims to help close the gap on understanding what Sensitive Data Protection capabilities the three main cloud vendors currently have as opposed to reviewing generic rationales for having DLP, bespoke solutions which are not commercial in nature or single cloud analysis.

3 Research Methodology

3.1 Comparative Framework

To support a full comparison of cloud Sensitive Data Capabilities a range of technical and operation considerations need to be examined. In this regard the research by López, Richardson, Carvajal mentioned above provides a starting point even though the framework they outline is for Endpoint DLP and not Sensitive Data Protection in public cloud. The paper looked at a framework that included all supported file types across various endpoint channels for both common file formats and those considered to be evasive (encrypted, base64 or compressed). It then expanded on agent versus agentless across Windows and Linux platforms, user experience, deployment flexibility, granularity of policy templates and finally a very high level look at sensitive information categories. A number of these categories are not applicable to Cloud infrastructure but the approach is useful.

Sensitive Data Protection capabilities in the Cloud, if broken down into component parts, can be considered to consist of the below elements:

Functional Overview:

- Data storage infrastructure covered by the scanning function
- Additional public cloud security support (for instance access control, backup and restore features etc)

- Functionality available through provider portal or API
- File types supported
- Additional functionality over and above information type scanning (masking, de-identification etc)
- Reporting capability
- Alert generation
- Scanning accuracy

Sensitive Information Types:

- Categories of sensitive information types (credentials, financial, health, PII (personal identifiable information)
- Numbers of Sensitive Information Types per category
- Range of out of the box sensitive information types
- Ability to generate custom identifiers through regex
- Geographical spread of identifiers (EMEA, APAC, LATAM, India, North America)
- Confidence rating – additional signals such as proximity keywords, checksums

Usability:

- Policy configuration
- Alert information (what information do the Sensitive Data Protection alerts contain)
- Portal access and functionality in regard to alert handling (workflows)
- Vendor roadmap
- Service cost

While each cloud vendor service can be assessed on a range of these elements individually other elements are less clear. For instance to compare scanning accuracy a range of applicable SITs which are common across vendors would be needed. Issues like the requirement for keywords, which is prominent in AWS and not as clear in GCP, would need to be considered. GCP points to the use of match likelihood being based on checksums and contextual clues but it is not clear when checksums are used. Azure defines a comprehensive list of sensitive information types and indicates when a checksum is used in the pattern match – a number of sensitive information types have checksums but are not included by Microsoft presumably because the calculation method is not public.

There is some overlap in common identifiers such as driving licences, names, social security numbers medical identification numbers and financial identifiers but these information types can also be bundled into managed identifiers that are catch all's for a range of similar sensitive information types. Other information types are not clear as to the extent of their coverage – for instance GCP has an information type called 'ENCRYPTION_KEY' while AWS defines a number of private keys for OpenSSH and PGP. It becomes important to ensure that a like for like comparison is being made with information type scanning which will necessitate choices being made to facilitate comparisons. Usability and cost are also key considerations for any organisation implementing these controls and not all organisational tolerances are the same. This research is focused on the efficacy of the Sensitive Data Protection capabilities and did not examine all listed elements but all need to be considered to decide on the correct configuration and usage.

Given the research will focus only on the efficacy of the Sensitive Data Protection elements the question becomes how should this be tested across the cloud environments.

3.2 Infrastructure as Code & Sensitive Data Protection Configuration

3.2.1 Infrastructure as Code

The intention of the research was to mimic real life public cloud implementations utilising Secure Data Protection functionality through alignment with common public cloud software deployment methods. Organisations can choose a number of approaches to software development but one of the most popular is the DevOps methodology. Statista have indicated that 47% of respondents in 2022 stated that they were using a DevOps or DevSecOps approach to software devolvement making this methodology by far the most popular in their survey (Statista, 2022). DevOps centres on a CI/CD (continuous integration / continuous deployment) model where CI involves integrating all code changes into the main branch of a shared code repository where its tested before the CD element handles the deployment into the production environment. CI/CD is particularly well suited to Cloud environments since infrastructure can be built (and decommissioned) easily, code can be deployed using automated tools and the entire process can be coded using cloud APIs. DevSecOps builds on the DevOps approach to ensure that security measures are 'shifted left' in the software development process and included from the outset in configuration and deployment cycles. This means that as infrastructure is coded

security measures should be added to ensure that data protection measures are always present and configured appropriately.

Infrastructure as Code is the mechanism used to allow infrastructure deployment to be coded as part of the CI/CD pipelines thus ensuring consistent infrastructure builds each time the pipeline deploys. There are a number of IaC frameworks depending on the cloud provider but Terraform was chosen as it is widely considered to be the most popular and can be used across the AWS, Azure and GCP platforms. Terraform provides infrastructure modules for each platform and each infrastructure element can be configured through the module as required. Once the infrastructure is built the Sensitive data Protection function can also be enabled through Terraform and required policies configured. Finally Terraform provides a mechanism to load test files into the configured storage item automatically each time the code is executed. Once given the destroy command all infrastructure, services and storage content is deleted and removed from the cloud platform. All Terraform code was written and executed from Visual Studio Code. The process can be summarised in the below diagram.

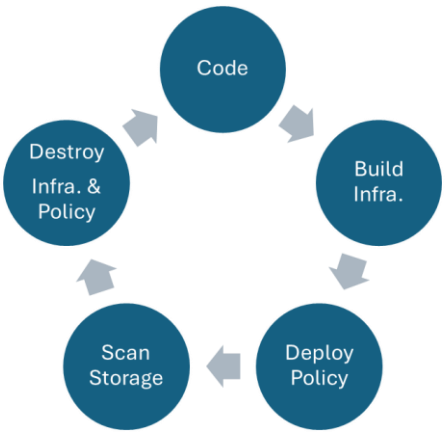

Diagram 1 – DevSecOps approach to Sensitive Data Protection testing using IaC

The following sections outline how the infrastructure and Sensitive Data Protection function were implemented through IaC for each cloud platform.

3.2.2 Azure Configuration

3.2.2.1 Terraform

Three Terraform files were written to facilitate the infrastructure and service enablement in Azure: providers.tf, main.tf and dlp.tf. Providers.tf defines the Cloud API that Terraform should use to connect to the public cloud instance. For Azure a command line login was used to authenticate to the Azure subscription which was created separately through the Azure portal. The main.tf file contained the code needed to create the resource group, virtual network, subnet, storage account, storage network rules (restrictions on access to storage), storage container and lastly the blob storage. Azure resource groups act as a container under which all associated infrastructure elements sit. Storage accounts are then defined to hold all storage objects in this instance storage containers which hold blob storage instances. The reason for selecting blob storage will be expanded on in the next section. The blob storage configuration included code to automatically upload all files from a named directory on creation which provided an automatic means to upload test objects. A separate dlp.tf file was created to enable the Microsoft Defender for Storage and Microsoft Defender for Cloud Security Posture Management which are the modules under Microsoft Defender for Cloud that provide the sensitive data threat protection capabilities.

3.2.2.2 Microsoft Defender for Cloud

Microsoft offers a wide range of data protection capabilities across its estate covering both Azure infrastructure and DLP functionality across M365 and external SaaS. It is important to note that this space is fluid and Microsoft are constantly making changes and releasing new functionality changing how Secure Data Protection functionality works and interacts with other MS features. Microsoft Defender for Cloud according to Microsoft is ‘a cloud-native application protection platform (CNAPP) with a set of security measures and practices designed to protect cloud-based applications from various cyber threats and vulnerabilities’ (Microsoft, 2023). Microsoft’s Secure Data Protection features for cloud infrastructure are contained in the Sensitive Data Threat

Detection function which itself is embedded in the Defender for Cloud Storage and Defender Cloud Security Posture Management modules which sit under Microsoft Defender for Cloud. Azure users have the ability to select various Defender for Cloud modules depending on their requirements. It is important to note that Microsoft Purview provides DLP/Secure Data Protection for the M365 estate and at the time of testing both functions are accessed separately.

Several important considerations should be noted on Defender for Cloud - the sensitive information types available to Defender for Cloud are a subset of those available through Purview. Custom sensitive information types (custom pattern matching using regex) defined in Purview could be imported into Defender for Cloud as could sensitivity labels defined in Purview. These features were not tested as the testing executed was on a standalone Azure subscription using out of the box sensitive information types for testing without a separate M365 account. Further at the time of testing the Sensitive Data Threat Detection function was not GA (general availability) and was in public preview thus the full feature was not available for testing. Given Sensitive Data Threat Detection is relatively new the actions available through Terraform were limited – the Defender for Storage and Defender for Cloud Security Posture Management modules could be enabled thus enabling the Sensitive Data Threat Detection features. However, there was no ability to create specific policies based on sensitive information types instead all sensitive information types were enabled through the portal.

In relation to the types of infrastructure supported - Sensitive Data Threat Detection is valid for Azure Blob storage accounts and also for Azure Data Lake storage which is built upon Blob storage. For the purposes of this testing only Blob storage accounts were tested through Sensitive Data Threat Detection which scans for a defined set of sensitive information types listed in the portal.

3.2.3 GCP Configuration

3.2.3.1 Terraform

Again for GCP configuration three Terraform files were created, as with Azure the providers.tf file defined the GCP API, GCP project (container for all defined infrastructure) and also the authentication details in a locally saved json file. The main.tf file contained the Terraform configuration to create a storage bucket for testing, a storage bucket for logging, a storage object which allows uploading of files from the named directory and a BigQuery dataset and table. BigQuery is defined as a data warehouse which allows SQL queries to process the data and in this content is a repository for the output of the Sensitive Data Protection policies which can then be searched as required. The dlp.tf file contains the enablement of the GCP data_loss_prevention_inspection function and also contains the specific sensitive information types that are applied (these can be applied individually or as a group). This means that each time the function is enabled specific sensitive information types can be included. GCP offers a number of options in relation to matching accuracy ranging from very unlikely to very likely – this is predicated on the confidence the platform has that the pattern match is correct. For this testing only likely (one or more strong signals for a given infoType) was used. Lastly in the data_loss_prevention_inspection template an arbitrary limit of 1000 was set for the number of findings per request. GCP requires a job trigger to be defined which stipulates how often the scan should run (time in seconds), the dataset to send the results to, the internal URL of the storage bucket and finally the applicable file types.

3.2.3.2 GCP Sensitive Data Protection (Previously GCP Data Loss Prevention)

GCP Sensitive Data Protection has three modes of operation - content, storage and hybrid. Both the content and hybrid options use the available DLP API to send different workloads to the Sensitive Data Protection function. For the content mode the results are retuned over the API but for hybrid they are sent to a dataset. The third mode of operation, and the one selected for testing, scans GCP storage types such as Cloud Storage, Big Query or Datastore directly for sensitive information types. GCP Cloud Storage was selected for this research given it provides an unstructured data repository in the form of a storage bucket providing a perfect mechanism to test scanning of unstructured data.

GCP Sensitive Data Protection provides three functions – sensitive data discovery, sensitive data inspection and sensitive data de-identification. Discovery creates profiles for data in BigQuery databases and reports these at the project, table and column levels. Inspection performs a deep scan of storage instances for sensitive information types and generates a report into every match. The final function, de-identification, allows the obfuscating of sensitive information once found through a number of methods such as masking, redaction and tokenisation. This research focused on the Inspection functionality which allowed the deep inspection of storage buckets for specific sensitive information types both out of the both and custom.

3.2.4 AWS Configuration

3.2.4.1 Terraform

As with GCP and Azure three Terraform files were created for the AWS configuration. Providers.tf contained the API info for AWS and also the directory of the credentials file with the secret access key and the AWS account identifier. The main.tf file contains the configuration of the VPC (virtual private cloud) under which all infrastructure sits, the subnet, S3 bucket definition and finally S3 object which allows files to be automatically uploaded to the storage. The dlp.tf file enables the Macie function which is the name of the Sensitive Data Protection function in AWS. A Macie job is also configured with the frequency of the scans (one-time, daily, weekly, monthly) and the bucket to be scanned. It should be noted that the configuration through Terraform can not specify specific sensitive information types and that the Macie configuration uses the default set of sensitive information types assigned in Macie. To add additional information types the infrastructure is still defined with terraform but the Macie job must be configured through he portal where specific sensitive information types not part of the default group can be added.

3.2.4.2 AWS Macie

Macie is described as a data security service and focuses on providing information on the security posture of Amazon Simple Storage Service (S3) instances. This can include monitoring access control to the buckets but critically also includes discovery and reporting of sensitive data within the S3 estate. It should be noted that Macie covers S3 only and no other storage is included. Alerts generated can be sent to AWS EventBridge or to AWS Security Hub where they can be actioned either manually or as part of some automated process. Macie has a number of out of the box sensitive information types and also allows configuration of customer identifiers through regex.

The following sections outline which sensitive information types were tested and how the tests were executed for each cloud platform.

3.3 Testing Methodology

3.3.1 Overview

The tests were targeted to identify if the Sensitive Data Protection capabilities could find sensitive data types located in files within storage – how accurate are the scans and are there differences between cloud vendor scanning capabilities. In order to test this a variety of sensitive information types were selected to represent the widest spread of the main sensitive information types and where possible geographic locations. It should be noted that a limitation exists in relation to sourcing publicly available test data. For instance it is difficult to find accurate sensitive information types in the health category. The concern becomes whether the testing will be carried out with data that is not representative and is not matched, correctly, by the scanning. A range of publicly accessible data loss protection sites listing sensitive information types for testing were used. The test data was validated against publicly available information on the format and with the Azure definitions which provide detail on each sensitive information type. Where used and where information was publicly available the checksum was also checked and validated. The sites used are listed with each sensitive information category below. In total forty five separate sensitive information types were tested using a range of common and supported file types.

3.3.2 Sensitive Information Types

SIT Category	**Individual SITs**
Financial (5)	IBAN GB, IBAN ES, IBAN DE, IBAN IT *(DLPTest, 2023, Nuapay Developer Hub, 2023)*
Social Security Numbers (2)	SSN US, SSN ES *(DLPTest, 2023, Feria de la Ciencia, 2012)*
Personal Information (3)	Name, DOB, Email *(DLPTest, 2023)*
Driving License (7)	Driving License US, Driving License UK, Driving License Australia, Driving License Japan, Driving License Italy, Driving License France, Driving License Spain *(Trellix, 2022)*
National Identification Number (9)	NIN India (AADHAAR), NIN France (CNI), NIN Brazil (CPF), NIN Brazil (RG), NIN Italy (Fiscal Code), NIN Indonesia (NIK), NIN Croatia (OIB), NIN China (Resident ID No), NIN Chile (CDI) *(BFI Finance, 2023, Cisco, 2022, Protecto, 2023, Trellix, 2022)*

Passports (5)	Passport UK, Passport US, Passport France, Passport Germany, Passport Canada *(Protecto, 2023)*
Tax Identification Number (5)	Tax No Australia, Tax No Brazil (CPF), Tax No Germany, Tax No India, UK National Insurance Number *(Protecto, 2023, Trellix, 2022)*
Credentials (7)	AWS secret key, Putty private key, OpenSSH private Key, PGP private key, Encryption Key, HTTP Basic Authentication header, Jason Web Token (JWT)
Vehicle Number (1)	Vehicle Identification Number *(MyVehicle, 2017)*
Customer Identifiers (1)	Custom 1

3.3.3 Testing Procedure

To test the Sensitive Data Protection scanning of each cloud vendor a number of standard file types where used - .docx, .xlsx, .csv, .pdf, .pptx and .txt. The file types chosen are popular file types, commonly in use and also those which would generally hold the sensitive information types chosen. The sensitive information types were divided into ten categories as outlined in section 4.3.2 and each of the files contained the same category of sensitive information. It is key to note that the same files with the same sensitive information were used across all three vendors – the data was exactly the same in each case with no modification. The only exception to this was for AWS where .pptx file types were not supported. It should also be noted that no attempt was made to obfuscate the information or to confuse the scanning every attempt was made to ensure that scanning requirements were met.

3.3.3.1 AWS

All tests used a standard S3 bucket into which the five files where automatically uploaded. A Macie job was configured to run a 'one time' job in order for the scanning to initiate immediately – this was purely to facilitate testing. As also mentioned Terraform does not facilitate defining individual sensitive information types in the Macie job profile. The job instead takes the default information types enabled and since this did not cover all sensitive information types needed some of the testing required the Macie job to be configured from the portal. All other infrastructure configuration was still through Terraform. Results were checked in the AWS portal via Macie.

For the vast majority of sensitive information types AWS requires keywords to be in close proximity (names, IBAN and credit card are exceptions). In relation to scanning structured data in the form of excel and csv required keywords must be in the field or in the name of the column. For unstructured data such as pdf, word, txt etc the keyword must be within 30 characters. All test files ensured that these conditions were met. Once the test was run the results were examined through the AWS Macie portal where findings are displayed for all Macie jobs.

3.3.3.2 GCP

All tests used a standard GCP cloud storage bucket into which six files were automatically uploaded. An inspection template was configured to initialise the scanning and which allowed specific sensitive information types to be configured each time. All results were sent to a GCP BigQuery dataset where they can be examined immediately or exported to Google Sheets for easy of analysis. GCP uses a likelihood measure to indicate how strong the pattern match should be and whether it should contain contextual clues. These clues would include proximity key words but there are no clear explanation on how or if keywords are used or how they should be defined in the files. A likelihood match of 'likely' was selected which includes contextual clues. The files used for GCP were identical to those used in AWS including placement of key words.

3.3.3.3 Azure

All tests used standard Blob storage bucket into which six files were automatically uploaded. There is no mechanism through Terraform to configure which sensitive information types should be scanned so all sensitive information types were activated through the Azure portal. A key finding of this research, which will be discussed further in the results section, is that the Azure Sensitive Data Protection function is materially different from either GCP or AWS and testing could not be completed in the same way. Each test case used multiple sensitive information type categories supported by Defender for Cloud - Sensitive Data Threat Detection. Microsoft Purview does list all sensitive information types and whether proximity keywords and checksums are applicable (Defender for Cloud uses a subset of these). In Purview this aligns to a confidence level which can be defined in policy. This ability is not present in Defender for Cloud and thus it was not clear

to what extent proximity keywords and checksums were used in the pattern matching. For all applicable test cases the same test files used with GCP and AWS were used with Azure.

4 Results & Discussion

4.1 Public Cloud Sensitive Information Type Overview

The research reviewed all sensitive information types (SIT) available through GCP Sensitive Data Protection, AWS Macie and Microsoft Defender for Cloud. Between the three vendors they covered 336 separate individual sensitive information types. With the exception of IBAN, managed sensitive information types that group similar individual information types together where excluded from the analysis. Figure 1 illustrates the geographical distribution of the SITs across the three vendors and as can be seen the predominant region is EMEA which has 153 (45%) SITs unique to the region. It should be noted that all but three of the SITs are from Europe with two from the middle east and one from Africa. APAC & North America represent 12 % and 9% respectively with LATAM and India on 4% and 1%. The remainder are classed as global and not tied to any geographical location.

With regard to the breakdown of SITs by category, illustrated in figure 2, the majority can be tagged as PII (personal identifiable information) which covers a multitude of identifiers from passports, driving licenses, tax numbers etc and represents 65% of the total. IT data types constitute 20% and encompass a wide range of technical identifiers with the majority being authentication credentials. Health and finance identifiers both represent 6% of the total with the final category 'other' containing miscellaneous identifiers not fitting into the main categories such as company/business numbers, dates etc and making up the remaining 3%.

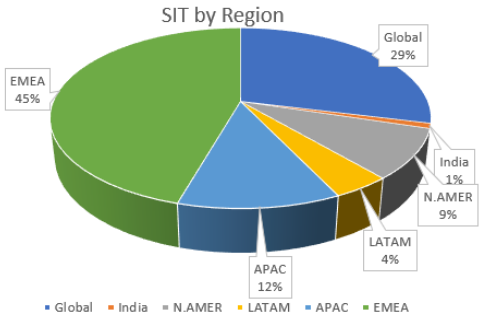

Figure 1 – Geographical breakdown of SITs

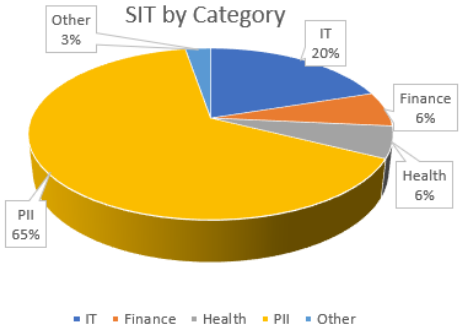

Figure 2 – SITS segregated by category

Figure 3 illustrates the breakdown between the cloud vendors themselves with Azure having the largest number of SITs and 44% of the total. GCP is next with 33% and finally AWS with 23%.

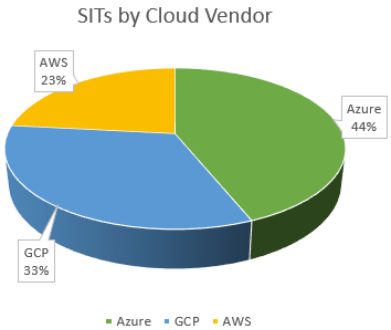

Figure 3 – SITs by Cloud Vendor

4.2 Results of SIT Testing

4.2.1 Overview

A full suite of test cases were run against AWS and GCP infrastructure and the results of this testing will be examined below. However, Azure did not provide the same level of functionality and could not be tested in the same manner leaving the Azure testing inconclusive. Azure will be discussed separately. A test was deemed successful once the SIT was correctly identified with no false positives. Each scan of an individual file type per individual SIT was considered a single test. The below table gives a high level view of the testing results for GCP and AWS.

Vendor	Test Cases Total	Pass	Partial Pass	Fail	Pass Rate %
AWS	181	126	6	49	**72%**
GCP	229	137	6	86	**62%***

Table 1 – Results of all test cases for GCP and AWS

SIT Category	AWS Accuracy	GCP Accuracy
Financial	100%	100%
Social Security Numbers	100%	50%
PII	100%	94%
Driving License	66%	36%
National Identification Numbers	60%	58%
Tax Identification Numbers	100%	83%
Passports	64%	63%
Credentials	60%	29%
Vehicle Identification Number	0%	100%
Custom SIT	100%	0%

Table 2 – Results of testing per SIT category for GCP and AWS (accuracy applied to supported SITs only)

The difference in the number of test cases run between the vendors is due to two reasons, firstly, AWS did not support eight of the selected forty five SITs two more than GCP. GCP did not support six of the SITs (all different from those not supported by AWS) and, secondly, GCP supports all six file types selected whereas AWS Macie does not support .pptx and thus for each SIT AWS has one less test. Partial pass indicated that the SIT was found but that the scan did not find all instances in the file – this was included in the pass rate given the correct trigger on one of the SITs is enough to generate an alert and thus remediation action. However, if the scan incorrectly tagged data as an information type (false positive) then the test was deemed to have failed. The AWS documentation states that keywords are to be 'within 30 characters (inclusively) of the data' but does not make it clear they should be before the SIT although testing indicated this to be the case (not a limitation for GCP) (AWS, 2023). Both GCP and AWS require keywords for structured file types (.xlsx, .csv) to be in the field or column header.

*When a like for like comparison is made for information types and file types that are supported by both vendors the respective pass rate for AWS and GCP respectively is 75% and 64%.

4.2.2 Financial

Two separate tests for financial sensitive types were conducted:

- Credit card numbers – five in each file with keywords (Visa, MC, Mastercard, Amex) in the column headers and within 30 characters of the numbers.
- IBANs seven in each file (UK x 2, ES x 2, DE x 2, IT x 1). AWS does not require proximity key words for IBAN although they are present but not within 30 characters for all numbers.

Result: Both GCP and AWS had a 100% success rate for both credit cards and IBANs across all file types

4.2.3 Social Security Numbers

American and Spanish social security numbers were used in each file with key words. The US SSN keyword was within 30 characters of the first number and then not repeated. The ES SSN had the keyword withing 30 characters of the numbers.

- GCP found all US SSNs in .xlsx and .csv files and found two of the ten in .doc, .pdf, .pptx and .txt
- AWS found all US SSNs in .xlsx and .csv files and found one of the ten in .doc, .pdf, .pptx and .txt

Result: Both GCP and AWS had a 100% success rate for US SSN given scanning boundaries.
The results suggest that GCP uses proximity key words for a confidence rating of 'likely' and that the key word search extends past 30 characters.

- GCP did not trigger on any ES SSN
- AWS had a 100% success rate for ES SSN

Result: GCP 0% pass rate for ES SSN and AWS had a 100% success rate for ES SSN.
For ES SSN the numbers where taken from publicly provided test data so it should be left open that the numbers are incorrect. However, this would mean that at least one of the vendors is scanning incorrectly given they trigger for AWS but not for GCP. It should be noted that in AWS testing the key word 'número de la seguridad social' was ineffective and only 'social security number' triggered despite both being given as valid keywords.

4.2.4 Personal Information

Full names, dates of birth and email addresses were used across all file types. AWS does not support email addresses so only name and date of birth were applicable. AWS also does not require keywords for names but does for data of birth (dob was used).

- AWS triggered with 100% accuracy for dob and names once keywords are before the sensitive information type.
- GCP executed successfully with the exception of the .csv which returned 92 names despite the file containing 50 – the data was exactly the same as for the other file types and the sensitive data identifier used in GCP was for full names. This was considered a fail.

Result: AWS 100% pass rate for names & DOB and GCP had 100% pass rate for DOB but 83% for names.
A further test for emails was executed for GCP which was successful.
Result: GCP 100% pass rate for emails.

4.2.5 Driving Licenses

Driving license numbers for the US, UK, AU, JP, IT, FR and ES were tested with separate keywords added for each license instance.

- GCP does not support identifiers for Italian or French driving licenses. The results for US licenses were mixed with the scan appearing to pick up every instance of the applicable key work even if not associated with a US driving license number. Four instances of US driving licenses were present but the scan found between 7 and 14 with only the .csv file correct these tests were deemed to have failed. The scans for UK and Australia were correct with the exception of the .csv file which did not trigger despite having the same data. The Japanese and Spanish driving license scans failed completely.
- AWS scanning was similarly mixed the .doc, .txt and .pdf files had 2 US licenses from 4 but did not over estimate the numbers. The .xlsx and .csv files tagged 9 and were deemed fails. The UK scan was fully successful and the Australia, Italy and French scans were also successful bar the .csv scan which failed. The Spanish scan also failed across all file types.

It should be noted that the Japanese and Spanish driving license numbers could not be verified by either vendor and further validation of these numbers would be required.
Result: GCP had a 36% success rate for driving license identifiers tested. AWS had a 66% pass rate for driving license identifiers tested.

4.2.6 National Identification Numbers

National Identification Numbers were tested for IN, FR, BR, IT, ID, HR, CN and CL. It should be noted that the identifiers come from public sources and could not be verified. Further a number of them could only be tested against one vendor. All applicable keywords were used.

- AWS had a 60% success rate against the identifiers tested which excluded Indonesia, Croatia, China and Chile. The Italian and Franch identifiers failed completely while the Indian and Brazil identifiers were fully successful.
- GCP had a 58% success rate against Indian, Franch, Brazilian, Italian, Indonesian, Croatian, Chinese and Chilian identifiers. The Croatian identifier did not trigger for .xlsx or .csv despite the same keyword (OIB) being in the column header. The Indonesian and Chinese identifiers also failed.

Result: GCP had a 58% success rate for national identity number identifiers tested. AWS had a 60% pass rate for same.

4.2.7 Passports

Passport numbers from UK, US, DE, FR and CA were tested with specific keywords per country.

- AWS results were inconsistent as for instance the UK identifier only triggered for .xlsx and .csv despite the same keyword and passport number being in all other file types. Canadian and US identifiers passed successfully while the French identifier failed only with .pdf. Finally the German identifier failed completely.
- GCP results were similarly inconsistent – the identifiers for UK, US, Germany and Canada all passed with the exception of .xlsx and .csv file types. France also failed for .pdf.

All passport numbers passed in one or both cloud platforms and the failures for .xlsx and .csv in GCP are inexplicable given the key words were all present in the column header. In a similar fashion the failures in AWS cannot be explained by a difference in the passport numbers or key words.

Result: GCP had a 63% pass rate for passports while AWS had a 64% pass rate on identifiers tested.

4.2.8 Tax Identification Numbers

Tax identification numbers for DE, BR, AU and IN were tested again with all specific keywords added.

- AWS had a 100% success rate for all tax identification numbers – AWS does not cover Indian PAN number.
- GCP had a 100% success rate for Australia and India but did not trigger at all for German tax number and only for .xlsx and .csv for Brazilian CPF number.

It should be stressed again that all test files are the same for both GCP and Azure.

Result: GCP had a 62% pass rate for tax identification numbers while AWS had a 100% pass rate on the same identifiers.

The UK National Insurance Number was also tested separately from the above tax identification numbers. This passed completely for AWS and also for GCP with the exception of .csv.

Result: GCP had a 83% pass rate for UK National Insurance Number AWS had a 100% pass rate on the same identifiers.

4.2.9 Credentials

A number of credentials were tested and all were generated for the testing and were legitimate data types. Of the seven chosen four were supported by GCP and six supported by AWS. The seven tested are AWS secret key, putty private key, Open SSH private key, PGP private key, Encryption key, HTTP Basic Auth Header and Jason Web Token.

- GCP successfully triggered on JSON Web Token for all but the .pdf file type, the encryption key information type triggered for .docx and .pptx while AWS secrets and HTTP Basic Auth Header both failed for all file types.
- For AWS HTTP Basic Auth header triggered for all file types as did the AWS secret key. OpenSSH triggered for .docx, .csv and .txt. PGP private key triggered for .docx and .txt. Finally JSON Web Token did not trigger for any file type.

All data for all sensitive information types was identical across all files and the same files were tested for both vendors.

Result: GCP had a 29% pass rate for Credentials AWS had a 60% pass rate on the same identifiers.

4.2.10 Vehicle Number

A single VIN number was tested for both platforms the test was unsuccessful for any file type for AWS and fully successful for GCP.

Result: GCP had a 100% pass rate for Vehicle Identification Number AWS had a 0% pass rate on the same identifier.

4.2.11 Custom SIT

Both GCP and AWS allow custom sensitive data identifiers to be created to match specific requirements users may have. These identifiers can be created with regex and used in scanning against files. For GCP the regex can be added to the identifiers defined in the Terraform code per inspection template. For GCP the regex is created in the AWS portal and added to Macie jobs when created from the portal. The regex pattern used is: ^C.+:\s[0-9]{6}.[0-9]{3} – the pattern looks for test starting with 'C' and then any characters to ':' it will then look for 6 digits then any character and finally another three digits. This regex is intended to catch test such as 'Contract No: 456897/123'. This pattern was placed in a .docx file and the scan run. For AWS this was successful but not so for GCP. The reason is not clear as the regex matches the expressions given by GCP (GCP,2023).

Result: GCP had a 0% pass rate for Custom Identifier AWS had a 100% pass rate on the same identifier.

4.3 Azure

The intention of the research was always to compare the Sensitive Data Protection capabilities of all three of the main cloud providers. However, the functionality as implemented by Azure did not allow test cases to be executed in a structured fashion. The Azure feature Sensitive Data Threat Detection is embedded in two of the Defender for Cloud modules (Defender for Cloud Security Posture and Defender for Storage). The first issue is that the feature was not released at the time of testing and was in public preview – thus it was unclear if fully functioning. The documentation states that once enabled results are generated within 24 hours and then it scans once a week. Even this cadence seemed to be incorrect and the scanning could only be triggered when the access level of the storage was changed to allow anonymous access (reflected in the alerts generated which highlight this access change). The second issue is that when it did scan it appeared not to scan all objects in the storage but a subset. The Microsoft documentation does state that 'Sensitive data threat detection is powered by the sensitive data discovery engine, an agentless engine that uses a smart sampling method to find resources with sensitive data' (Microsoft, 2023).

There is no way to configure the feature it is simply enabled when either of the two Defender for Cloud modules are enabled. The portal allows you to choose the sensitive information types that are scanned and there are options to import custom identifiers and sensitivity labels from Purview but nothing else. The net result is that it is impossible to run a series of structured test cases – the Terraform code can turn the modules on and this enables the Sensitive Data Threat Detection feature but no specific test cases can be configured. It may be that the function is intended to simply highlight that sensitive information is in the storage container and nothing more.

In order to try and test the function all information types were placed in the Blob storage and a number of configuration changes made to get the scan to trigger over numerous attempts. While haphazard the storage did scan and the identified sensitive information types are listed below (the file types they were found in were also listed in the portal):
Passports: correctly identified - CA, FR, US and UK. Incorrectly identified – HR, FI, UA, HU, SK, BE, TW, RU, BG, IE, DK, IT & RO.
Finance: correctly identified - IBAN & Credit Cards. Incorrectly identified EU Debit Cards
Drivers License: correctly identified – UK, JP, IT, FR & ES. Incorrectly identified – SK, RO, CY, MT.

None of the national identity numbers, social security numbers, credentials or tax identification numbers triggered. Given the number of false positives, which outweigh the actual legitimate numbers in the files, it can be concluded that the lowest confidence rating is being used. It also would appear that the storage resource is not fully scanned and a partial scan returned. Microsoft intend implementing data mapping capability into Purview in future releases and it may be that this functionality provides a more like for like comparison with the Sensitive Data Protection capabilities of GCP and AWS. For the moment it is assumed that Azure Sensitive Data Protection provides a general warning as the presence of sensitive data in storage under certain conditions. For the purposes of this research Azure Sensitive Data Protection will be excluded.

5 Conclusion & Future Work

While not one of the main focus areas of the research a number of interesting findings relate to the number and type of sensitive information types the vendors offered. Of the 336 individual information types covered by all three vendors only 33 were covered jointly by the three – representing common agreement on just 9% of the total number of information types. Further if the information types considered global are removed then 64% of the remainder are specific to EMEA (16% APAC, 13% N.America, 5% LATAM, 1% India). The regional basis towards EMEA is clear with other regions under represented and the Middle East and Africa particularly ignored. This leads to the question – is there a nominal distribution of information types needed to offer what would be termed a comprehensive offering? Clearly for some entities in particular regions and with particular data types the offered functionality may not be sufficient. The option to use custom identifiers remains but this also has a high overhead in that these identifiers will need to be created, tested and maintained.

The research wanted to utilise the Sensitive Data Protection capabilities of the three vendors by using a DevOps approach. Terraform was used to configure all infrastructure and DLP policies. For the most part this was successful and validated that all three vendors supported DevSecOps and that data protection can be embedded into DevOps workflows from inception. The only caveats where that only default sensitive information types (SITs) could be enabled through Terraform for AWS and that Azure did not allow specific SITs to be selected. In the main this can be deemed a success.

The main body of the research tested the efficacy of the Sensitive Data Protection capabilities of all three and here conclusions are mixed. Structured testing was not possible for Azure, the function should scan for all new storage but only ever triggered when storage access was changed. Further it missed a wide range of SITs and gave false positives for a range of others which neither AWS or GCP did. When it did trigger it did not list the files the SITs triggered in but did give the file types. The Azure functionality could not be deemed sufficient to act as a Sensitive Data Protection control and could be considered of limited value for Sensitive Data Protection. For GCP and AWS the conclusions are more nuanced. GCP provided a wider range of SITs and greater functionality through the Terraform configuration but had a much lower success rate. With a 62% success rate organisations need to decide if this represents a sufficient level of accuracy to meet their risk appetite. For some it will be too low and not sufficient for a stand alone control. AWS, which covered less SITs and had limitations in what can be configured through Terraform, had a 72% success rate which is considerably better. Both platforms cover a number of SITs with 100% accuracy and if limited to these would prove a strong stand alone control. There is no formal conclusion to the research other than as it stands Azure is clearly not sufficient for a stand alone Sensitive Data Protection control while GCP and AWS may be acceptable depending on risk appetite and SIT dependency. No Sensitive Data Protection control can claim to be 100% all of the time but it is up to organisations to define their requirements and what they believe is acceptable – in order to do this they need clarity on the controls available and it is this the research has tried to contribute to.

As noted in the Test Methodology section the test data leveraged publicly available sources. This is clearly a gap in the research in that real data was not used and for those tests that failed it could be legitimately argued that the issue is with the data. As a counter argument only one of the 45 SITs failed to trigger across all three vendors or two if we restrict the analysis to GCP and AWS. Ideally future research would validate all SITs against real word data to verify the results of this research and the pass rates. In addition all three vendors offer a range of controls that may fit into a wider security framework which could be designed and tested. For instance AWS covers the security posture of S3 buckets and GCP offers mask and de-identification features. What sequence of controls and infrastructure offers the best security for cloud deployments given the a myriad options and configurations? Software could also be included as a critical component in the DevOps pipelines, how are images and repositories secured, how is application code checked and what additional security controls are required. How secure are the controls is this space? Finally the area of what is acceptable risk – given the legislative requirements on organisation holding sensitive data, what constitutes an acceptable level of risk?

References

Frawley, C. (2023) *Efficacy of native public cloud DLP capabilities* National College of Ireland. Unpublished essay.

European Data Protection Supervisor (2023) *European Data Protection Supervisor.* Available at: https://edps.europa.eu/data-protection/data-protection_en [Accessed: 13 July 2023].

ICO (2023) *About the Guide to Data Protection* Available at: https://ico.org.uk/for-organisations/ [Accessed: 13 July 2023]

CPRA (2023) *California Privacy Rights Act* Available at: https://www.caprivacy.org/cpra-resource-center/ [Accessed: 15 July 2023]

IAPP (2023) *US State Privacy Legislation Tracker* Available at: https://iapp.org/resources/article/us-state-privacy-legislation-tracker/ [Accessed 15 July 2023]

Thales (2023) *Beyond GDPR: Data Protection Around The World* Available at: https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/magazine/beyond-gdpr-data-protection-around-world [Accessed 15 July 2023]

UNCTAD (2023) *Data Protection and Privacy Legislation Worldwide*. Available at: https://bit.ly/3pGuVYf [Accessed: 13 July 2023].

Gartner (2022) *Gartner identifies top five trends in privacy through 2024 (2022) Gartner*. Available at: https://bit.ly/46ODqB1 [Accessed: 12 July 2023].

Azure (2023) *What is a public cloud?* Available at: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-public-cloud [Accessed 24 July 2023]

Accenture (2023) *The race to cloud: Reaching the inflection point to long sought value* Available at: https://www.accenture.com/ie-en/insights/cloud/cloud-outcomes-perspective [Accessed 24 July 2023]

Gartner (2023) *Worldwide Public Cloud End-User Spending to Reach Nearly $600 Billion in 2023* Available at: https://www.gartner.com/en/newsroom/press-releases/2023-04-19-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-nearly-600-billion-in-2023 [Accessed 24 July 2023]

Forbes (2023) *The Cloud Is Still a Multibillion-Dollar Opportunity. Here's Why* Available at: https://www.forbes.com/sites/glennsolomon/2023/01/04/the-cloud-is-still-a-multibillion-dollar-opportunity-heres-why/ [Accessed 24 July 2023]

NIST (2023) *Data Loss Prevention - Glossary: CSRC, CSRC Content Editor*. Available at: https://csrc.nist.gov/glossary/term/data_loss_prevention [Accessed: 10 July 2023].

IBM (2023) *Why is risk management important?* Available at: https://www.ibm.com/topics/risk-management (Accessed: 1 August 2023).

Anchar (2022) *Cloud Computing Security for Multi-Cloud Service Providers: Controls and Techniques in our Modern Threat Landscape*, World Academy of Science, Engineering and Technology International Journal of Computer and Systems Engineering, Vol:16, No:9, 2022 (Cited 20 times Google Scholar)

Alsuwaie, M. Gladyshev, P. and Habibnia, B. (2021) *Data Leakage Prevention Adoption Model & DLP Maturity Level Assessment* in 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC). Rome, Italy 12-14 November 2021IEEE Xplore DOI: 10.1109/ISCSIC54682.2021.00077 (Cited 3 times Google Scholar)

Gupta, I. Singh, A. (2022) *A Holistic View On Data Protection For Sharing, Communicating, And Computing Environments: Taxonomy And Future Directions*, arXiv:2202.11965 https://DOI.org/10.48550/arXiv.2202.11965 (Cited 16 times Google Scholar)

Paracha, M et al. (2022*) Implementation of Two Layered DLP Strategies* in 2022 International Conference on Cyber Warfare and Security (ICCWS). Islamabad, Pakistan 7-8 December DOI: 10.1109/ICCWS56285.2022.9998436

Han, P. Liu, C. Cao, J. Duan, S. Pan, H. Cao, Z. Fang, B. (2020) *CloudDLP: Transparent and Scalable Data Sanitization for Browser-Based Cloud Storage*, IEEEAccess DOI: 10.1109/ACCESS.2020.2985870 (Cited 11 times in Google Scholar)

Ahmed, A. Haq, A. Sheeraz, M and Durad, M. (2022) *Design and Development of Cloud based QR Coded Watermarking DLP system* in 2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST). Islamabad, Pakistan 16-20 August. DOI: 10.1109/IBCAST54850.2022.9990327 (peer reviewed in Discovery)

Sharma, D. Dhote, C. and Potey, M. (2016) *Managed Data Loss Prevention Security Service In Cloud* in 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences (EEECOS 2016). Tadepalligudem, India 1-2 June 2016 IEEE Xplore DOI: 10.1049/cp.2016.1502 (Cited 3 times Google Scholar)

Grunewald, E. and Schurbert, L. (2022) *Scalable Discovery and Continuous Inventory of Personal Data at Rest in Cloud Native Systems* in 20th International Conference on Service-Oriented Computing (ICSOC 2022). Seville, Spain 29 Nov – 2 Dec 2022. https://doi.org/10.1007/978-3-031-20984-0 (Core Conference Rank A)

Diaz, B. (2022) *Deployment Of a Lab Environment To Identify And Protect Sensitive Data In The Cloud*". BSc Thesis. Barcelona, Spain. Universitat Politecnica De Catalunya.

López, G. Richardson, N. Carvajal, J.(2015) *Methodology for Data Loss Prevention Technology Evaluation for Protecting Sensitive Information* in Vol. 36 Núm. 3 (2015): Revista Politécnica 30th September 2015 https://revistapolitecnica.epn.edu.ec/ojs2/index.php/revista_politecnica2/article/view/582

Networkworld (2010) *Data loss prevention comes of age* Available at: https://www.networkworld.com/article/2205740/data-loss-prevention-comes-of-age.html [Accessed 10 August 2023]

Statista (2022) *Breakdown of software development methodologies practiced worldwide in 2022* Available at: https://www.statista.com/statistics/1233917/software-development-methodologies-practiced/ [Accessed 18 November 2023]

Microsoft (2023) *What is Microsoft Defender for Cloud?* Available at: https://learn.microsoft.com/en-us/azure/defender-for-cloud/defender-for-cloud-introduction [Accessed 18 November 2023]

DLPTEST (2023) *Test Data* Available at: https://dlptest.com/sample-data/ [Accessed 18 November 2023]

NuaPay Developer Hub (2023) *Test IBANs* https://developer.nuapay.com/np_testibans.html [Accessed 18 November 2023]

Feria de la Ciencia (2012) *DÍGITOS DE CONTROL DEL NÚMERO DE AFILIACIÓN A LA SEGURIDAD SOCIAL* Available at: https://www.grupoalquerque.es/ferias/2012/archivos/digitos/codigo_seguridad_social.pdf [Accessed 18 November 2023]

Trellix (2022) *Trellix Data Loss Prevention 11.x.x Classification Definitions Reference Guide - December 2022* Available at: https://docs.trellix.com/bundle/data-loss-prevention-11.10.x-classification-definitions-reference-guide/page/GUID-96141374-4DFC-4FAC-9B27-CDAF2D03005A.html [Accessed 22 November 2023]

BFI Finance (2023) *How to Check NIK Online Easily* Available at: https://www.bfi.co.id/en/blog/cara-cek-nik-online-dengan-mudah [Accessed 22 November 2023]

Cisco (2022) *Data Loss Prevention (DLP): Test Sample Data for Built-In Data Identifiers* Available at: https://support.umbrella.com/hc/en-us/articles/4402023980692-Data-Loss-Prevention-DLP-Test-Sample-Data-for-Built-In-Data-Identifiers [Accessed 22 November 2023]

Protecto (2023) *Download personal data for testing* Available at: https://www.protecto.ai/download-personal-data-for-testing [Accessed 22 November 2023]

MyVehicle (2017) WHAT IS A VIN NUMBER? Available at: https://www.myvehicle.ie/car-news/vehicle-identification-number-vin#:~:text=For%20example%2C%20in%20this%20VIN,World%20Manufacturer%20Identifier%20(WMI). [Accessed 22 November 2023]

AWS (2023) *Keyword requirements for Amazon Macie managed data identifiers.* Available at: https://docs.aws.amazon.com/macie/latest/user/managed-data-identifiers-keywords.html [Accessed 22 November 2023]

GCP (2023) *Syntax for Regular Expressions* Available at: https://support.google.com/a/answer/1371415?hl=en [Accessed 22 November 2023]

Microsoft (2023) *Detect threats to sensitive data* Available at: https://learn.microsoft.com/en-us/azure/defender-for-cloud/defender-for-storage-data-sensitivity [Accessed 26 November 2023]