National College of Ireland

Designing the Architecture of an Efficient Cloud-based Data Security Posture Management System

MSc Research Project
MSc Cyber Security

Amiket Kumar Srivastava
Student ID: X22119451

School of Computing
National College of Ireland

Supervisor: Dr Vanessa Ayala-Rivera

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Amiket Kumar Srivastava
Student ID:	22119451
Programme:	MSCCYB1 **Year:** 1
Module:	MSc Research Project
Supervisor:	Dr Vanessa Ayala-Rivera
Submission Due Date:	14th December, 2023
Project Title:	Designing the Architecture of an Efficient Cloud-based Data Security Posture Management System
Word Count:	6405 **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: Amiket Kumar Srivastava

Date: 13/12/2023

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):	

Designing the Architecture of an Efficient Cloud-based Data Security Posture Management System

Amiket Kumar Srivastava
22119451

Abstract

The rise in data processing and subsequent cloud adoption within the industry has raised fresh concerns about data security. Sensitive data exposure in leaks and breaches has become a regular occurrence nowadays leading to heavy monetary and reputational damages. We need to identify this sensitive data and apply strict security controls to protect it. Existing solutions like Cloud Security Posture Management system (CSPM) are cloud-centric which focus on perimeter security. Data security is usually a limited sub-module which is costly to implement in these solutions. This leaves our systems vulnerable to data leaks and breaches. This paper proposes to design an efficient cloud-based Data Security Posture Management (DSPM) system which can identify sensitive Personal Identifiable Information (PII), calculate its risk of disclosure, assign sensitivity labels, and deliver cost-effective security controls. Using this vision, we implemented and evaluated the different components of our DSPM system through seven experiments where we, classified and ingested non-standard data-sources (improvement from its predecessor AURUM that required high expertise to ingest non-standard sources and external classification mechanism), implemented the principle of least privilege and geo-fencing, created dynamic data-masking rules and dynamic data-backups for tables, published residency compliance report and installed a self-hosted agent to automate workloads. Throughout our experiments, we observed high cost-effectiveness, efficiency, and potency for tackling data security problems. Our contribution involves deepening the understanding of data-related challenges and their effective resolution through the development of an efficient, data-centric cloud-based DSPM system. We believe that our research promotes data security within the open-source community.

1 Introduction

In the current information technology landscape, there is a race between competing organizations in the industry for acquiring and processing all kinds of data. This has made data the most valuable asset for an organization (Bento, Neto, & Côrte-Real, 2022). To meet these high demands, most organizations are adopting cloud computing technology. Data processing requires focus on its key properties like accuracy, completeness, accessibility, relevance, validity, and consistency (Shaikh & Sasikumar, 2015). However, it is difficult to do so in cloud data stores. We have observed that sensitive data exposure in data leakages or breaches from cloud services is a recurring theme nowadays. Furthermore, Data is regulated by both international and domestic laws. Some of these like General Data Protection Regulation (GDPR) define data residency requirements as well (Europa, 2022). Scalable Cloud environments are constantly changing, which adds to the difficulty of tracking sensitive data while staying compliant. Failure to manage data security results in fines, penalties, and lawsuits. Hackers are aware of these problems and are actively trying to use this favourable battleground to steal sensitive data for monetary gains. The consequences of such attacks are

devastating and usually mean heavy financial losses in addition to damage to a company's image (Meisner, 2017). The estimated cost of data breaches in the United States of America was around $9.44 million in the year 2022 and around 422 million users were affected in some form (Petrosyan, Average cost of a data breach worldwide as of 2022, by country or region, 2022) (Petrosyan, Annual number of data compromises and individuals impacted in the United States from 2005 to 2022, 2023). As more companies move towards using cloud services, they become prone to these risks.

Past approaches to fill this gap in security have resulted in many different solutions with the most popular being Cloud Security Posture Management (CSPM) (Enriquez, 2021). However, these solutions prioritize cloud security first and data security is an afterthought in many of the cases (Enriquez, 2021) (Sawhney, Kaur, & Deorari, 2022). Others implement data security in some form but fail to segregate and treat sensitive information with special care with respect to public information. The lack of identification and classification of personal identifiable information (PII) could result in data breaches borne out of negligence.

Data Security Posture Management (DSPM) which is the approach we are proposing is a data-centric approach to manage an organisation's security. It is a continuous systematic process of scanning, analysing, assessing, remediating, and establishing different aspects of security through a feedback mechanism. It identifies and classifies sensitive data found in an environment. The classifications are based on the risk associated with the disclosure of the data and these classifications can then be used for prioritizing security. Data with higher levels of risk on disclosure will be secured first with strictest security controls and then onto the next level in a descending order with public data the last one to be secured with the most lenient policies. There are some DSPM solutions available in the market today (Normalyze, 2023) (Zscaler, 2023) (IBM, What is DSPM?, 2023), but they are not mature (still under development), costly and require infrastructure investments as well on the user end.

This problem motivates the following research question: "How to design a cost-effective, efficient, data-centric and cloud-based Security Posture Management System?"

The aim of this research is to help solve the above problem by designing an efficient, cost-effective and cloud based DSPM system which can help fill the gap in data security in the industry. The major contribution of this research is a novel design for the DSPM system with focus on data security while minimizing cost and increasing efficiency. The design contribution promotes data security in the open-source community.

This paper discusses related work that focuses on past research work done in the data security space in section 2. Section 3 describes the research methodology used in this research. Section 4 discusses the design specification for this research. Section 5 discusses the implementation of this research. Section 6 discusses the experiments involved in this research and the evaluation of the results and section 7 discusses the conclusion and scope for future work of this research.

2 Related Work

There has been rising interest in creating a Data Security Posture Management System. We will discuss past work in this domain which can be used by us in this research in the subsequent subsections.

2.1 Rise in Cloud Adoption and use of CSPM

(Sawhney, Kaur, & Deorari, 2022) discusses the growing popularity of cloud computing and its advantages like reduced costs, scalability, and availability but it also stresses on the lack of creativity in data security controls to govern it. (Gupta & Narayan, 2023) concurs about the rise in cloud adoption while discussing the latest trends in technologies. Both discuss common security problems and the use of CSPM systems to detect threats within the organisation. (Gupta & Narayan, 2023) provide survey results that show 96.7% of the participants acknowledging the importance of cloud, 25.8% participants do not trust that cloud is safe, and 30% participants wanted a cloud-less network. These results suggest that even though cloud is considered important today, but the people are still hesitant to transition as they do not feel safe.

(Khalid EI Makkaoui, Beni-Hssane, & Motamed, 2016) gives details on how to use CSPM to handle attacks. Similarly, (Enriquez, 2021) tells us how we can use a CSPM system to secure our cloud environment. All these approaches are very cloud-centric, and they fail to answer problems related to data security especially in areas where public and sensitive PII data are both stored together but need to be managed with separate security controls. (Khalid EI Makkaoui, Beni-Hssane, & Motamed, 2016) mentions in its future work about countermeasures in the Data layer and Access and Privilege Management Controls.

2.2 Data Discovery and Catalogue

(Terzo, Ruiu, Bucci, & Xhafa, 2013) discusses the DaaS (Data as a service) model. Its data discovery service with its ability to find datasets and adaptation to finding new data formats is of interest to us. A good discover service should find both data and metadata as per our requirement and should follow open standards for easy integration with cloud services. (Fernandez, et al., 2018) discusses AURUM and its data discovery. It relies on enterprise knowledge graphs (EKGs) that comprise of relationships between different entities. It makes use of resource efficient signature sampling (RESS) to preserve these graphs. Both these approaches are data-centric, and so we can find inspiration from them in our research. The discovered data can be stored in a data catalogue like Apache Atlas (Rodrigues, Almeida, Guimarães, & Santos, 2022). The shortcoming of both these services is that they consume a lot of resources.

2.3 Data Classification and Risk Rating

(DIAO Zhe & SU Naizheng, 2017) discusses the technical, management and legal risks of using cloud storage. It recommends that the cloud providers should take responsibility for securing the data. This contradicts (Joshi, Raturi, Kumar, Dumka, & Singh, 2022) directly as it shares concerns about third-party service integrations and how they are vulnerable points our security posture. It also discusses the risks associated with each stage of data storage, right from its ingestion to deletion. (Shaikh & Sasikumar, 2015) mentions data classification as the phenomenon of sorting data into groups based on their sensitivity level. This level is directly proportional to the risk associated with the data's disclosure. Commonly used levels of sensitivity are public, internal, confidential, or restricted. (Hasan, et al., 2023) summarises cloud security issues with the statement "data security does not imply data integrity". It specifies the various levels of threats at each segment of cloud computing. It suggests a broader perspective by showing different levels of classified attacks and urges us to classify our data based on sensitivity, so we are prepared for the different levels of attack.

2.4 Data Governance and Informed Decisions

(Bento, Neto, & Côrte-Real, 2022) discusses the study of eleven critical success factors (CSFs) from five data governance frameworks and an assessment matrix. It concludes by promoting the use of assessment matrix with the CSFs and different frameworks to grow data governance maturity of organizations. (Saed, Aziz, Ramadhani, & Hassan, 2018) discusses breaches and attacks that shattered the industry. Statistics from its survey reveal that only 3 out of 10 organisations employed specialised data governance teams. It exposes the fact that 42% respondents were clueless about where to start with data governance for their organisation. It concludes by suggesting that data security controls should be applied at all levels. It implores the employees to have a sense of duty towards data security. It suggests an analytical approach to framing governance policies. (AlGhamdi, Win, & Vlahu-Gjorgievska, 2020) show us how to develop CSFs for our own organization. (Rahimi, Maimaiti, & Zincir-Heywood, 2014) discusses the use of geo-fencing and access management in securing an environment.

2.5 Novelty of my proposed approach and Conclusion

From the above discussions, we can observe that cloud adoption is increasing but people are still not completely sure about the safety of data in cloud. Another observation is that most of the past work in this field is cloud-centric and very less significance has been given to data security controls. Some interesting research was conducted around cloud security posture management (CSPM), but it is very different to our approach. It is focused on cloud instead of data and only prioritizes protection against internal and external threats and vulnerabilities. In most of the past approaches, they have failed to segregate the data based on its sensitivity.

Our approach is unique as we aim to design an efficient, cost-effective, data-centric, and cloud-based data security posture management system (DSPM). It is centred on identifying and classifying sensitive PII information with appropriate sensitivity value which would be directly proportional to the risk associated with that information's disclosure. We can then make use of the classified data to create data security policies and procedures. They key novel aspect of our research comes from the fact that we are combining the best parts of many distinct and different data research approaches such as data discovery, data classification, data catalogue and risk assessments to create a data-centric model (DSPM). This will enable us to have a holistic view of our data stores in cloud and govern them with well-informed procedures and policies which will in turn, help us grow the security posture of our organisation through data-driven decisions.

3 Research Methodology

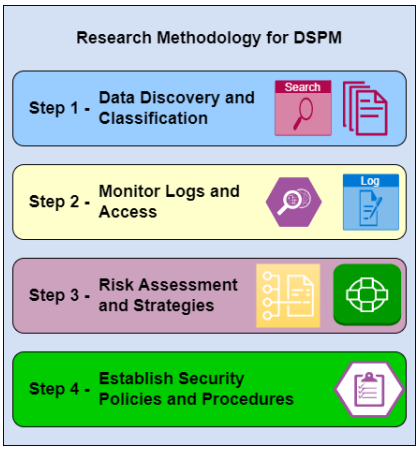


Figure 1 - Research Methodology for DSPM with steps

Figure 1 shows the research methodology used for our DSPM system. It comprises of 4 steps namely Data Discovery and Classification, Monitor Logs and Access (Principle of Least Privilege), Risk Assessment and Strategies, and Establish Security Policies and Procedures.

The first step is Data Discovery and Classification. The ability to find data sources, ingest metadata information and perform autoclassification is key to building our DSPM model. Not only does this process allow us to find data stored in different sources across our cloud platform but it also eases subsequent risk assessments as we have a baseline of classification values to choose from.

The second step is implementing the Principle of Least Privilege by continuous monitoring of resource logs and access information. This is a critical component in securing data containers as it reduces the chances of accidental data disclosure due to inept identity and access management policies. It also helps in quickly identifying and rectifying any unwanted access attempts by analysing the logs and performing blacklisting and geo-fencing techniques.

The third step is performing risk assessment and using different strategies to deal the issues found. This is done to ensure the safety of all the components in our system. The results of the risk assessment help us in identifying and defects within our critical components and dealing with them swiftly and securely. There are 4 strategies to handle risk: mitigate, avoid, accept, or transfer.

The fourth step is establishing security policies and procedures. This is the most important step in our process as it helps us drive our DSPM system by generating automated security controls. These security controls help us prevent data breaches. The policies created in this step make our overall process of securing the architecture efficient and cost effective as we are leveraging on the information and analysis conducted in the previous steps to regulate how information flows within and outside our environment.

4 Design Specification

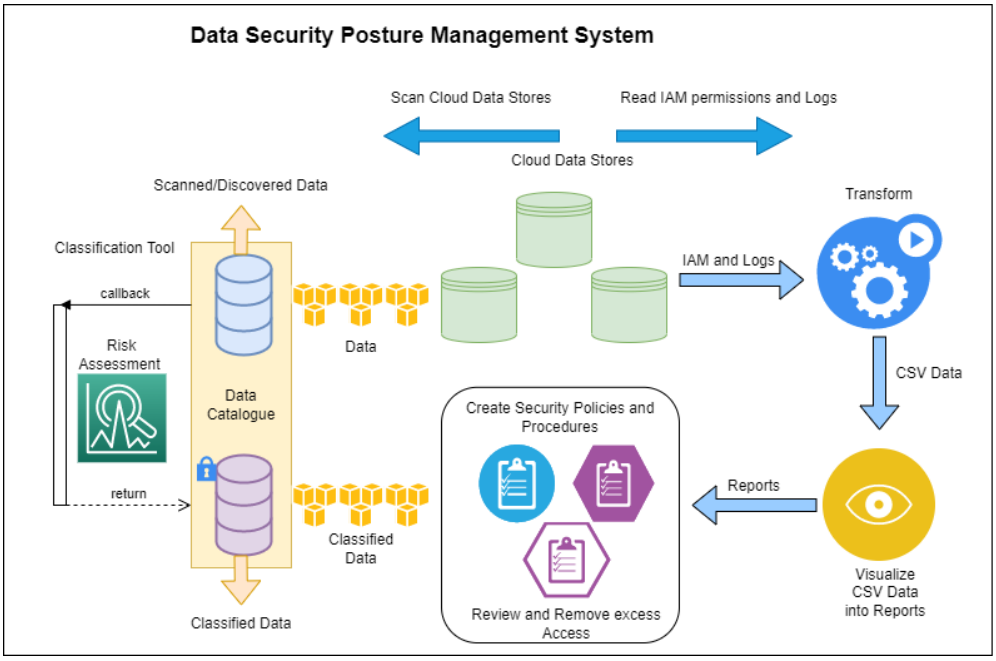


Figure 2 - Proposed Architecture Diagram of DSPM

Figure 2 shows the proposed architecture for DSPM. It combines data discovery, classification and risk assessment which allows us to evaluate the risk level associated with different resources in our environment and create security policy and controls dynamically. We take a closer look at the designs for different parts of our DSPM architecture in the below subsections.

4.1 Data Discovery and Classification

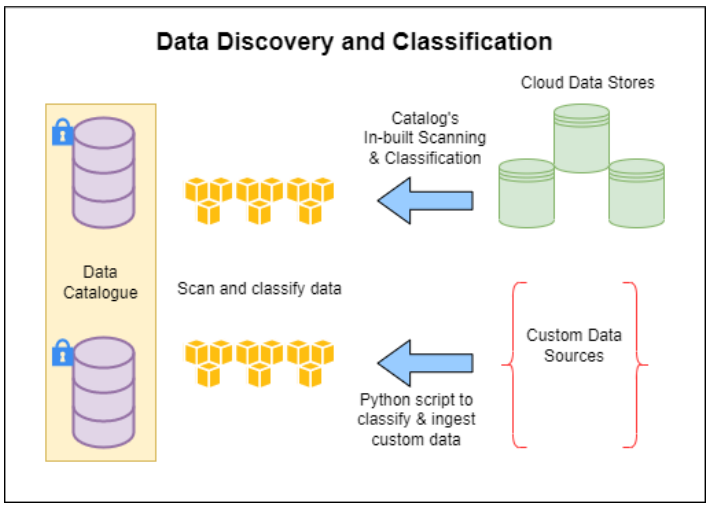


Figure 3 - Data Discovery and Classification

Figure 3 depicts the scanning of cloud data inventory and other custom data sources. The data catalogue is authorized to search through the cloud sources using managed identities and service principals. The scans can be scheduled after registering the sources in the data catalogue as shown in Figure 4.

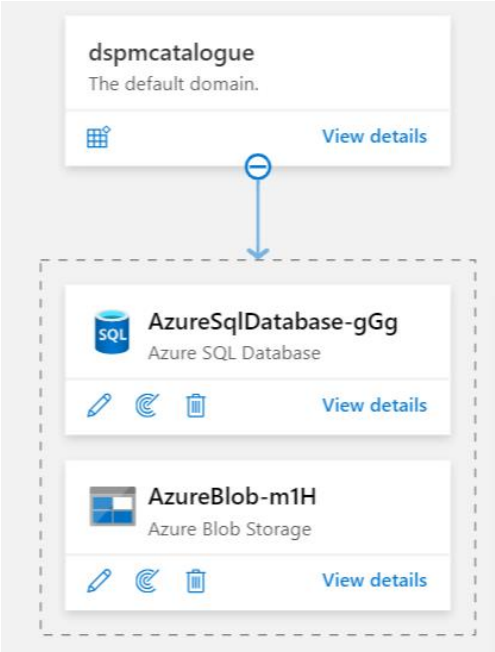


Figure 4 - Data Source Registration and Scan Scheduling

It searches through these different data stores like data lakes, data shares or data dictionaries for metadata information and then identifies and classifies the sensitive Personal Identifiable Information (PII) present in them based on classification rules defined in the data catalogue as shown in Figure 5.

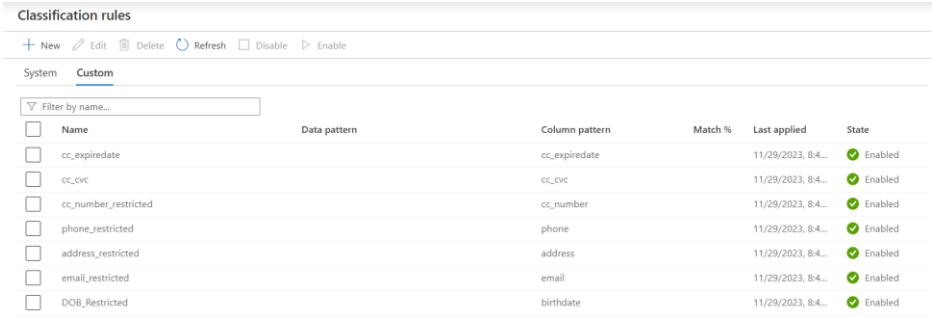


Figure 5 - Classification Rules in Data Catalogue

For custom data types, a python script and dictionary file are used to classify the data before ingesting it into the data catalogue. The dictionary file is to be maintained by the user of the DSPM in the future. The classification values are determined based on the risk associated with the disclosure of the respective data. The data catalogue is used to map of all the organisation's data stored in cloud and custom data sources (public and sensitive) to get a holistic view of the entire inventory. Figure 6 shows different assets and their classification values in the catalogue.

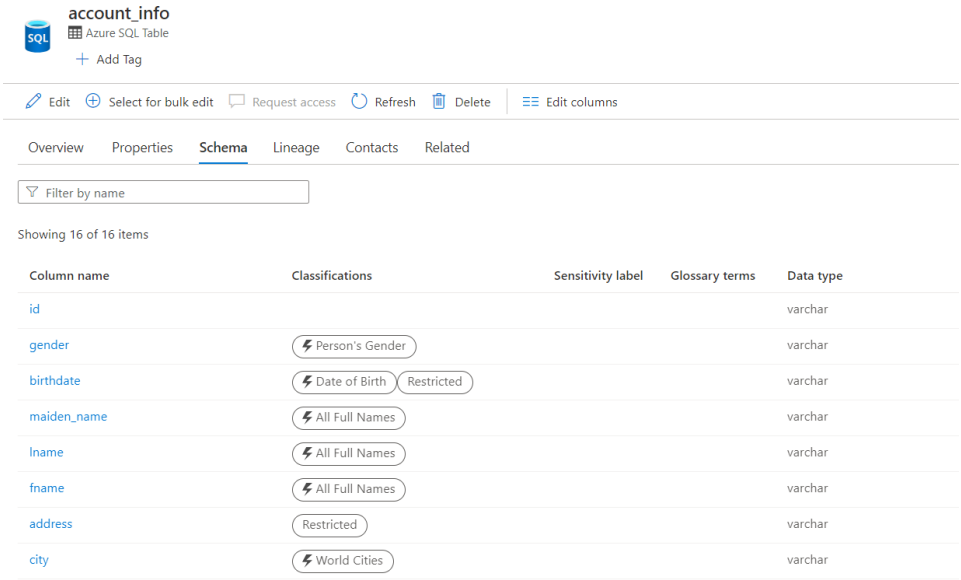


Figure 6 - Assets and their Classifications in the Data Catalogue

4.2 Monitor Logs and Access (Principle of Least Privilege)

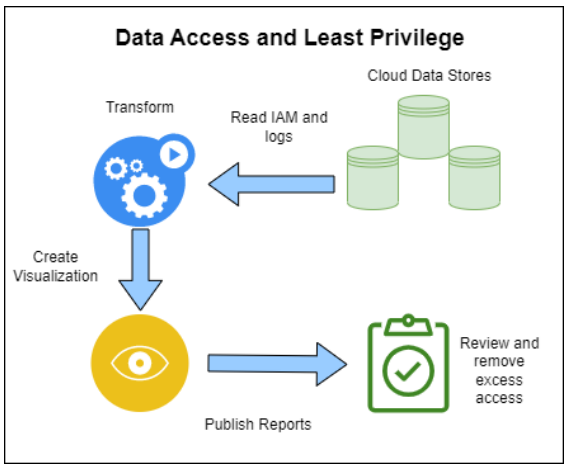


Figure 7 - Monitor Data Access and Implement Principle of Least Privilege

Figure 7 illustrates the process of monitoring logs and access permissions. We use automated pipelines to regularly download the IAM permissions and access logs of different cloud components via API and use them to create reports with the required visualizations as shown in Figure 8 below.

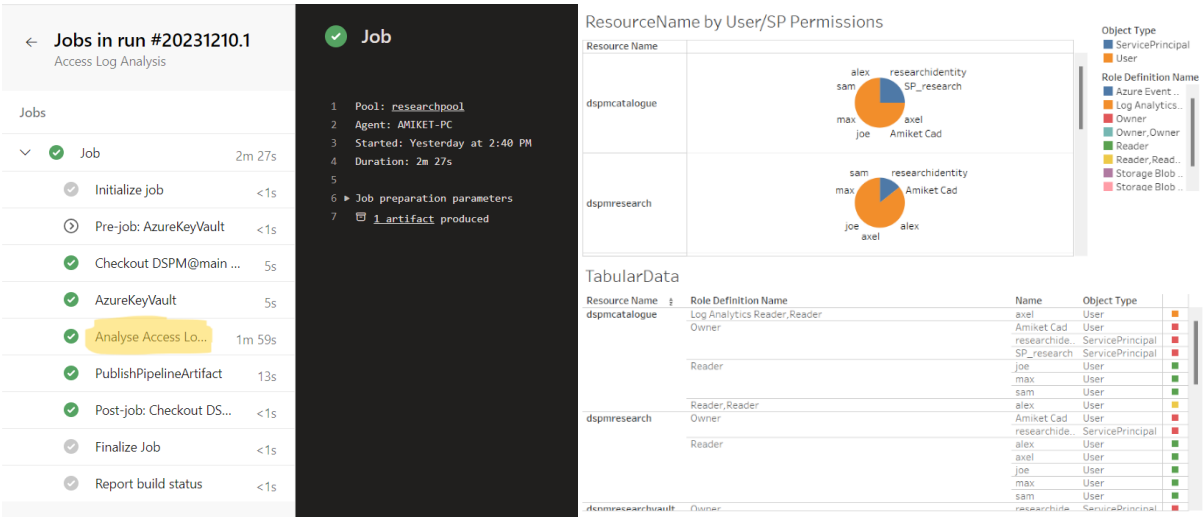


Figure 8 – Automated Pipeline to Monitor Logs and Access and corresponding Visual Report

We then look up the individual experts or owners of each resource (as shown in Figure 9) in the Data Catalogue and send these reports to them for manual review. The purpose of this activity is to restrict and remove any unnecessary access to any of the resources. This is important as often some employees are assigned access to resources for specific activities, but the access is not removed upon activity completion. It can also help remediate any access that was erroneously provided to someone (account copied from existing employee with elevated privileges). We can read the access logs of a component and report when it was accessed and by whom. We can also show the requestor's IP Address and country of origin and use these details to implement geo-fencing in our organization.

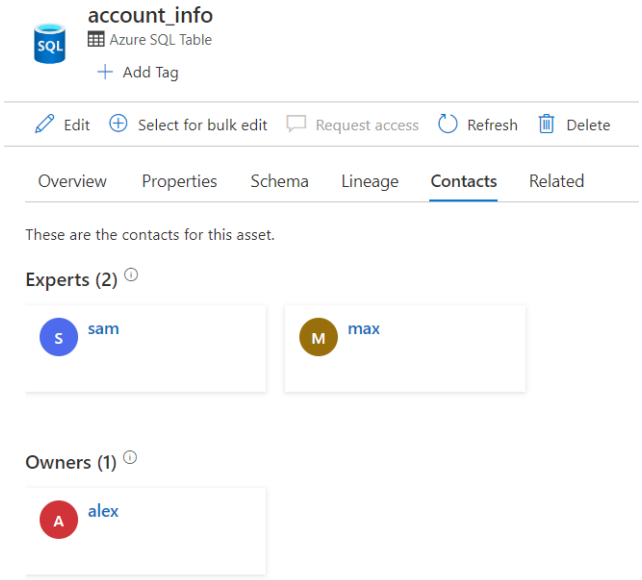


Figure 9 - Experts and Owners for a resource in the Data Catalogue

4.3 Risk Assessment and Strategies

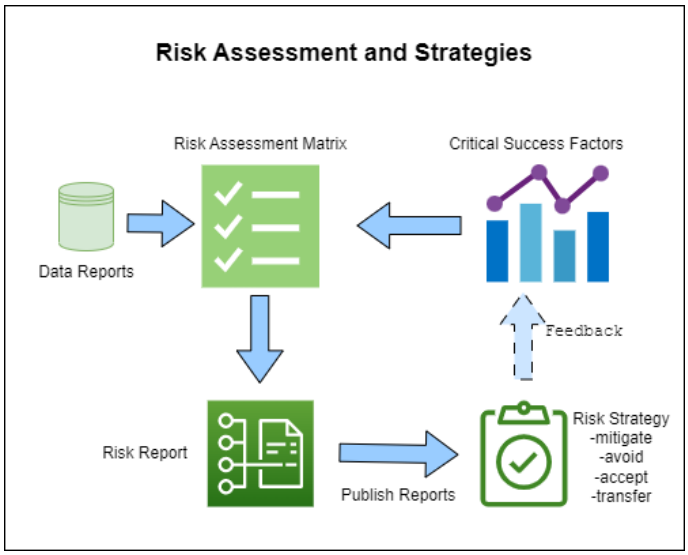


Figure 10 - Risk Assessment and Strategies

Figure 10 shows an illustration of risk assessment and strategies process using critical success factors. This is a two-step process. First step involves conducting a manual risk assessment of each cloud resource using the assessment matrix and critical success factors. Second step involves contacting the resource owner and providing them with the risk report so that they can work on the findings. Risk can be treated by the owners using any of the following Risk strategies – mitigate, avoid, accept, and transfer. We are using a point-based system for our risk assessment. The CSFs we have defined for our organization and their corresponding risk ratings are shown in Figure 11 below. These have been carefully calibrated to the assigned points value to ensure fair risk assessments of all cloud sources.

Critical Success Factors	Points
Sensitivity Label?	Restricted = 5 points
	Confidential = 6 points
	Internal = 4 points
	Public = 0 points
Residency Requirement?	Yes = 1 point
	No = 0 point
Azure Resource Type?	StorageAccount=1 points
	BlobStorage=1 points
	SQLServer=1 points
	SQLDatabase=1 points
	Other=0 point
Data Masking?	Yes = -3 point
	No = 0 point
Data Backup?	Yes = -1 point
	No = 0 point
Residency Compliant?	Yes = -1 point
	No = 0 point

Points	Risk Rating
>=7	High
>=6	Medium/High
>=5	Medium
>=4	Low/Medium
<4	Low

Figure 11 – (a) Critical Success Factors and (b) Risk Ratings

Using the above CSFs, we have defined the risk assessment matrix as shown in Figure 12. This matrix is used for determining the risk rating of a resource. The formulas applied to this matrix are directly derived from our CSFs and risk ratings.

Resource Name	Sensitivity Label?	Residency	Azure Resource Type?	Data Masking?	Data Backup?	Residency Compliant?	Points	Risk Rating
researchtest	Confidential	Yes	SQLDatabase	Yes	Yes	Yes	3	Low
researchanothercatalogue	Restricted	Yes	StorageAccount	No	No	No	7	High
							0	Low
							0	Low
							0	Low

Figure 12 - Risk Assessment Matrix

For example, in Figure 12 we can see that risk assessment is being carried out for two resources, one structured query language (SQL) database "researchtest" and another storage account "researchanothercatalogue". Let's first look at "researchtest" db. As per our CSFs, this resource is given 6 points for the Confidential sensitivity level. It is given another point for the residency requirement condition and another point because it is a database type resource. So, 6+1+1=8 points for this resource. Next, 3 points are deducted as the data is masked on this DB. Another point is deducted as the backups are also complete and another point is deducted as the DB is residency compliant. So, 8-3-1-1=3 which is why it is assigned a "Low" risk rating.

Now let's look at the second example for resource "researchanothercatalogue", it is first given 5 points for the Restricted sensitivity level and then 1 point each for residency requirement and azure resource type. However, no points are deducted as this resource is missing masking policy and backups and residency compliance. So, 5+1+1-0-0-0=7 which is why it is assigned a "High" risk rating. Once the risk ratings are determined, the owners/experts are looked up in the data catalogue and are contacted if their resource is at a non-acceptable risk level. They are then tasked with risk treatment for their resource.

4.4 Establish Security Policies and Procedures

This is the key step in unlocking the full potential of a DSPM system. Using the information collected in the previous steps, we can evaluate the current needs of our environment. We can then draft policies and procedures to fulfil those needs using code hosted in automated pipelines. These policies and procedures are also continuously monitored by the DSPM system and feedback is used to improve them in a continued fashion which helps in improving the overall security posture of our environment.

For example, we can create an access review and geo-fencing policy by reading resource logs and mapping requesting IP addresses to their country of origin. We can then whitelist, or blacklist IP addresses based on geo-location to implement our geo-fencing policy. We can remove user or service principal's access to resources based on periodic access reviews as a part of the same policy. Figure 13 illustrates this process in detail.

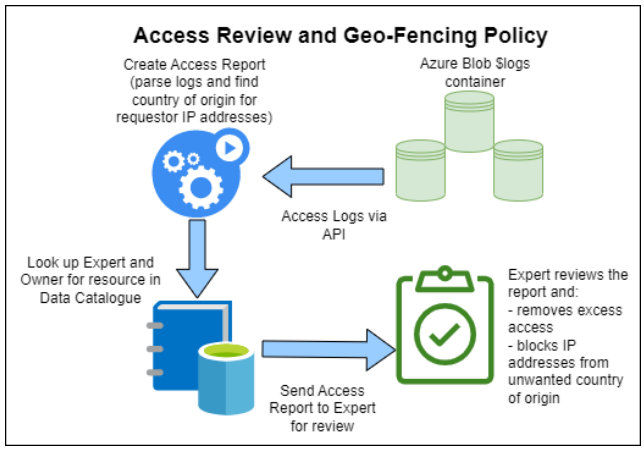


Figure 13 - Access Review and Geo-fencing Policy

5 Implementation

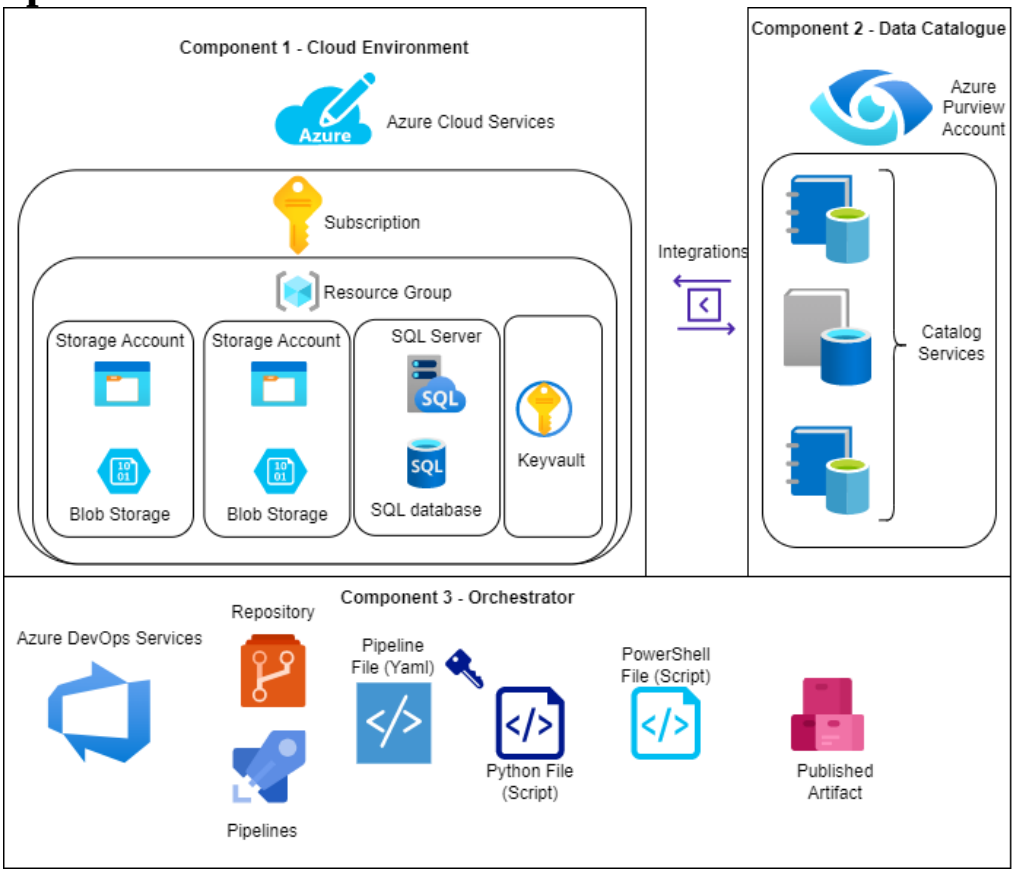


Figure 14 – Technical Diagram of implemented DSPM Architecture

Figure 14 shows a technical diagram of the implemented DSPM architecture. It consists of 3 key components – a Cloud Environment, a Data Catalogue, and an Orchestrator. For the first component we are using an Azure cloud environment where we have deployed multiple data resources such as storage accounts and SQL databases. A key vault was also provisioned to securely manage secrets and keys. We made the choice to use Azure as it is the most popular choice for personal and commercial use.

For the second component, we deployed a Purview account which will provide us with catalogue services. Purview has built-in data scanning and classification capabilities and has integrations with most common cloud services. We added custom classification rules in Purview so it could identify and classify sensitive PII data. Purview was selected as it is highly configurable and flexible and easily available in Azure.

We used Azure DevOps orchestrator as the third component of our DSPM system. We installed a self-hosted agent to run our pipeline jobs. This enabled us to run the pipeline for free. We chose Azure DevOps for its easy user interface in addition to the free cost of operation.

Once all the components were set up, we set up scans in purview using manged identities and service principals to acquire metadata information about the cloud resources. For external non-standardized sources, we used API endpoints for ingestion. We created Python and PowerShell scripts to automate the workflows and used Azure DevOps pipelines to run these scripts. The pipeline fetched service principal details from the key vault and supplied them to the scripts so they could work securely.

6 Evaluation

The above implementation required a series of experiments which are well documented in the below subsections. It consists of a comprehensive report of the major finding of each experiment and the results and conclusions from them.

6.1 Experiment 1: Ingest Custom Data Source into Data Catalogue

The aim of this experiment is to ingest custom data schema from external sources. This procedure will help users of DSPM in managing the security of non-standardized data sources. Data sources are scanned and metadata information about them is ingested and stored into the data catalogue using python scripts hosted in automated pipelines. For this experiment, we used the following sample data as shown in Figure 15.

id	gender	birthdate	maiden_name	lname	fname	address	city	state	zip	phone	email	cc_type	cc_number	cc_cvc	cc_expiredate
172-32-1176	m	21-04-1958	Smith	White	Johnson	10932 Bigge Rd	Menlo Park	CA	94025	408 496-7223	jwhite@domain.com	m	5270 4267 6450 5516	123	25-06-2010
514-14-8905	f	22-12-1944	Amaker	Borden	Ashley	4469 Sherman Street	Goff	KS	66428	785-939-6046	aborden@domain.com	m	5370 4638 8881 3020	713	01-02-2011
213-46-8915	f	21-04-1958	Pinson	Green	Marjorie	309 63rd St. #411	Oakland	CA	94618	415 986-7020	mgreen@domain.com	v	4916 9766 5240 6147	258	25-02-2009
524-02-7657	m	25-03-1962	Hall	Munsch	Jerome	2183 Roy Alley	Centennial	CO	80112	303-901-6123	jmunsch@domain.com	m	5180 3807 3679 8221	612	01-03-2010
489-36-8350	m	06-09-1964	Porter	Aragon	Robert	3181 White Oak Drive	Kansas City	MO	66215	816-645-6936	raragon@domain.com	v	4929 3813 3266 4295	911	01-12-2011
514-30-2668	f	27-05-1986	Nicholson	Russell	Jacki	3097 Better Street	Kansas City	MO	66215	913-227-6106	jrussell@domain.com	a	3.4539E+14	232	01-01-2010
505-88-5714	f	23-09-1963	Mcclain	Venson	Lillian	539 Kyle Street	Wood River	NE	68883	308-583-8759	lvenson@domain.com	d	3.02049E+13	471	01-12-2011
690-05-5315	m	02-10-1969	Kings	Conley	Thomas	570 Nancy Street	Morrisville	NC	27560	919-656-6779	tconley@domain.com	v	4916 4811 5814 8111	731	01-10-2010
646-44-9061	M	12-01-1978	Kurtz	Jackson	Charles	1074 Small Street	New York	NY	10011	212-847-4915	cjackson@domain.com	m	5218 0144 2703 9266	892	01-11-2011
421-37-1396	f	09-04-1980	Linden	Davis	Susan	4222 Bedford Street	Jasper	AL	35501	205-221-9156	sdavis@domain.com	v	4916 4034 9269 8783	33	01-04-2011
461-97-5660	f	04-01-1975	Kingdon	Watson	Gail	3414 Gore Street	Houston	TX	77002	713-547-3414	gwatson@domain.com	v	4532 1753 6071 1112	694	01-09-2011
660-03-8360	f	11-07-1953	Onwunli	Garrison	Lisa	515 Hillside Drive	Lake Charles	LA	70629	337-965-2982	lgarrison@domain.com	v	4539 5385 7425 5825	680	01-06-2011
751-01-2327	f	16-02-1968	Simpson	Renfro	Julie	4032 Arron Smith Drive	Kaunakakai	HI	96748	808-560-1638	jrenfro@domain.com	m	5325 3256 9519 6624	238	01-03-2009
559-81-1301	m	20-01-1952	Mcafee	Heard	James	2865 Driftwood Road	San Jose	CA	95129	408-370-0031	jheard@domain.com	v	4532 4220 6922 9909	311	01-09-2010
624-84-9181	m	16-01-1980	Frazier	Reyes	Danny	3500 Diane Street	San Luis Obispo	CA	93401	805-369-0464	dreyes@domain.com	v	4532 0065 1968 5602	713	01-11-2009
449-48-3135	m	14-06-1982	Feusier	Hall	Mark	4986 Chapel Street	Houston	TX	77077	281-597-5517	mhall@domain.com	v	4556 0072 1294 7415	953	01-05-2010
477-36-0282	m	10-03-1961	Vasquez	Mceachern	Monte	456 Oral Lake Road	Minneapolis	MN	55401	952-412-3707	mmceachern@domain.com	m	5527 1247 5046 7780	889	01-03-2009
458-02-6124	m	20-09-1955	Pennebaker	Diaz	Christopher	582 Thrash Trail	Dallas	TX	75247	903-624-9156	cdiaz@domain.com	m	5299 1561 5689 1938	584	01-08-2011
044-34-6954	m	28-05-1967	Simpson	Lowe	Tim	1620 Maxwell Street	East Hartford	CT	6108	860-755-0293	tlowe@domain.com	m	5144 8691 2776 1108	616	01-10-2011
587-03-2682	f	24-10-1958	Dickerson	Oyola	Lynette	2489 O Conner Street	Pascagoula	MS	39567	228-938-2056	loyola@domain.com	v	4532 9929 3036 9308	991	01-07-2011
421-90-3440	f	17-07-1953	Kroeger	Morrison	Adriane	4696 Retreat Avenue	Birmingham	AL	35209	205-276-1807	amorrison@domain.com	v	4539 0031 3703 0728	322	01-12-2009
451-80-3526	m	09-06-1950	Parmer	Santos	Thomas	173 Lunetta Street	Fort Worth	TX	76104	940-859-1393	tsantos@domain.com	v	4716 6984 4983 6160	767	01-09-2011
300-62-3266	m	10-02-1965	Spain	Faulkner	Victor	1843 Olive Street	Toledo	OH	43602	419-340-3832	vfaulkner@domain.com	m	5548 0246 6336 5664	276	01-02-2010
322-84-2281	m	19-08-1977	Miley	Iorio	Albert	4899 University Hill Road	Springfield	IL	62701	217-615-6419	aiorio@domain.com	v	4916 6734 7572 5015	347	01-02-2010

Figure 15 - Custom Data to be Ingested into Data Catalogue

Data is assigned classification values by another python script that uses a dictionary file which can be modified as needed by the user while doing manual risk review. The dictionary file and corresponding classified source file are shown in Figure 16.

Dictionary File

Classification	Field
Restricted	cc_expiredate
Restricted	cc_cvc
Restricted	cc_number
Restricted	phone
Restricted	address
Restricted	email
Restricted	birthdate

Classified Data

cc_type	cc_number	cc_cvc	cc_expiredate	Classification
m	5270 4267 6450 5516	123	25-06-2010	Restricted
m	5370 4638 8881 3020	713	01-02-2011	Restricted
v	4916 9766 5240 6147	258	25-02-2009	Restricted
m	5180 3807 3679 8221	612	01-03-2010	Restricted
v	4929 3813 3266 4295	911	01-12-2011	Restricted
a	3.4539E+14	232	01-01-2010	Restricted
d	3.02049E+13	471	01-12-2011	Restricted
v	4916 4811 5814 8111	731	01-10-2010	Restricted
m	5218 0144 2703 9266	892	01-11-2011	Restricted
v	4916 4034 9269 8783	33	01-04-2011	Restricted
v	4532 1753 6071 1112	694	01-09-2011	Restricted
v	4539 5385 7425 5825	680	01-06-2011	Restricted
m	5325 3256 9519 6624	238	01-03-2009	Restricted

Figure 16 - Dictionary File and Classified Data

To ingest the data shown in Figure 16, we define and upload the custom schema using a combination of JavaScript Object Notation (JSON) and python files in automated pipelines. This expands the data catalogue schema so that it can accept our custom data. Figure 17 shows the custom data ingested via API in the data catalogue.

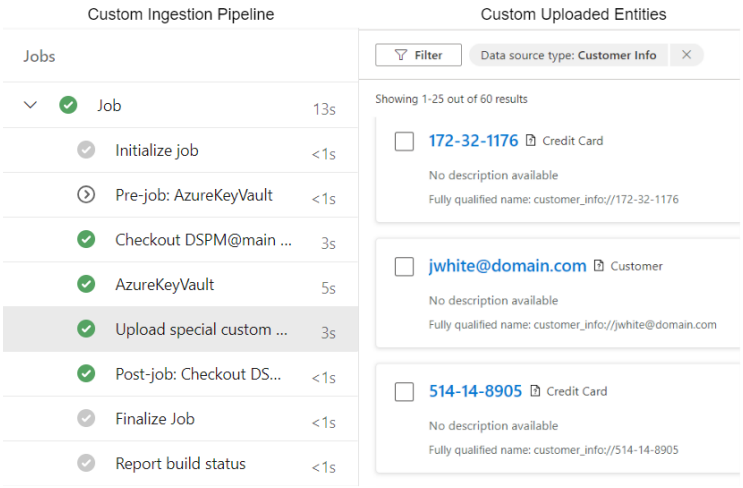


Figure 17 – Custom Data ingested via Automated Pipeline into Data Catalogue

This result indicates that the users of DSPM are now able to manage the security of non-standardized data sources. Other observations from this experiment were that the solution is efficient and cost-effective as its run time is 13 seconds. Despite the positives, it should be noted that users do need to define the JSON schema template file for their custom sources manually and the execution time will increase as the size of the data to be ingested will increase.

6.2 Experiment 2: Review IAM permissions and Send to Owners/Experts

The aim of this experiment is to implement the principle of least privilege through periodic review of IAM permissions by owners of data sources. This policy will prevent stagnant and unnecessary access to resources which could lead to sensitive data exposure by unauthorized users. A combination of scripts first generates a list of all the resources in cloud. IAM Permissions are then downloaded for each resource via API and reports are generated for them. The owners for these resources are looked up in the data catalogue and reports are then sent to them for manual review. Figure 18 shows an example of a visual report.

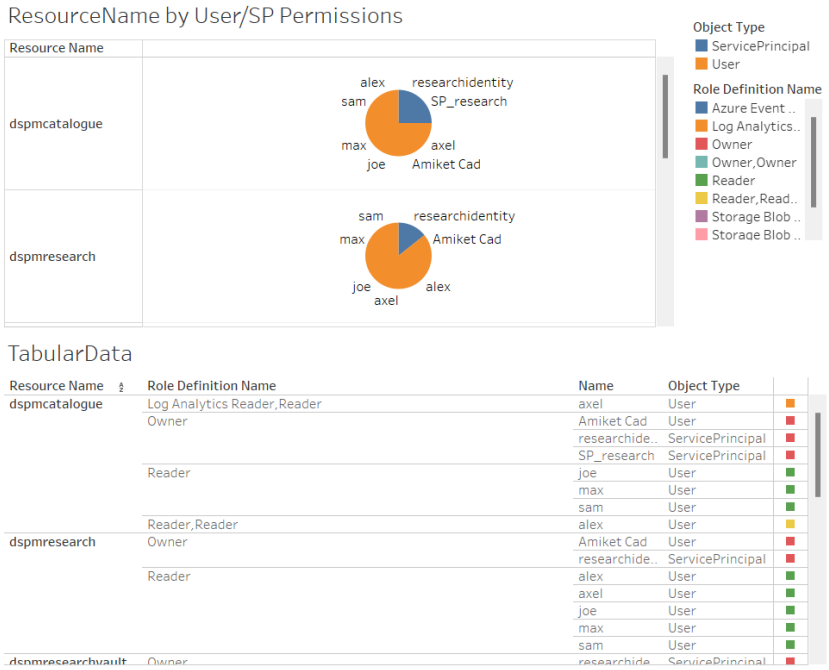


Figure 18 - Access Review Report for Different Cloud Resources

The result of this experiment is that the principle of least privilege was enforced on the cloud resources using continuous monitoring. Other observations from this experiment were that the solution is cost-effective and efficient with a run time of 1 minute 47 seconds. The downside to this experiment is that we need to manually create templates for visual reports for different sources and reviewing the report to remove excess access is a manual process for the owners.

6.3 Experiment 3: Implement Geo-fencing around Cloud Resources

The aim of this experiment is to implement Geo-fencing policy around cloud resources. This policy will prevent unwanted access to cloud resources by blocking IP addresses that originate from unauthorized locations. Access logs are downloaded via API and locations are mapped to the requestor IP addresses using scripts. Visual reports are then published which can be used to determine which IP addresses need to be blocked. Figure 19 shows such a visual report.

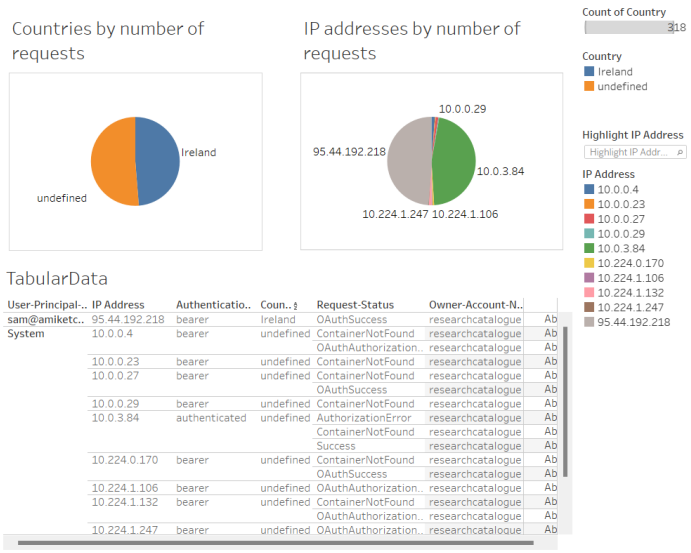


Figure 19 – Visual Report for Implementing Geo-Fencing

This result indicates that the Geo-fencing policy was successfully implemented. Other observations from this experiment were that while solution is cost-effective and efficient as the run time is 2 minutes 27 seconds, the creation of visualization template for each individual cloud resource is a tedious one-time job. The IP addresses can then be blocked manually by the users, but this design can be improved as the blocking of IP addresses can be automated if a pre-built list of acceptable geo-locations is supplied to the script.

6.4 Experiment 4: Dynamic Data Masking

The aim of this experiment is to create data masking policy automatically based on the classification value of a cloud resource. This policy will prevent sensitive data exposure to unauthorized users. Data sources are scanned and metadata information about them is ingested and stored into the data catalogue. The sources are auto assigned classification values based on classification rules set by us. Classification value for a resource is shown in the Figure 20.

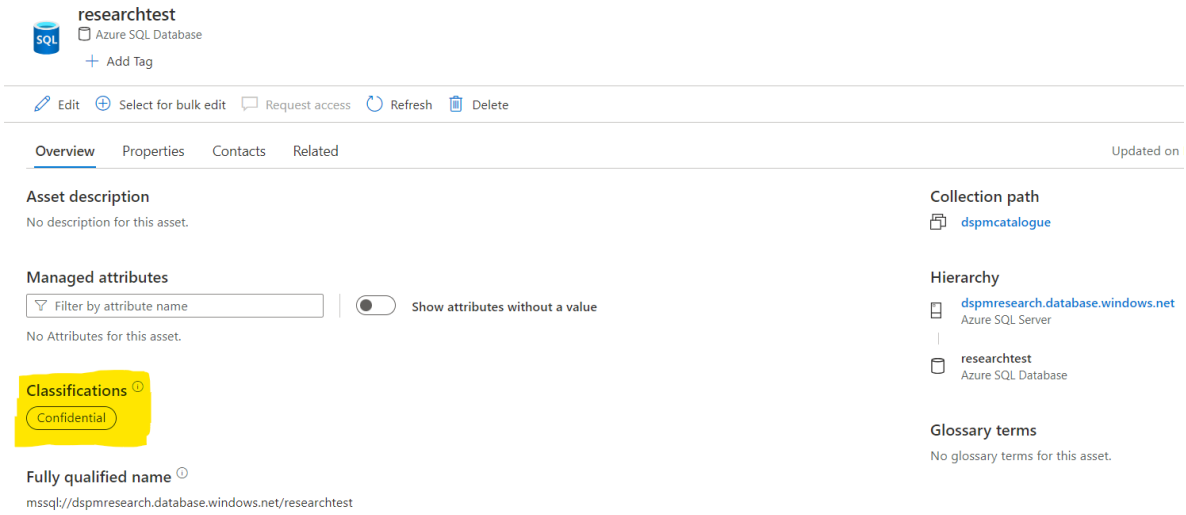


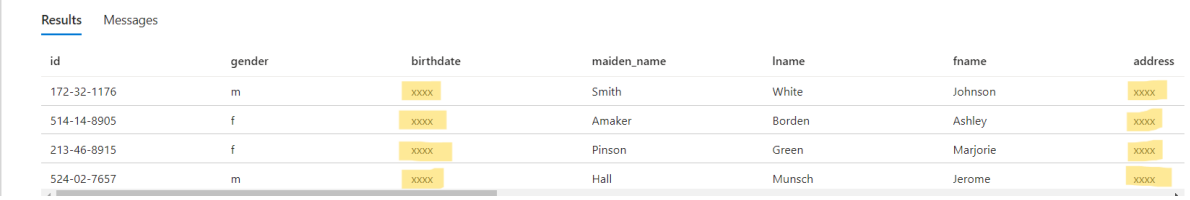
Figure 20 - Classification Value for Database

A combination of scripts is run in automated pipelines which query cloud for a list of all the resources and then query the data catalogue for their classification values. If the classification value of a resource is confidential, restricted, or internal, then the script creates data masking rules for that resource using API endpoints. Figure 21 shows masking rules created in a run.

Schema	Table	Column	Mask Function
dbo	account_info	birthdate	Default value (0, xxxx, 01-01-1900)
dbo	account_info	address	Default value (0, xxxx, 01-01-1900)
dbo	account_info	phone	Default value (0, xxxx, 01-01-1900)
dbo	account_info	email	Email (aXXX@XXXX.com)
dbo	account_info	cc_number	Credit card value (xxxx-xxxx-xxxx-1234)
dbo	account_info	cc_cvc	Default value (0, xxxx, 01-01-1900)
dbo	account_info	cc_expiredate	Default value (0, xxxx, 01-01-1900)

Figure 21 - Dynamic Data Masking Rules created by Pipeline

Once the rules are setup then if any unauthorized user tries to access the database, the sensitive columns are masked for that user which prevents unwanted information disclosure. This is shown in Figure 22 below where a non-authorized user Max tries to access the database table. The values of birthdate and address (highlighted in Figure 22) are sensitive and so masked and not visible to Max.



id	gender	birthdate	maiden_name	lname	fname	address
172-32-1176	m	xxxx	Smith	White	Johnson	xxxx
514-14-8905	f	xxxx	Amaker	Borden	Ashley	xxxx
213-46-8915	f	xxxx	Pinson	Green	Marjorie	xxxx
524-02-7657	m	xxxx	Hall	Munsch	Jerome	xxxx

Figure 22 - Sensitive Data is masked in the table for unauthorized user Max

This result indicates that our dynamic data masking policy was successfully implemented. Other observations from this experiment were that the solution is cost-effective as we just need to run the scripts via an orchestrator and the process is fairly efficient as execution time of the script in one run is 1 minute 37 seconds.

6.5 Experiment 5: Dynamic Data Backup as per Risk Assessment

The aim of this experiment is to create data backups dynamically based on the classification value of a cloud resource. This policy will create redundancy for sensitive data sources. A list of resources and their classification values are fetched as described in the previous experiment. If sensitive classification value is found for a resource, then the script creates a backup of that source's data in an Azure Blob container using cloud API endpoints. The pipeline run is shown in Figure 23 below:

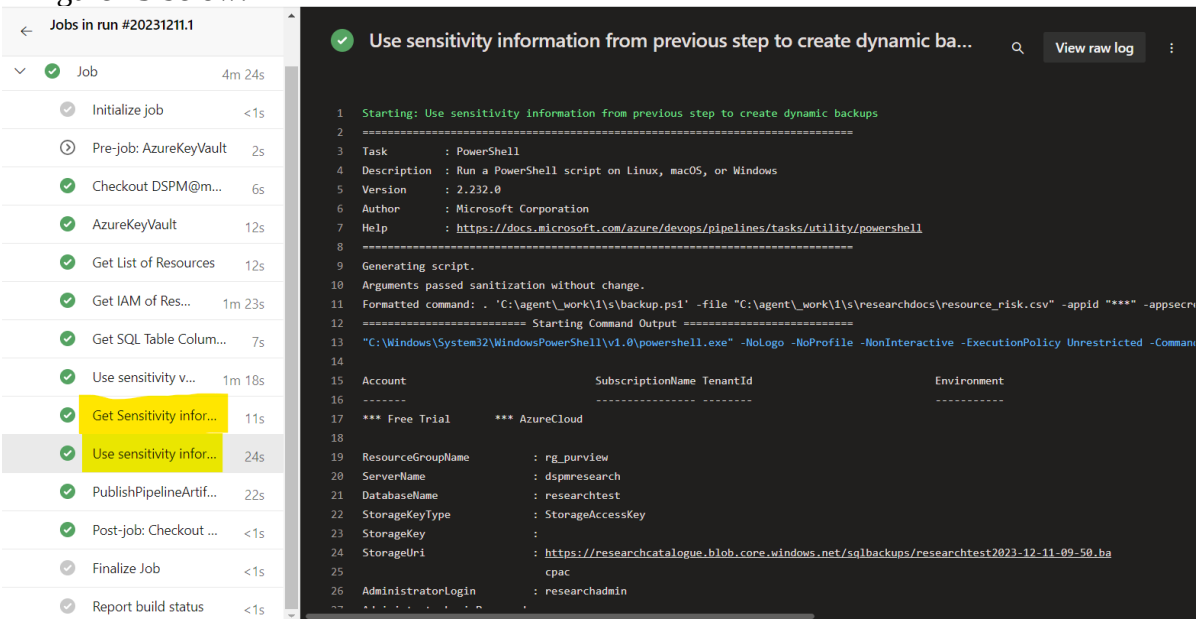


Figure 23 – Dynamic Data Backup Pipeline

Once the script finishes running, it creates a backup of the cloud resource in the Azure Blob container as seen highlighted in Figure 24 below. The script can create a daily backup in this manner ensuring the safety of sensitive data.

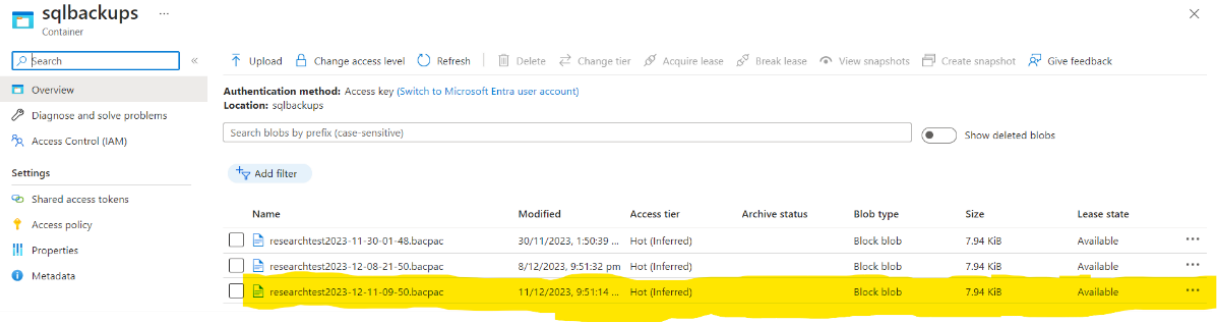


Figure 24 – Dynamic Data Backups created in Azure Blob Container

This result indicates that the dynamic data backup policy was successfully implemented. This solution was observed to be cost-effective and efficient with a run time of 47 seconds. Another observation to note is that the execution time of the script will increase as the size of the data to backup grows.

6.6 Experiment 6: Residency Requirement Audit Policy

The aim of this experiment is to audit cloud sources and produce a report of their data residency compliance. This policy will help us avoid fines by staying compliant with applicable laws and policies. The residency requirements are flagged by assigning appropriate glossary terms in the data catalogue to cloud resources. Residency term assignment is shown in Figure 25 below.

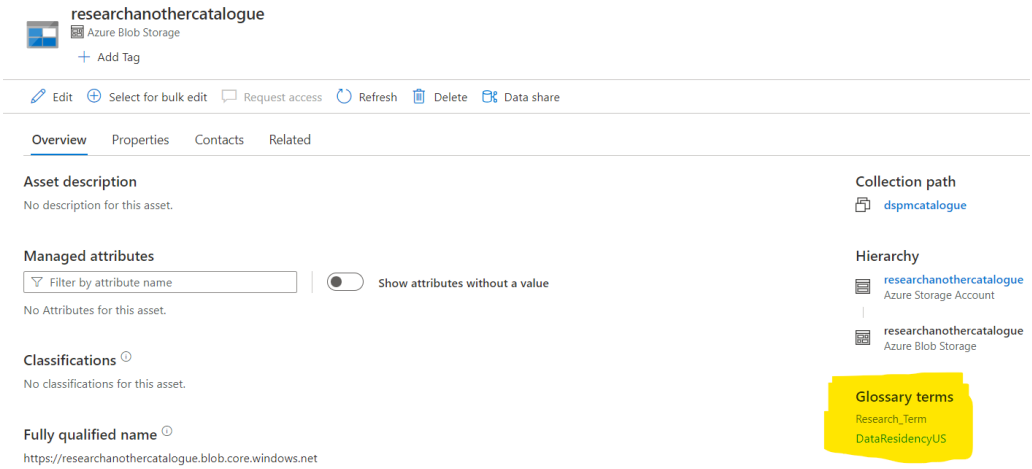


Figure 25 – Data Residency Value for Blob Storage

A combination of scripts is run in automated pipelines which query cloud for a list of all the resources and then query the data catalogue for their residency requirement. The residency requirement is then checked against the location of the resource and if the values match the resource gets a "pass" rating but if it does not match then the resource gets a "failure" rating. The pipeline run is shown in Figure 26 below:

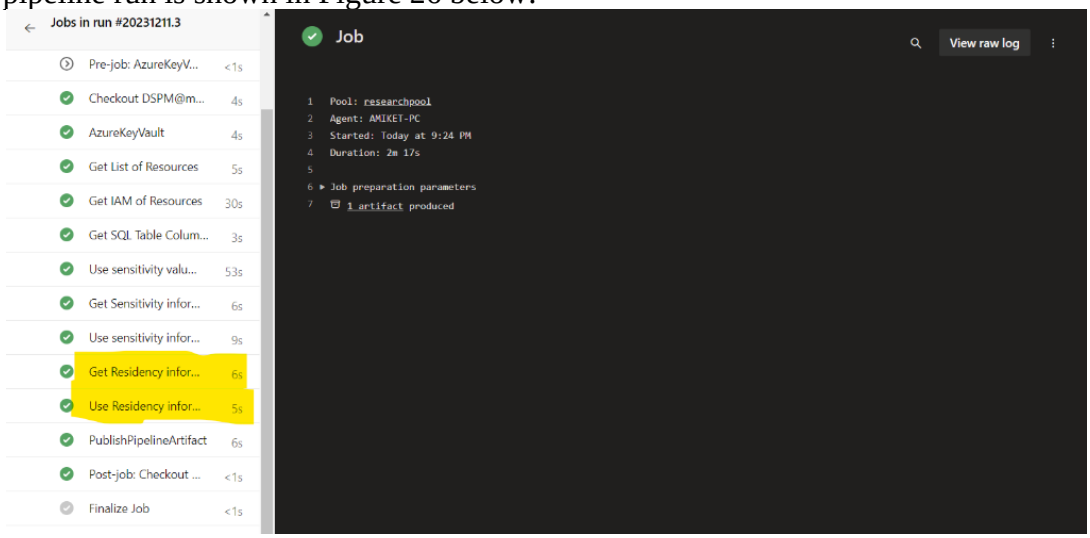


Figure 26 - Data Residency Audit Pipeline

The results are then published in a csv file as shown in Figure 27 and owners whose resources are failing the compliance with residency requirements are sent failure reports.

name	auditstatus
dspmresearch	Pass
researchtest	Pass
researchanothercatalogue	Fail
researchanothercatalogue	Pass
researchcatalogue	Pass
researchcatalogue	Pass

Figure 27 - Compliance Report for Data Residency Requirements

This result indicates that data residency audit policy was successfully implemented. Other observations from this experiment were that the solution is cost-effective and efficient with a run time of 15 seconds. The only downside to this process is that it is not fully automated as audit can be done automatically but fixing the issue needs manual effort from the owners.

6.7 Experiment 7: Set up Automated Pipelines using Self-Hosted Agent

The aim of this experiment is to install a self-hosted Azure DevOps agent onto our local machine and run automated pipelines using this agent. This will allow us to run 1 free parallel job in our Azure DevOps environment. We can use this job to run our automated pipelines for free and this will help bring down the overall costs associated with the DSPM system. The installed agent is shown in Figure 28 below:

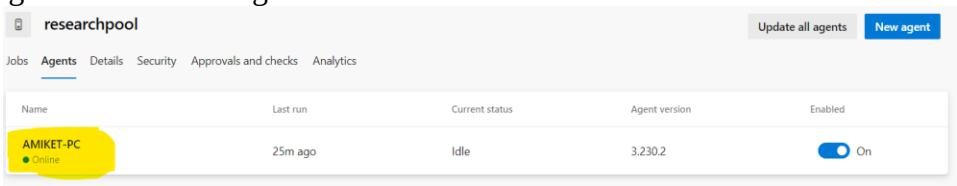


Figure 28 – Self-hosted Installed Agent Details

Once the agent was installed under the agent pool called "researchpool", we defined this agent pool explicitly in all the pipelines as shown in Figure 29. The pipelines ran successfully.

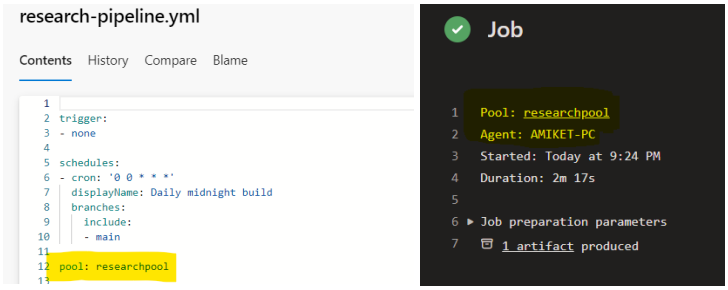


Figure 29 - Explicitly defining a Pool in a Pipeline and status of a Sample Run

This result indicates that our self-hosted agent was successfully installed and can run pipelines successfully. Other observations from this experiment were that one can use Azure DevOps orchestrator for free if they have a local machine to host the agent. This service is limited to a single free parallel job. For bigger environments, the cost of running multiple parallel jobs would be added to the DSPM system costs.

6.8 Result and Discussion

Through the various experiments described above, we were able to generate proof of concepts for different components of our DSPM system. A key observation is that the performance of the systems was consistent across experiments with high efficiency and cost-effectiveness. IBM's Security Guardium starts at $1000/month (IBM, Guardium Insights SaaS pricing for data security and compliance, 2023). Other solutions did not disclose their prices, but Figure 30 shows a comparison of the costing models across different DSPM solutions available in the market (Team, 2023).

Vendor	Pricing model
Laminar	Subscription based per asset
Dig security	Not disclosed
Polar security	Tiered based: offers three pricing tiers: free, premium, and enterprise.
Normalyze	Tiered based
Lookout	Subscription based
Nightfall	Pay-as-you-go
Securiti.ai	Tiered based

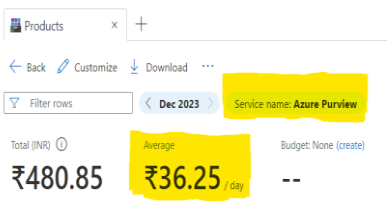


Figure 30 – (a) Comparison of Costing Models across different DSPM Solutions (Team, 2023)
(b) Average Cost per day of Data Catalogue Services

In comparison, our solution only uses the Data Catalogue services which cost us around ₹36.25/day ~ $0.43/day ~ $13/month as shown in Figure 31. This result indicates that our solution can provide a cheap and efficient alternative to the market.

7 Conclusion and Future Work

The aim of this research was to create a cost-friendly, efficient, data-centric, and cloud-based security posture management system which could plug the gap in data security in the industry. This research proposes a novel approach to fix the above problem by implementing a DSPM system with the help of data catalogue services, coding, visualisations, and pipeline automation. The key finding from the evaluation of the DSPM model is that system was cheap to implement and performed consistently and efficiently across the various experiments. The performance and efficiency can be attributed to the use of efficient code while the cost-effectiveness is a result of the installation and use of a self-hosted Azure DevOps agent so that automation pipelines could be run for free. The result shows promise for the development of a free and open-source DSPM model for the industry. It could gain popularity in the open-source community as the rival solutions in the market are costly to purchase and require large investments in the infrastructure to implement.

For future work, the research can be extended by automating the manual components of the DSPM model such as risk assessments and enforcement of data residency policy. It can even be evolved into a hybrid model to include some aspects of CSPM models which will enable the solution to be both cloud-centric and data-centric at the same time.

References

AlGhamdi, S., Win, K. T., & Vlahu-Gjorgievska, E. (2020). Information security governance challenges and critical success factors: Systematic review. *Elsevier* (p. 39). Australia: Elsevier.

Bento, P., Neto, M., & Côrte-Real, N. (2022). How data governance frameworks can leverage data-driven decision making: A sustainable approach for data governance in organizations. *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 01-05). Madrid: IEEE.

DIAO Zhe, W., & SU Naizheng, Z. (2017). *Study on Data Security Policy Based On Cloud Storage.* Beijing: International Conference on Cyber Resilience (ICCR).

Enriquez, R. (2021). *Cloud Security Posture Management (CSPM) in Azure.* Helsinki: Metropolia. Retrieved June 16, 2023, from https://www.theseus.fi/bitstream/handle/10024/504136/Cloud%20Security%20Posture%20Management.pdf?sequence=2&isAllowed=y

Europa. (2022, January 26). *Data protection under GDPR*. Retrieved July 29, 2023, from Europa: https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_en.htm

Fernandez, R. C., Abedjan, Z., Koko, F., Yuan, G., Madden, S., & Stonebraker, M. (2018). Aurum: A Data Discovery System. *IEEE* (p. 12). Paris: IEEE 34th International Conference on Data Engineering.

Gupta, Y., & Narayan, N. (2023). *Data Security in Cloud Computation.* Noida: 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

Hasan, M. Z., Hussain, Z. M., Mubarak, Z., Siddiqui, A. A., Qureshi, A. M., & Ismail, I. (2023). *Data security and Integrity in Cloud Computing.* Goa: International Conference for Advancement in Technology (ICONAT).

IBM. (2023, January 26). *Guardium Insights SaaS pricing for data security and compliance*. Retrieved December 11, 2023, from IBM: https://www.ibm.com/products/guardium-insights/pricing

IBM. (2023, January 26). *What is DSPM?* Retrieved December 10, 2023, from IBM: https://www.ibm.com/topics/data-security-posture-management

Joshi, A., Raturi, A., Kumar, S., Dumka, D., & Singh, D. P. (2022). *Improved Security and Privacy in Cloud Data Security and Privacy: Measures and Attacks.* Uttarakhand: International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP).

Khalid EI Makkaoui, A., Beni-Hssane, A., & Motamed, C. (2016). *Cloud security and privacy model for providing secure cloud services.* Marrakech: 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech).

Meisner, M. (2017). *Financial Consequences of Cyber Attacks leading to Data Breaches in Health Sectors* (6(3) ed.). Torun: Copernican Journal of Finance & Accounting.

Normalyze. (2023, January 26). *An Everything Guide to Data Security Posture Management (DSPM)*. Retrieved December 10, 2023, from Normalyze: https://normalyze.ai/what-is-dspm/

Petrosyan, A. (2022, September 6). *Average cost of a data breach worldwide as of 2022, by country or region*. (Statista) Retrieved July 31, 2023, from Statista: https://www.statista.com/statistics/463714/cost-data-breach-by-country-or-region/

Petrosyan, A. (2023, August 29). *Annual number of data compromises and individuals impacted in the United States from 2005 to 2022.* Retrieved July 31, 2023, from Statista: https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/

Rahimi, H., Maimaiti, T., & Zincir-Heywood, A. N. (2014). A Case Study for a Secure and Robust Geo-fencing. *IEEE* (p. 8). Krakow: 2014 IEEE Network Operations and Management Symposium (NOMS).

Rodrigues, D., Almeida, M., Guimarães, P., & Santos, M. Y. (2022). DataHub and Apache Atlas: A Comparative Analysis of Data Catalog Tools. *CAPSI* (p. 21). Portugal: CAPSI 2022 Proceedings.

Saed, K. A., Aziz, N., Ramadhani, A. W., & Hassan, N. H. (2018). *Data Governance Cloud Security Assessment at Data Center.* Kuala Lumpur: 4th International Conference on Computer and Information Sciences (ICCOINS).

Sawhney, G., Kaur, G., & Deorari, R. (2022). CSPM: A secure Cloud Computing Performance Management Model. *IEEE* (p. 5). Dubai: International Conference on Cyber Resilience (ICCR).

Shaikh, R., & Sasikumar, D. (2015). Data Classification for achieving Security in cloud computing. *Elsevier* (p. 6). Mumbai: ICACTA 2015: International Conference on Advanced Computing Technologies and Applications.

Team, L. S. (2023, January 26). *How to choose the best DSPM solution for your organization: comparison of features, benefits, and pricing models of different DSPM vendors*. Retrieved December 11, 2023, from Laminar Security: https://laminarsecurity.com/blog/how-to-choose-a-dspm-solution-vendor-comparison-guide/

Terzo, O., Ruiu, P., Bucci, E., & Xhafa, F. (2013). Data as a Service (DaaS) for Sharing and Processing of Large Data Collections in the Cloud. *IEEE* (p. 6). Taichung: Seventh International Conference on Complex, Intelligent, and Software Intensive Systems.

Zscaler. (2023, January 26). *What Is Data Security Posture Management (DSPM)?* Retrieved December 10, 2023, from Zscaler: https://www.zscaler.com/zpedia/what-is-data-security-posture-management