National College of Ireland

Data Exposure Analysis of Misconfigured S3 Buckets: A Quantitative Approach

MSc Research Project
Cloud Computing

Franklin Ebuka Onyia
Student ID: 21221600

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Franklin Ebuka Onyia
Student ID:	21221600
Programme:	Cloud Computing
Year:	2023
Module:	MSc Research Project
Supervisor:	Rejwanul Haque
Submission Due Date:	14/12/2023
Project Title:	Data Exposure Analysis of Misconfigured S3 Buckets: A Quantitative Approach
Word Count:	5125
Page Count:	19

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:	Franklin Ebuka Onyia
Date:	29th January 2024

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):	

Data Exposure Analysis of Misconfigured S3 Buckets: A Quantitative Approach

Franklin Ebuka Onyia
21221600

Abstract

This research paper evaluates the risk of exposure of AWS S3 bucket as regards to poor configuration of its security setting. Collection of datasets (exposed AWS S3 Buckets) were done using Grayhat Warfare API, with focus on AWS S3 Bucket. This was followed by data cleaning and transformation using Python Pandas library. Visualization tools were utilized for data interpretation, while statistical methods were used to analyse the content type and exposure duration. Correlation, regression analysis, as well as heatmaps generation were done using Scikit-learn. Study, made sure it followed ethical standards by ensuring data accuracy and confidentiality. The study also focused on the general trends rather than datapoints.

Our findings showed that insecure buckets are numerous and remain a serious problem if close attention is not paid towards it. It also showed that many of the S3 Buckets held sensitive files of which greater quantity are "PDF" files that were left unsecured due to negligence. It further showed that many data were exposed for many days, while some of the files remained unchanged for thousands of days with regards to the data from the "last modified" date analysis. In addition, files size analysis with respect to time of exposure, showed that large files are not necessarily at risky when exposed over a long period of time, taken from the low negative correlation result between the file size and the exposure time analysis. However, this study reveals the need for improvement in cloud storage security.

Development of automatic tools or software for the identification and resolution of misconfigurations in cloud-based storage and the exploration of machine learning for predicting risk analysis is recommended for future work of this research work. Also, Businesses would thrive in this era of cloud technology growth if advanced cloud security solutions are built.

1 Introduction

Cloud storage systems such as the Amazon Simple storage service(S3) have made a great change in the data storage ways, creating scalable and efficient solutions for both small and medium-sized organizations or individuals. In addition, the scalability and flexibility of S3, most times comes at the cost of security as regards to when access controls are misconfigured or not properly configured Continella et al. (2018). This section would delve into the issues of misconfigured S3 buckets and their significance or implications.

1.1 Background

In the year 2006, Amazon S3(Simple Storage Service) was introduced,, It offers its users a storage like box called "Buckets", where the data that is contained in the buckets are called "Objects" *Amazon S3* (n.d.). Amazon S3, Currently, it has existed for more than a decade , but it has been challenged with numerous threat or data leaks, caused as a result of misconfiguration issues in access control Jäger (2021). Current studies, has shown that the effect of these threats are too many and severe , and ranges from normal leakage of data to a more serious issue such as malicious code injections Guffey and Li (2023).

1.2 Problem Definition

Misconfigurations of S3 buckets, makes open secret and confidential data to be exposed to the public, such data include confidential company documents, payment details and Personal identifiable information (PII) Continella et al. (2018). Cable et al. (2021) gave instance of an issue which arose from combinations of guessable bucket names and complex security configurations. Some of these variables aid unauthorized access, which therefore creates room for malicious activities to occur.

1.3 Aims and Objectives

During this research process, we would investigate and evaluate the issues that comes from misconfigured S3 buckets, with a focus on these major areas below:

1. The Character of Exposed Data: The primary objective is to know the nature of the compromised data, thus, this would provide insights into the weight of each exposure incident Continella et al. (2018).

2.Exposure Timeframe Analysis: To understand the effect of exposing data for a long time , which can possibly show a view of the extent of damage. Cable et al. (2021)

3. Recommendations and Best Practices: To provide comprehensive recommendations and best practices based on these research findings and evidence. This, in turn, would give both large and medium organizations useful guidance to address and prevent similar issues from occurring in the future.

1.4 Research Question(s)

1. What types of data or files are most left exposed due to misconfigured S3 (Simple Storage Service) buckets?

2. How serious are the consequences of exposing data for a long time in a misconfigured S3 Bucket?

3. Based on the findings, what recommendations and best practices can be proposed to prevent and address similar incidents in the future?

1.5 Rationale of Research

We are currently in a digital world , where cloud services are mostly used in sectors like finance and health care Jäger (2021). So therefore , there is need for these sectors data to be stored securely protected at all cost. This research would contribute the base knowledge of S3 buckets misconfigurations, this would help organizations and large enterprises to easily reduce risks.

1.6 Significance of Study

This research offers these values. First, it gives insight on the scale and nature of misconfiguration problem so that organizations and individuals would be able to see the need to ensure proper configuration having learned the risk involved. It can also provide methods or base frameworks that organizations can use to check and resolve vulnerabilities in their S3 buckets. The serious threats eating deep into cloud misconfigurations calls for proactive and continuous monitoring of systems, for instance, with recent innovations like CSBAuditor shows a detection rate of 98% for real-time misconfigurationsTorkura et al. (2021)

1.7 Summary of section

This section introduces the problems faced by cloud services, of which Misconfigured S3 buckets represent a significant part of vulnerability in the cloud storage space. Therefore, understanding the scope and implications of this issue and by deploying efficient tools, organizations or individuals can reduce risks, ensuring that cloud storage remains both efficient and secure.

The next section would explore the Literature review of existing academic works that are important to misconfigured S3 buckets and data exposure.

2 Related Work

This Literature review looked into various aspects of cloud storage security, with emphasis on S3 Bucket misconfigurations and their impact on data exposure. The related work would have references from other research works, it aims to improve the understanding of certain vulnerabilities in relation with S3 buckets, with great insight on cloud storage risks. This analysis would involve analysis of the time the data were exposed ,the nature of the exposed data and effective security measures, so as to provide answers to the research questions, as well as provide recommendations that would improve security practices for cloud storage.

2.1 Evolution and Significance of Cloud Storage and S3 Buckets

The coming of cloud storage, most especially Amazom AWS S3 Buckets, represents of a shift in storage of data and its management. These buckets, found in the Amazon Web service, offer scalable and flexible storage solutions. Furthermore, the use of AWS S3 buckets goes along side with its security issues, primarily from misconfigurations. Wood and Pereira (2011) and Galibus et al. (2016), noted the occurance of S3 buckets in data breaches, with emphasis on the need for accurate configuration and improved security practices. This section explored the path of cloud storage development, with a focus on the functionality and security implications of S3 buckets. A good understanding of the benefits of S3 buckets and their potential security risks is important for making cloud storage solutions better.

2.2 Incidence and Implications of Misconfigurations in Cloud Storage

Misconfigurations in cloud storage, especially the Amazon S3 buckets, has become a serious security concerns. Cable et al. (2021) and Continella et al. (2018) in their studies, showed that a combination of predictable naming conventions and wrong security configurations, lead to serious data breaches. These misconfigurations not only result in unauthorized data access but also in wider security implications for organizations relying on cloud storage. This analysis lay emphasis on the urgent need for improved security practices in managing cloud storage, with a focus on the right configuration and regular audits to reduce risks linked with such misconfigurations, thereby protecting data privacy and integrity.

2.3 Case Studies and Research on S3 Bucket Vulnerabilities

In cloud storage space, S3 buckets vulnerabilities have created serious risk. Research by Kolevski et al. (2021) and Chen et al. (2021), showed the extent of cloud data breaches. These breaches most times, involve personal and proprietary information, whose major source comes from misconfigurations and lack of security protocols. These scenario educates individuals on the need for strict security measures and careful management of cloud resources. The growth on dependence on cloud services proves that there is need for proper understanding of these vulnerabilities, and thus push for improved strategies to secure data integrity and confidentiality. The understanding of these vulnerabilities and their implications is important towards shaping the future of security frameworks in cloud storage, ensuring the protection of sensitive information as data increases generally in this digital world.

2.4 Strategies for Enhancing Security in Cloud Storage

Security improvement in cloud storage calls for various approaches with a combination of good protocol and industrial best practices. Emphasis on the importance of enterprise security in cloud computing was laid by Ramachandran and Chang (2014) , who supported strategies to improve trust and address security issues. They further recommended an in-depth analysis and modeling of cloud organizational security, with a focus on data and storage technologies. Tunc et al. (2017), in his work, he proposed a cloud security automation framework, which addressed the challenges of manual security configurations. Alongside, their methodology followed the compliance with NIST SP 800-53 security controls and continuous monitoring of cloud systems.

In addition to this framework,Alavizadeh et al. (2019) discussed an automated security analysis framework that would be handy in the detection and control of security weakness in cloud deployments. An et al. (2019) introduced a tool called Cloudsafe, which was used to automate security analysis. This tool emphasized the need for steady security evaluation in cloud computing. Furthermore, Baviskar (2022) explored automated encryption methods that was used to prevent S3 bucket exposure, with emphasis on the evolving nature of cloud security threats.

Chen et al. (2021) checked the effect of data breaches in organizations , after which he laid emphasis on the need for stronger security measures in cloud storage. This idea was supported by Galibus et al. (2016), who gave insights into cloud storage security concepts

and optimized practices. However, Improvement of cloud storage security continues to push further strongly towards addressing security issues as contributions from industrial and academic experts increases.

2.5 Advancements and Future Directions in Cloud Storage Security

Advancements in cloud storage security are of upmost importance towards addressing new challenges as technology growth continues. Cloudsafe, a tool for automating for automating cloud security analysis in cloud computing was illustrated by An et al. (2019), as a tool which improves security reports and assessments. This innovation plays an important role in understanding and reduction of vulnerabilities in cloud environments.

In Andrei-Cristian et al. (2021) study on Security vulnerabilities in cloud deployments, Iosif highlighted the need for proper security guidelines and the need for the use of raw data in the development of secured cloud systems. These findings would create room for continuous innovation in cloud security strategies that would match the rise cyber traits.

This section, however, looks into the technologies and methodologies in cloud storage security, taking a look on how they shape the future of cloud computing. It further looked into the existing literature and combined insights from various research works to provide a view of the current state and future prospects of cloud storage security. However, this details above would help to address the issues in these existing field of study.

2.6 Identifying Research Gaps and Future Research Directions

Asides various research carried out on the field of cloud security, I Identified some gaps in quantitative analysis, with emphasis on data exposure due to misconfigured S3 buckets. Previous studies have emphasized on qualitative assessments of cloud security issues. In attempt to address this gap, this project practically evaluated the nature, scope, and duration of data exposure incidents that are specific to S3 bucket misconfigurations. This is unique because it concentrates on a small, randomly selected sample of open S3 buckets gotten from Grayhat Warfare. This method was chosen because of the limitation found in other methods as this approach promises a depth analysis. This quantitative approach would be used to identify the characteristics and exposure duration of data within these buckets. It would also provide a detailed view on cloud storage vulnerabilities. Therefore, the outcome would contribute to practical and academic fields because it offers a quantifiable insight and recommendation, that would be used to address some challenges faced by S3 buckets.

2.7 Section Summary

This literature review above, carefully shows various aspects of cloud storage security, with specific emphasis on the vulnerabilities of Amazon S3 buckets. The review first looked at the evolution and significance of cloud storage, then, it reviewed the incidence and implications of misconfigurations in general, with emphasis on its impact on data security. It went further to highlight a real-world effect of S3 Buckets vulnerabilities. In the quest to improve cloud storage security for future improvement, innovative and current practices would be combined in other to get a good result. Having identified the research gap, this review did not do enough work on the quantitative analysis for

understanding and mitigating the risks that comes with S3 Buckets. This review provides base foundation for the development of healthy security measures in cloud storage, with emphasis on the need to be at alert in this era of continued digital threats.

3 Methodology

This section carefully states the methodology used for the analysis of the misconfigured Amazon Web Services (AWS) S3 buckets. The approach is designed to provide a framework to help us understand the issues of cloud storage security. The methodology includes theoretical concept as noted by Bendat and Piersol (1987) to ensure an accurate result. The methodology would be used to addressing the objective of this study, which includes the identification of common misconfigurations in S3 buckets, assessing the potential risks that is associated with these misconfigurations, and providing possible solutions to reduce these risks.

3.1 Collection of Data

This is the first step in our research as it involves the collection of data. Grayhat Warfare API was used for the collection of datasets (AWS open Buckets) from a pool of large datasets. Grayhat Warfare tool was chosen as a tool for our data collection because the tool makes sure we have varieties of data that represent various Misconfigured AWS S3 buckets that would be used for the analysis Sawant and Bacchelli (2015).Also this API was picked because of its ability to gather real-world data that would be a representative of the current practice and configurations in cloud storage.

3.2 Data Cleaning and Transformation

This is the next stage of the data processing process after the data collection, during which the data undergoes some cleaning and transformation process. During this stage, the data which is in JSON format is converted to a CSV format, as this format would aid the handling and analysis of the dataset Broman and Woo (2018). Python's Pandas Library was used for manipulation of the dataset. Furthermore, Data transformation process, which includes standardization and extraction of various attribute that would aid the analysis.

3.3 Analysis of Data

Next is the analysis of the data. Statistical methods were used in the analysis process, where Matplotlib and Seaborn from the python library were important in the visualization and interpretation of complex dataset Embarak (2018). In addition, Python's Scikit-learn library helped to aid the correlation and regression analysis, it further gave insight into the relationship between the attribute of the data Massaron and Boschetti (2016).

3.4 Testing and Evaluation

Evaluation and testing of is an important part of this process, because it helps us to validate the outcome or findings. The analysis continuously repeated the test validation to confirm the accuracy of the results. Such testing process involves the cross-checking

of results to verify the accuracy and reliability which is in tune with statistical methods as said by Ziegel and Ott (1977).

3.5 Ethical Considerations

Ethical consideration is very important in any research work, and they were strictly followed to ensure confidentiality and integrity of data used throughout this project. Also, this study carefully considered the ethics, while working with real-world data, paying attention to the general trends rather than the individual data so as to maintain confidentiality of data Chu et al. (2016).

3.6 Summary of the Section

This section outlined the procedure followed in this research. This procedure would ensure smooth analysis that would make way towards the understanding of security issues that comes with cloud storage, most especially in S3 buckets. This methodology, follows the best practice for data analysis according to Walker (2020);Alavizadeh et al. (2019).

4 Design Specification

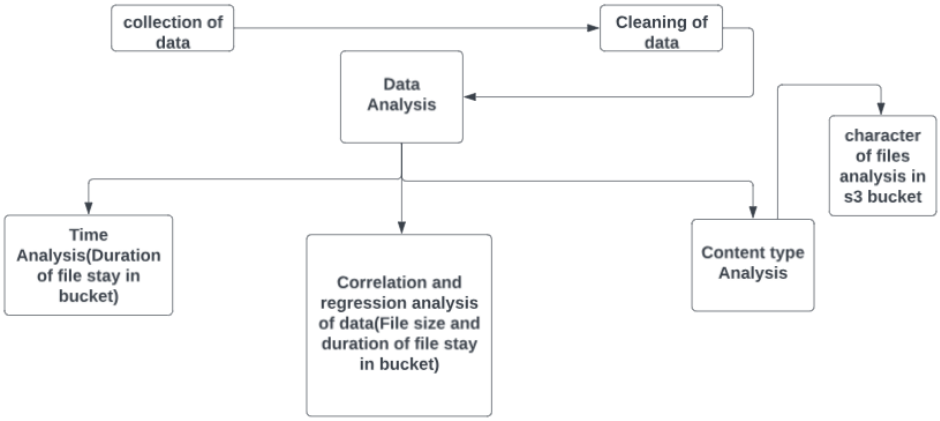

Figure 1: Design process Diagram

This section describes the design specifications used for the data analysis process, with emphasis on the analysis of the exposed misconfigured S3 buckets. It also gives details about the methodologies and frameworks that were used in this work. Thus, this would lead us to good understanding of this implementation stage.

4.1 Collection and Storage of data

This work started with the collection of data from Grayhat Warfare API. We took this approach in other to have access to large dataset that represents the present state of S3 Bucket configuration. Furthermore, the API was configured to filter all AWS buckets where the limit set to 1000 results. This data collection phase was important to make sure we have a quality dataset that would be used for analysis.

Furthermore, after data collection, the data was stored in JSON format. This is because the JSON format aligns with different data types and structures. Thereafter, the data was converted to a CSV format so it can be easily used during analysis of data because we worked with a large dataset. CSV files were used because it is a generally accepted standard format for data storage, such as in data analysis works Broman and Woo (2018).

4.2 Cleaning and transformation of data

Data Cleaning and transformation is important for insightful analysis. This transformation and cleaning process was done using the Python Pandas Library, as it is well known for its simple way of data manipulation ability Molin and Jee (2021). The following steps were taken in the transformation process:

• Conversion of Last modified timestamps from Unix format to human-readable date format. In addition, this process includes extracting file extensions from filenames, this is important to help categorize and analyze data based on file types.

4.3 Techniques Used for the data Analysis

The analysis technique utilized statistical methods to analyse the data.

It starts with the content type analysis, which involved grouping the data by its file extension and their frequencies were calculated. This was done in order to get the most common types of file in the dataset. This procedure provided insight into the character of data that is commonly stored in the Misconfigured S3 Buckets.

Next, is the time analysis. Statistical calculations such as mode, mean and median were done for this calculation as it provided a quantitative measure of time that period that elapsed since the last modification of the files. This gives an insight of the total time period these files has been exposed in the misconfigured S3 buckets

Also, tools such as the pythons's Matplotlib and Seaborn were important , because it helped to facilitate the visualization of our findings in the form of scatter plots, histogram and boxplots, these visual forms shows both the distribution and relationships within the data Embarak (2018).

4.4 Correlation and Regression Analysis

Again, Correlation and Regression statistical analysis was used to get the relationship within the data. First, the Pearson's correlation was used to check the relationship between the file size and total time that was accumulated since the last modification of file. Later, Regression analysis (Linear Regression Model) was used to predict the exposure risk factors, based on these variables (Lemenkova, 2019; Massaron and Boschetti, 2016). Also, calculations and data handling were carried out using Pandas and Numpy which are found in Python Scikit-learn library.

4.5 Requirements

Requirements for this work include a Python environment with the support of Google Colab for easy view and calculations. It also required the use of Python libraries such as Scikit-learn and Numpy which were used for the analysis. In addition, Pandas were used for data wrangling, as well as Matplotlib and Seaborn for visualization.

4.6 Summary

The design specifications used in this work were aligned with the project's objectives. Also, the tools selected would be helpful in the detection of the character and exposure risk associated with this ill configured S3 buckets. Next is the implementation state, where these tools, specifications and methodologies were combined to achieve a successful implementation.

5 Implementation

Implementation stage is an important stage in this research as it brought to reality, the real design, and project's objectives. In Jäger (2021), he noted that accurate implementation of cloud storage analysis is important in fishing out security lapses. Therefore, this section implements the intended design with respect to its goals and objectives.

5.1 Preparation of data

Data preparation involves getting the data ready for analysis. During this process, our data is converted from Unix timestamps to a readable format as recommended byChu et al. (2016). Thereafter, the standardization process followed next, after which the file was extracted to help make the analysis process easy as noted by Broman and Woo (2018).

5.2 Character of Exposed Data Analysis Implementation

During the content type analysis, the files that were exposed in the misconfigured buckets were categorized by their file extension after which the frequency of these file types were calculated as guided by McKinney (2017). This approach helped us to note the most common file type that were exposed in the Misconfigured S3 buckets. Again, the bucket naming analysis was carried out manually to ensure accuracy of the analysis, with respect to the project scope and dataset. This bucket naming analysis were carried out to specifically rate the guessability of the bucket names on a scale of 1 to 10, indicating categories that go by a generic name to those with a common name, thus suggesting high sensitive data or highly guessable data for those with a common name. Output from this analysis was presented in the evaluation section, which gave us insight on the exposure patterns.

5.3 Exposure Timeframe Analysis Implementation

During the timeframe Analysis, the "Last Modified" dates of file, played an important role as it was utilized to calculate the exposure durations of these files, and this estimation of time of exposure involved a statistical methods that aligned with Ziegel and Ott (1977) recommendations. Embarak (2018) says that Time series plot is an important tool that is utilized in visualization, thus, this was used to find out temporal trends in file modification dates.

5.4 Size Analysis Implementation

The file size analysis was carried out to identify how the file size relates to exposure risk specifically identifying the amount of the time that has elapsed since the oldest modification which was performed on a file within a specific folder or directory. This analysis would add towards understanding risk factors in cloud storage security."

5.5 Statistical Analysis Implementation

A combination of statistical analysis which includes correlation and regression analysis were very important in this research as these methods provided a base for this research to check the relationships between the file characteristics, of which Lemenkova (2019) explained the theory behind the application of correlation and regression analysis. Thus, this relationship between files, would help us predict exposure risk factors.

5.6 Recommendations and Best Practices Development

For this research work, our recommendation of best practice was based on our findings.Our finding might include knowing a common threat patterns as suggested by Continella et al. (2018) and Torkura et al. (2021). Based on this, we would go further to make recommendations and best practices.

5.7 Summary of the Section

This section successfully narrated the implementation stage of this research. This would not have been possible without the use of the tools and methodology described above. Next, we discuss the evaluation, findings from our implementation.

6 Results and Evaluation

This section evaluates research's results, with emphasis on the misconfigurations in S3 Buckets that lead to the exposure of data. It includes insight from existing literature, with emphasis on the character of exposed data, the duration of exposure and the effectiveness of the proposed recommendations and best practices. This analysis findings and results were compared with previous research to make clear the studies contributions to understanding and reduction of cloud storage threats.

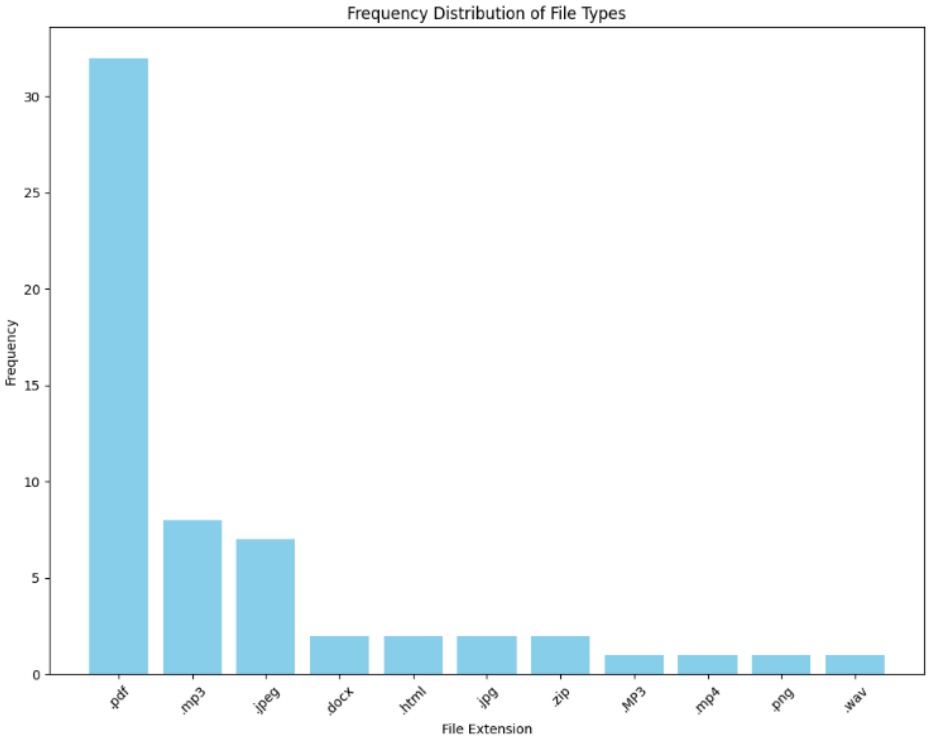

Figure 2: Frequency Distribution of File Types

Bucket Name	File Names	Bucket Naming Pattern	Bucket Guessability Score	Bucket Reason for Score	File Naming Pattern	File Guessability Score	File Reason for Score
mg-cdn.s3.amazonaws.com	index.html	Generic	3	Generic name	Common	7	Common word
mg-cdn.s3.amazonaws.com	login.html	Generic	3	Generic name	Common	7	Common word
bucket-01.s3.amazonaws.com	Annual-Report-2016.pdf	Generic	3	No specific pattern identified	Organizational data	4	Presence of "Annual-Report" suggests organizational data

Figure 3: A Subset of the Result from the Bucket Naming Analysis

6.1 Character of Exposed Data Analysis Implementation

The content type analysis, illustrated in the histogram (see Figure 2), shows that 'pdf' files are mostly present within misconfigured S3 buckets, giving insight that documents

containing sensitive data, are the most frequent file type exposed. In comparison, multimedia files like 'mp3' and 'jpeg' are less seen in the misconfigured S3 Buckets. Most buckets in the analysis were named in a simple way, of which most people might not guess easily, showing a risk that is not fully recognized or a risk that is hard to crack (see fig 3). This situation, where 'pdf' files are common but the bucket names are plain , shows a gap between what people think is safe and the real risk, underlining the need for better security measures and awareness. This outcome meets the study's goal to find out what kind of data gets mostly exposed, indicating that 'pdf' files, even though they might look less likely to be guessed, are often part of big security mistakes.

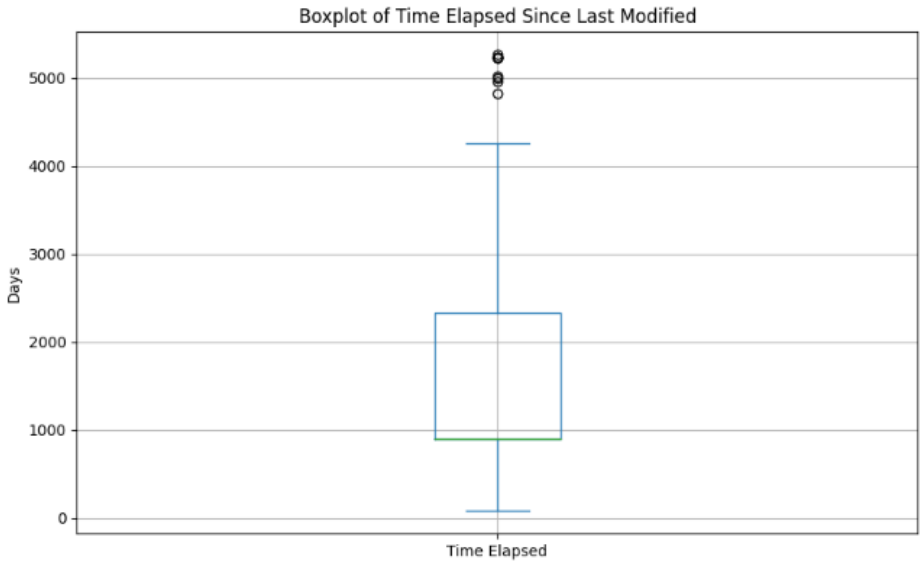

Figure 4: Boxplot of Time Elapsed Since Last Modified

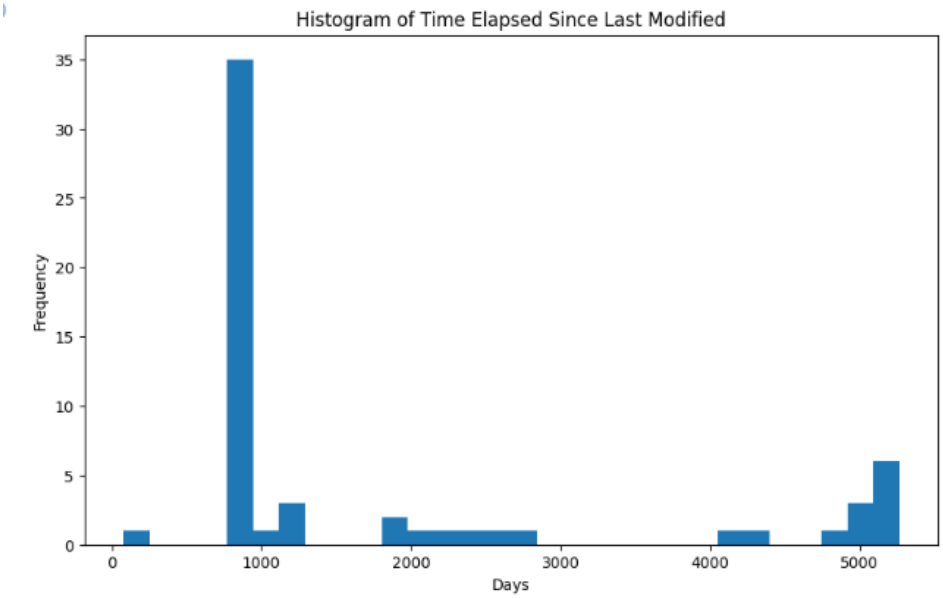

Figure 5: Boxplot of Time Elapsed Since Last Modified

Statistic	Time Elapsed
Mean	1893 days 13:38:09.853033216
Median	898 days 22:07:52.497100992
Mode	77 days 11:00:23.497101

Figure 6: Last Modified Date Analysis

6.2 Exposure Timeframe Analysis Execution

During the exposure Timeframe analysis, The 'Last Modified' date column were analyzed and the study showed some critical insights. The boxplot in fig 4 and histogram in fig 5 presented above, shows an abnormal distribution with a long tail, suggest that a majority of files were modified of recently,while few of the files, remain unmodified for a long period of time, exceeding thousands of days. This difference shows that a handful of data are constantly at risk, showing the importance of regular updates and monitoring as recommended by Torkura et al. (2021), Guffey and Li (2023). These findings shows the importance of proper auditing, which is inline with the study's goal to understand the impact of longtime exposure of data in data security.

Statistic	Value
Count	59
Mean	19,496,090 bytes
Std Dev	64,747,520 bytes
Min	6 bytes
25%	561,464.5 bytes
50% (Median)	1,899,102 bytes
75%	8,026,673 bytes
Max	372,884,200 bytes

Figure 7: Size Analysis (Correlation coefficient)

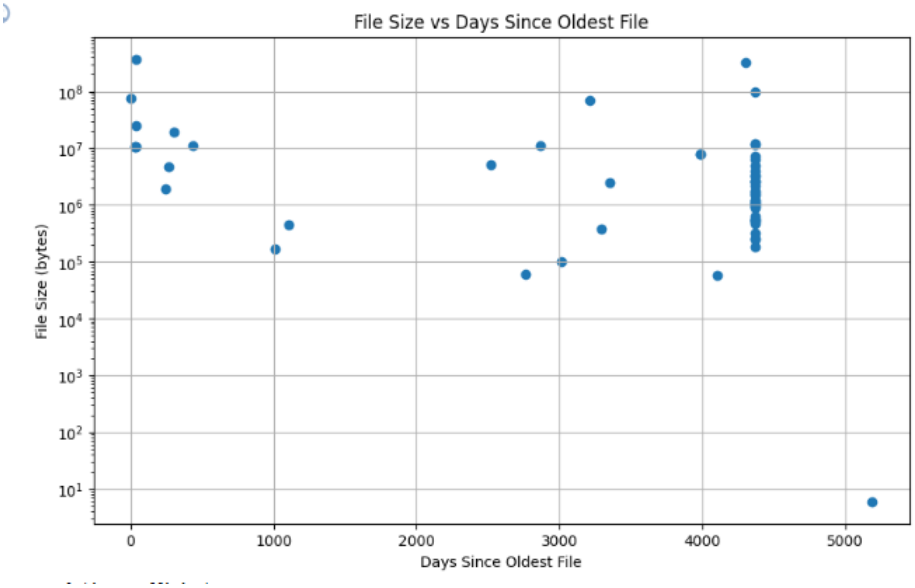

Figure 8: File Size vs Days Since Oldest File

6.3 Size Analysis

Size analysis, the third objective which captured 'file size' to check if it is a factor responsible for exposure risk, considering file sizes of systems and time of its oldest file modification.

The size analysis shown in Figure in fig 8 and Table in fig 7 shows a negative correlation between file size and exposure time, showing that larger files may not necessarily be at higher risk of long-term exposure, as the correlation coefficient shows a negative value of -0.21622658195101557. This finding goes against common assumptions,There by showing the need for proper understanding of risk factors in cloud storage security. In addition, this analysis and findings , aligns with the conclusion of Torkura et al. (2021) on the importance of continuous threat detection. Therefore this addresses the research question by showing that file size alone does not predict exposure risk. So therefore , this would add a little contribution to the development of targeted security measures.

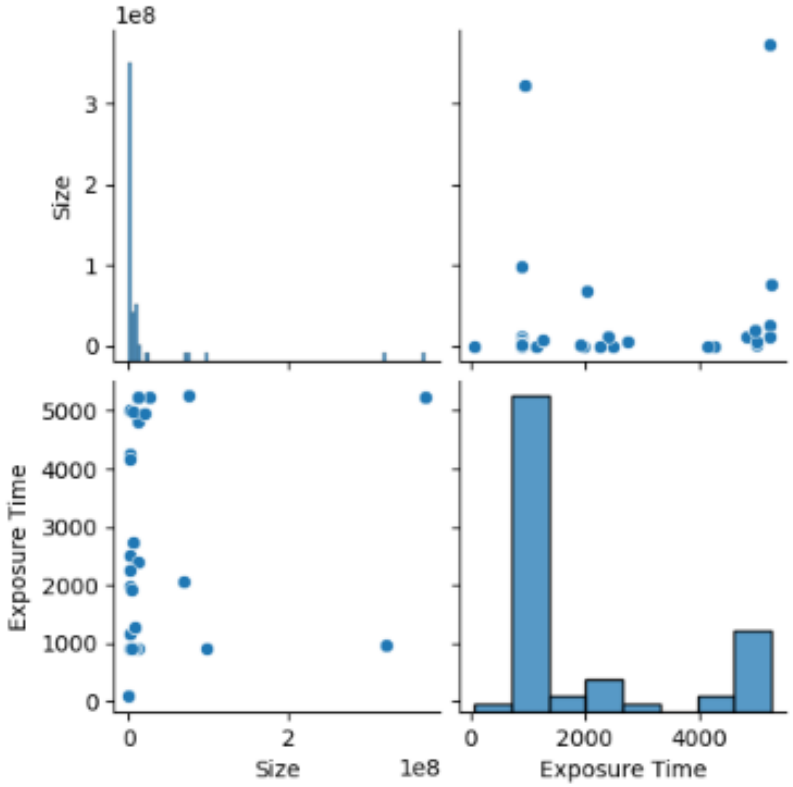

Figure 9: Size Analysis (Correlation Analysis)

	File ID	Bucket ID	Size	Exposure Time
count	59.0000	59.0000	5.9000e+01	59.0000
mean	648.711864	55.966102	1.949609e+07	1892.728814
std	99.81155	0.182521	6.474752e+07	1644.540370
min	127.0000	55.0000	6.000000e+00	77.0000
25%	651.5000	56.0000	5.614645e+05	898.0000
50%	666.0000	56.0000	1.899102e+06	898.0000
75%	680.5000	56.0000	8.026673e+06	2328.0000
max	695.0000	56.0000	3.728842e+08	5265.0000

Figure 10: Correlation analysis

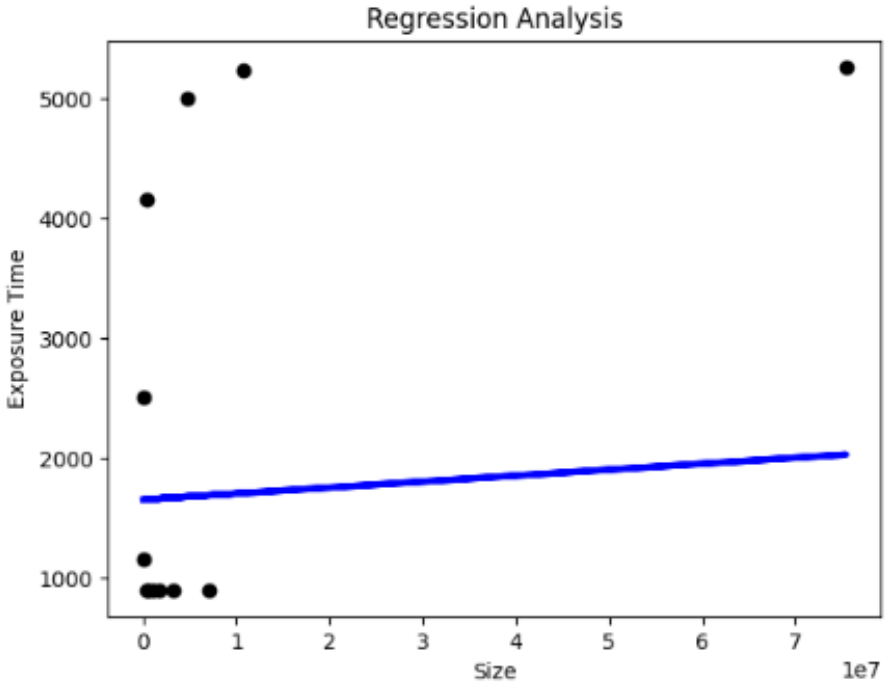

Figure 11: File Size vs Days Since Oldest File

	File ID	Bucket ID	Size	Exposure Time
File ID	1.000000	0.986559	0.086584	0.252295
Bucket ID	0.986559	1.000000	0.056804	0.146326
Size	0.086584	0.056804	1.000000	0.216247
Exposure Time	0.252295	0.146326	0.216247	1.000000

Figure 12: Regression Analysis

6.4 Statistical Analysis

Here, Correlation and Regression statistical method were used to ascertain the association and relationship between the file characteristics and the exposure risk. The analysis, showed a non-linear relationship between the file characteristics and exposure risk as shown in the negative correlation table in Fig 10, and Figure 9 and the regression analysis in Table 12 and Figure 11, shows that Larger file are not liable to long time exposure risk. Also, the regression model's negative R-squared value (-0.0857110992876181) and high RMSE (1929.936508414454) value, shows that file size is not the only criteria that determines risk, as this is inline with the research from Cable et al. (2021);Continella et al. (2018); Guffey and Li (2023)).Therefore , this calls for the need to use various ways or approach to reduce risk in Misconfigured S3 buckets

6.4.1 Summary of the Section

This section looked into S3 bucket misconfigurations and their impact on data exposure. It contrasts empirical findings with existing literature, suggesting a need for improved security approaches. Next Section states possible recommendation strategies

7 Conclusion and Future Work

Finally, this research explored the security concerns related to AWS S3 buckets due to Misconfiguration Issues, focusing on implications of these misconfigurations. The aims and objectives were to identify common misconfigurations, assess related risks, and suggest recommendations based on our findings. Through a systematic method, data collection,data transformation and analysis, the research was able to highlight the key aspects of S3 bucket vulnerabilities, giving insights into cloud storage security. Furthermore, we came up with some recommendations and the best practices based on our research findings. The recommendations are stated below:

1. **Regular Auditing and Monitoring**: A steady auditing and monitoring, as noted by Torkura et al. (2021), are required for identifying and resolving misconfiguration issues in S3 buckets.

2. **Proper Security Training**: There is need to educate individuals and staff in organization on best security practices, as noted by Guffey and Li (2023). This should include training them on how to create less guessable bucket names and sensitize them on the reason behind this approach, most especially in the creation of "PDF" file format as shown by this research .

3. **Implementation of Advanced Security Tools**: It is recommended to use advanced tools to detect threat as this can reduce most risk related issues with large file that has been exposed for along time frame.

4. **Multi-Layered Security Approach**: It is also recommended to adopt strict security measures that would include physical, network, and application security layers, to guard against these threats.

5. **Regular Policy Review**: It is also recommended that a continuous update and review of access policy would reduce these increasing security threat.

The study's scope focused primarily on S3 bucket misconfigurations. It suggests the need for wide research in areas of cloud services and configurations. In future, I recommend a look into strong and reliable tools in real-world scenarios or the use of machine learning for preventing and predicting the exposure of data in cloud storage space.

In conclusion, this research has carefully addressed the research questions with respect to the nature of data exposed due to S3 bucket misconfigurations, the relation between exposure duration and its consequences. It has also provided a set of recommendations. Overall, this would serve as a stepping stone for individuals and organization , towards improving cloud storage security in general in other to reduce the exposure risk.

References

Alavizadeh, H., Alavizadeh, H., Kim, D. S., Jang-Jaccard, J. and Torshiz, M. N. (2019). An automated security analysis framework and implementation for cloud.

Amazon S3 (n.d.). `https://aws.amazon.com/s3/`. Accessed: 2023-11-22.

An, S., Eom, T., Park, J. S., Hong, J. B., Nhlabatsi, A., Fetais, N., Khan, K. M. and Kim, D. S. (2019). Cloudsafe: A tool for an automated security analysis for cloud computing.

Andrei-Cristian, I., Gasiba, T. E., Zhao, T., Lechner, U. and Pinto-Albuquerque, M. (2021). A large-scale study on the security vulnerabilities of cloud deployments, *International Conference on Ubiquitous Security*.
URL: *https://api.semanticscholar.org/CorpusID:249047978*

Baviskar, C. R. (2022). Cloud based automated encryption approach to prevent s3 bucket leakage using aws lambda.

Bendat, J. S. and Piersol, A. G. (1987). Random data: Analysis and measurement procedures.
URL: *https://api.semanticscholar.org/CorpusID:109797040*

Broman, K. W. and Woo, K. H. (2018). Data organization in spreadsheets, *The American Statistician* **72**(1): 2–10.
URL: *https://doi.org/10.1080/00031305.2017.1375989*

Cable, J., Gregory, D., Izhikevich, L. and Durumeric, Z. (2021). Stratosphere: Finding vulnerable cloud storage buckets, pp. 399–411.

Chen, D., Chowdhury, M. M. and Latif, S. (2021). Data breaches in corporate setting, *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 01–06.

Chu, X., Ilyas, I. F., Krishnan, S. and Wang, J. (2016). Data cleaning: Overview and emerging challenges, *Proceedings of the 2016 International Conference on Management of Data* .
URL: *https://api.semanticscholar.org/CorpusID:11192413*

Continella, A., Polino, M., Pogliani, M. and Zanero, S. (2018). There's a hole in that bucket!: A large-scale analysis of misconfigured s3 buckets, *Proceedings of the 34th Annual Computer Security Applications Conference* .
URL: *https://api.semanticscholar.org/CorpusID:54445722*

Embarak, O. H. (2018). Data analysis and visualization using python: Analyze data to create visualizations for bi systems.
URL: *https://api.semanticscholar.org/CorpusID:219960700*

Galibus, T., Krasnoproshin, V. V., Albuquerque, R. d. O. and de Freitas, E. P. (2016). *Elements of Cloud Storage Security: Concepts, Designs and Optimized Practices*, 1st edn, Springer Publishing Company, Incorporated.

Guffey, J. and Li, Y. (2023). Cloud service misconfigurations: Emerging threats, enterprise data breaches and solutions, pp. 0806–0812.

Jäger, A. (2021). Finding and evaluating the effects of improper access control in the cloud.
URL: *https://api.semanticscholar.org/CorpusID:263709007*

Kolevski, D., Michael, K., Abbas, R. and Freeman, M. B. (2021). Cloud data breach disclosures: the consumer and their personally identifiable information (pii)?, *2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW)* pp. 1–9.
URL: *https://api.semanticscholar.org/CorpusID:237518916*

Lemenkova, P. (2019). Computing and plotting correlograms by python and r libraries for correlation analysis of the environmental data in marine geomorphology, *Jeomorfolojik Araştırmalar Dergisi* **3**: 1–16.

Massaron, L. and Boschetti, A. (2016). Regression analysis with python.
URL: *https://api.semanticscholar.org/CorpusID:62124581*

McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd edn, O'Reilly Media, Inc.

Molin, S. and Jee, K. (2021).

Ramachandran, M. and Chang, V. (2014). Recommendations and best practices for cloud enterprise security, *Proceedings of the 2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, CLOUDCOM '14, IEEE Computer Society, USA, p. 983–988.
URL: *https://doi.org/10.1109/CloudCom.2014.105*

Sawant, A. A. and Bacchelli, A. (2015). A dataset for api usage, *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories* pp. 506–509.
URL: *https://api.semanticscholar.org/CorpusID:7685675*

Torkura, K., Sukmana, M. I., Cheng, F. and Meinel, C. (2021). Continuous auditing and threat detection in multi-cloud infrastructure, *Computers Security* **102**: 102124.
URL: *https://www.sciencedirect.com/science/article/pii/S0167404820303977*

Tunc, C., Hariri, S., Merzouki, M., Mahmoudi, C., Vaulx, F., Chbili, J., Bohn, R. and Battou, A. (2017). Cloud security automation framework, *Proceedings - 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems, FAS*W 2017*, Proceedings - 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems, FAS*W 2017, Institute of Electrical and Electronics Engineers Inc., pp. 307–312.

Walker, M. (2020). *Python Data Cleaning Cookbook: Modern techniques and Python tools to detect and remove dirty data and extract key insights.*., Packt Publishing Ltd.

Wood, K. and Pereira, E. G. (2011). Impact of misconfiguration in cloud – investigation into security challenges.
URL: *https://api.semanticscholar.org/CorpusID:56357207*

Ziegel, E. R. and Ott, L. (1977). An introduction to statistical methods and data analysis.
URL: *https://api.semanticscholar.org/CorpusID:123347451*