

Alleviating Cloud Forensic Backlog Using Machine Learning Models



By
Ummer Farooq
(Registration No: 00000318257)

Supervisor
Dr. Arslan Shaukat

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD, PAKISTAN

September 2023

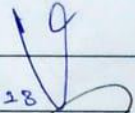
THESIS ACCEPTANCE CERTIFICATE

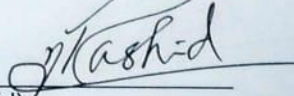
Certified that final copy of MS/MPhil thesis written by **NS Ummer Farooq** Registration No. 00000318257, of College of E&ME has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the thesis.

Signature : 

Name of Supervisor: Dr Arslan Shaukat

Date: 19-09-2023

Signature of HOD: 
(Dr Usman Qamar)
Date: 19-09-2023

Signature of Dean: 
(Brig Dr Nasir Rashid)
Date: 19 SEP 2023

Boundless gratitude to my parents, whose unwavering love and support have been my guiding light throughout this journey. And to my adored siblings, whose constant encouragement and unwavering belief in me have helped me in my determination to succeed.

And to my esteemed Thesis Supervisor, your mentorship, wisdom, and unwavering guidance have been instrumental in shaping this thesis.

This work is dedicated to all of you, for without your support, this achievement would not have been possible. Thank you for your guidance and for being my pillars of strength.

ACKNOWLEDGEMENTS

I am extremely thankful to ALLAH Almighty for his continuous blessings throughout my work. It was quite an arduous exercise that could not have been completed without the help of God Almighty and the strength that he bestowed upon me.

I would like to express my heartfelt gratitude to my thesis supervisor, Dr. Arslan Shaukat, for their guidance, invaluable insights, and patient encouragement throughout this research. Your mentorship has been instrumental in shaping my academic and research skills.

I am also thankful to other thesis committee members Dr. Wasi Haider Butt and Dr. Ali Hassan, for their expertise and feedback that enriched the quality of this work. Their constructive criticism and valuable suggestions have been pivotal in refining my ideas and methodologies.

To my dear friends and colleagues, thank you for being a source of inspiration and motivation. Your discussions, camaraderie, and shared experiences have made this journey both intellectually stimulating and enjoyable.

Lastly, I owe a debt of gratitude to my family, especially my parents, for their unconditional love, support, and encouragement throughout my academic pursuits. Your belief in me has been my greatest strength.

This thesis is a product of the collective efforts of these individuals, and I am deeply thankful for their contributions. While it is impossible to name everyone who has played a part, please know that your influence has not gone unnoticed or unappreciated.

ABSTRACT

Cloud computing has revolutionized the way data is stored and managed, providing unparalleled scalability and accessibility. However, the rapid adoption of cloud services has led to an escalating challenge in digital forensic investigations, resulting in a considerable backlog of cases. In response to this critical issue, cloud forensic constraints are defined, and then a Cloud Forensic Framework is designed to alleviate the burden of this forensic backlog. Building upon cloud constraints and cloud forensic framework, a streamlined Cloud Forensic Process Flow is established. To address the issue of data duplication that contributes to the forensic backlog, we reduce it by using hashing. By doing so it optimizes storage utilization and minimizes redundancy, thereby expediting investigation processes. And in the context of fraud detection within cloud-stored email data, we focus on relevant data extraction and prioritization. Our framework offers an approach to identify pertinent information efficiently, enhancing effectiveness of subsequent analysis. Specifically, we employ Topic Modelling using Latent Dirichlet Allocation (LDA) for the detection of fraudulent emails, facilitating rapid fraud identification. To further augment information extraction, Named Entity Recognition (NER) powered by BERT is employed to identify entities of interest from the email text data. Additionally, we described Relation Extraction at the end to uncover connections between entities, aiding in the identification of different named entities and relations between them to help in our investigative purposes. The results of our experimentation found out that BERT model gives exceptional results as compared to rule-based approach and CRF model. Further it is revealed that by using data deduplication, by using relevant data extraction, and prioritization within our Forensic Framework significantly reduces investigation time, storage, and backlog.

Keywords: *Cloud Forensics, Cloud Forensic Constraints, Cloud Forensic Framework, Cloud Forensic Backlog, Topic Modelling, Latent Dirichlet Allocation, Named Entity Recognition, Relation Extraction, BERT Model.*

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	i
ABSTRACT	ii
TABLE OF CONTENTS	iii
LIST OF FIGURES	vii
LIST OF TABLES	ix
LIST OF ABBREVIATIONS	x
Chapter 1. Introduction	1
1.1 Overview	1
1.2 Challenges in Cloud Forensics	1
1.3 Motivation behind Research	3
1.4 Problem Statement.....	3
1.5 Objective of Research.....	4
1.6 Scope of Proposed Work	4
1.7 Significance of Research	5
1.8 Methodology of Research.....	6
1.9 Research Process	6
Chapter 2. Literature Review	9
2.1 Introduction	9
2.2 Research Questions.....	10
2.3 Research Methodology	10
2.3.1 Define Category	10
2.3.2 Developing a Review Protocol.....	10
2.3.2.1 Selection criteria	10
2.3.2.2 Search strategy.....	11
2.3.2.3 Quality Assessment	11
2.3.2.4 Data Extraction and Data Synthesis	12
2.4 Results	13
2.4.1 Cloud Forensic Frameworks (Research Question 1).....	15
2.4.2 Cloud Forensic Tools (Research Question 2).....	16
2.4.3 Challenges & Limitations in Cloud Forensics (Research Question 3)	16
2.5 Discussion	17
Chapter 3. Cloud Forensics and its Concepts	18
3.1 Introduction	18
3.2 Types of Forensics.....	19
3.3 General Cloud Forensic Process Flow	20
3.4 Digital Forensics.....	21
3.4.1 Digital Forensic and Its Background.....	21
3.4.2 Digital Evidence and Modes of Digital Investigation	22
3.5 Cloud Forensics	22

3.5.1	Digital Evidence in Cloud Environment	23
3.6	Discussion	24
Chapter 4. Methodology		26
4.1	Introduction	26
4.2	Centralised Cloud Forensic Evidence System	26
4.2.1	Cloud Forensic Evidence Collection.....	26
4.2.2	Disk Image Generation	27
4.2.3	Classification of Pertinent Data	28
4.3	Cloud Forensic Analysis Techniques to Reduce Backlog.....	28
4.3.1	Known Data Acquisition.....	28
4.3.2	Unknown Data Relevancy and Prioritization	28
4.4	Alleviating Cloud Forensic Backlog Methodology	29
4.4.1	Data Deduplication using Hashing.....	29
4.4.2	Relevant Data Extraction and Prioritization.....	29
4.4.3	Information Extraction	30
4.5	Experimentation Design	30
4.6	Discussion	31
Chapter 5. Cloud Forensic Framework for Reducing Backlog (CFFRB)		32
5.1	Introduction	32
5.2	Cloud Forensic Investigation Process Flow.....	33
5.2.1	Incident Identification and Reporting	33
5.2.2	Case Initiation and Planning	34
5.2.3	Evidence Identification and Preservation.....	35
5.2.4	Cloud Service Provider Cooperation.....	35
5.2.5	Evidence Collection and Acquisition.....	36
5.2.6	Evidence Examination and Analysis.....	37
5.2.7	Artifact Interpretation and Extraction	37
5.2.8	Document Validation and Reporting.....	38
5.2.9	Presentation and Legal Proceedings.....	39
5.3	Cloud Forensic Constraints for Reducing Backlog.....	39
5.4	Cloud Forensic Framework for Reducing Backlog (CFFRB).....	41
5.5	Discussion	43
Chapter 6. Data Deduplication of Cloud Forensic Evidence		44
6.1	Introduction	44
6.2	Data Deduplication to Reduce Storage in Cloud Forensics	45
6.3	Data Deduplication Process	46
6.4	Data Deduplication Process Implementation	47
6.5	Evaluating Generated Test Disk Images	47
6.6	Actual & Effective Acquisition Speed & Disk Size Comparison	50
6.7	Discussion	57
Chapter 7: Relevant Data Extraction and Prioritization		58
7.1	Introduction	58
7.2	Forensic Evidence Information and Dataset Details	59
7.3	Testing Environment for Experimentation	59

7.4	Data Preparation Stages.....	59
7.5	Relevant Fraudulent Emails Detection Methods.....	61
7.6	Fraudulent Email Detection using Topic Modelling.....	62
7.6.1	Preprocessing the Email Text Data.....	63
7.6.2	Creating the LDA Model.....	63
7.6.2.1	Latent Dirichlet Allocation Model (LDA).....	63
7.6.3	Extracting Fraud-Related Topics.....	66
7.6.4	Classification and Scoring.....	68
7.6.4.1	Logistic Regression.....	69
7.6.4.2	Linear Support Vector Classifier.....	69
7.6.4.3	Bernoulli Naive Bayes.....	70
7.6.4.4	K-Nearest Neighbours.....	70
7.6.4.5	Random Forest Classifier.....	71
7.6.4.6	Gradient Boosting.....	71
7.6.4.7	Decision Tree Model.....	72
7.6.5	Model Evaluation and Improvement.....	73
7.7	Prioritizing Fraudulent Email Data to Reduce Forensic Backlog.....	73
7.7.1	LDA Model and Fraud Detection.....	73
7.7.2	Flagging Fraudulent Emails.....	73
7.7.3	Prioritization for Testing.....	73
7.7.4	Resource Optimization.....	74
7.7.5	Risk Mitigation and Efficiency.....	74
7.7.6	Post-Testing Steps and Continuous Monitoring and Adaptation.....	74
7.8	Discussion.....	74
Chapter 8: Name Entity Recognition Using BERT Model & Relation Extraction.....		75
8.1	Introduction.....	75
8.2	Name Entity Recognition & Relation Extraction.....	75
8.2.1	Name Entity Recognition (NER).....	75
8.2.2	Relation Extraction (RE).....	76
8.2.3	NER and RE After Topic Modelling.....	77
8.2.4	Name Entity Recognition & Relation Extraction Process.....	78
8.2.4.1	Preprocessing Before NER.....	78
8.2.4.2	Named Entity Recognition (NER).....	78
8.2.4.3	Relation Extraction (RE).....	79
8.2.4.4	Postprocessing and Analysis.....	79
8.3	Name Entity Recognition Using BERT.....	80
8.3.1	Preprocessing Data.....	80
8.3.2	Model Initialization & Training Setup.....	82
8.3.3	Fine-tuning BERT Model.....	82
8.3.4	Token-level Predictions.....	83
8.3.5	Post-processing.....	83
8.3.6	Performance Evaluation.....	83
8.4	Relation Extraction.....	86
8.4.1	Data Preprocessing.....	86

8.4.2	Entity Pair Identification.....	87
8.4.3	Feature Extraction.....	87
8.4.4	Apriori Algorithm.....	87
8.4.5	Association Rules Generation.....	87
8.4.6	Rule Filtering and Evaluation.....	87
8.5	Comparison Between NER using BERT and NER using Combination of Rule-based Approach & CRF Model ...	87
8.5.1	Metrics Results.....	88
8.5.2	Differences in Approach.....	89
8.6	Discussion.....	90
Chapter 9: Conclusion and Future Directions		91
9.1	Introduction.....	91
9.2	Data Deduplication of Cloud Forensic Evidence.....	91
9.3	Relevant Data Extraction and Prioritization of Cloud Forensic Evidence.....	92
9.4	Data Analysis of Collected Evidence Using NER and RE.....	92
9.5	Conclusion and Research Summary.....	93
9.6	Future Directions.....	93
References		95

LIST OF FIGURES

Figure 2.1: Search Strategy	12
Figure 2.2: Cloud Forensics Process Flow	13
Figure 3.1: NIST Visual Model Representation	18
Figure 3.2: Types of Forensics	19
Figure 3.3: Stages of General Cloud Forensic Process Flow	20
Figure 3.4: GAO Report for Cyber Related Incidents for Fiscal Year 2021	23
Figure 4.1: Centralized Cloud Forensic Evidence System.....	27
Figure 4.2: Alleviating Cloud Forensic Backlog Method	30
Figure 5.1: Cloud Forensics Process Flow	33
Figure 5.2: Cloud Forensics Framework for Reducing Backlog (CFFRB).....	42
Figure 6.1: Data Deduplication Process Flow	46
Figure 6.2: Disk Images with Different Sizes	48
Figure 6.3: Creating Disk Image	49
Figure 6.4: Number of Files and Folders in Disk Image	49
Figure 6.5: Disk Image Summary.....	50
Figure 6.6: Number of Duplicates in Disk Images	52
Figure 6.7: Experimentation of Disk Image file 5.....	52
Figure 6.8: Initial and Final Size Comparison after Deduplication.....	53
Figure 6.9: Initial and Final Size Comparison w.r.t Area Graph.....	54
Figure 6.10: Initial and Final Size Comparison w.r.t Time Series Plot.....	54
Figure 6.11: Actual and Effective Speed Comparison	55
Figure 6.12: Actual and Effective Speed Comparison w.r.t Area Graph	56
Figure 6.13: Actual and Effective Speed Comparison w.r.t Time Series Plot	57
Figure 7.1: Enron Dataset with Columns file & message	60
Figure 7.2: Top 20 Employees Who Sent Most Mails	60
Figure 7.3: Days of Week and Hours in Which Emails were Sent	61
Figure 7.4: Combining subject & body Column Figure.....	63
Figure 7.5: Cleaned Data Column After Text Cleaning	63
Figure 7.6: The LDA model	65
Figure 7.7: Four Topics are Printed Each with 5 No. of Words.....	65
Figure 7.8: Visual Representation of LDA model.....	66
Figure 7.9: Data Frame after Flagging Topic 1 as Fraudulent	67
Figure 7.10: Counts of Fraudulent Emails.....	67
Figure 7.11: Fraudulent Email Detection using LDA Model.....	67
Figure 7.12: Comparison of Manual and Email Detection using LDA Model	68
Figure 7.13: Training & Testing Data	68
Figure 7.14: Confusion Matrix & Classification Report using Logistic Regression.....	69
Figure 7.15: Confusion Matrix & Classification Report using Linear SVC	69
Figure 7.16: Confusion Matrix & Classification Report using Bernoulli Naive Bayes	70
Figure 7.17: Confusion Matrix & Classification Report using K-Nearest Neighbours	70
Figure 7.18: Confusion Matrix & Classification Report using Random Forest.....	71
Figure 7.19: Confusion Matrix & Classification Report using Gradient Boosting	71
Figure 7.20: Confusion Matrix & Classification Report using Decision Tree.....	72
Figure 7.21: Comparison of Accuracy of Different ML Algorithms	72
Figure 7.22: Data Frame after Flagging Topic 1 as Fraudulent	73
Figure 8.1: NER and RE After Topic Modelling	76

Figure 8.2: NER and RE Process.....	79
Figure 8.3: Bidirectional Encoder Representations from Transformers Model	80
Figure 8.4: Fraudulent Email Text Data After Topic Modelling Using LDA Model	81
Figure 8.5: Reshuffled body Column	81
Figure 8.6: New word Column with sentence_no	81
Figure 8.7: New word Column with POS & NER tags	81
Figure 8.8: Last 2 Epoch with Average Train Loss.....	83
Figure 8.9: Confusion Matrix for NER using BERT.....	84
Figure 8.10: Classification Report for NER using BERT	85
Figure 8.11: Learning Curve About Training and Validation Loss	86
Figure 8.12: Comparison of NER using BERT and NER using Combination of Rule-based Approach & CRF Model	88
Figure 8.13: Comparison of F1-Score using BERT & Combination of Rule-based Approach & CRF.....	88
Figure 8.14: Comparison of Precision using BERT & Combination of Rule-based Approach & CRF.....	89
Figure 8.15: Comparison of Recall using BERT & Combination of Rule-based Approach & CRF.....	89

LIST OF TABLES

Table 2.1: Search Terms and Database Details.....	12
Table 2.2: Data Extraction.....	12
Table 2.3: Data Synthesis	12
Table 2.4: Digital Library Details with Research Reference.....	13
Table 2.5: Category Details with Research Studies	13
Table 2.6: Cloud Forensic Frameworks, Models, Processes	15
Table 2.7: Cloud Forensic Tools	16
Table 2.8: Challenges in Cloud Forensics	17
Table 6.1: Image Files with information	48
Table 6.2: Image Files with Information After Data Deduplication	51

LIST OF ABBREVIATIONS

LDA	Latent Dirichlet Allocation
NER	Named Entity Recognition
RE	Relation Extraction
BERT	Bidirectional Encoder Representations from Transformers
IoTs	Internet of Things
CSP	Cloud Service Provider
SLR	Systematic Literature Review
AWS	Amazon Web Services
NIST	National Institute of Standards and Technology
DFRW	Digital Forensic Research Workshop
LEA	Law Enforcement Agents
GAO	Government Accountability Office
CFFRB	Cloud Forensic Framework for Reducing Backlog
IDS	Intrusion Detection Systems
ADI	Autodesk Device Interface
NTFS	New Technology File System
SVC	Support Vector Machine
NLP	Natural Language Processing
DBSCAN	Density-based spatial clustering of application
POS	Part of Speech
LSTM	Long Short-Term Memory
CRF	Conditional Random Fields

Chapter 1. Introduction

1.1 Overview

Cloud forensics is a specialized field within digital forensics that addresses the unique challenges associated with investigating digital incidents and criminal activities in cloud computing environments. As organizations increasingly adopt cloud services to store, process, and manage their data, the need for effective methods to preserve, analyse, and present digital evidence in cloud-related cases has become paramount. Cloud forensics involves the application of traditional forensic principles and techniques to cloud environments, considering factors such as virtualization, multi-tenancy, dynamic resource allocation, and complex data storage models. This field encompasses the development of methodologies, tools, and best practices that enable forensic investigators to navigate the complexities of cloud ecosystems, ensuring the integrity of evidence and facilitating successful investigations in cases ranging from data breaches and fraud to intellectual property theft and cyberattacks.

Cloud forensic backlog refers to the accumulation of pending digital forensic cases that require investigation within cloud computing environments. As the adoption of cloud services continues to rise, the volume of potential cases demanding analysis has also increased, resulting in a backlog of unresolved investigations. This backlog can stem from various factors, including the intricate nature of cloud environments, the challenges associated with acquiring and analysing cloud-based evidence, and the evolving landscape of digital threats. The backlog can have detrimental effects on the effectiveness and efficiency of digital forensic processes, leading to delayed response times, compromised evidence integrity, and hindered legal proceedings. Addressing cloud forensic backlog necessitates the development of innovative strategies, tools, and methodologies that streamline investigation workflows, enhance evidence preservation, and expedite the resolution of cases, ultimately ensuring the integrity of the investigative process in the realm of cloud computing.

1.2 Challenges in Cloud Forensics

Cloud computing has introduced numerous benefits in terms of scalability, accessibility, and cost efficiency. However, along with these advantages, it has also brought about a set of unique challenges in the realm of digital forensics.

- **Data Location and Jurisdiction:** Data Location and Jurisdiction challenge is because cloud data is distributed across various physical locations and

jurisdictions. Determining the precise location of data and which legal regulations apply can be complex & can potentially impact admissibility of evidence in court.

- **Data Ownership and Multi-Tenancy:** And cloud services often involve multiple clients sharing the same infrastructure. Establishing data ownership and isolating evidence relevant to a specific case can be challenging due to the shared nature of resources.
- **Virtualization and Abstraction:** Cloud environments utilize virtualization and abstraction technologies, making it difficult to directly access and analyse the underlying hardware. Traditional forensic tools may not be effective in these scenarios.
- **Dynamic and Elastic Nature:** Cloud resources can be dynamically provisioned and scaled, leading to constant changes in the infrastructure. These dynamic changes can complicate the preservation of evidence and the reconstruction of events.
- **Encryption and Access Control:** One another challenge is the strong encryption and access control mechanisms that are commonly employed in cloud environments to ensure data security. While this is beneficial for protecting data, it can hinder forensic investigator's ability to access and analyse relevant information.
- **Logging and Audit Trails:** Cloud service providers often maintain extensive logs and audit trails, but these logs can be dispersed across different services and may not provide a comprehensive view of events. Aggregating and interpreting logs can be a significant challenge.
- **Data Deletion and Retention:** Cloud data may be replicated and stored in multiple locations, making complete data deletion complex. Additionally, different cloud providers have varying policies on data retention, which can affect the availability of historical data.
- **Cross-Boundary Investigations:** Cloud environments transcend geographical boundaries, and investigations may involve data stored in different countries. Coordinating international legal processes and adhering to diverse regulatory frameworks can be arduous.

- **Integrity and Authenticity:** Ensuring the integrity and authenticity of cloud-stored evidence is challenging due to the dynamic nature of cloud resources and the potential for tampering or alteration. Forensic investigations within cloud environments may face resource limitations imposed by cloud providers, affecting the ability to analyse data effectively and in a timely manner.
- **Resource Constraints:** The lack of standardized procedures, tools, and protocols for cloud forensics can lead to inconsistencies and difficulties in collaboration among investigators.
- **Complexity of Evidence Collection:** Collecting evidence from cloud services can involve a mix of traditional and cloud-specific methods. The variety of sources and formats can make evidence collection complex and time-consuming.

1.3 Motivation behind Research

The motivation behind research in reducing Cloud Forensic Backlog stems from the critical need to address the mounting backlog of digital forensic cases within cloud environments, coupled with the inherent challenges posed by the dynamic and distributed nature of cloud computing. The surge in cloud adoption and rapid proliferation of cloud services has led to a growing accumulation of cases awaiting investigation, posing challenges to effective and timely resolution. This research seeks to address this pressing issue by proposing a comprehensive Cloud Forensic Framework that optimize the investigative process, that integrates innovative strategies such as data deduplication, prioritized data extraction, and advanced analysis methods. By mitigating the backlog through enhanced efficiency, the study aspires to not only expedite forensic procedures but also to contribute to the overall effectiveness of cloud forensic practices in cloud computing scenarios.

1.4 Problem Statement

The problem of cloud forensic backlog arises from the increasing adoption of cloud computing services, leading to a significant accumulation of pending digital forensic cases within cloud environments. This backlog stems from the intricate challenges associated with investigating digital incidents and criminal activities in the cloud, including data fragmentation, dynamic resource allocation, virtualization complexities, and jurisdictional issues. The backlog negatively impacts the timeliness and effectiveness of investigations, potentially compromising evidence integrity, impeding the pursuit of justice, and hampering legal proceedings. Addressing the cloud forensic backlog requires the development of novel approaches, methodologies, and tools that can streamline investigation processes, enhance

evidence preservation, and expedite case resolution, ultimately ensuring the efficient and reliable administration of justice in cloud-related cases.

1.5 Objective of Research

The primary objective of this research is to design, develop, and validate a comprehensive set of methodologies, techniques, and tools aimed at effectively alleviating the cloud forensic backlog. By focusing on the intricate challenges posed by forensic investigations within cloud environments, this research seeks to streamline investigation workflows, enhance evidence collection and preservation methods, and expedite the resolution of pending cases. The research aims to address issues that contribute to the backlog. Through the proposed strategies, the research aims to establish a framework that not only improves the efficiency and timeliness of cloud forensic investigations but also upholds the integrity of evidence and the accuracy of findings, thus contributing to the overall advancement of the digital forensic field in cloud computing scenarios.

1.6 Scope of Proposed Work

The proposed work will focus on investigating the feasibility and effectiveness of reducing cloud evidence backlog by implementing deduplication and by relevant data extraction and prioritization for cloud forensic investigations. The research focuses on designing and implementing solutions that enhance the efficiency and effectiveness of cloud forensic investigations:

- The scope includes the creation of a specialized cloud forensic framework tailored to the challenges of investigating incidents within cloud environments. This framework will integrate best practices, methodologies, and tools for seamless evidence acquisition, preservation, and analysis.
- Addressing the challenge of data duplication contributing to the backlog, the research scope encompasses the exploration of advanced data deduplication techniques. By identifying and eliminating redundant data instances through hashing and comparison algorithms, the investigation process can be expedited.
- The proposed work involves the development of methods to efficiently extract pertinent information from voluminous cloud datasets. Through techniques such as keyword analysis, pattern recognition, and machine learning, the aim is to prioritize and extract relevant data, thereby streamlining analysis efforts.
- The scope includes the utilization of advanced analysis methodologies such as topic modelling using LDA (Latent Dirichlet Allocation) to uncover patterns,

trends, and themes within cloud-stored data. Additionally, Named Entity Recognition (NER) powered by BERT and relation extraction techniques will be employed to further enhance investigation depth.

1.7 Significance of Research

The significance of this research focused on addressing the cloud forensic backlog lies in its potential to revolutionize the efficiency and effectiveness of digital forensic investigations within cloud environments. The research not only acknowledges pressing issue of mounting investigative cases but also offers innovative solutions that have wide-reaching implications:

- **Timely Case Resolution:** By streamlining the investigative process through the proposed framework, the research significantly reduces the time required for case resolution. This is crucial in the realm of digital forensics, where timely responses are essential to preserving evidence integrity and ensuring justice.
- **Enhanced Evidence Preservation:** The research's focus on data deduplication and relevant data extraction ensures that investigators work with high-quality, pertinent evidence. This enhances evidence preservation, minimizing the risk of data loss, tampering, or corruption that can occur when dealing with large & complex cloud datasets.
- **Efficient Resource Utilization:** Implementing data deduplication and prioritized data extraction optimizes the utilization of resources, both in terms of storage capacity and investigator effort. This efficiency translates to cost savings and improved resource allocation within investigative teams.
- **Legal Admissibility:** The systematic process flow and innovative techniques proposed in the research contribute to the establishment of best practices in cloud forensics. This enhances the credibility and admissibility of evidence in legal proceedings, bolstering the case's chances of success.
- **Advancement of Cloud Forensic Practices:** Research introduces novel methodologies & strategies tailored to cloud environments, contributing to the growth and development of cloud forensic practices. As cloud adoption continues to expand, these advancements are vital in addressing emerging challenges.
- **Cross-Disciplinary Impact:** The research has potential implications beyond the field of cloud forensics. The innovative techniques, such as data deduplication and relevant data extraction, can also find applications in data management, cybersecurity, and information retrieval domains.

- **Future Research and Collaboration:** The proposed framework and techniques open doors for further research and collaboration in cloud forensics. The research can inspire other scholars and practitioners to build upon its foundation, leading to a continuous.

1.8 Methodology of Research

The research methodology employed for addressing the challenge of cloud forensic backlog encompasses a systematic and iterative approach designed to develop effective strategies for expediting investigations within cloud environments. The methodology begins with a comprehensive review of existing literature on cloud forensics, digital investigation techniques, and backlog reduction strategies to identify gaps and establish a foundation for innovation. Subsequently, data collection and analysis are conducted to gain insights into the nature of cloud forensic backlogs and the specific challenges they pose. Based on the identified challenges, cloud forensic constraints are defined, and a specialized framework is proposed to guide investigators through evidence acquisition, analysis, and reporting stages. Techniques for data deduplication using hashing are developed to optimize storage resources and reduce redundancy, while methods for relevant data extraction and prioritization are designed to expedite investigation workflows. To enhance information extraction, Named Entity Recognition (NER) with BERT and relation extraction techniques are applied. The methodology concludes with rigorous evaluation and validation of the proposed strategies using real-world cloud forensic scenarios and datasets, culminating in a comprehensive analysis of results and their implications. This methodology serves as a structured framework for tackling the cloud forensic backlog, ultimately contributing to the advancement of cloud forensic practices and the efficient resolution of pending cases within cloud computing environments.

1.9 Research Process

The research process for addressing the cloud forensic backlog involves a series of interconnected steps that systematically build upon one another. This process is designed to develop effective strategies and methodologies for streamlining investigations and reducing the accumulation of pending cases within cloud environments. The key stages of the research process are as follows:

- **Problem Identification and Scope Definition:** Identify the problem of cloud forensic backlog and its implications. Define the scope of the research, including the specific challenges and constraints within cloud environments.

- **Literature Review:** Conduct an in-depth review of existing literature related to cloud forensics, digital investigation techniques, and methods for handling investigative backlogs. Identify gaps in the current knowledge and techniques applicable to cloud forensic backlog reduction.
- **Data Collection and Analysis:** Gather relevant data sources, including cloud forensic case studies, datasets, and cloud system architectures. Analyse the data to understand the nature of cloud forensic backlogs, the underlying causes, and the specific challenges faced.
- **Constraint Identification and Framework Proposal:** Identify the constraints unique to cloud environments that impact forensic investigations. Propose a cloud forensic framework that considers these constraints, providing a structured approach to handling cloud forensic cases.
- **Process Flow Design:** Develop a detailed process flow that guides investigators through evidence acquisition, analysis, and reporting stages within cloud environments. Address the dynamic nature of cloud resources and emphasize proper evidence preservation.
- **Data Deduplication Strategy:** Design and implement a data deduplication strategy using hashing techniques to identify and eliminate duplicate data instances. Evaluate the effectiveness of this strategy in reducing data redundancy and optimizing storage resources.
- **Relevant Data Extraction and Prioritization:** Develop techniques to efficiently extract relevant data from large cloud datasets. Prioritize critical data for investigation using methods such as keyword analysis and pattern recognition.
- **Advanced Information Extraction:** Implement Named Entity Recognition (NER) using BERT to identify entities within the data. Explore and implement relation extraction techniques to uncover connections between entities.
- **Evaluation and Validation:** Test the proposed framework, process flow, and techniques using real-world cloud forensic scenarios and datasets. Evaluate effectiveness of developed.
- **Results Analysis and Conclusion:** Analyse the results obtained from the evaluation to assess the impact of the proposed strategies on reducing the cloud

forensic backlog. Draw conclusions on the effectiveness of the developed methods and their implications for cloud forensic practice.

- **Discussion and Future Work:** Discuss the findings in the context of existing literature and implications for the broader field of cloud forensics. Identify areas for further research and improvements in tackling the cloud forensic backlog more effectively.

The research process shows our systematic journey from problem identification to practical implementation and validation, culminating in insights and strategies that contribute to reducing the cloud forensic backlog and enhancing the efficiency of investigations in cloud computing environments.

Chapter 2. Literature Review

2.1 Introduction

Cloud computing is widely used in this era of IoTs. Cloud users use cloud computing to get various cloud services. The defects in cloud services are exploited by the suspicious actors. On the other hand, cloud forensic help against such suspicious actors. It is found on many occasions that cloud services were not well designed. Cloud services are exposed to cyber threats due to bad implementation of cloud services which helps suspicious actor to exploit vulnerabilities in cloud environment. Main objective is to identify different cloud forensics frameworks. We have used the method of systematic literature review to find & analyse different research studies which mainly are published between 2010-2022. We've discovered 36 different cloud forensic frameworks and tools. We also mentioned some limitations of cloud forensics. The systematic literature review shows major cloud forensic frameworks and tools related to it and highlight some of the challenges of cloud forensics.

Developing world is of big data and (IoT) i.e., Internet of Things, almost every device uses cloud to get its services and to run different applications around the world. And Cloud datacentres are stored in centralized locations, in which networking and different computing equipment works to store, collect, distribute, process, and allow access to large data stored in them. Cloud computing has developed too much in information technology. Users use Cloud one way or another and Cloud services users are rising. Due to frequent use and ease of access cloud services are in demand. It is estimated that Cloud Services users are increasing day by day with many switching to cloud. With increasing cloud services cloud vulnerabilities are also on the rise and suspicious actors are exploiting such defects present in cloud environment.

Developers on the other hand do not pay attention nor they follow certain criteria to avoid the suspicious actor attacking the cloud services. Developing a service is one thing but protecting it from attackers is also important. Developers do not follow Forensic enabled criteria to develop and to implement Cloud services framework and they do not consider cloud forensic basics needs. Which impact on cloud forensic investigation because forensic investigation becomes difficult. If design and implementation standards are met, this will in turn ensure sound cloud forensic investigation. If cloud service providers design cloud forensic services in a standardized way, it will allow the investigation agencies to solve the cyber-crimes in a big way. To design and to develop cloud forensic services developers should follow a certain framework for development and necessary requirements and processes. The main purpose of this SLR is identification of recent research in the field of cloud forensics.

2.2 Research Questions

We developed three research questions that will help in the identification of recent research in the field of cloud forensic frameworks and recent tools that are being used and cloud forensic challenges will also be discussed. Following research questions have been discussed in this literature review.

- **Research Question 1:** Which latest frameworks, methods are being used to develop Cloud Forensics since 2010 to 2022?
- **Research Question 2:** Which kind of Tools are being used in cloud forensics since 2010 to 2022?
- **Research Question 3:** What are important challenges in Cloud Forensics?

2.3 Research Methodology

A researcher [1], In methodology for systematic literature review we will first define category, then we discuss review protocol, then we will discuss quality assessment.

2.3.1 Define Category

Our research which is about cloud forensics has three main categories shown as follow.

- **Cloud Forensic Framework Category:** Studies related to Cloud Frameworks.
- **Cloud Forensic Tools Category:** Studies related to Cloud Forensic.
- **Cloud Forensic Challenges/ Limitations Category:** Studies related to Limitations of Cloud Forensic.

2.3.2 Developing a Review Protocol

Development has four part which are:

2.3.2.1 Selection criteria

In selection criteria we will be looking at five major parts which includes subject relevance, years selected for research, specific publisher repositories, effectiveness, result oriented.

- **Subject Relevance:** We have selected only those research papers which are relevant to our field that is cloud forensic frameworks. Because relevancy will play important role in answering our research questions and we have rejected those research papers which are not relevant and does not help in our research questions.

- **2010-2020:** We have selected only those research papers which are mainly from 2010-2022. We have included the latest work and not included irrelevant research which are older than 2010.
- **Databases:** We have limited our research to major databases. We are only selecting four top publication databases that are IEEE, Elsevier, ACM, Springer, and the relevant research related to other databases are also included.
- **Crucial Effects:** Only those research work is included which is crucial and important and have encouraging effect in the field of cloud forensic framework.
- **Result Oriented:** We only selected those research papers in which framework related to cloud forensics are being implemented and didn't select research work which does not yield any framework.

2.3.2.2 Search strategy

A search process is mainly composed of information related to databases, keywords or search terms used to extract the desire information. The search or keywords used for extracting information is discussed in the Table 2.1. Keywords and search terms are used in multiple ways to find the desired information. To make the research effective Boolean operators “AND” and “OR” are used with search techniques. We use different search terms with the help of operators to find the relevant studies. The keywords used for extracting the relevant information are cloud forensic framework, cloud forensic services, cloud forensic tools, digital cloud forensic framework. Our search strategy has four steps in which we narrow down our findings. In our selection criteria we use inclusion and exclusion criteria to find the relevant research papers in the field of frameworks for developing cloud forensic frameworks and related to tools used for cloud forensics. The research papers which we collected are from four major scientific repositories that are IEEE, ACM, ELSEVIER, SPRINGER while Figure 2.1 shown below show us the search strategy of this SLR paper.

2.3.2.3 Quality Assessment

In this part we will be identifying different types of quality assessment criteria. Which include purpose of this research is to search the latest frameworks, tools for the development in cloud forensics. The research papers selected from the databases are latest and are related to field & are from the year 2010 to 2022. All the selected research papers are in English language and Duplication of research papers is removed. Research papers discussing limitations are also included.

Keywords	B/O	Number of Search Results				
		IEEE	ACM	Elsevier	Springer	Other
Cloud Forensic Framework	AND	77	521	1020	1537	80
	OR	246,766	148,557	844,537	878,559	79,300
Cloud Forensic Tools	AND	82	601	1234	1766	20
	OR	236,488	169,926	1,230,763	997,334	88,200
Cloud Forensic Services	AND	157	640	1233	1831	25
	OR	291,500	122,454	839,423	867,877	106,000
Digital	AND	59	485	693	1001	30
Cloud Forensic Framework	OR	396,906	257,780	1,211,354	1,106,107	51,100

Table 2.1: Search Terms and Database Details.

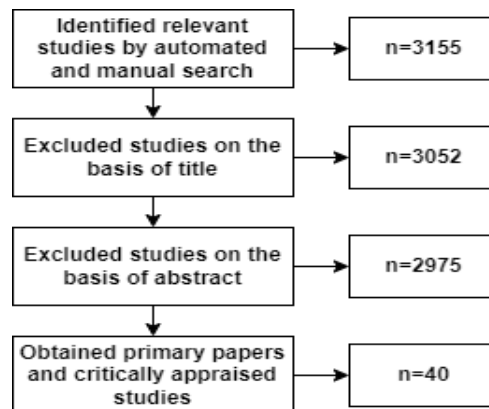


Figure 2.1: Search Strategy.

2.3.2.4 Data Extraction and Data Synthesis

When data extraction is done, we analyse the data collected to find the different frameworks and tools of the cloud forensic services. Digital library with research studies is shown in Table 2.4 and type of research is also given. While, in Table 2.2 we showed details of information extracted and in Table 2.3 information of synthesis of data is given.

Data Extraction Details		
Sr #	Description	Details
1	Bibliography	Research title, authors, publisher info, Year of the publication.
2	Overview	Info about selected research studies.
3	Results	Result of selected research studies.

Table 2.2: Data Extraction.

Data Synthesis Details		
Sr #	Description	Details
1	Frameworks for Cloud Forensics	Frameworks related to cloud forensic will be discussed. (Table 5)
2	Cloud Forensic Tools	Tools related to cloud forensic will be discussed. (Table 6)
3	Limitations of cloud forensics	Limitations related to cloud forensic will be discussed. (Table 7)

Table 2.3: Data Synthesis.

Digital Library	No. of Papers	Type	Selected Research
IEEE	13	Journal	[14][23]
		Conference	[8][13][15][16][18][21][22][25][26][29][37]
Elsevier	02	Journal	[27][34]
		Conference	-/-
Springer	07	Journal	[35][40]
		Conference	[4][11][28][32][33]
ACM	01	Journal	[3]
		Conference	-/-
Others	16	Journal	[2][9][12]
		Conference	[5][6][7][10][17][19][20][24][30][31][36][38][39]

Table 2.4: Digital Library Details with Research Reference.

2.4 Results

In the results section we will discuss the results of selected research studies w.r.t to our research questions. which we have tried to answer in our systematic literature review. In this SLR we have selected about ‘36’ research studies. 13 of these studies are from IEEE, 02 from Elsevier, 07 from Springer, 01 from ACM, 16 of these studies are from other databases as shown in Table 2.5.

Category	No. of Papers	Type	Selected Research
Cloud Forensic Frameworks	36	J	[2][3][9][12][14][23][27][34][35][40]
		C	[4][5][6][7][8][10][11][13][15][16][17][18][19][20][21][22][24][25][26][28][29][30][31][32][33][36][37][38][39]
Cloud Forensic Tools	01	J	[2]
		C	-/-
Cloud Forensic Challenges & Limitations	02	J	[3]
		C	[4]

Table 2.5: Category Details with Research Studies.

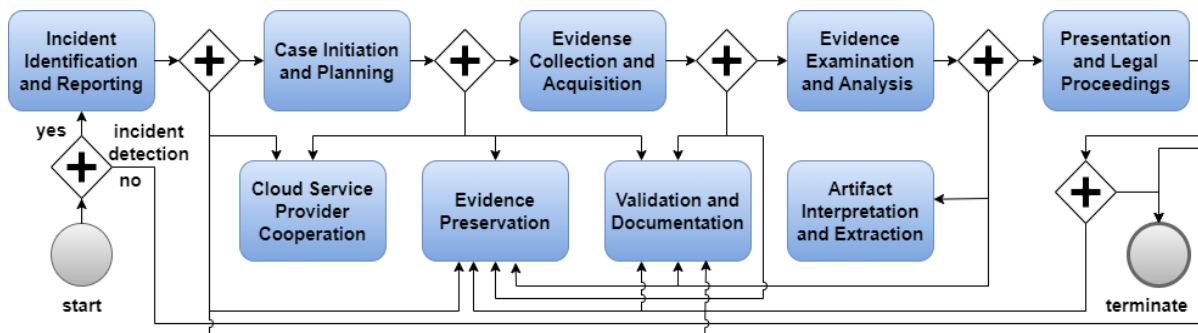


Figure 2.2: Cloud Forensics Process Flow.

Sr/ No	Frameworks/ model/ process	Ref.	No. of Stages	Incident Identification & reporting	Case Initiation & Planning	Evidence Collection & Acquisition	Evidence Examination & Analysis	Presentation & Legal Proceeding
01	Forensic computing process. (McKemmish, 1999)	[5]	4 stages	identification	x	preservation	analysis	presentation
02	Forensic process by NIST. (Kent et al., 2006)	[6]	4 stages	x	x	collection	examination, analysis	reporting
03	Forensic investigations process. (Guo et al., 2012)	[7]	3 stages	identification	x	preservation, collection	x	x
04	Cloud forensics process. (Chen et al., 2012)	[8]	3 stages	identification	x	preservation, collection	x	x
05	Integrated conceptual DFF for cloud computing. (Martini et al., 2012)	[9]	4 stages	identification	x	preservation, collection	examination, analysis	reporting, presentation
06	Live digital forensic framework for cloud environment. (Sibiya et al., 2012)	[10]	4 stages	x	monitoring tests	live data collection, memory, logs, cache, user data	Log mining, data extraction, find relationships	Results presentation
07	Cloud forensics maturity model. (Ruan et al., 2013)	[11]	4 stages	pre-investigative readiness, investigative interface	x	proactive & reactive data-coll, evidence management	core-forensic process (examination, analysis)	supportive process (case management)
08	Advanced data acquisition model. (Adams, 2012)	[12]	4 stages	Preparation, notification, awareness	onsite survey	preservation, collection, documentation	x	x
09	OpenStack cloud framework. (Saibharath et al., 2014)	[13]	3 stages	x	x	Data seizure, acquisition	analysis	x
10	Digital forensic framework for cloud. (Shah et al., 2014)	[14]	4 stages	cloud stack identification	x	Live/static data acquisition	Data mining, evidence analysis	presentation
11	Logging framework for cloud. (Pătrașcu et al., 2014)	[15]	5 stages	x	manage, enable, cloud deploy, virtual, logging	raw data gathering	analyzing, ordering, processing, aggregating	result storage and presentation
12	Framework for analyzing IaaS cloud. (Ahmad et al., 2015)	[16]	8 stages	IaaS formation, detection	validate incident response	capturing, examination	analysis, extraction	reporting
13	Cloud forensic framework for IaaS. (Banas, 2015)	[17]	5 stages	x	x	media collection	data examination and analysis	reporting evidence
14	Open cloud forensics. (Zawoad et al., 2015)	[18]	6 stages	identification	x	preservation, collection	organization, (examin & analy)	presentation, verification
15	Cloud forensics logging framework. (Faldu, 2016)	[19]	5 stages	x	cloud management module	virtualization, logging module	raw data, processing layer	final data
16	Framework for data iden & collection in mob cloud. (Faheem et al., 2016)	[20]	7 stages	forensic log info, identification	x	preservation, collection	potential evidence, correlation	reporting
17	Open and continuous cloud forensic process flow. (Datta et al., 2016)	[21]	4 stages	identification	x	preservation, collection	organization, verification	presentation
18	Mobile cloud forensic framework. (Faheem et al., 2016)	[22]	5 stages	identification	x	preservation, collection	osnit evidence correlation	reporting
19	Framework for cyber physical cloud system. (Ab Rahman et al., 2016)	[23]	5 stages	identify potential evidence sources	plan pre incident coll & analysis, plan detection	define storage, evidence handling	x	x
20	Comparison framework for digital and cloud forensic. (Simou et al., 2016)	[24]	4 stages	identification	x	preservation, collection, doc	examination, analysis	presentation
21	Cloud forensic readiness framework for organizations (Alenezi et al., 2017)	[25]	2 stages	x	x	data collection from literature, industry standards	evaluate, analyse CFR factors, rem duplications	x
22	Cloud centric framework for isolating Bigdata forensic evidence from IoT. (Kebande et al., 2017)	[26]	11 stages	Observe, identify	Deploy agent-based solution	isolate, extract, cluster evidence, preserve, store	commence, acquire, investigate	x
23	Log aggregation forensic analysis	[27]	5 stages	x	x	log acquisition	correlation,	x

	framework. (Ahmed Khan et al., 2017)					.and integration,	sequencing, analysis, and reporting	
24	Fuzzy data mining-based framework. (Santra et al., 2018)	[28]	4 stages	identification of source	x	data collection from source	examination, analysis	present evidence
25	Forensic recovery of cloud evidence. (Sampana et al., 2019)	[29]	6 stages	x	preparation and isolation	collection and storage	analysis	reporting
26	Heterogeneous joint cloud framework. (Umar et al., 2019)	[30]	6 stages	identification	x	preservation, collection	examination, analysis	presentation
27	Private cloud investigation framework. (Sudyana et al., 2019)	[31]	5 stages	identification	x	collection, acquisition	investigation	presentation
28	Framework for users in virtual environment of cloud. (Pandi Jain et al., 2020)	[32]	6 stages	incident, identification	x	preservation, collection, storage	examination, org, analysis	verification, presentation
29	Dependable framework for forensic readiness in cloud. (Bhatia et al., 2020)	[33]	10 stages	detection, connection establishment	strategy, policy making, ready for execution	artifact identification, collection, and acquisition	org artifacts, investigation and analysis	outcome, report, closure, preservation
30	Forensics using intelligent edge computing. (Razaque et al., 2021)	[34]	8 stages	detection	response	acquisition, record, control, extraction, preservation	forensic analysis report	forensic user presentation
31	Framework for anti-forensic attacks in the cloud. (Rani et al., 2021)	[35]	3 stages	identification of suspected packet	x	packet marking	traceback	x
32	Cloud forensic readiness framework. (Fadilla et al., 2022)	[36]	5 stages	resource identification	policy and procedure	technical readiness	forensic response	evaluation and reporting
33	Multi source-based cloud forensic. (Kumari et al., 2022)	[37]	11 stages	awareness, identification	preparation	preservation, collection, distribution	pre analysis, comparison, final analysis	Result improvement, reporting, presentation
34	Forensic framework validation and cloud forensic readiness. (Simou et al., 2022)	[38]	5 stages	incident confirmation, identification	training and planning	preserve, update, collection, acquisition	examination, analysis	presentation
35	A tamper proof cloud forensic framework. (Ye et al., 2022)	[39]	4 stages	identify tampered evidence	x	provenance data gen, data collection from node	noise data, evidence verification	data release to EVC and online
36	Cloud-based framework for digital forensic investigation. (Prakash et al., 2022)	[40]	7 stages	identification	survey	collection, preservation, investigator	examination, analysis, reconstruction	reporting, presentation

Table 2.6: Cloud Forensic Frameworks, Models, Processes.

2.4.1 Cloud Forensic Frameworks (Research Question 1)

After going through several research (Simou et al., 2016) proposed a general comparison cloud forensic process, which has 4 stages. But to accommodate latest developments, proposed cloud forensic process flow which has 5 stages is compared here with other latest cloud frameworks. The proposed cloud forensic process flow is shown in Figure 2.2. It has stages that are Incident Identification & reporting, Case Initiation & Planning, Evidence Collection & Acquisition, Evidence Examination & Analysis, and Presentation and Legal Proceedings and 4 others are concurrent stages that are CSP Cooperation, Evidence Preservation, Validation & Documentation, Artifact Interpretation and Extraction. In this SLR we have collected the information of 36 cloud forensic frameworks which are compared with respect to their stages. Table 2.6 shows different framework details with compassion to given proposed cloud forensic process model. The proposed cloud forensic process flow is

discussed in detail in chapter 4 which is of cloud forensic framework in which first forensic constraints are defined then a cloud forensic framework is proposed and after that cloud forensic process flow is explained.

2.4.2 Cloud Forensic Tools (Research Question 2)

We have found about eight cloud forensic tools, which are being used in cloud environment which are shown with their description in Table 2.7. Cloud tools are given in column “tools”. Research study [2] also discussed some cloud forensic tools and describe their purpose of use. These tools include EnCase tool, Diffy tool, FTK tool, FROST tool, Oxygen forensic suit tool, SIFT tool, AW-IR tool and UFED cloud analyser tool.

Sr/ No	Tools	Description
1	EnCase	This tool related to cloud forensic, IaaS based used to collect data remotely from guest operating system layer of cloud.
2	Diffy	Diffy is cloud-based tool. Used to help digital forensic and incident response team to find suspicious host and cloud instances during incident.
3	FTK	Forensic tool used to extract the desire information that is present in the layer of guest operating system of cloud, and it is used to scan the hard drive and looking for evidence.
4	FROST	Cloud, OpenStack, IaaS based tool used to find the Api’s logs and Virtual disk and guest firewall logs
5	Oxygen Forensics Suit	This tool helps in digital evidence collection from cloud services used on smartphones.
6	SIFT	Ubuntu based tool, SIFT or SANS is used for forensic analysis and incident response study.
7	AWS-IR	It is python command line interface. It has two functions key compromise, instance compromise.
8	UFED cloud analyser	Cloud based tools used for analysing cloud data and meta-data.

Table 2.7: Cloud Forensic Tools.

2.4.3 Challenges & Limitations in Cloud Forensics (Research Question 3)

Cloud frameworks also have some limitations as well shown in Table 2.8. Challenges can be of physical location, or it may be of SLA based or data issue. Challenges category with description & recommendation is shown. When we have a challenge of physical location of servers. The CSP must make available recourses for forensic investigation.

Challenges	Recommendations
Lack of forensic tools	By hypervisor which allow live forensic.
Cloud service provider dependence	Collect forensic data outside of cloud.
Logging issue	We can remove it by the help of proper log-based resources and framework.
Cloud Forensic enabled services	By using cloud forensic enabled frameworks shown in table 6.
Lack of forensic capability and readiness	Cloud forensic readiness in organization-based framework can be used.
Trust issue	Can be remove by proper connection of VM and cloud platform

	through reservoir.
Identification of malicious actor	By using frameworks which control network traffic and identify such actors.
Architecture based	By using the framework which supported IaaS, SaaS, and PaaS.
Collection of evidence	By using frameworks which store information related to security.
Location based	Cloud service provider should give resources without location dependency.
Data related	Data must be encrypted, and duplication must be removed.

Table 2.8: Challenges in Cloud Forensics.

2.5 Discussion

We are doing this SLR to identify cloud forensic frameworks and tools that are used in cloud forensics and point out some limitations and challenges of the cloud forensic frameworks. Our aim is to collect information that is recent, that is why we include research studies that are recent in the field. To find different kind of frameworks and tools used for the development and finding defects of cloud forensic, we answered our research question and found many development frameworks that can be used to develop the cloud services. By using these frameworks, cloud service providers can make cloud services forensically investigate able and can solve many cybercrimes related to cloud environment. Research papers which were related to our field are identified and synthesized. We only add research studies that are from four major databases, and we narrow down our research to only English research studies. We can add more results from other digital libraries to strengthen our research.

Chapter 3. Cloud Forensics and its Concepts

3.1 Introduction

Cloud computing refers to the practice of using remote servers hosted on the internet to store, manage, and process data instead of using local servers or personal computers. It involves the delivery of various computing services, including servers, storage, databases, networking, software, analytics, and more, over the internet. In cloud computing, users can access and utilize computing resources on-demand and pay only for the resources they use. These resources are typically provided by cloud service providers, such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform, who own and manage the infrastructure required to deliver these services. (NIST) [41] defines cloud computing as "*Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.*" NIST definition is shown in Figure 3.1, having five essential characteristics, three services models, and four deployment models.

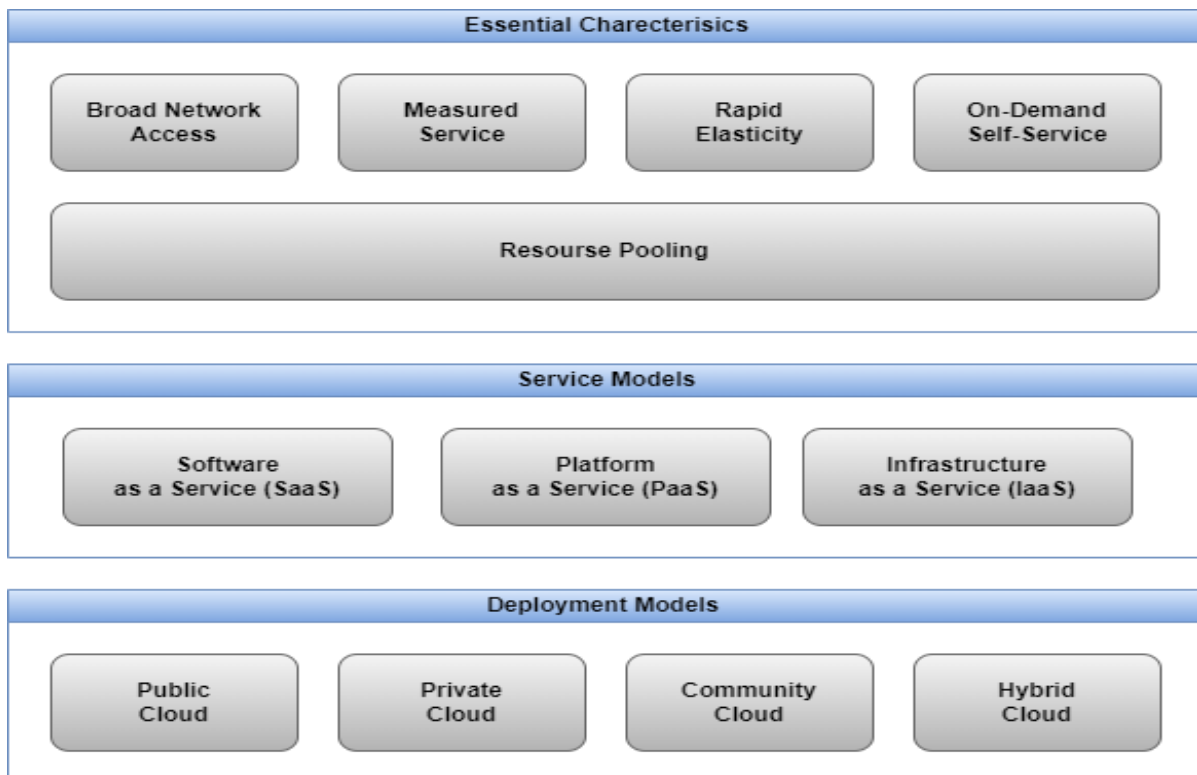


Figure 3.1: NIST Visual Model Representation.

3.2 Types of Forensics

When an incident occurs, forensics is done to identify it, then evidence is collected after that, then examination is done on that collected evidence, data related to that incident is preserved and then a result is concluded in the presentation and reporting phase. There are different types of forensics that the investigator needs in order to find source of evidence as shown in Figure 3.2. Types of forensics are as follows.

- Digital forensics, also known as computer forensics, is a branch of forensic science that involves the investigation and analysis of digital devices and electronic data to uncover evidence for legal proceedings. It is concerned with the identification, preservation, extraction, interpretation, and documentation of digital evidence.
- Network forensics is a branch of digital forensics that focuses on the investigation and analysis of network traffic & communication data to uncover evidence related to cybercrimes or security incidents. It involves retrieving data from network ports, & capturing, inspecting, & interpreting network packets to reconstruct events, identify malicious activities, & gather evidence for legal proceedings.
- Web forensics, also known as web-based forensics or web application forensics, is a branch of digital forensics that focuses on investigation and analysis of web-based evidence. It involves examination of web servers, web applications, web browsers, & related technologies to uncover digital evidence for legal proceedings.
- Cloud forensics is a specialized field of digital forensics that focuses on the investigation and analysis of digital evidence in cloud computing environments. It involves the collection, preservation, and examination of data stored, processed, or transmitted through cloud services and platforms.
- Mobile forensics, also known as mobile device forensics or mobile phone forensics, is a branch of digital forensics that focuses on the investigation and analysis of digital evidence from mobile devices such as smartphones, tablets etc.

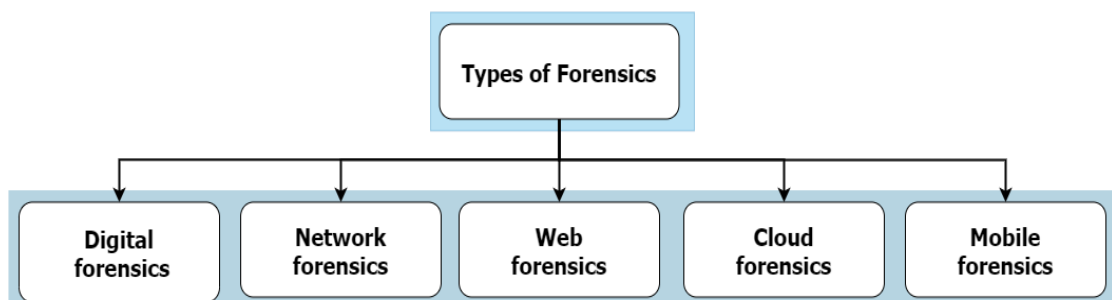


Figure 3.2: Types of Forensics.

3.3 General Cloud Forensic Process Flow

After going through several research [4],[42], [43], Simou proposed a general cloud forensic process, involving several stages, including identification, collection and preservation, examination and analysis, and presentation and reporting. Other two stages which are constant throughout the process are chain of custody and the documentation stage. As shown in Figure 3.3, here's an overview of typical steps involved in cloud forensic investigations:

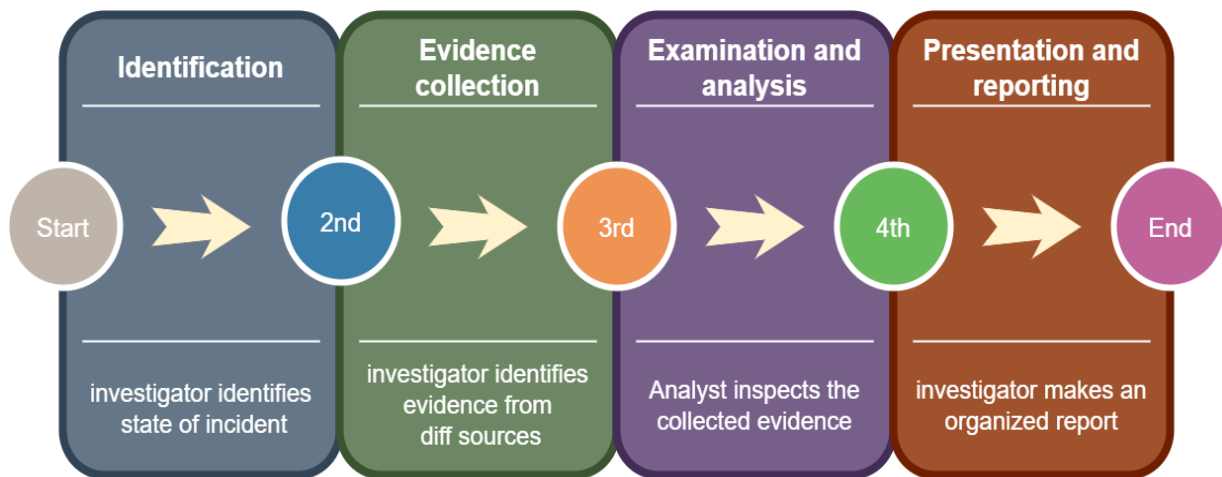


Figure 3.3: Stages of General Cloud Forensic Process Flow.

- **Identification:** Identify the state of incident and then identify cloud service or provider involved in the investigation. Determine the type of cloud deployment (public, private, hybrid), the specific services used, and the relevant legal and contractual agreements.
- **Preservation and Collection:** Collect relevant data and evidence from the cloud environment. This can include data stored in cloud storage, logs, virtual machines, network traffic, user accounts, etc. Ensure that collection methods adhere to legal and privacy requirements and document the chain of custody.
- **Examination and Analysis:** Analyst will examine the collected data and evidence using forensic techniques and tools. This may involve recovering deleted files, analysing metadata, reconstructing user activities, and identifying potential sources of evidence within the cloud environment. Analyst will analyse the collected evidence to extract relevant information and identify patterns, anomalies, or potential indicators of malicious activity. This may involve correlating data from different sources, reconstructing timelines, and identifying potential sources of compromise or unauthorized access.

- **Presentation and Reporting:** Cloud investigator document and report the findings of the investigation. Cloud investigator, prepare a comprehensive forensic report that outlines the methodology, findings, and conclusions. Clearly present the digital evidence and provide expert opinions to support the investigation's outcomes.

3.4 Digital Forensics

In digital forensics the investigation process is done after the incident while in the case of cyber security it deals with the prevention of cyber-attacks beforehand and deals with forming such systems which are secure in nature. So, we can say that after a failure from cyber security, when an incident happened then digital forensics is used for investigating the incident. Its examples include cases like fraud, theft i.e., stealing valuable information etc. Investigator main concern is the excess of data that is present across different locations and different devices complicated the forensic investigation process. Excess of data in turn result into increase in acquisition speed, the amount of data stored which will require a lengthy time to test and analyse. This will in case will result in complication of the investigation process thus creating backlog. It will give us multiple cases without substantial evidence in our backlog to process and the investigation can prolong for months to come [44].

3.4.1 Digital Forensic and Its Background

The development of internet technology also enabled different companies to find vulnerabilities in their systems by creating different forensic tools. These tools helped them to identify hidden evidence in case of an incident. This type of forensics deals with the identification of digital evidence where crime had occurred. In digital forensics the investigators use a process to find evidence which is called as the digital forensic process, it has four stages. The use this process to find evidence so that they can present this evidence in any court of law. A DFRW i.e., digital forensic research workshop has defined digital forensics as *“The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation & presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations”* [45]”. There were other definitions too to describe digital forensics such as [46], which have defined digital forensics as *“the study of evidence from attacks on computer systems in order to learn what has occurred, how to prevent it from recurring, and the extent of the damage”*. For identification of digital evidence, the researchers have developed many processes but NIST [6] which has four phases that are collection phases, second is the examination phase, analysis is the third, and reporting is last.

3.4.2 Digital Evidence and Modes of Digital Investigation

Digital forensics is done on evidence called as digital evidence, it is a data which is present on digital devices which include documents, data files, audios, videos etc. Evidence can be used for crime investigation in case of an incidents like child abuse, data theft, drugs dealings, network related attacks etc. Digital forensic investigation of several types which depend on digital device type, data types etc. Many digital investigation process models have been proposed, widely known model is NIST model. Which have several investigation phases like collection, examination, analysis, and reporting. Forensic investigation has two modes which are live and static mode. [47], [48], [49], [50]. Static forensic investigation is described as an investigation which is done in a forensic environment, all the evidence is collected, and the investigation is done in a specifically designed forensic environment. In live digital forensic investigation, the investigation is done on a live device on which the incident happened. This type of forensic is more difficult than the traditional forensic which is static [48]. It includes snaps of data and live analysis of data and investigators keep a copy of data which they are investigating because of fear of losing data when the system they are working is turned off. Existing tools present to investigate digital forensic investigation is mostly based on the static mode of investigation i.e., data collected from storage media.

3.5 Cloud Forensics

In last decade, we all know that cloud computing has developed and spread a lot in the information technology. Cloud computing offers different kind of cloud related services to its users. In a study which was conducted in year 2016, it was found that average organization uses 1427 cloud services, which shows an increased no of services by 23.7% over a span of one year [51]. As shown in Figure 3.4, GAO [52] report on cyber related incidents is shown. Criminals uses cloud platform to gain access to the data stored on cloud by finding any vulnerabilities. Cloud platform can also be used by criminals to distribute false or doctored information to deceive others, they do so by concealing their identity, so that the law enforcement agents or the (LEA) cannot find them.

Cloud forensics is a subset of digital forensics, different definitions and terms defining cloud forensics exist, a survey was conducted by Ruan [53]. it was found that cloud forensic is basically comprised of traditional forensics and their application in cloud environment, it is not a new area of research. He presented three perspectives which include technical, organizational, and legal perspective of cloud forensic. In the technical perspective processes and procedures were described which include data identification of incident, live forensics, evidence collection, cloud environment information. In the organizational perspective people, which are related to cloud forensics are discussed, and in the legal perspective service level agreements and multi tenancy information, jurisdiction information was discussed.

Ruan [54] also defined cloud forensics as “*Cloud forensics is the application of digital forensic science in cloud computing environments. Technically, it consists of a hybrid forensic approach (e.g., remote, virtual, network, live, large-scale, thin-client, thick-client) towards the generation of digital evidence. Organizationally it involves interactions among cloud actors (i.e., cloud provider, cloud consumer, cloud broker, cloud carrier, cloud auditor) for the purpose of facilitating both internal and external investigations. Legally it often implies multi- jurisdictional and multi-tenant situations*”. Whereas NIST i.e., National Institute of Standards and Technology [55] defined cloud forensics as “*the application of scientific principles, technological practices and derived and proven methods to reconstruct past cloud computing events through identification, collection, preservation, examination, interpretation and reporting of digital evidence*”.

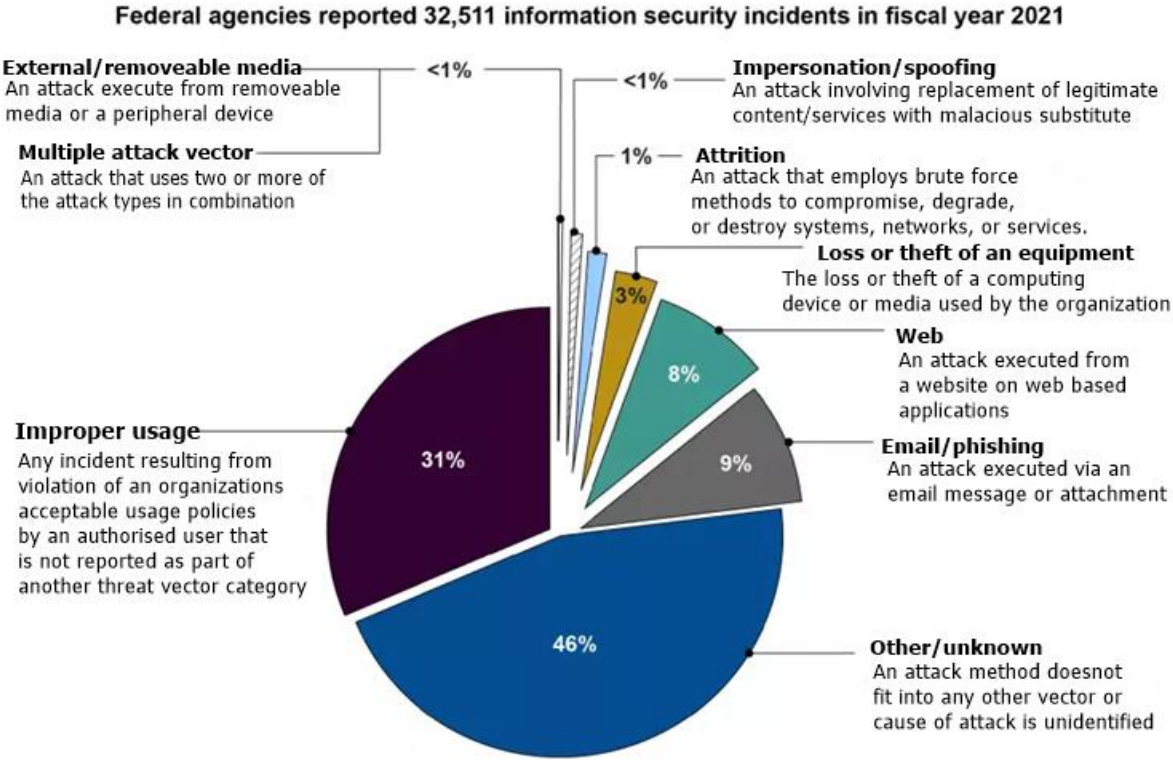


Figure 3.4: GAO Report for Cyber Related Incidents for Fiscal Year 2021.

3.5.1 Digital Evidence in Cloud Environment

Digital evidence in cloud computing refers to electronic data or information that is stored, processed, or transmitted through cloud services and is stored on distributed datacentres. Cloud computing involves the use of remote servers and networks to store and manage data, accessed over the internet. When legal cases involve cloud services, digital evidence may need to be collected and analysed from these cloud environments. Here are some key aspects related to digital evidence in the context of cloud computing:

- **Cloud Storage and Data Preservation:** Cloud service providers offer storage solutions where users can store their data remotely. Digital evidence may exist within these cloud storage platforms, such as files, documents, or backups. It's important to understand the terms of service and data retention policies of the cloud provider to ensure the availability and preservation of digital evidence.
- **Data Privacy and Jurisdiction:** Cloud computing often involves the storage and processing of data in various geographical locations, which can impact the legal aspects of digital evidence. Data privacy laws & jurisdictional considerations may come into play when accessing and collecting digital evidence from cloud services, especially when dealing with cross-border data transfers.
- **Accessing and Collecting Digital Evidence:** Accessing and collecting digital evidence from cloud services typically requires legal processes and cooperation with the cloud service provider. Law enforcement agencies or legal professionals may need to follow proper legal procedures, such as obtaining subpoenas, search warrants, or court orders, to access the relevant data stored in the cloud.
- **Metadata and Audit Logs:** Cloud services often maintain metadata and audit logs that can be valuable for digital evidence. Metadata, such as timestamps, access logs, and user activity records, can provide crucial context and help establish the authenticity and integrity of the evidence. These records may be used to track user actions, data transfers, or system relevant activities.
- **Chain of Custody and Authentication:** Maintaining the chain of custody and ensuring the authenticity of digital evidence collected from the cloud is essential. Proper documentation, including timestamps, log files, and secure handling practices, must be followed to establish the integrity and the admissibility of the evidence in cloud.
- **Compliance and Data Security:** Cloud service providers often implement security measures and comply with industry standards and regulations to protect customer data. The security and compliance posture of the cloud service provider should be considered when dealing with digital evidence in the cloud.

3.6 Discussion

In the Cloud Forensic Concepts chapter, we delved into the foundational principles of forensic investigation, establishing a comprehensive understanding of its historical evolution and critical importance in the digital age. We explored various types of forensics, from the

traditional realms to the emerging digital frontier, emphasizing the unique challenges posed by cloud computing environments. In this context, digital forensics took center stage, with its methodologies and techniques dissected for uncovering digital evidence. Finally, our journey led us to the evolving discipline of cloud forensics, where we acknowledged the necessity of adapting traditional forensic practices to the dynamic and complex nature of cloud-based technologies, setting the stage for a deeper exploration of this crucial field.

Chapter 4. Methodology

4.1 Introduction

The methodology introduces a cloud forensic framework, which will address the challenge of cloud forensic backlog which will enhance efficiency of digital investigations within cloud environments. The methodology defines cloud forensic framework, and integrates advanced techniques for data deduplication, and facilitate relevant data extraction by using topic modelling and check its performance using different machine learning models, and also helps prioritize those test cases which are of great priority, thus resulting in timely and fast investigation and for further analysis of data collected, we can use information extraction. Basically, this methodology offers a structured roadmap to solve the challenges posed by cloud forensic backlogs, ultimately ensuring the timely and proficient resolution of investigations in the dynamic landscape of cloud computing. This methodology gives a structured framework encompassing procedures, techniques, and tools tailored to navigate the intricacies of cloud-based digital evidence retrieval, preservation, analysis, and presentation. By providing a comprehensive roadmap for investigators and digital forensics experts, this methodology offers a strategic guide to mitigate complexities of cloud-related investigations and enhance accuracy and reliability of findings, thus contributing significantly to evolution of modern forensic practices.

4.2 Centralised Cloud Forensic Evidence System

A Centralized Cloud Forensic Evidence System represents a significant leap forward in streamlining the complex and often overwhelming task of cloud forensic investigations. This system employs a well-structured approach to reduce the cloud forensic backlog, employing a series of critical forensic processing stages: Cloud Forensic Evidence Collection, Disk Image Generation, and Classification of Pertinent Data as shown in Figure 4.1. This centralized system optimizes cloud forensic procedures by employing data deduplication, efficient disk image reconstruction, and machine learning-driven relevance assessment. By leveraging these techniques, it not only minimizes backlog but also enhances the overall efficiency and effectiveness of cloud forensic investigations, ultimately aiding in the pursuit of justice and security in cloud computing environments.

4.2.1 Cloud Forensic Evidence Collection

In the first stage, the Centralized Cloud Forensic Evidence Processing System initiates the collection of digital evidence from cloud-based sources. It utilizes a range of specialized tools and protocols to ensure the secure and reliable retrieval of data from various cloud platforms.

At this crucial stage, hash values are generated for each file obtained. These hash values serve a dual purpose - they not only confirm the integrity of the acquired data but also enable efficient data deduplication. The use of hash values facilitates the recognition of known illegal or benign files, effectively reducing redundancy in the forensic dataset.

4.2.2 Disk Image Generation

After the evidence acquisition process, system proceeds to reconstruct or generate forensic disk images. This step is vital for creating a coherent and forensically sound representation of the cloud environment under investigation. For virtualized cloud environments, employ authorized methods to create snapshots or images of virtual machines. Use forensically sound imaging tools to capture the entire state of the cloud instance, preserving volatile data and memory contents when possible. Verify the accuracy and completeness of disk image generation from cloud instances or virtual machines. Assess whether the generated disk images are faithful replicas of the original data in the cloud environment. Compare the extracted data with the original cloud data to confirm the integrity of the imaging process. By reconstructing disk images, the system ensures that all relevant data, including file structures and metadata, is preserved for analysis. This meticulous approach is fundamental in guaranteeing the completeness and accuracy of the investigation.

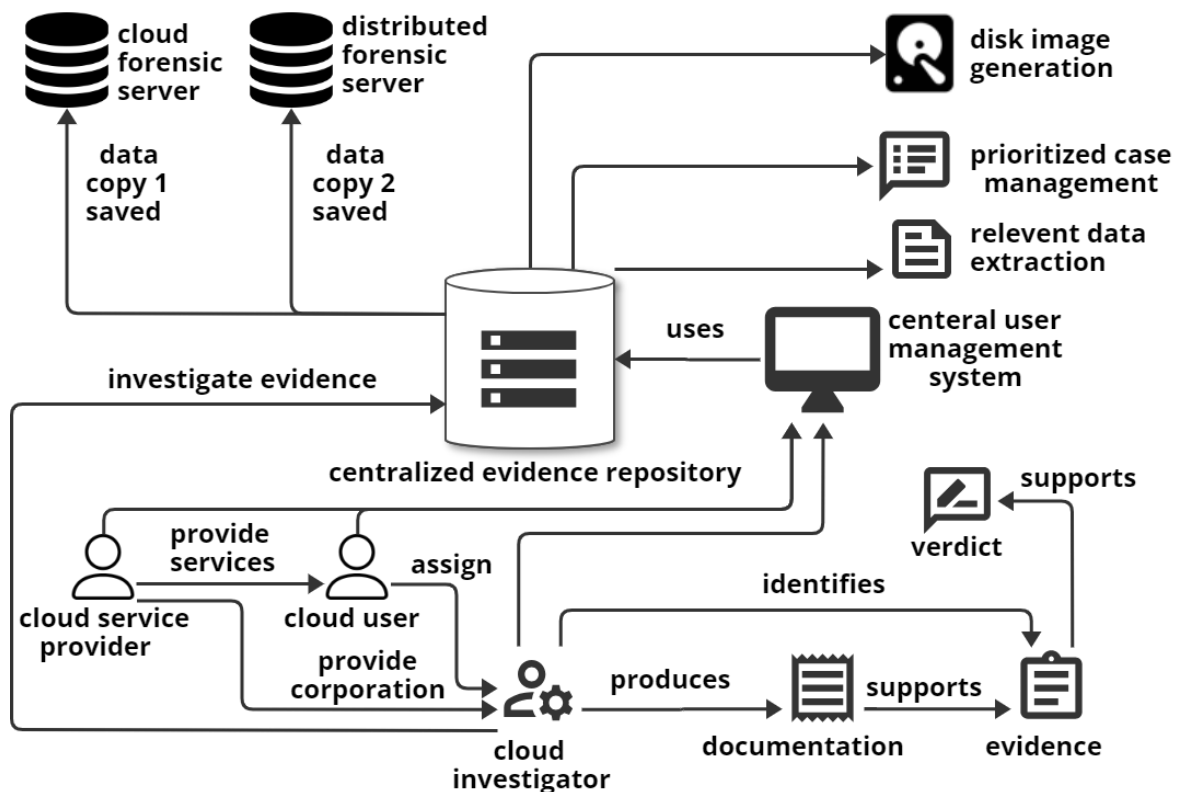


Figure 4.1: Centralized Cloud Forensic Evidence System.

4.2.3 Classification of Pertinent Data

The heart of the Centralized Cloud Forensic Evidence Processing System lies in its ability to extract relevant data efficiently. Here, the previously generated hash values play a pivotal role. Known files, identified through their hash values, serve as a foundational dataset for training machine learning models. These models, once trained, possess the capability to classify and prioritize files that are likely to be more pertinent to the investigation. This intelligent classification significantly reduces workload for the forensic analysts, enabling them to focus their attention on crucial evidence, expediting the entire forensic process.

4.3 Cloud Forensic Analysis Techniques to Reduce Backlog

There are different types of cloud forensic related analysis software in the market. By using these software's, we can alleviate cloud forensic process particularly when we are dealing with substantial forensic data for investigations. We can collect cloud forensic evidence by two ways, one is extracting the deduplicated data, other is prioritizing forensic data. The processes of Known Data Acquisition and Unknown Data Relevancy and Prioritization are crucial steps that significantly enhance the efficiency and effectiveness of the cloud forensic investigations.

4.3.1 Known Data Acquisition

This process employs hashing techniques to eliminate duplicate data within the forensic dataset. Duplicate data can often clutter investigations and waste valuable time. By generating hash values for each file and comparing them, the system can quickly identify and eliminate redundant copies of files. Moreover, it also checks the existence of files within a central storage repository which is known file database, which contains records of files that are already recognized as illegal files or benign files. By referencing this database, the system can promptly classify files, further reducing the forensic workload. This deduplication process not only conserves storage space but also streamlines subsequent analysis.

4.3.2 Unknown Data Relevancy and Prioritization

After the deduplication phase, the focus shifts to extracting and prioritizing relevant data. This data typically starts as unlabelled and unstructured, making it challenging for investigators. However, through machine learning techniques, this data can be converted into labelled data. Machine learning models are trained on known files, including those flagged as illegal during previous investigations. These trained models can then be applied to the unlabelled data to identify files that are likely to be relevant to the current investigation. Importantly, this process involves using different detection models to flag potentially illegal files. These flagged files are then prioritized for closer examination in the cloud forensic

investigation. Prioritization is a key strategy for reducing backlog since it allows investigators to focus their efforts on the most critical and suspicious files first. Moreover, the data extracted during this phase, both labelled and flagged, can be used as valuable input for model training, continuously improving the system's ability to classify and prioritize new files in future investigations.

4.4 Alleviating Cloud Forensic Backlog Methodology

The Cloud Forensic Framework implemented as part of this research here is a comprehensive approach designed to effectively tackle the challenges associated with cloud-based digital investigations and reduce the backlog that often accumulates in such scenarios. This framework encompasses a series of interconnected processes that collectively streamline the investigation process and enhance its efficiency. The Cloud Forensic Process Flow constitutes the foundation of this framework, outlining a step-by-step guide to conducting cloud-related investigations, from evidence identification and collection to analysis and presentation. Alleviating Cloud Forensic Backlog Method is shown in Figure 4.2.

4.4.1 Data Deduplication using Hashing

A pivotal element of this framework is data deduplication by hashing, a technique that ensures the elimination of redundant data, thereby optimizing storage and expediting analysis. By employing cryptographic hashing algorithms, duplicate files within the cloud environment can be identified and removed, leading to a more streamlined investigation process. Data deduplication using hashing is a technique employed in digital forensics and data management to identify and eliminate redundant copies of data within a storage system, thereby optimizing storage space and improving efficiency. This process involves generating a unique hash value for each piece of data and comparing these hash values to identify duplicates. Hashing algorithms, such as MD5, SHA-1, and SHA-256, are commonly used for this purpose.

4.4.2 Relevant Data Extraction and Prioritization

To focus on the relevant data extraction and prioritization, the framework incorporates advanced data mining strategies. This involves intelligent data filtering techniques that sift through the data corpus, identifying and prioritizing information based on predefined criteria. This targeted approach not only accelerates the investigation but also enhances the precision of findings. Relevant data extraction and prioritization in cloud forensic investigations involve the systematic process of identifying, extracting, and organizing data that is pertinent to the investigation at hand. Given the vast amount of data stored in cloud environments, this process is crucial to streamline investigations, reduce backlog, & focus resources on most critical information.

4.4.3 Information Extraction

By using Named Entity Recognition (NER) and Relation Extraction (RE) is a sophisticated approach employed in natural language processing and text analysis to identify specific entities within text data and understand the relationships between them. This technique has valuable applications in various domains, including cloud forensic investigations. To further enhance the information extraction phase, the framework integrates Named Entity Recognition (NER) utilizing BERT (Bidirectional Encoder Representations from Transformers). This natural language processing technique enables the identification of key entities within textual data, thereby aiding in the extraction of valuable information. Building upon NER, the framework incorporates relation extraction, a process that uncovers meaningful connections between entities within data. By identifying relationships, patterns, and interactions, this stage adds depth & context to investigation, potentially revealing hidden insights crucial to the case.

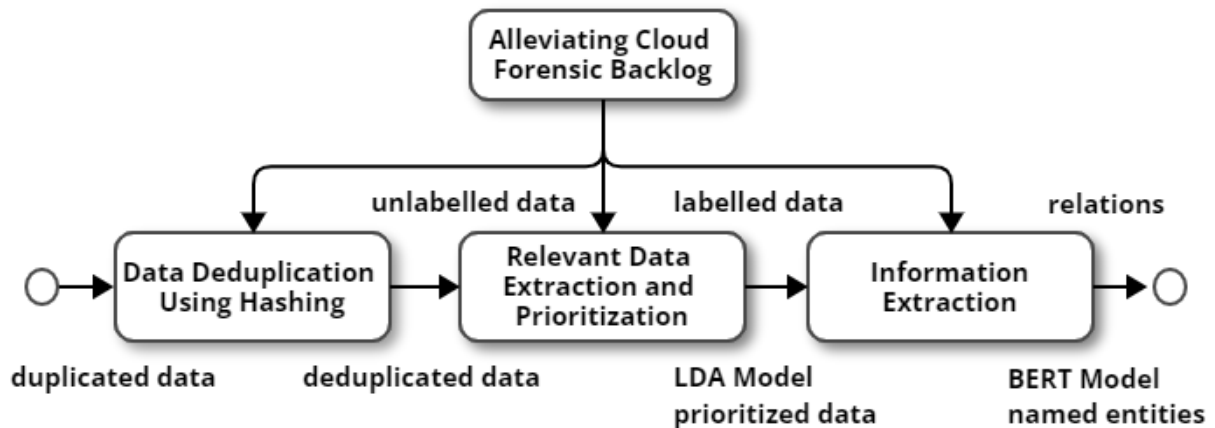


Figure 4.2: Alleviating Cloud Forensic Backlog Method.

4.5 Experimentation Design

Designing an effective experimentation plan for the four processes involved in reducing cloud forensic backlog which include Disk Image Generation, Deduplication, Illegal File Detection, and Information Extraction. Here's a structured experimentation design for each of these processes:

- **Disk Image Generation using FTK Imager:** Selected dataset includes cloud storage services, file types, and sizes. Before disk image generation, measure the initial forensic backlog by recording the total data volume and number of files within the chosen dataset. Generate disk images from the selected cloud environments using FTK Imager.
- **Deduplication using Hashing:** We will use disk images generated in the first experiment. Then we will calculate hash values for all files in the dataset. And by

applying deduplication process using hashing we will identify and eliminate duplicate files. We will measure the reduction in data volume, the number of duplicate files removed, and time taken i.e., effective, and actual acquisition time for deduplication. We also verify that deduplication does not compromise data integrity by confirming that hash values match.

- **Illegal File Detection using LDA Model and Flagging Relevant Data:** Prepare a labelled dataset containing known illegal and benign files. This dataset should be representative of cloud storage environments. Train the LDA model using the prepared dataset for classification of files. Apply the trained LDA model to the deduplicated dataset from the previous experiment. Flag files identified as potentially illegal. Measure the model's accuracy, precision, recall, and F1-score for illegal file detection. Also, track number of files flagged as relevant. Investigate flagged files to determine actual relevance to investigation.
- **Information Extraction using BERT and RE:** For analysis apply BERT-based NER to extract named entities from the dataset. And using association rule mining to discover relationships between entities in the data. Evaluate the precision, recall, and F1-score for NER and assess quality of discovered associations for RE.

4.6 Discussion

The Centralised Cloud Forensic Evidence System represents a crucial advancement in the digital forensics in cloud environment, offering a centralized approach that harnesses a combination of Cloud Forensic Analysis Techniques to Reduce Backlog and an Alleviating Cloud Forensic Backlog Methodology. This integrated approach not only streamlines evidence processing but also enhances the effectiveness of investigations. By leveraging techniques such as data deduplication, efficient evidence acquisition, machine learning-based illegal file detection, and advanced information extraction, this system significantly reduces the backlog of cloud forensic cases, allowing investigators to focus their resources on the most relevant and critical tasks, ultimately improving the efficiency and accuracy of cloud forensic analysis.

Chapter 5. Cloud Forensic Framework for Reducing Backlog (CFFRB)

5.1 Introduction

Cloud investigation process refers to the systematic and methodical examination of digital evidence within cloud computing environments. It involves the identification, collection, preservation, analysis, and reporting of evidence related to security incidents, data breaches, unauthorized access, or other malicious activities occurring in cloud-based systems. The process typically begins with incident detection and reporting, followed by evidence identification and preservation, data collection, analysis, and finally, the generation of a comprehensive forensic report. Cloud investigation is necessary to uncover the truth behind cyber incidents, identify responsible parties, mitigate risks, and support legal proceedings. It helps organizations understand the extent of the compromise, assess the impact, and take appropriate actions to prevent future incidents. Cloud investigation process specifically focuses on digital evidence within cloud computing environments, considering the distributed nature, dynamic characteristics, and legal complexities associated with cloud systems. It requires specialized knowledge, collaboration with CSPs, and an understanding of cloud-specific technologies to effectively investigate and analyse evidence in a forensically sound manner.

While cloud investigation is a subset of digital forensics, it has some unique characteristics and considerations that differentiate it from traditional digital forensic processes. One key difference lies in the distributed nature of cloud computing. Cloud environments are composed of multiple servers, storage systems, and networks spread across different locations and managed by cloud service providers (CSPs). This requires investigators to understand and navigate the complex infrastructure and collaboration with CSPs to obtain access to relevant data and logs.

Moreover, dynamic nature of cloud computing adds complexity to the investigation process. Virtual machines & resources can be provisioned, deprovisioned, or migrated, potentially affecting integrity and availability of evidence. Investigators need to consider potential volatility of cloud resources and ensure that evidence is properly preserved and is collected in a timely manner. Additionally, privacy & legal challenges related to jurisdiction and data protection arise in cloud investigations. Data may reside in different geographic locations or be subject to different laws & regulations. Investigators must work closely with legal teams to ensure compliance with applicable laws and regulations, obtain necessary permissions, & handle cross-border data transfer and storage related issues.

5.2 Cloud Forensic Investigation Process Flow

The process flow of cloud forensics involves several key steps to effectively investigate and analyse digital evidence in a cloud computing environment. Throughout the process, collaboration with the CSP, legal teams, and other stakeholders is crucial. Effective communication, adherence to legal requirements, and proper documentation are essential to ensure a thorough and reliable cloud forensic investigation. The specific steps and techniques employed may vary depending on the nature of the incident, the cloud environment, and the available resources and tools. Cloud forensics requires a combination of technical expertise, knowledge of cloud computing architectures, and proficiency in digital forensic techniques to effectively investigate and analyse digital evidence within a cloud computing environment. After going through several research (Simou et al., 2014b, Simou et al., 2015, Simou et al., 2016b) proposed a general cloud forensic process, which has 4 stages that were identification, Collection/Acquisition, Examination/Analysis, and presentation which is discussed in the cloud forensics and its concept chapter. But with our nature of investigation, we propose a cloud forensic process flow which have nine stages. Five stages of the process are key stages, which are incident identification and reporting stage, case initiation and planning stage, evidence collection and acquisition stage, evidence examination and analysis stage, and the last stage is presentation and legal proceedings. The cloud forensic process flow is depicted as fallow in Figure 5.1:

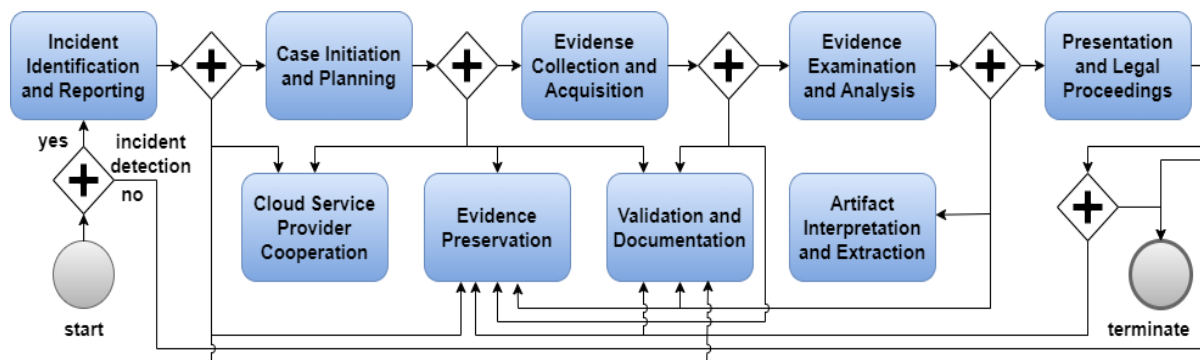


Figure 5.1: Cloud Forensics Process Flow.

5.2.1 Incident Identification and Reporting

Incident identification and reporting is the initial step in the process for cloud forensic investigation. It involves recognizing and documenting the incident or suspicious activity that requires investigation within the cloud environment. Incident identification & reporting sets the foundation for a successful cloud forensic investigation. Timely detection and reporting of incidents enable rapid response and mitigation measures to minimize potential damage. Proper documentation and communication during this phase ensure that all relevant

information is captured and shared with the necessary parties, setting the stage for subsequent steps in the investigation process.

Cloud forensic investigations often begin with the detection of an anomaly or an event that raises suspicion of unauthorized access, data breach, system compromise, or any other security incident. Detection mechanisms may include intrusion detection systems (IDS), security monitoring tools, log analysis, or reports from users or system administrators. Once an anomaly or incident is detected, it is important to triage the incident to determine its severity and prioritize investigation efforts. Initial assessment involves understanding the nature of the incident, potential impact, and any immediate actions required to contain or mitigate the situation. The incident needs to be promptly reported to the appropriate stakeholders, including the cloud service provider (CSP) and internal security teams. A formal incident report should be prepared, documenting key details such as the date and time of the incident, a brief description of the event, and any available evidence or indicators. Gather as much information as possible about the incident, including logs, system alerts, user reports, network traffic captures, or any other relevant data sources. Document the sources and locations of the potential evidence to ensure it can be later retrieved and analysed during the investigation. Notify all necessary parties involved in the investigation process, such as the incident response team, legal and compliance teams, and senior management. Establish clear lines of communication and coordination to ensure effective collaboration among different stakeholders. Take immediate steps to preserve the integrity of potential digital evidence. Secure the affected cloud resources, systems, or accounts to prevent further compromise or alteration of evidence. Work closely with the CSP to ensure that necessary data and logs are preserved and not overwritten or deleted.

5.2.2 Case Initiation and Planning

Case initiation and planning is a crucial phase in the process for cloud forensic investigation. It involves setting the foundation for the investigation by defining the scope, objectives, and timeline, as well as identifying the resources and strategies required to conduct a thorough examination. Case initiation marks the official start of the cloud forensic investigation after an incident has been identified and reported. An investigator or forensic examiner is assigned to lead the investigation and coordinate the activities involved. The investigator gathers initial information about the incident, reviews the incident report, and identifies key stakeholders.

The planning phase involves developing a comprehensive strategy to guide the investigation process. In this we define the scope and objectives of the investigation, including the specific areas, systems, or cloud resources to be examined. Identifying the legal and regulatory requirements that apply to the investigation, ensuring compliance with privacy laws and obtaining necessary permissions. Also determining the available resources, such as personnel,

tools, and budget, needed to conduct the investigation effectively. And establishing a timeline or schedule, by considering any critical deadlines, dependencies, or constraints.

5.2.3 Evidence Identification and Preservation

Evidence identification and preservation is a crucial phase in the process for cloud forensic investigation. It involves identifying potential sources of digital evidence within the cloud environment, such as virtual machines, storage accounts, logs, and network data. And collaborating with the CSP to ensure the preservation of relevant evidence by securing access to the affected resources and taking forensic copies or snapshots of the data. And establishing a proper chain of custody to maintain the integrity of the evidence. By properly identifying and preserving evidence, investigators can lay the foundation for subsequent analysis and interpretation, contributing to the overall success of the investigation process.

The identification of evidence begins by determining the potential sources of digital evidence within the cloud environment. This includes identifying relevant cloud resources, such as virtual machines, storage accounts, databases, logs, and network data. Investigate potential artifacts that may contain evidence, such as user accounts, access logs, system configurations, or communication records. Once potential evidence is identified, preservation measures must be implemented to maintain its integrity and prevent tampering. Collaborate with the cloud service provider (CSP) to ensure the preservation of relevant evidence. Securing access to affected cloud resources to prevent unauthorized modifications or deletion of data. And taking forensic copies or snapshots of the data to create a replica for analysis, while ensuring the original evidence remains untouched. And lastly maintain a strict chain of custody to track the handling and movement of evidence, documenting each transfer to ensure its admissibility in legal proceedings. We should keep in mind that cloud environments may present unique challenges for evidence identification and preservation due to their distributed and shared nature. Consider the dynamic nature of cloud resources, where data can be replicated or moved across different data centres or regions and data encryption and decryption processes, as well as the availability and retention policies set by the CSP. Also, considering any legal or contractual obligations regarding data preservation and privacy.

5.2.4 Cloud Service Provider Cooperation

In the ongoing cloud forensic process, we establish communication and collaboration with the cloud service provider to understand their infrastructure, logging mechanisms, and available resources. Request and collect relevant information from the provider, such as access logs, configuration details, and virtual machine snapshots. Cloud Service Provider (CSP) cooperation plays a vital role in the process of cloud forensic investigation. As cloud

environments are managed and controlled by CSPs, their active participation and cooperation are essential to access and retrieve relevant information.

CSPs have a responsibility to comply with legal and regulatory requirements regarding data access, retention, and disclosure. Cooperation from CSPs ensures that investigators can obtain the necessary permissions and access rights to conduct the investigation within the boundaries of the law. CSPs control the infrastructure and resources within the cloud environment, including virtual machines, storage, networks, and logs. They allow investigators to gain authorized access to these resources, enabling the collection of relevant evidence. CSPs play a critical role in preserving and retaining data within their cloud infrastructure. CSPs ensure that the relevant data is not inadvertently modified, deleted, or overwritten during the investigation process. CSPs possess valuable technical knowledge and expertise about their cloud platforms and services. CSPs can provide investigators with insights, guidance, and technical assistance in navigating the cloud environment and retrieving evidence effectively. Effective communication and prompt response from CSPs are crucial for the smooth progress of the investigation. CSPs can respond to queries, providing necessary documentation, and assisting in resolving technical issues expedites the investigation process. CSPs' cooperation in maintaining the chain of custody and preserving the forensic integrity of evidence is essential. Their assistance in documenting the handling, transfer, and storage of evidence ensures its admissibility and reliability in legal proceedings.

5.2.5 Evidence Collection and Acquisition

Evidence collection and acquisition in the process of cloud forensic investigation is a crucial step that involves gathering relevant digital evidence from cloud resources. We employ appropriate forensic tools and techniques to collect data from the identified cloud resources and capturing network traffic, system logs, and user activities within the cloud environment and collecting relevant metadata, including timestamps, user identifiers, and file attributes associated with the evidence.

Data collection and acquisition focus on identifying and retrieving electronic evidence from various cloud sources, such as virtual machines, storage systems, databases, and network logs. Forensic investigators employ specialized tools and techniques to acquire the data in a forensically sound manner, ensuring its integrity and preserving the chain of custody. The methods used for data collection may include creating forensic copies or snapshots of cloud resources, extracting relevant files and metadata, and capturing network traffic. Collaboration with cloud service providers (CSPs) is necessary to gain access to the cloud resources and obtain the required permissions and credentials. Proper documentation of the data collection process, including timestamps, file attributes, and relevant metadata, ensures the admissibility and reliability of the acquired evidence. Data collection and acquisition phase are essential

for building a comprehensive and accurate picture of the digital evidence within the cloud environment. It provides the foundation for subsequent analysis and interpretation, enabling investigators to uncover insights, establish facts, and support legal proceedings. By employing robust methodologies and working closely with CSPs, forensic investigators can effectively collect and acquire the necessary data for a thorough cloud forensic investigation.

5.2.6 Evidence Examination and Analysis

Evidence examination and analysis in cloud forensic investigation require a combination of technical expertise, analytical skills, and a deep understanding of cloud technologies. It enables investigators to derive meaningful insights from the collected evidence, reconstruct events, and provide valuable information for legal proceedings or incident response. By effectively analysing and reconstructing the data, investigators can draw conclusions, make informed decisions, and contribute to the resolution of the cloud-related incident or crime at hand. It analyses the collected data to reconstruct events and identify potential digital artifacts relevant to the investigation and examine log files, access records, and user activity logs to trace the activities within the cloud environment to correlate different data sources to establish timelines and relationships between cloud resources and user actions.

It involves examining and interpreting the collected digital evidence to reconstruct events and uncover relevant information. Data analysis and reconstruction aim to make sense of the acquired digital evidence and identify patterns, correlations, and relationships among different data points. Investigators employ various techniques and tools to analyze the data, such as data mining, keyword searches, timeline analysis, and correlation analysis. The analysis may involve reconstructing activities, timelines, and user interactions within the cloud environment to understand the sequence of events and the actions taken by relevant parties. Investigators may also utilize forensic techniques to recover deleted or modified data, decrypt encrypted information, and extract hidden or obscured information. The findings from the data analysis phase help in establishing facts, identifying potential culprits or malicious activities, and supporting the overall investigation process.

5.2.7 Artifact Interpretation and Extraction

Artifact interpretation and extraction require a deep understanding of cloud technologies, file systems, network protocols, and system configurations. Investigators need to be proficient in using forensic tools and techniques to accurately extract, analyse, and interpret the artifacts within the cloud environment. The insights gained from this phase contribute to building a comprehensive understanding of the digital evidence and provide valuable information for further investigation, incident response, or legal proceedings. We extract and interpret relevant digital artifacts, such as files, emails, databases, or application data, from the

collected evidence and recovering deleted or hidden data using appropriate forensic techniques. Also analysing encryption mechanisms, if applicable, and attempt to recover encrypted data, if necessary.

It involves analysing and extracting relevant artifacts from the digital evidence collected during the investigation. Artifact interpretation and extraction focus on identifying and interpreting various artifacts present within the digital evidence, such as files, logs, metadata, system configurations, and user activities. Investigators utilize specialized tools and techniques to extract and analyse these artifacts, uncovering valuable information related to user actions, system events, communication patterns, and potential security breaches. The interpretation of artifacts helps investigators understand the context, significance, and potential implications of the evidence within the cloud environment. Common artifacts that are examined and interpreted include file metadata, email headers, browser history, system logs, network traffic logs, and user account information. By extracting and interpreting artifacts, investigators can reconstruct events, establish timelines, identify key actors, and gather evidence to support their findings.

5.2.8 Document Validation and Reporting

Document the findings, analysis processes, and techniques used during the investigation. Prepare a comprehensive forensic report that presents the evidence in a clear and concise manner. Include relevant details such as timestamps, activities, findings, and any other information that supports the investigation. It involves documenting the entire investigation process and preparing comprehensive reports to present the findings and conclusions. Investigators maintain detailed records of the evidence collected, the analysis performed, and the actions taken throughout the investigation process. Validation in cloud forensic investigation involves verifying the accuracy and integrity of the collected evidence, analysis results, and conclusions drawn from the investigation. Validation helps ensure that the evidence and analysis are reliable, consistent, and free from errors or biases.

Comprehensive reports are prepared to summarize the investigation process, present the findings, and provide a clear overview of the evidence and its significance. The reports include a description of the investigation objectives, the methodologies used, the analysis performed, and the conclusions drawn based on the findings. Additionally, the reports may include recommendations for future actions, suggestions for improving security measures, and any legal or regulatory implications that arise from the investigation. Effective documentation and reporting ensure transparency, accountability, and traceability throughout the cloud forensic investigation. It provides a clear and organized record of the investigation process, allowing stakeholders to understand the methodology used and the validity of the findings. Furthermore, the documentation and reports serve as essential artifacts for legal

proceedings, enabling the presentation of the investigation's findings and supporting the case in a court of law if necessary.

5.2.9 Presentation and Legal Proceedings

Present the forensic findings as expert testimony, if required, in legal proceedings. Collaborate with legal professionals to ensure the admissibility of the digital evidence in court. Assist in the preparation of legal strategies based on the forensic analysis. It's important to note that the specific steps and techniques may vary depending on the cloud environment, the nature of the investigation, and the available resources and tools. Cloud forensics requires a combination of technical expertise, knowledge of cloud computing architectures, and proficiency in digital forensic techniques to effectively investigate and analyse digital evidence within a cloud computing environment. This phase is critical for the effective utilization of the investigation results and the pursuit of legal actions, if necessary. By presenting the findings in a compelling and accurate manner, investigators facilitate decision-making, support legal proceedings, and contribute to the resolution of the cloud-related incident or crime. Effective collaboration with legal teams ensures that the investigation results are properly interpreted and can be effectively used in the legal process.

Investigators communicate results, analysis, and conclusions to stakeholders, such as management, legal teams, or law enforcement agencies. Investigators prepare and deliver presentations that summarize the investigation process, highlight key findings, and provide a comprehensive overview of the evidence and its significance. The presentation may include visual aids, such as charts, graphs, or timelines, to effectively convey complex information and make it accessible to non-technical audiences. Clear and concise communication of the findings ensures that stakeholders understand the implications, can make informed decisions, and take appropriate actions based on the investigation results. In legal proceedings, investigators may be required to present their findings and provide expert testimony to support the case. Investigators collaborate with legal teams to ensure that the investigation findings and evidence comply with legal requirements, regulations, and standards.

5.3 Cloud Forensic Constraints for Reducing Backlog

Cloud forensic constraints play a pivotal role in shaping strategies to reduce backlog in cloud forensic investigations. These constraints stem from the unique characteristics of cloud environments and influence the methodologies and techniques used in tackling the backlog. Some key cloud forensic constraints relevant to reducing the backlog include:

- **Virtualization and Abstraction:** The virtualized nature of cloud resources complicates direct access to physical hardware, hindering traditional forensic

practices. Investigators must navigate the abstraction layers introduced by virtualization while preserving evidence integrity.

- **Data Distribution and Fragmentation:** Cloud data can be dispersed across various geographical locations and storage nodes, making evidence collection complex. Strategies to efficiently gather fragmented data while maintaining a coherent investigative trail are crucial.
- **Multi-Tenancy:** Cloud environments often involve multiple clients sharing the same infrastructure. Investigators must navigate data segregation challenges to ensure the integrity of collected evidence and prevent cross-contamination.
- **Dynamic Resource Allocation:** Cloud resources can be dynamically allocated and de-allocated, impacting the stability of evidence over time. Developing methods to capture the state of resources at a specific point and preserve evidence through dynamic changes is essential.
- **Encryption and Access Control:** Strong encryption and access controls are common in cloud services. While enhancing data security, these mechanisms challenge investigators' access to relevant data. Overcoming encryption barriers without compromising security is a constraint that demands innovative solutions.
- **Jurisdictional Complexity:** Cloud data may be stored in various jurisdictions, necessitating compliance with different legal frameworks. Managing cross-border investigations and adhering to varying regulations can be complex and time-consuming.
- **Evidence Integrity:** Ensuring the integrity and authenticity of evidence in a highly distributed and dynamic cloud environment poses a significant challenge. Strategies for maintaining evidence integrity despite changes in resource allocation are essential.
- **Lack of Standardization:** The lack of standardized cloud forensics procedures and tools can lead to inconsistencies in investigations. Developing adaptable approaches that account for variations among cloud providers is critical.
- **Scale and Volume:** Cloud environments handle vast amounts of data, leading to challenges in collecting, processing, and analysing large datasets efficiently. Scalable techniques for evidence handling and analysis are crucial.
- **Logging and Audit Trails:** Cloud providers maintain extensive logs and audit trails, but these can be dispersed across services. Extracting, aggregating, and

interpreting these logs efficiently to reconstruct events is a constraint that requires specialized techniques.

- **Resource Limitations:** Cloud service providers often impose limitations on investigative resources, affecting the scope and efficiency of analysis. Developing resource-efficient methods while maintaining investigation quality is essential.

Addressing these constraints through innovative methodologies and techniques is essential to effectively reduce the backlog of cloud forensic cases. Each constraint presents a unique challenge that requires tailored solutions to ensure the timely and accurate resolution of investigations.

5.4 Cloud Forensic Framework for Reducing Backlog (CFFRB)

Cloud forensic investigations involve the collection, analysis, and preservation of digital evidence from cloud environments to uncover security breaches, data breaches, and other cybercrimes. As the volume of digital data generated in cloud environments continues to grow, a backlog of forensic cases can accumulate, leading to delays in investigations, potential loss of evidence, and compromised security. To address this challenge, a Cloud Forensic Framework for Reducing Backlog (CFFRB) has been developed.

The Cloud Forensic Framework for Reducing Backlog is a systematic approach designed to streamline and speed up conducting forensic investigations in cloud environments as shown in Figure 5.2. It combines various techniques, tools, and methodologies to enhance the efficiency and effectiveness of cloud forensic activities while minimizing the backlog of pending cases. The framework encompasses the following key elements:

- **Automated Data Collection:** Traditional forensic processes often involve manual data collection, which can be time-consuming and error prone. CFFRB emphasizes use of automated data collection tools & scripts to gather relevant evidence from cloud platforms. This reduces time required for data acquisition and minimizes chances of human errors.
- **Scalable Analysis:** Cloud environments generate massive amounts of data. CFFRB promotes the use of scalable analysis techniques, such as parallel processing and distributed computing, to accelerate the examination of evidence. By leveraging inherent scalability of cloud resources, investigators can analyse data faster and reduce backlog.
- **Prioritized Case Management:** Not all forensic cases have same level of urgency. CFFRB implements case prioritization mechanism that ensures critical

- **Machine Learning and AI:** Leveraging machine learning and artificial intelligence (AI) can significantly speed up the analysis process by automating tasks such as pattern recognition, anomaly detection i.e., our case we detected fraudulent emails by using LDA, and correlation of evidence. CFFRB integrates these technologies to assist investigators in identifying relevant information more rapidly.
- **Real-time Monitoring:** Implementing real-time monitoring and alerts within the cloud environment helps detect and respond to security incidents promptly. By addressing potential threats early, framework reduces the number of cases that may lead to backlogs.
- **Collaborative Workflows:** CFFRB promotes collaboration among various stakeholders, including forensic analysts, legal teams, and IT personnel. This collaborative approach ensures that investigations proceed smoothly, with insights from different perspectives contributing to more comprehensive results.

Cloud Forensic Framework for Reducing Backlog (CFFRB) is a comprehensive approach that is designed to tackle the growing backlog of cloud forensic cases. By incorporating automation, scalability, prioritization, centralized storage, advanced technologies like machine learning, and collaborative workflows, the framework aims to expedite investigations, enhance accuracy, and improve overall cloud security. As cloud computing continues to evolve, CFFRB serves as a crucial tool to address the challenges of digital forensics in the cloud era.

5.5 Discussion

The chapter outlined a structured Cloud Forensic Investigation Process Flow, providing a step-by-step guide to navigate the complexities of cloud-based environments efficiently. We also explored the unique constraints posed by cloud forensics in the context of backlog reduction, emphasizing the need for adaptive methodologies and specialized tools. The Cloud Forensic Framework for Reducing Backlog (CFFRB) emerged as a pivotal solution, orchestrating the integration of deduplication, efficient data extraction, machine learning-based prioritization, and intelligent information extraction techniques. By addressing these critical aspects, CFFRB offers a holistic strategy to enhance the effectiveness and timeliness of cloud forensic investigations while reducing the backlog, ultimately contributing to improved security and accountability in cloud computing environments.

Chapter 6. Data Deduplication of Cloud Forensic Evidence

6.1 Introduction

Data deduplication is a technique used to reduce storage needs by identifying and eliminating duplicate data. It involves analysing data to identify identical data blocks and storing only one copy of each block, with subsequent references pointing to the original copy. This can help to save storage space and reduce costs, particularly in large-scale storage environments such as data centres. There are several different types of data deduplication techniques, including:

- **File-level deduplication:** Identifying & removing duplicate, regardless of their contents.
- **Block-level deduplication:** Identifying & removing duplicate blocks of data within files.
- **Inline deduplication:** Performing deduplication in real-time as data is being written.
- **Post-process deduplication:** Performing deduplication after data written to storage.

Data deduplication is widely used in backup and disaster recovery applications, as it can help to reduce backup times and storage requirements. It is also used in virtualized environments, where multiple virtual machines may share common data blocks.

However, it is important to note that data deduplication can be resource-intensive and may have an impact on system performance. It is therefore important to carefully consider the use of data deduplication and to choose a deduplication approach that is appropriate for the specific storage environment and workload.

The main goal of data deduplication is to reduce the amount of storage space required to store data by identifying and eliminating duplicate data. This can be achieved by using algorithms to compare data and identifying duplicate copies. Once duplicates are identified, only one copy of the data is stored, and subsequent references to that data point to the single copy. Data deduplication has become increasingly important in recent years due to the explosion of digital and cloud data, which has led to growing storage needs and increased costs. By reducing the amount of data that needs to be stored, deduplication can help organizations

save on storage costs, improve backup and disaster recovery times, and increase the overall efficiency of their data management processes.

In addition to its storage benefits, data deduplication can also help to improve data integrity, as it ensures that only one copy of a piece of data is stored, eliminating the risk of conflicting or inconsistent copies of data. This can be particularly important in fields such as healthcare and finance, where data accuracy and consistency are critical. Overall, the main goal of data deduplication is to improve the efficiency and effectiveness of data storage and management processes by reducing storage requirements and improving data integrity.

6.2 Data Deduplication to Reduce Storage in Cloud Forensics

Data deduplication can be an effective way to reduce storage requirements of forensic evidence. In cloud forensics, cloud investigators often deal with large amounts of data, including disk images, mobile phone backups, and other types of digital evidence extracted from cloud environment. Deduplication can help to reduce the amount of storage required for this data by identifying and removing duplicate data. There are several advantages to using data deduplication in cloud forensics:

- **Reduced storage requirements:** By identifying and removing duplicate data, deduplication can significantly reduce the amount of storage required to store digital evidence. This can save time and money in terms of hardware and storage costs.
- **Faster data processing:** Deduplication can also speed up the data processing and analysis phase of digital forensics. With less data to process, investigators can analyse the evidence more quickly and efficiently.
- **Improved accuracy:** By removing duplicate data, investigators can be sure that they are only analysing unique data, which can improve the accuracy of their findings.
- **Preservation of original evidence:** Deduplication can help to preserve the original evidence by removing duplicate copies. This can be important in legal cases, where the authenticity and integrity of the evidence must be maintained.

However, it is important to note that data deduplication should be carried out carefully to avoid the loss of important evidence. Investigators should use reliable and accurate deduplication tools and should have a clear understanding of the data being processed. They should also ensure that the original evidence is preserved and that any duplicates that are removed are properly documented.

6.3 Data Deduplication Process

Deduplication process is to reduce the storage footprint of data while maintaining its integrity and accessibility. The specific method and implementation of deduplication may vary depending on factors such as the type of data, the available storage resources, the desired level of deduplication, and the performance and security requirements of the system. The process of deduplication as shown in Figure 6.1, which typically involves the following steps:

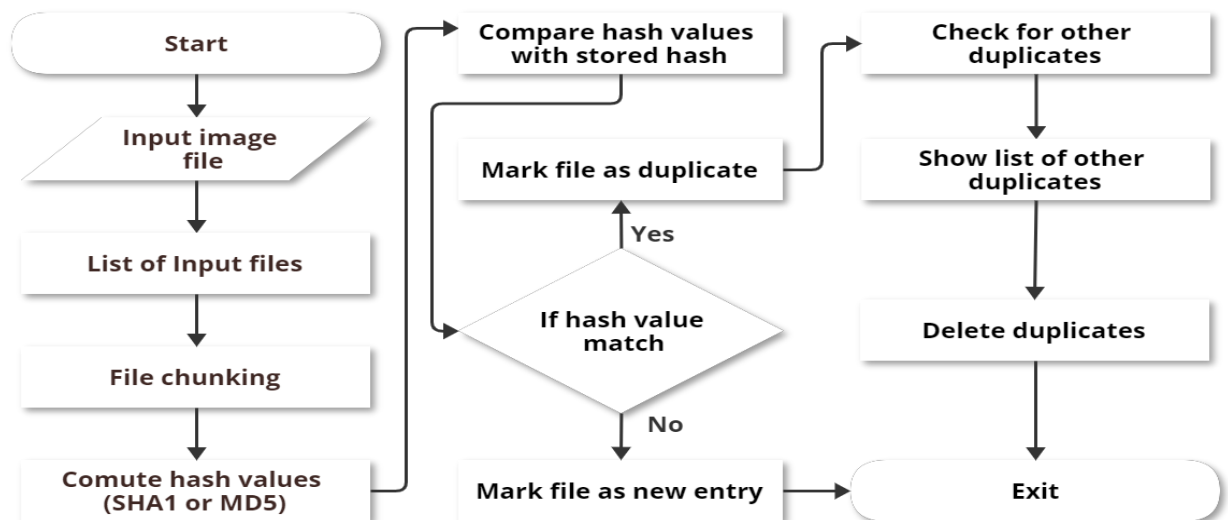


Figure 6.1: Data Deduplication Process Flow.

- **Identify the data to be deduplicated:** This can include files, databases, backups, archives, or other types of data that contain duplicate information.
- **Create a hash of each data block:** A hash function is used to create a fixed-size digital fingerprint of each block of data. This fingerprint is unique to the content of the block and is used to identify duplicates.
- **Compare the hashes:** The hashes are compared to identify duplicate blocks. This can be done using a hash table or other data structure that allows fast lookup and comparison.
- **Replace duplicate blocks with pointers:** When a duplicate block is found, it is replaced with a pointer or reference to the original block. This eliminates redundant data and saves storage space.
- **Maintain integrity and security:** Deduplication can introduce potential risks to data integrity and security, such as data loss, corruption, or unauthorized access. Therefore, it is important to implement safeguards such as data backup, encryption, access control, and error checking to ensure that the deduplication process is reliable and secure.

- **Monitor and optimize performance:** Deduplication can also affect system performance, particularly during the initial deduplication phase or when processing large amounts of data. Therefore, it is important to monitor performance metrics such as CPU usage, memory consumption, and I/O throughput, and to optimize the deduplication process as needed to minimize impact on system performance.

6.4 Data Deduplication Process Implementation

The deduplication process involves a combination of analysis, hashing, comparison, and replacement techniques to eliminate duplicate content and optimize storage efficiency. The specific implementation and configuration of the process may vary depending on factors such as the type and volume of data, the available storage resources, and the performance and security requirements of the system. To calculate the number of duplicate files in a deduplication process using Python, we can follow these general steps:

- **Traverse the file system:** Use a library like ‘os’ or ‘glob’ to recursively traverse the file system and identify all the files.
- **Compute hash values:** Compute hash values for each file using hash algorithm like SHA256 or MD5. You can use the ‘hashlib’ module in Python for this task.
- **Store hash values:** Store the hash values in a dictionary or a list to keep track of the number of times each hash value occurs.
- **Count duplicates:** Iterate through the hash values and count the number of occurrences of each hash value. If the hash value occurs more than once, it indicates that there are duplicate files.
- **Measure acquisition speed:** Measure the time it takes to traverse the file system and compute the hash values for each file. You can use the ‘time’ module in Python to measure the time taken by a particular section of code.

6.5 Evaluating Generated Test Disk Images

FTK Imager is a digital forensic tool that is often used to create forensic images of storage media, such as hard drives and USB drives. When you use FTK Imager to create an image of a drive, it typically creates two ADI (Autodesk Device Interface) files as output: a data file and a metadata file. The data file contains a bit-for-bit copy of the data on the drive, while the metadata file contains information about the drive and the image itself, such as the date and time of creation, the type of drive, the size of the image, and so on. The metadata file also includes a hash value, which is a digital fingerprint of the image that can be used to verify its integrity. Table 6.1 provides a list of information related to data acquisition and deduplication

from various image files. Each entry includes details about the initial size of the files, the number of deduplicated files removed, the size after deduplication, actual acquisition speed, effective acquisition speed, CPU execution speed, and the acquisition start and finish times.

Sr.no.	Image File	Image Size	Total Files	File System	Operating System
1	Imagefile1.ad1	2.78 GB	5,984 Files, 160 Folders	NTFS	Windows 11, 64-bit OS
2	Imagefile2.ad1	22.24 GB	4,385 Files, 350 Folders	NTFS	Windows 11, 64-bit OS
3	Imagefile3.ad1	1.60 GB	47,227 Files, 8,831 Folders	NTFS	Windows 11, 64-bit OS
4	Imagefile4.ad1	7.42 GB	2,410 Files, 416 Folders	NTFS	Windows 11, 64-bit OS
5	Imagefile5.ad1	18.12 GB	6,998 Files, 205 Folders	NTFS	Windows 11, 64-bit OS
6	Imagefile6.ad1	0.74 GB	745 Files, 175 Folders	NTFS	Windows 11, 64-bit OS
7	Imagefile7.ad1	4.93 GB	944 Files, 89 Folders	NTFS	Windows 11, 64-bit OS

Table 6.1: Image Files with information.

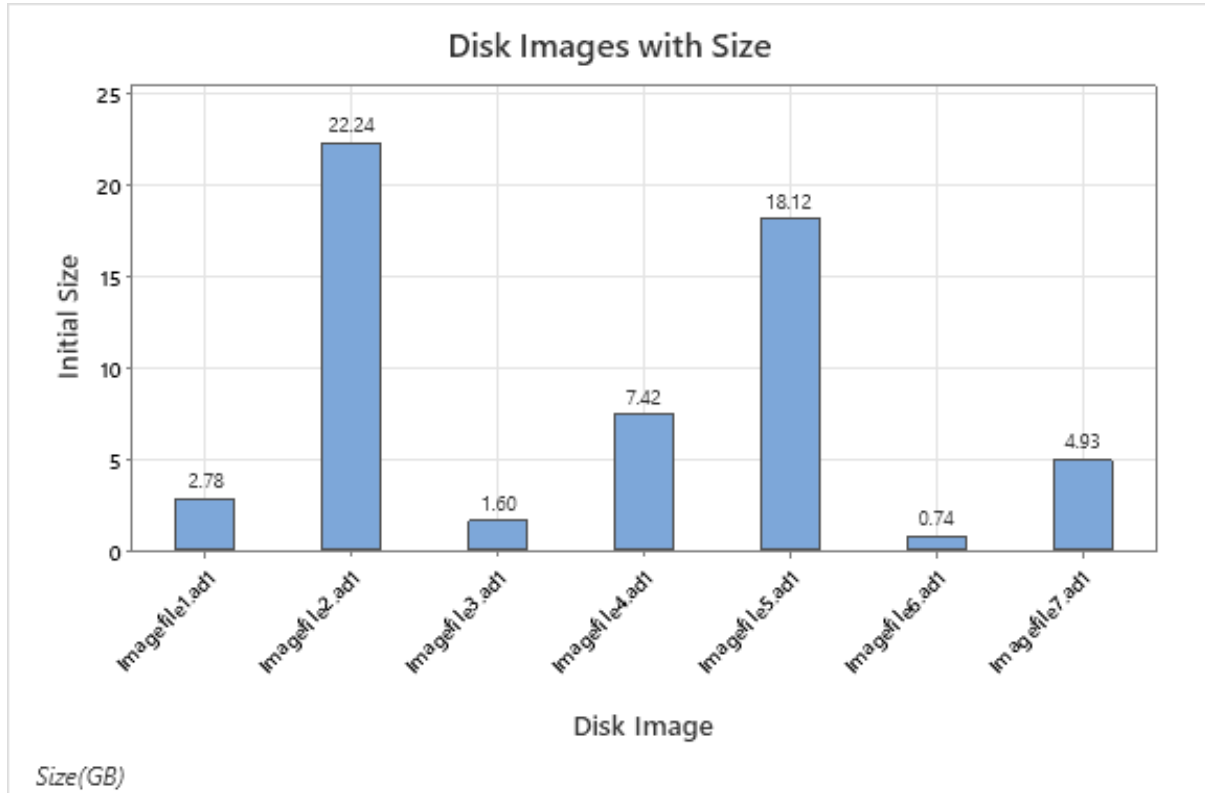


Figure 6.2: Disk Images with Different Sizes.

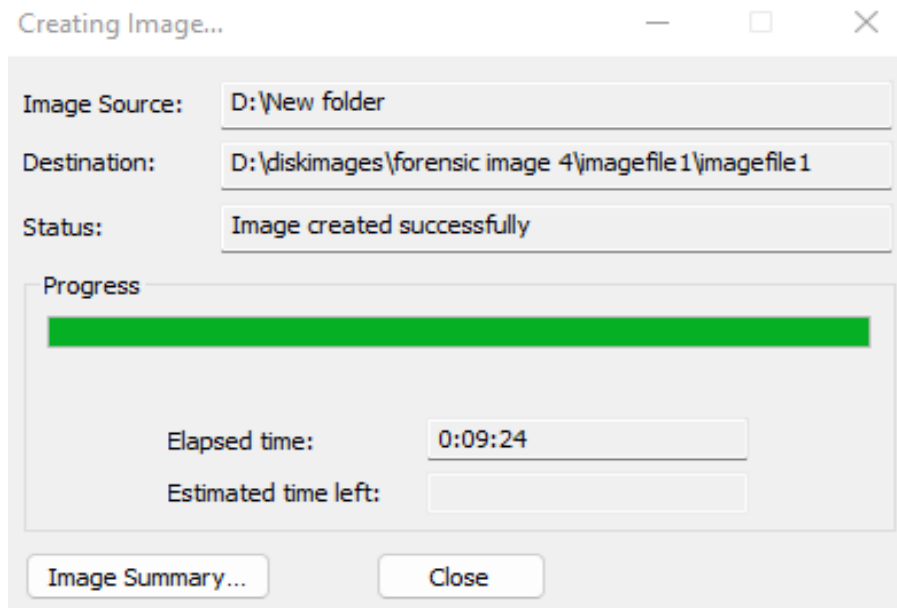


Figure 6.3: Creating Disk Image.

The above Table 6.1, shows image files created with different sizes as also shown in Figure 6.2, and also about files and folders every image files contains, as shown in Figure 6.4 and tell us what is the system type, in this case we are on NTFS file system, and our operating system is windows operating system, which is windows 11, with 64-bit operating system, imagefile1.ad1 that is an image file while it can have several meta files with all the information stored in them. Above, it is shown how a disk image is created using FTK imager in Figure 6.3.

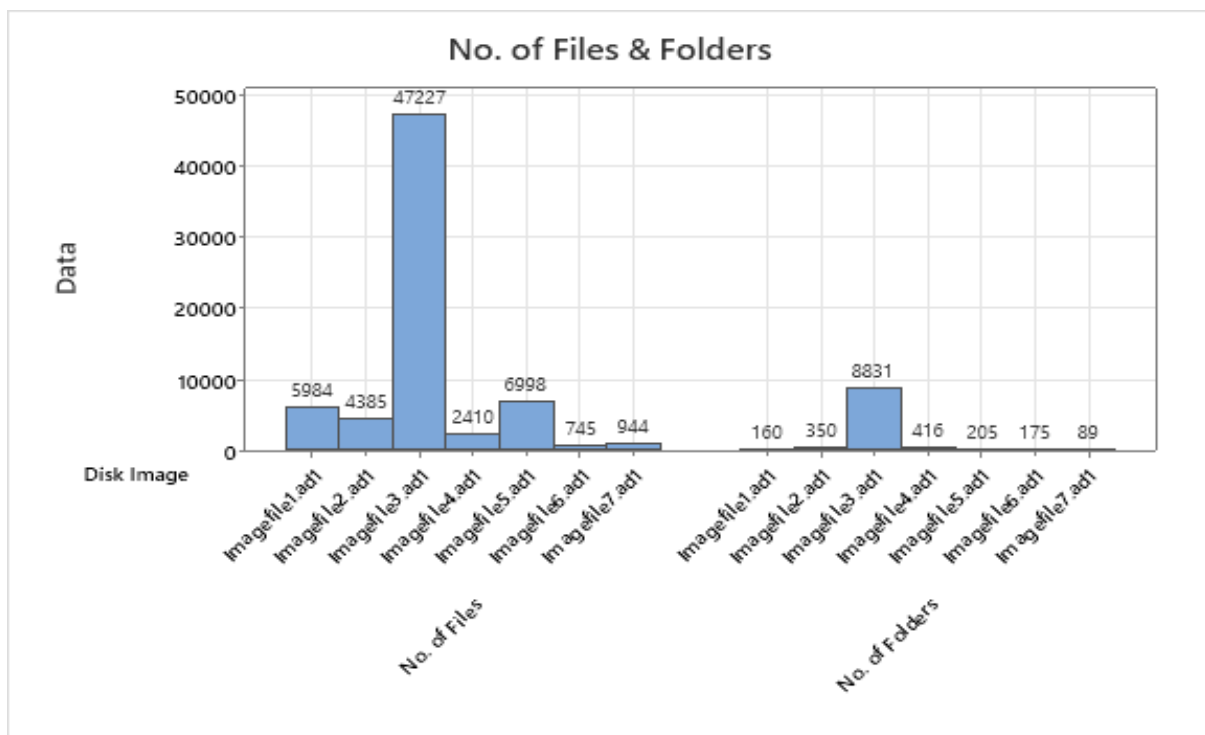


Figure 6.4: Number of Files and Folders in Disk Image.

In FTK imager the created disk image is checked before and after the disk creation to ensure that it has same files, and its integrity is confirmed. So, after the disk image creation it verifies results by comparing two hashes and match the results. In figure 6.5, image summary is shown, case information with evidence number and image file name, the information related to examiner is mentioned, a unique MD5, SHA1 hash of image file is created, and above image information is mentioned with image files name and disk image type as shown in Figure 6.5.

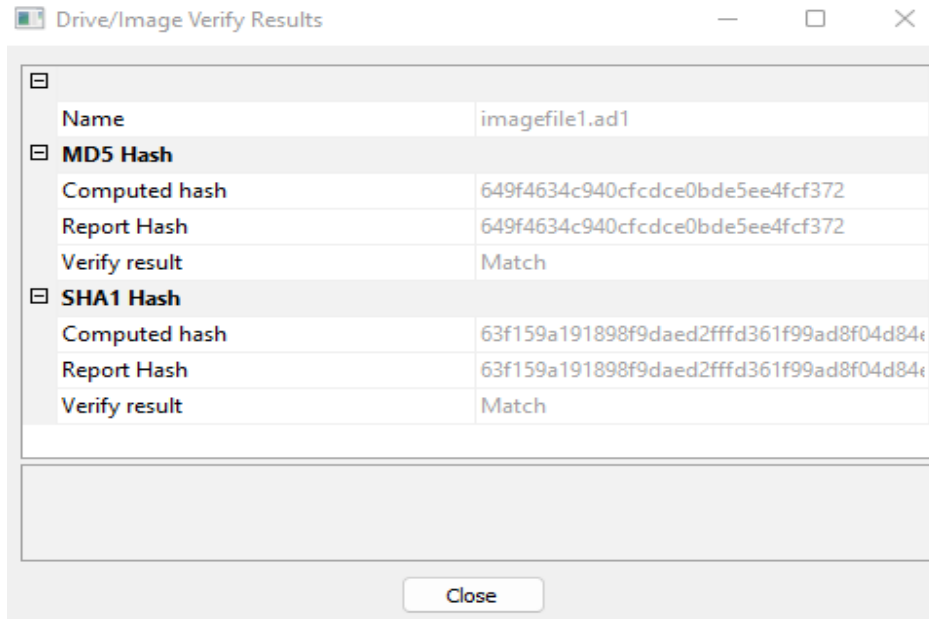


Figure 6.5: Disk Image Summary.

6.6 Actual & Effective Acquisition Speed & Disk Size Comparison

Table gives below is an overview of how the deduplication process performed for each the image file, which we selected as evident from the Table 6.2, including details about the reduction in file size, the number of duplicate files removed, and the speed of the process. The acquisition start and finish times give an idea of the duration of the process for each file. The effective acquisition speed considers the initial size of the files and provides insight into how efficiently the deduplication process worked. Following are formulas for calculation. Following are some equations (1), (2), (3) that we used for calculations.

$$acquisition_time = end_time - start_time \tag{1}$$

$$actual_speed = \frac{final_size}{acquisition_time} \tag{2}$$

$$effective_speed = \frac{initial_size}{acquisition_time} \tag{3}$$

Sr.no.	Initial Image Information After Deduplication
1	Imagefile1.ad1 / Initial size of the files: 2.78 GB / Deduplicate files removed: 142 files. Size after deduplication: 2.64 GB / Actual acquisition speed: 21.81 MB/s. Effective acquisition speed: 23.00 MB/s / CPU execution speed: 17.11 MB/s. Acquisition started time: 13:51:27 / Acquisition finished time: 27 13:53:22.
2	Imagefile2.ad1 / Initial size of the files: 22.24 GB / Deduplicate files removed: 906 files. Size after deduplication: 19.59 GB / Actual acquisition speed: 41.05 MB/s Effective acquisition speed: 46.60 MB/s / CPU execution speed: 120.56 MB/s Acquisition started time: 14:27:39 / Acquisition finished time: 14:35:14
3	Imagefile3.ad1 / Initial size of the files: 1.60 GB / Deduplicate files removed:24365files. Size after deduplication: 1.23 GB / Actual acquisition speed: 0.91 MB/s Effective acquisition speed: 1.18 MB/s / CPU execution speed: 56.84 MB/s Acquisition started time: 03:32:23 / Acquisition finished time: 03:53:55
4	Imagefile4.ad1 / Initial size of the files: 7.42 GB / Deduplicate files removed: 204 files. Size after deduplication: 7.34 GB / Actual acquisition speed: 32.95 MB/s Effective acquisition speed: 33.32 MB/s / CPU execution speed: 43.77 MB/s Acquisition started time: 23:04:40 / Acquisition finished time: 23:08:13
5	Imagefile5.ad1 / Initial size of the files: 18.12 GB / Deduplicate files removed: 371files. Size after deduplication: 18.07 GB / Actual acquisition speed: 43.01 MB/s Effective acquisition speed: 43.11 MB/s / CPU execution speed: 89.39 MB/s Acquisition started time: 23:34:37 / Acquisition finished time: 23:41:18
6	Imagefile6.ad1 / Initial size of the files: 0.74 GB / Deduplicate files removed: 491 files. Size after deduplication: 0.38 GB / Actual acquisition speed: 19.31 MB/s Effective acquisition speed: 37.52 MB/s / CPU execution speed: 4.58 MB/s Acquisition started time: 03:05:39 / Acquisition finished time: 03:05:58
7	Imagefile7.ad1 / Initial size of the files: 4.93 GB / Deduplicate files removed: 6 files. Size after deduplication: 4.93 GB / Actual acquisition speed: 56.19 MB/s Effective acquisition speed: 56.19 MB/s / CPU execution speed: 28.33 MB/s Acquisition started time: 03:10:27 / Acquisition finished time: 03:11:51

Table 6.2: Image Files with Information After Data Deduplication.

In the dataset of acquired image files, deduplication played a crucial role in optimizing storage efficiency and reducing redundancy. Across the different image files, a varying number of duplicate files were identified and subsequently removed. For instance, in the case of 'Imagefile1.ad1,' a total of 142 duplicate files were detected and eliminated as shown in Figure 6.6. Similarly, 'Imagefile2.ad1' had a substantial 906 duplicates removed, highlighting the potential for data redundancy in large-scale acquisitions. The process significantly impacted the sizes of these image files. For 'Imagefile1.ad1,' the initial size of 2.78 GB was

reduced to 2.64 GB after deduplication. Similarly, 'Imagefile2.ad1' saw its initial size of 22.24 GB shrink to 19.59 GB. This emphasizes the value of deduplication in freeing up valuable storage space and streamlining data organization as shown in Figure 6.8.

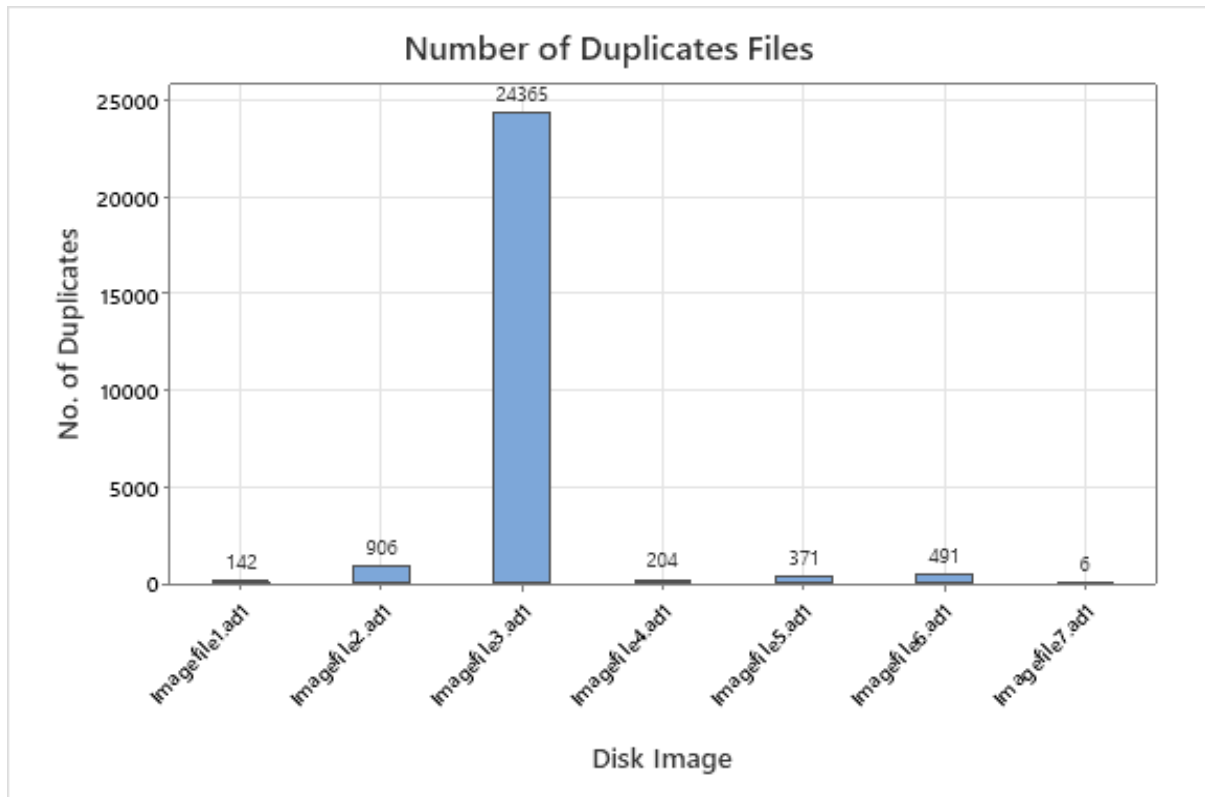


Figure 6.6: Number of Duplicates in Disk Images.

Beyond the first two image files, the impact of deduplication was also evident in the subsequent data sets. 'Imagefile3.ad1,' for instance, experienced a remarkable reduction in size from an initial 1.60 GB to a final 1.23 GB, because of eliminating a staggering 24,365 duplicate files. The pattern persisted in 'Imagefile4.ad1,' which witnessed a decrease from 7.42 GB to 7.34 GB after the removal of 204 duplicate files. Similarly, 'Imagefile5.ad1' and 'Imagefile6.ad1' showcased the effectiveness of deduplication, as the initial sizes of 18.12 GB and 0.74 GB respectively, were substantially reduced to 18.07 GB and 0.38 GB as shown in Figure 6.8. Following Figure 6.7 shows experimentation of Imagefile5.ad1.

```

===== RESTART: D:\Experimentation\deduplication.py
Initial size of the files      : 18.12 GB
Deduplicate files removed    : 371 files
Size after deduplication      : 18.07 GB
Actual acquisition speed      : 43.01 MB/s
Effective acquisition speed    : 43.11 MB/s
CPU execution speed           : 89.39 MB/s
Acquisition started time     : 2023-07-27 23:34:37
Acquisition finished time    : 2023-07-27 23:41:18

```

Figure 6.7: Experimentation of Disk Image file 5.

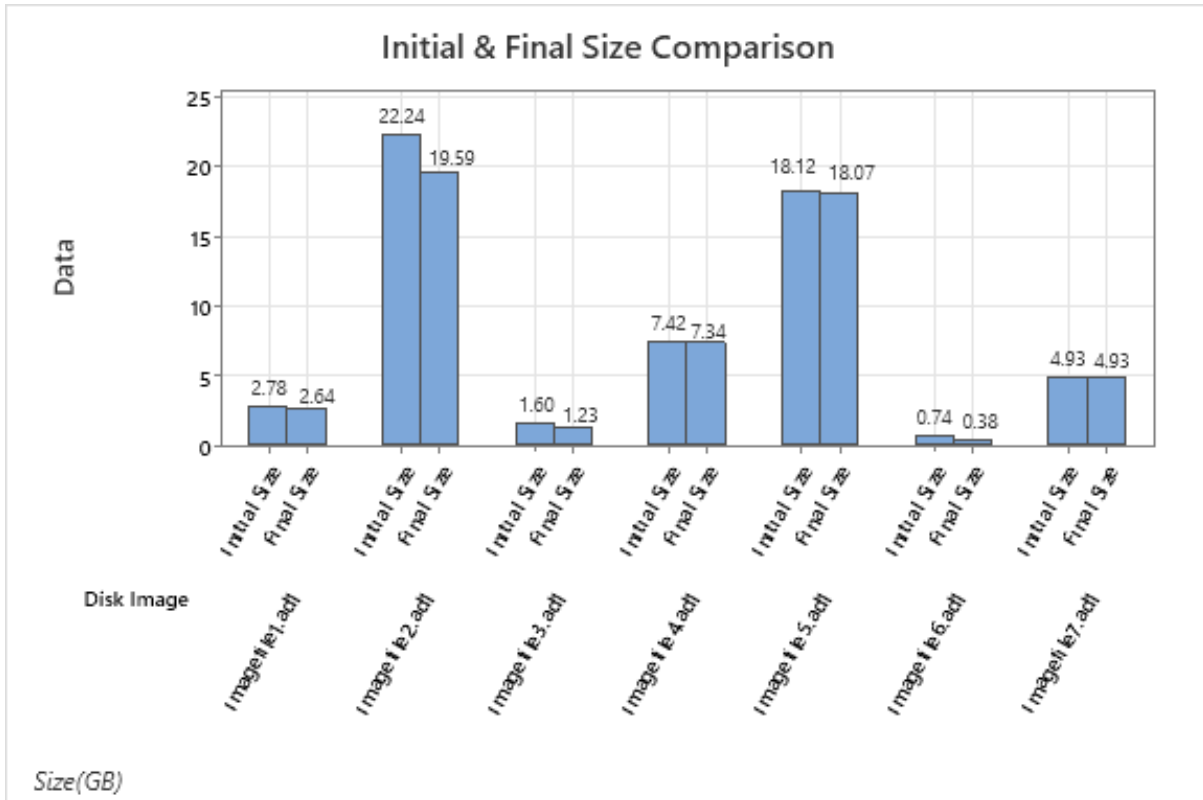


Figure 6.8: Initial and Final Size Comparison after Deduplication.

Beyond the reduction in file sizes, the process of deduplication also had a discernible impact on the acquisition speed. The effective acquisition speed considers not only the actual acquisition speed but also the CPU execution speed. Notably, 'Imagefile2.ad1' showcased a remarkable effective acquisition speed of 46.60 MB/s, indicating the efficiency achieved by removing duplicates during the acquisition process. This speed improvement, alongside the reduction in file sizes, underscores the significance of deduplication in optimizing both storage utilization and data transfer efficiency. Data provided highlights the positive outcomes of deduplication, ranging from storage space savings shown in Figure 6.8, to enhanced acquisition speeds shown in Figure 6.11, reinforcing its role in efficient data management. Beyond the first two image files, impact of deduplication was also evident in the subsequent data sets. 'Imagefile3.ad1,' for instance, experienced a remarkable reduction in size from an initial 1.60 GB to a final 1.23 GB, because of eliminating a staggering 24,365 duplicate files. The pattern persisted in 'Imagefile4.ad1,' which witnessed a decrease from 7.42 GB to 7.34 GB after the removal of 204 duplicate files. Similarly, 'Imagefile5.ad1' and 'Imagefile6.ad1' showcased the effectiveness of deduplication, as the initial sizes of 18.12 GB and 0.74 GB respectively, were substantially reduced to 18.07 GB and 0.38 GB. If we plot size comparison of disk images in a time series graph shown in Figure 6.10, it is shown that disk size decreases with the removal of duplicates. Area graph also shown the size comparison, blue area shows the initial size of disk image while red shows final size after

deduplication. By removing duplicates less data will be used during cloud investigations, thus resulting in reducing cloud forensic backlog as shown in Figure 6.9.

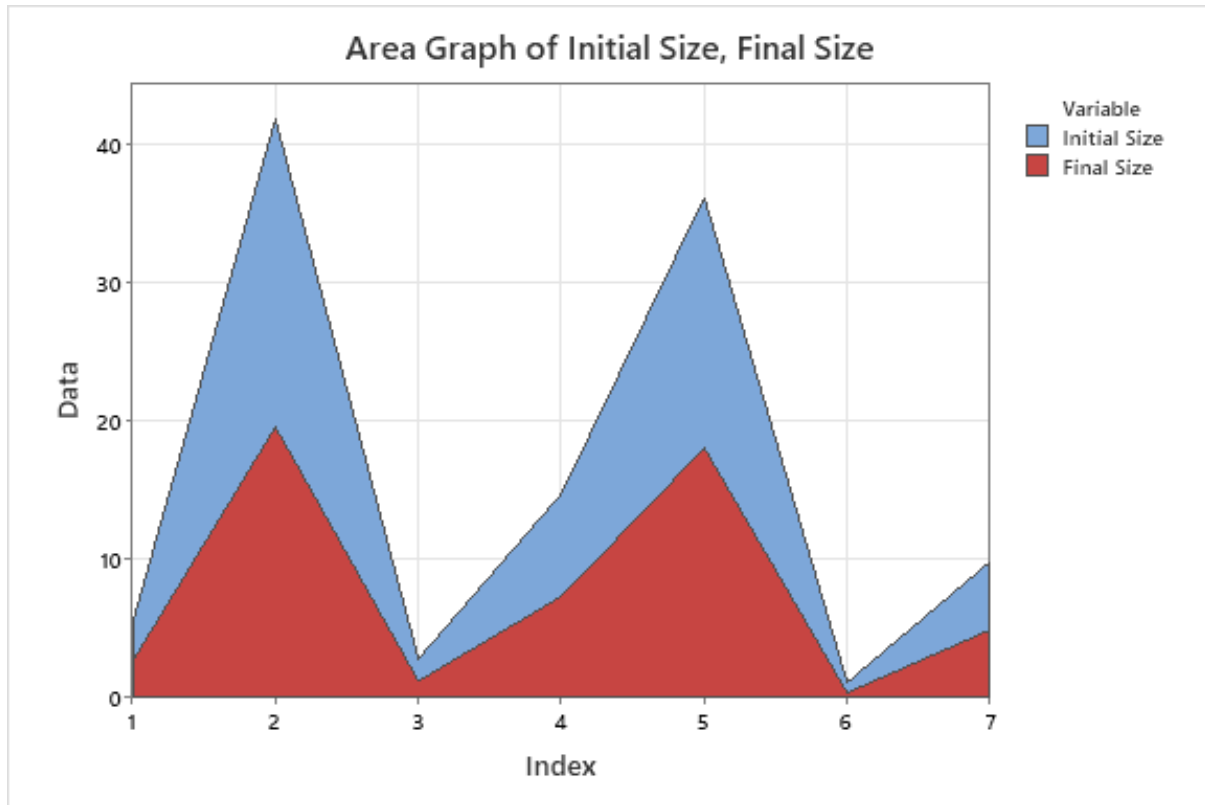


Figure 6.9: Initial and Final Size Comparison w.r.t Area Graph.



Figure 6.10: Initial and Final Size Comparison w.r.t Time Series Plot.

The provided data highlights the actual and effective acquisition speeds for multiple image files after the process of deduplication. The actual acquisition speed represents the rate at which data is collected during the acquisition process, measured in megabytes per second (MB/s). On other hand, effective acquisition speed considers impact of deduplication & CPU execution speed on overall acquisition process. It provides an accurate representation of speed at which unique data is acquired & processed.

Looking at the dataset, we can observe variations in both actual and effective acquisition speeds across different image files as shown in Figure 6.11. For instance, in the case of "Imagefile2.ad1," the initial actual acquisition speed is noted at 41.05 MB/s. However, after deduplication & considering influence of CPU execution speed, effective acquisition speed increases to 46.60 MB/s. This increase highlights efficiency gained through removal of duplicate files & optimization of CPU.

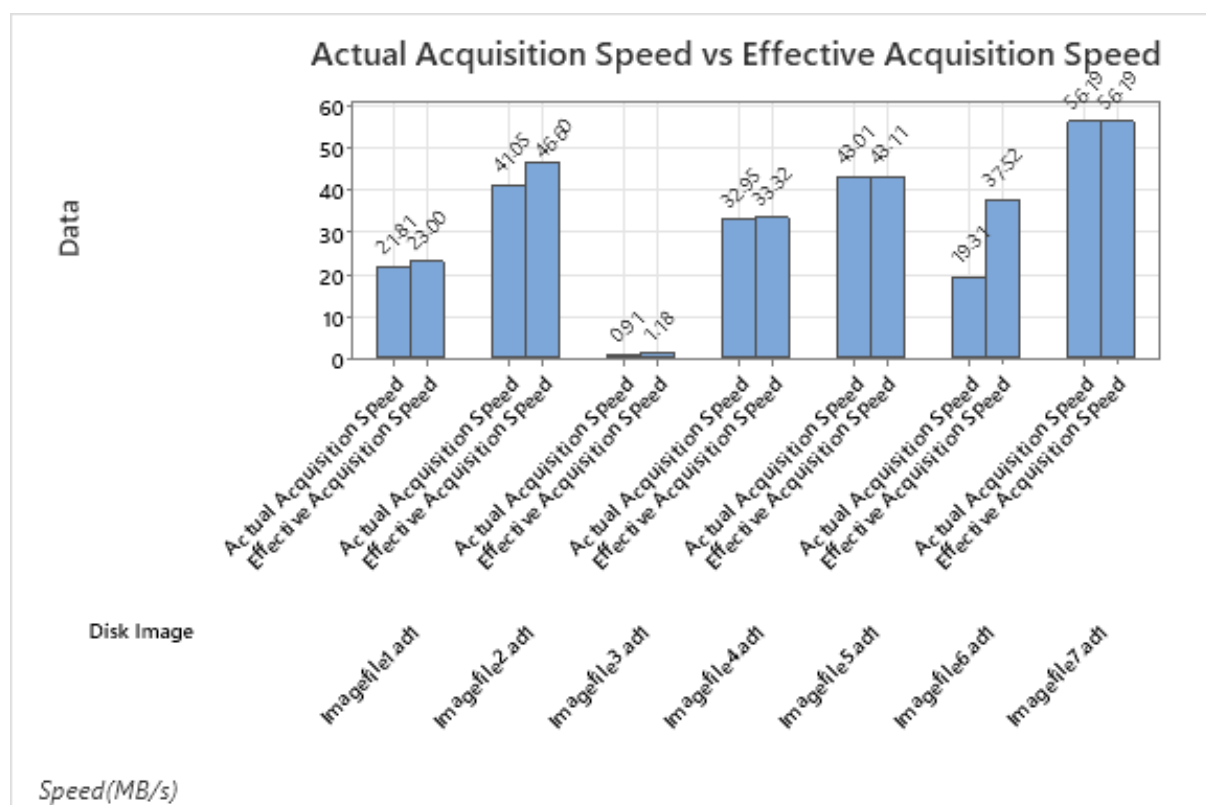


Figure 6.11: Actual and Effective Speed Comparison.

In contrast, some image files exhibit a less significant difference between their actual and effective acquisition speeds. For instance, in "Imagefile7.ad1," the actual acquisition speed is already relatively high at 56.19 MB/s, and the effective acquisition speed remains almost the same, indicating minimal influence from deduplication and CPU execution speed. The variance between actual and effective acquisition speeds can be attributed to several factors.

Deduplication plays a crucial role in reducing the amount of data that needs to be acquired and processed. Files that exhibit higher rates of duplication will likely show a more substantial increase in effective acquisition speed compared to their actual speed. To visually compare the actual and effective speed variations over time, a time series plot and an area graph can be useful tools. These graphs will provide a clear representation of how these speeds change throughout the acquisition process for the given image files.

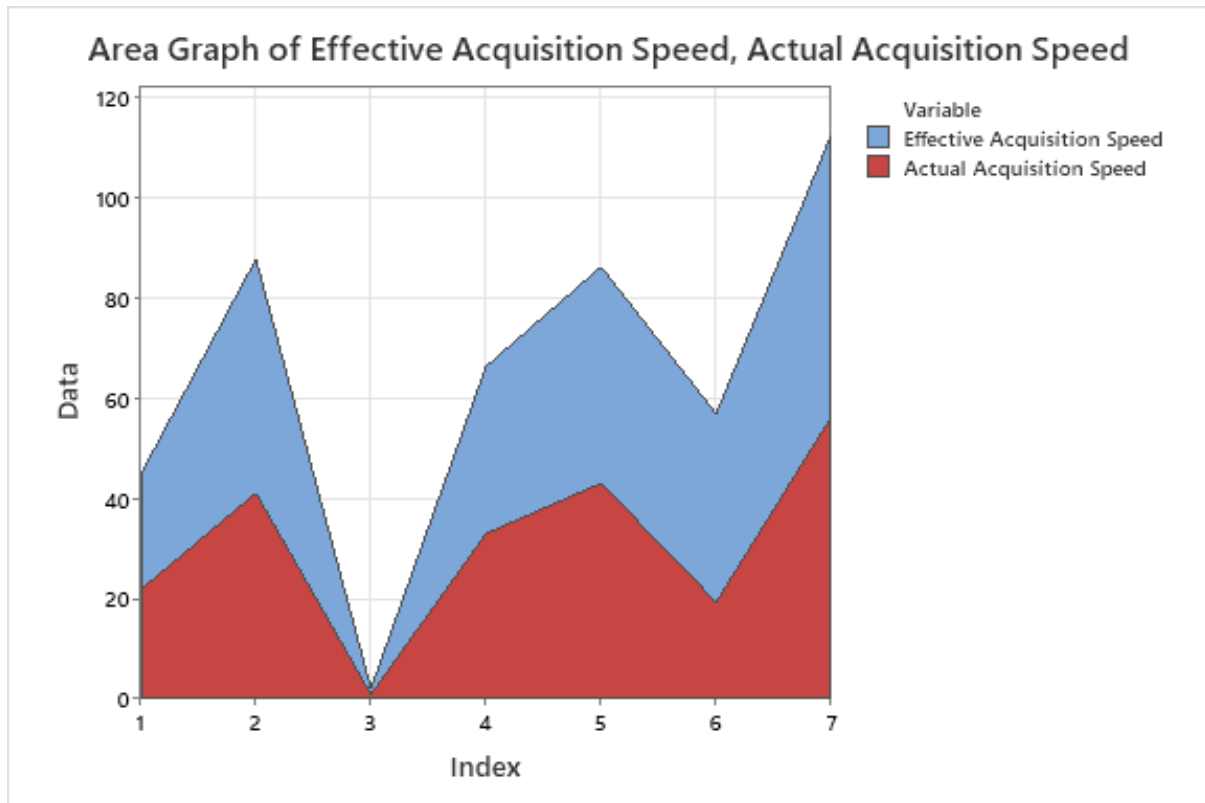


Figure 6.12: Actual and Effective Speed Comparison w.r.t Area Graph.

An area graph is another useful way to compare actual and effective speeds over time as shown in Figure 6.12. It's especially effective when you want to emphasize the cumulative impact of speed changes. In this case, you can use the area graph to show the cumulative actual and effective speeds over time for each image file, the blue area shows effective speed which cover more area as compared to actual speed which cover less area.

A time series plot is a common choice for displaying data trends over time. In this case, it can be used to show how both actual and effective acquisition speeds change during the acquisition process. Figure 6.13 shows actual and effective acquisition speed comparison, blue colour shows actual speed which decreases after duplication, and effective speed increases as shown in red.

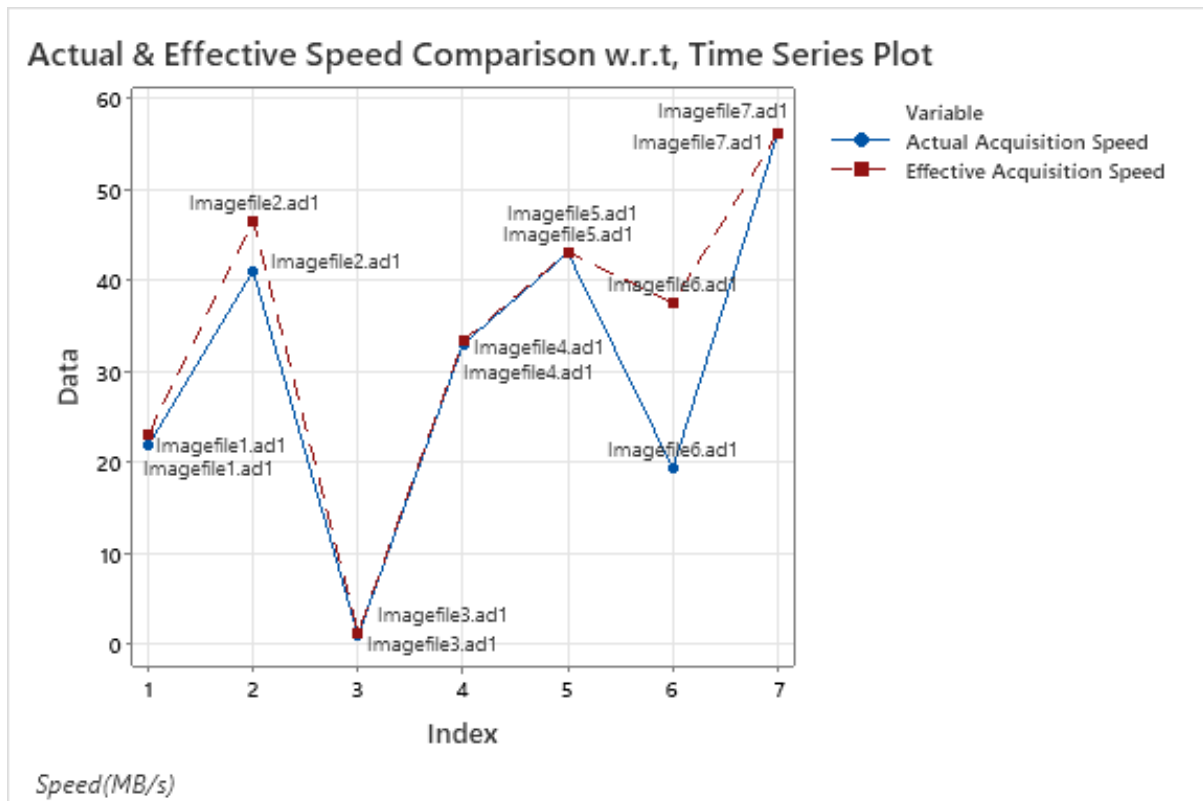


Figure 6.13: Actual and Effective Speed Comparison w.r.t Time Series Plot.

6.7 Discussion

In this chapter we discussed what is deduplication in cloud forensics, the introduction stage highlighted the importance of efficient data management within cloud forensics. It emphasizes the significance of reducing storage requirements while ensuring data integrity and accessibility. and then in Deduplication to Reduce Storage in Cloud Forensics our focus was the role of data deduplication in addressing storage challenges within cloud forensics. It underscores the potential benefits of deduplication, including optimized storage utilization and enhanced retrieval efficiency. Then in data deduplication process we give overview of the data deduplication process & explains the methodology of identifying and eliminating duplicate data segments, emphasizing the use of hashing algorithms to determine content uniqueness. Next, implementation of the data deduplication process is discussed. The script presented highlights the steps involved in calculating hash values, detecting duplicates, and removing redundant files to free up storage space. In evaluating generated test disk images focus is on the assessment of the generated test disk images post-deduplication. It covers the reduction in image sizes, the number of duplicate files removed, and how these outcomes reflect the efficiency of the deduplication process. In the end calculation of actual and effective acquisition speeds takes centre stage. The metrics are examined to gauge the speed at which data is processed during deduplication, both considering the real-time acquisition and the relative efficiency compared to initial data sizes.

Chapter 7: Relevant Data Extraction and Prioritization

7.1 Introduction

In cloud forensic investigations, the process of relevant data extraction and prioritization plays a pivotal role in uncovering digital evidence crucial for understanding and resolving digital incidents. Cloud environments, with their dynamic and distributed nature, introduce unique challenges to the forensic process. We will be extracting pertinent data from cloud sources while judiciously prioritizing the extracted information for efficient analysis.

Forensic investigations often involve vast amounts of cloud data spread across diverse cloud service providers and their associated platforms. In this context, relevant data extraction refers to the systematic identification and retrieval of data that holds potential evidentiary value. It involves deciphering the complex web of cloud resources, applications, and interactions to isolate information pertinent to the investigation.

Prioritization, on the other hand, involves ranking the extracted data based on its potential significance, relevance, and context within the investigation. Not all extracted data holds equal importance; some pieces might provide critical insights into the incident's timeline, causality, or actors involved. Effective prioritization ensures that limited investigative resources are allocated judiciously to areas that are likely to yield the most valuable results.

Different methodologies & tools will be discussed for data extraction in cloud environments, including techniques to retrieve data from cloud storage, virtual machines, logs, and network traffic. Additionally, it explores the criteria and considerations used for prioritizing data, which may encompass factors such as the timeline of events, the nature of the incident, legal requirements, and potential impact on the organization. LEAs need strategies necessary to navigate the complexities of relevant data extraction and prioritization in cloud forensic investigations. By mastering these techniques, investigators can enhance their ability to unearth crucial evidence efficiently and effectively, contributing to more accurate and conclusive digital investigations in cloud-based scenarios.

In case of Fraudulent email data, Relevant Data Extraction and Prioritization are paramount in forensic investigations of fraudulent emails. This process entails systematically identifying and retrieving pertinent electronic evidence from cloud-based sources. In cases of fraudulent emails, this methodology aids in isolating key communication threads, attachments, and metadata crucial for establishing fraudulent intent. Prioritization further aids investigators by focusing efforts on the most incriminating data, optimizing resource allocation, and

expediting the identification of suspects and patterns. In cloud forensic scenarios, these practices are pivotal for efficiently uncovering digital trails and strengthening the foundation for legal actions against fraudulent activities.

7.2 Forensic Evidence Information and Dataset Details

As cloud data related to emails are hard to find, due to the distributed nature of the cloud. We will be using famous ENRON email data because it is easily available, and it will help us in our study. Emails were from 150 employees of Enron Corporation. Enron employees covered up bad financial position of company, by keeping stock price artificially high. The shape of the data is as it has 2 columns file and message with 517401 rows. Further information of the dataset is given as follow.

- For relevant data extraction & prioritization [Enron email dataset] is used.
- The Enron email dataset contains approximately 500k emails generated by employees of the Enron Corporation.
- This dataset is collected from Kaggle repository.
- Dataset which we are using, is the May 7, 2015, Version which is published at <https://www.cs.cmu.edu/~./enron/>.
- This dataset was investigated by LEAs for stock fraud.

7.3 Testing Environment for Experimentation

For testing the Enron email dataset, we set up the following testing environment.

- Python programming language is used to extract data.
- Training and testing of data are performed on Jupiter Notebook tool.
- The computer system used is Intel(R) Core (TM) i5-4200U, CPU @ 1.60 GHz 2.30 GHz with 4 GB RAM.
- Installed operating system installed is Windows 11 Education, 21H2 version.

7.4 Data Preparation Stages

Data preparation stages form a critical foundation for any data-driven analysis or machine learning project. Properly executing each stage is essential to ensure that the final model is accurate, reliable, and capable of delivering valuable insights or predictions. Each stage serves a specific purpose and contributes to the overall quality and effectiveness of the final model. Let's briefly explain each of these stages:

- **Data Retrieval/Acquisition:** This is the initial step where the required data is collected from Kaggle. In our case, the dataset is the 2015 version of the Enron dataset, used widely for studying email related fraud. As data related to email fraud is hard to find, so we are considering this data as data collected from cloud storage for testing and studying forensic evidence acquisition.

	file	message
0	allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...
1	allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...
2	allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.e...
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e...
4	allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.e...

Figure 7.1: Enron Dataset with Columns file & message.

The dataset is of about stock fraud, and the email data is of 150 employees of the company. In figure 7.1, it is shown basic shape of data. Figure 7.2, is showing top 20 emails senders of the organization. Whereas, figure 7.3 shows hours and days of week on which emails were sent.

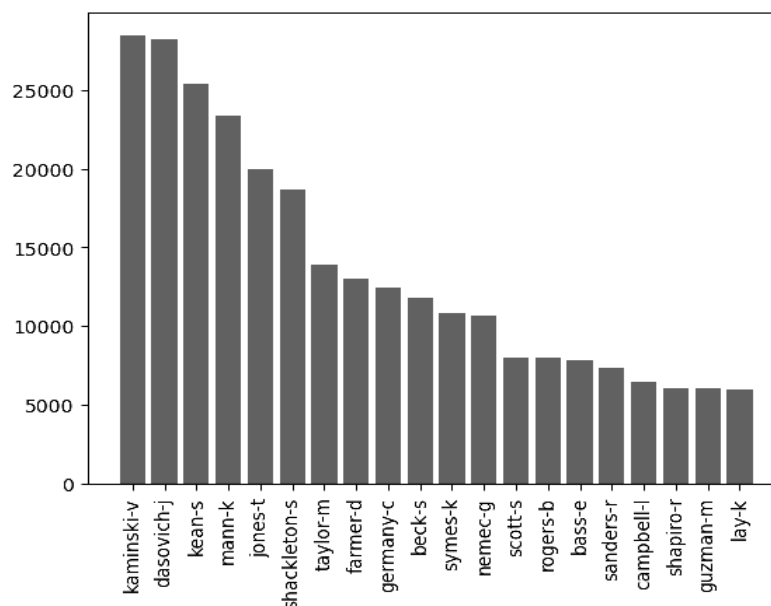


Figure 7.2: Top 20 Employees Who Sent Most Mails.

- **Data Cleaning and Transformation:** Once the data is collected, it often requires cleaning to remove inconsistencies, errors, duplicates, missing values, and outliers. Remove stop words, lemmatize, do stemming, the split email text data into sentences and then convert into tokenized words. Transform involves text data into lowercase format for analysis, remove punctuation, remove regular expressions. This stage ensures data reliability & readiness for further processing.

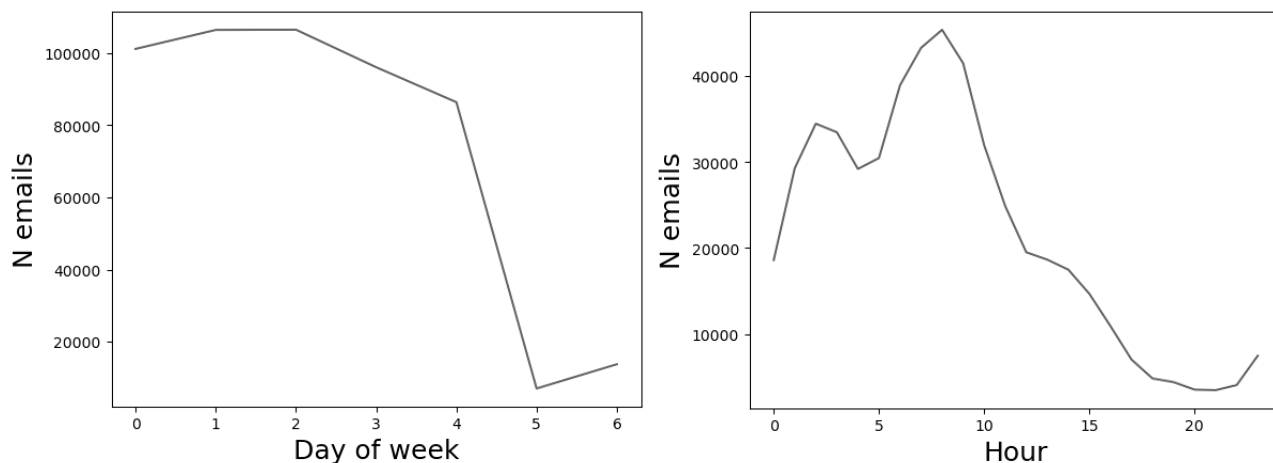


Figure 7.3: Days of Week and Hours in Which Emails were Sent.

- **Data Processing:** In this stage, the cleaned and transformed data is organized and structured in a way that it facilitates our analysis. This may involve reshaping the data, aggregating information, and creating appropriate data structures like for email fraud detection, we only need two columns, subject and body column.
- **Feature Extraction/Engineering:** Features are the variables or attributes that the model uses to make predictions or classifications. Feature extraction involves selecting relevant features from the dataset, for our fraud detection we only want fraudulent email data, we will prioritize this data to reduce forensic backlog.
- **Data Modelling:** Once features are defined, the labelled data is then used to build machine learning or statistical models. These models learn patterns from the data and can be used to make predictions, classifications, or other analyses.
- **Evaluation & Deployment:** After training the model, we will evaluate its performance using metrics that reflect its accuracy, precision, recall etc. If the model meets the desired performance, it can be deployed to start making predictions on new, unseen data. Deployment might involve integrating model into software applications, systems, or other operational contexts.

7.5 Relevant Fraudulent Emails Detection Methods

It was found that an organization typically losses about five percent of its revenue to fraudulent activities. There are various ways to detect fraud in a data. We will use python to detect data fraud. We can detect fraud manually and by using labelled data and by using unlabelled data as well, and lastly it is also detected by using text data. The labelled data approach harnesses historical fraud instances, unlabelled data approach explores deviations from the norm, and text data analysis extracts insights from the textual content of emails.

- **Manual Fraudulent Email Detection by Using text data:** Email content often holds vital clues for fraud detection. This approach canters on analysing the text within emails using natural language processing (NLP) techniques. By extracting features from the text, such as keywords, sentiment, or linguistic patterns, machine learning models can be trained to recognize language-based indicators of fraudulent intent. This approach is particularly valuable for detecting phishing or social engineering attempts.
- **Fraudulent Email Detection by Using Labelled Data:** This approach involves training machine learning models using labelled data, which consists of historical instances of both fraudulent and legitimate emails. By using supervised learning, we can flag fraudulent emails. By exposing the model to these labelled examples, it learns to distinguish patterns and characteristics associated with fraud. As a result, when new incoming emails are assessed, the model can accurately classify them as potentially fraudulent or legitimate based on the learned patterns. We can use classification, linear SVC, logistic regression, neural networks, decision trees, random forests and by comparing these methods to find most efficient detection model.
- **Fraudulent Email Detection by Using Unlabelled Data:** By using unsupervised learning techniques, and in scenarios where, labelled data is scarce or expensive to obtain, the unlabelled data approach becomes valuable. Unlabelled data refers to a dataset lacking explicit fraud labels. Through techniques like anomaly detection, clustering, or semi-supervised learning, patterns that deviate from the norm can be identified. This can potentially highlight instances of email fraud without requiring labelled examples of fraud explicitly. We can use K-mean clustering, DBSCAN, SVD, PCA, Apriori, FP-growth, Markov model to flag data.

7.6 Fraudulent Email Detection using Topic Modelling

Fraudulent email detection is a critical challenge in today's digital landscape. Leveraging advanced techniques like Topic Modelling, particularly the Latent Dirichlet Allocation (LDA) model, can provide a unique perspective on identifying fraudulent activities within emails. Topic modelling is used to discover topics in a text data, here we have text data, it basically tells us about the text data. Topic modelling i.e., LDA model is like clustering conceptually. We identify fraudulent emails by flagging them. To run a topic model, we must do the following.

7.6.1 Preprocessing the Email Text Data

Before applying the LDA model, the email text must be pre-processed. We have combined [subject] & [body] column of email dataset & then forming a new data column [completed_text] as shown in Figure 7.4. Preprocessing involves tasks like tokenization, removal of stopwords, punctuations, lemmatizing & stemming and removing excess spaces & by removing regular expressions we stored normalized text in new column [clean_text] as shown in Figure 7.5. Preprocessing ensures that the text is in a suitable tokenized format for further testing.

```
# There are 2 text variables in the model: subject a
# we want to join all text data in one single column

df["completed_text"]=df["subject"]+' '+df["body"]
```

Figure 7.4: Combining subject & body Column.

```
# Completed_text:
sample_df['completed_text'][480539]
```

```
'Re: US entities Regarding the following list, please make the following observation\n\n1. Pacific
ration \n-the parent entity\n2. AOL is subinvestment grade\n3. Quest Communications is wrongly spe
Q\n\nThanks and Regards\nMoazzam Khoja\n\n\n\nMartin McDermott@ECT\n02/16/2000 12:11 PM\nTo: Ro
rackett/HOU/ECT@ECT, William S Bradford/HOU/ECT@ECT\ncc: Bryan Seyfried/LON/ECT@ECT, Moazzam Khoj
oulkes \n\nSubject: US entities\n\n\n'
```

```
# clean_text:
print(sample_df['clean_text'][480539])
```

```
['entity', 'regarding', 'following', 'list', 'make', 'following', 'observation', 'pacific', 'elect
ity', 'subinvestment', 'grade', 'quest', 'communication', 'wrongly', 'spelled', 'qwest', 'communi
'mcdermott', 'nelson', 'mark', 'taylor', 'debbie', 'brackett', 'william', 'bradford', 'bryan', 's
lavya', 'sareen', 'stuart', 'ffoulkes', 'entity']
```

Figure 7.5: Cleaned Data Column After Text Cleaning.

7.6.2 Creating the LDA Model

The LDA model is built by processing the pre-processed email text. Previously we split emails text into tokenized words, now we can apply topic model. The model aims to assign topics to each document and keywords to each topic. Through iterative processes, the model learns to allocate topics to documents and words to topics.

7.6.2.1 Latent Dirichlet Allocation Model (LDA)

Unveiling Topics in Textual Data, the Latent Dirichlet Allocation (LDA) model is a widely used technique in natural language processing and topic modelling. It provides a framework for uncovering the hidden thematic structure within a collection of documents, making it

particularly useful for tasks such as text analysis, content categorization, and understanding the underlying patterns in textual data. Here's an overview of how the LDA model works:

- **Intuition and Assumptions:** LDA model is based on assumptions that each document in a corpus is a mixture of topics, and each topic is a mixture of words. The model aims to reverse-engineer this process by identifying the topics and the distribution of words within those topics.
- **Components of the LDA Model:** LDA model consist of documents, topics, words. Each topic is a distribution of words and each topic also have various words which belongs to it. Words within the documents contribute to the topics. Main purpose is to find topics a document belongs to, based on words in it as shown in Figure 7.6.
- **LDA Algorithm:** The LDA algorithm goes through an iterative process to assign words to topics and topics to documents. The key idea is that for each word in a document, the algorithm estimates the probability of it belonging to each topic and assigns words to a topic accordingly. Similarly, for each topic, the algorithm estimates the probability distribution of words associated with that topic.
- **Model Learning:** During training, LDA iteratively adjusts the topic assignments of words to find the best fit for the given documents. This involves optimizing the topic-word distributions and document-topic distributions to minimize the difference between the observed words and the reconstructed words based on the topics.
- **Application in Topic Modelling:** The LDA model results in a set of topics, each represented as a distribution of words. These topics are discovered without prior knowledge of what they might be, making them useful for uncovering the underlying themes in a collection of texts.
- **Use Cases:** LDA finds applications in various domains, including topic modelling in which it identifies themes within large sets of documents. And also, in content recommendations in understanding user preferences based on the topics they engage with. And also, in sentiment aanalysis in analysing the sentiment associated with different topics. And also, in information retrieval in improving search results by considering the topic relevance.

- **Limitations:** While LDA is powerful, it has some limitations, such as its sensitivity to the number of topics chosen and the complexity of real-world documents. Interpretation of the topics can also be subjective.

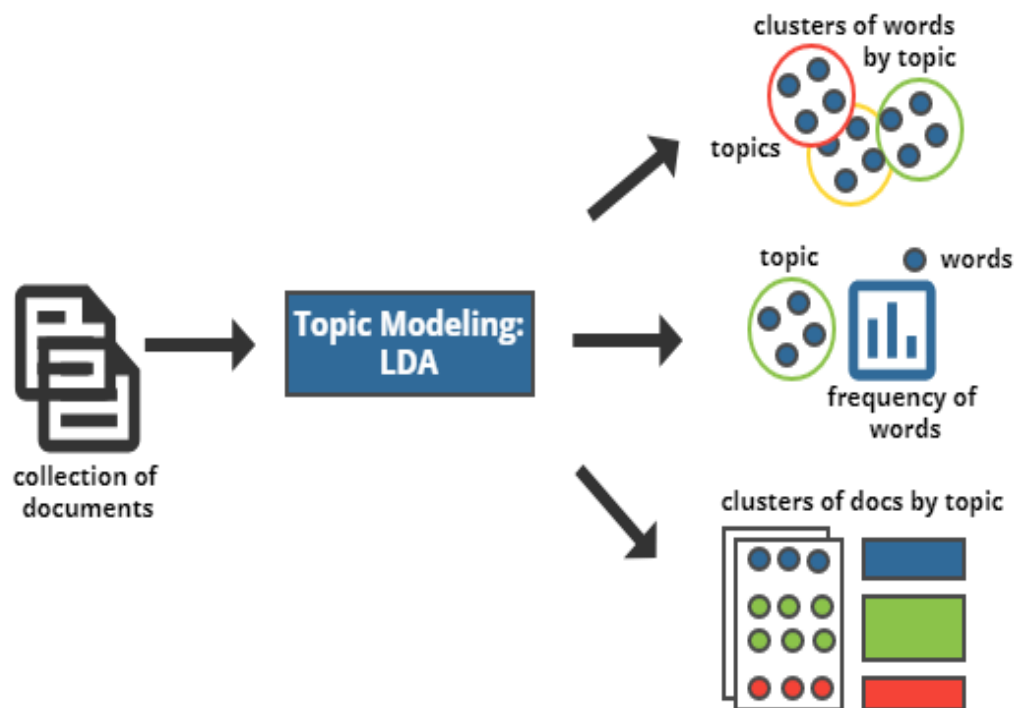


Figure 7.6: The LDA model.

When LDA model is built as shown below in figure 7.7, we printed 4 topics each containing 5 no. of words/ bag of words. An email document may contain multiple topics, with multiple bags of words. Next, we will visualize LDA model by using genism & find out which topics are prevalent in the email text data.

```
(0, '0.011*"company" + 0.008*"market" + 0.007*"enron" + 0.006*"business" + 0.006*"time"')
(1, '0.075*"enron" + 0.009*"would" + 0.009*"corp" + 0.007*"know" + 0.007*"deal"')
(2, '0.032*"enron" + 0.013*"data" + 0.009*"corp" + 0.009*"operation" + 0.009*"agreement"')
(3, '0.035*"font" + 0.017*"california" + 0.016*"power" + 0.015*"price" + 0.012*"state"')
```

Figure 7.7: Four Topics are Printed Each with 5 No. of Words.

In Figure 7.8, each bubble on left side, represents a topic. If the bubble is larger, the more prevalent that topic will be. We can get details by clicking on the relevant topic. Words form a topic, for a good topic model, it will have big bubble, which is not overlapping. A model which has a greater number of overlaps, and have small sized bubbles, which are clustered in one area has too many topics.

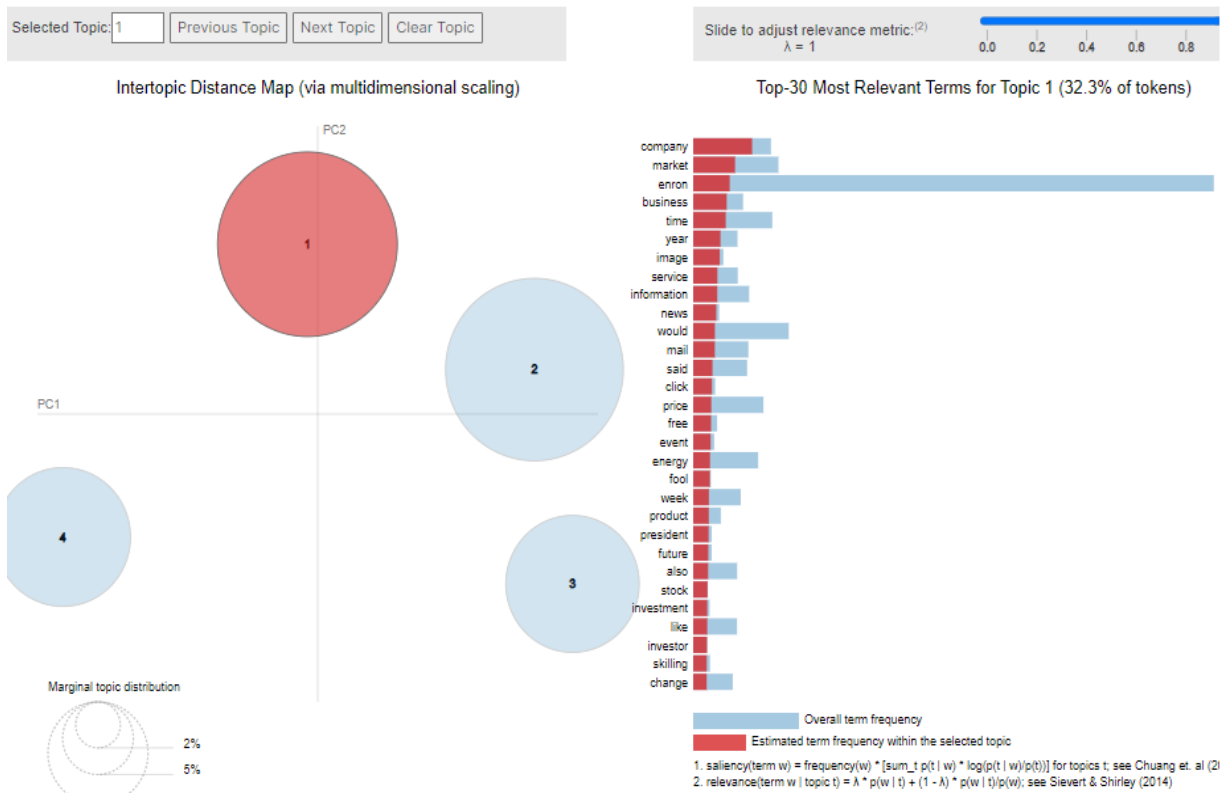


Figure 7.8: Visual Representation of LDA model.

7.6.3 Extracting Fraud-Related Topics

Once the LDA model is trained, we can analyse the topics it has generated. By identifying topics that seem to be closely related to fraudulent activities, we can discern patterns in the content of fraudulent emails. For Enron email data, a suspicious topic would be one where employees are discussing stock bonuses, selling stock, stock price, and perhaps mentions of accounting or weak financials.

In the case of the model above, topic 1 is suspicious topic which is describing the fraudulent behaviour. So, we will flag that topic as fraudulent as show in Figure 7.9. Now we will assign topics to our original data, now we will flag all our data where topic 1 as a fraudulent topic.

After flagging data that seems to be fraudulent, now we can easily identify which emails are of fraudulent nature and Figure 7.10 shows counts of fraudulent topic, this will help us to use this labelled data as a filter on top of many supervised machine learning models.

	Dominant_Topic	% Score	Original text	fraud_lda_model
0	1.0	0.843944	[entity, regarding, following, list, make, fol...	1
1	1.0	0.689911	[california, department, water, resource, rece...	1
2	2.0	0.500938	[enron, employee, meeting, notice, enron, empl...	0
3	2.0	0.988783	[york, contact, find, permanent, place, workin...	0
4	2.0	0.597396	[elite, broker, danm, sound, like, full, legal...	0

Figure 7.9: Data Frame after Flagging Topic 1 as Fraudulent.

```
count = topic_df['fraud_lda_model'].value_counts()
print(count)
```

```
0    986
1    294
Name: fraud_lda_model, dtype: int64
```

Figure 7.10: Counts of Fraudulent Emails.

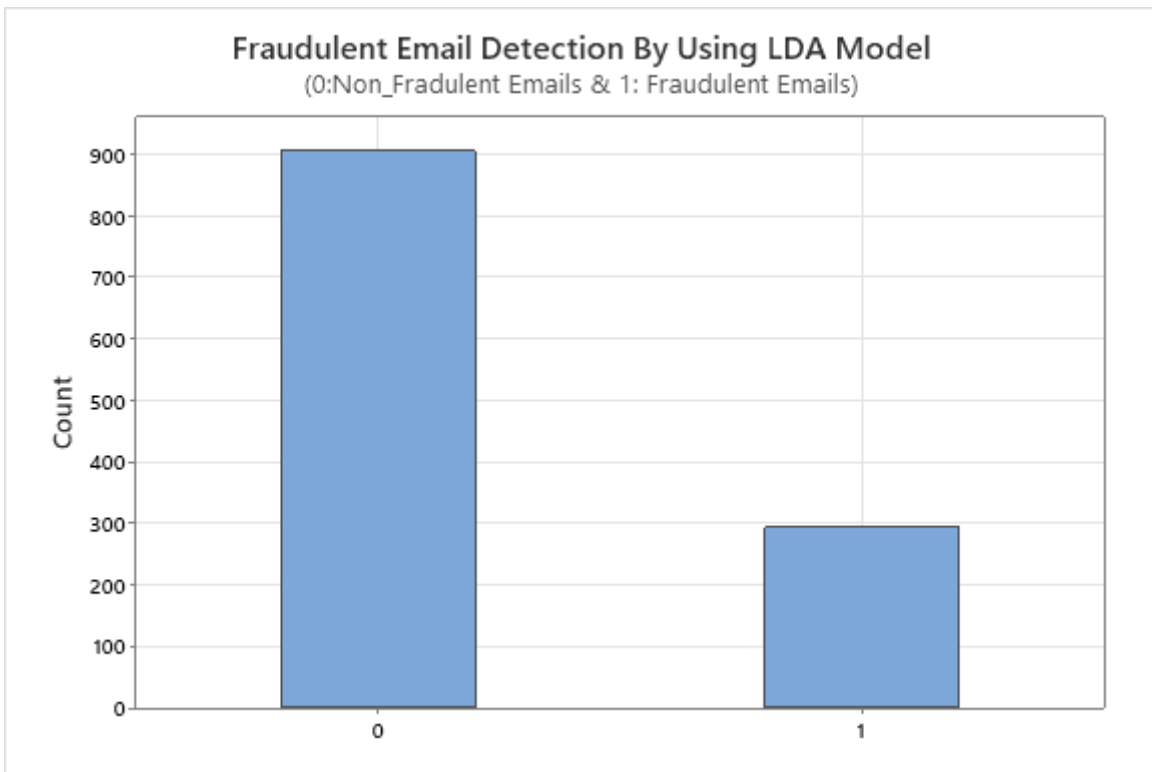


Figure 7.11: Fraudulent Email Detection using LDA Model.

If we compare the manual fraudulent email detection by using list of terms with LDA model which is a technique of topic modelling as shown above in Figure 7.11, it was found that by using LDA model we were able to detect more emails which are of fraudulent nature. In case of manual detection by using list of terms less emails were detected. In comparison results as follow, manual detection found 124 fraudulent emails out of 1200 selected whereas in case of

LDA model whereas in case of LDA model it detected 294 fraudulent emails out of 1200 selected emails. So, LDA perform well as shown below in Figure 7.12.



Figure 7.12: Comparison of Manual and Email Detection using LDA Model.

7.6.4 Classification and Scoring

The topics extracted from the LDA model can be used as features for fraud detection. Machine learning classifiers can be trained using these topics as input, alongside other relevant features. The classifier learns to differentiate between legitimate and fraudulent emails based on the identified topics. Now that the data is labelled it can be used as a feature in a machine learning model. And, also as filter on top of a machine learning model, as shown in Figure 7.13.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, shuffle=True, random_state=42)
print(f'Split successful')
print(f'X_Train: {len(X_train)}\nX_Test: {len(X_test)}\ny_Train: {len(y_train)}\ny_Test: {len(y_test)}')
```

```
Split successful
X_Train: 960
X_Test: 240
y_Train: 960
y_Test: 240
```

Figure 7.13: Training & Testing Data.

In the context of using topics extracted from the LDA model for fraud detection in emails, several machine learning models can be employed to achieve accurate classification and scoring. The choice of the model depends on factors such as the nature of the data, the complexity of the problem, and the desired interpretability of the results. The machine learning model should be based on experimentation and thorough evaluation using appropriate metrics like precision, recall, F1-score, and ROC curves. Moreover, feature engineering, including the incorporation of LDA-derived topics, can significantly impact the performance of ML models. Here are some common ML models used for fraud detection.

7.6.4.1 Logistic Regression

Logistic Regression is a simple yet effective linear classification algorithm. It's interpretable when relationship between features & target is relatively straight forward as in Figure 7.14.

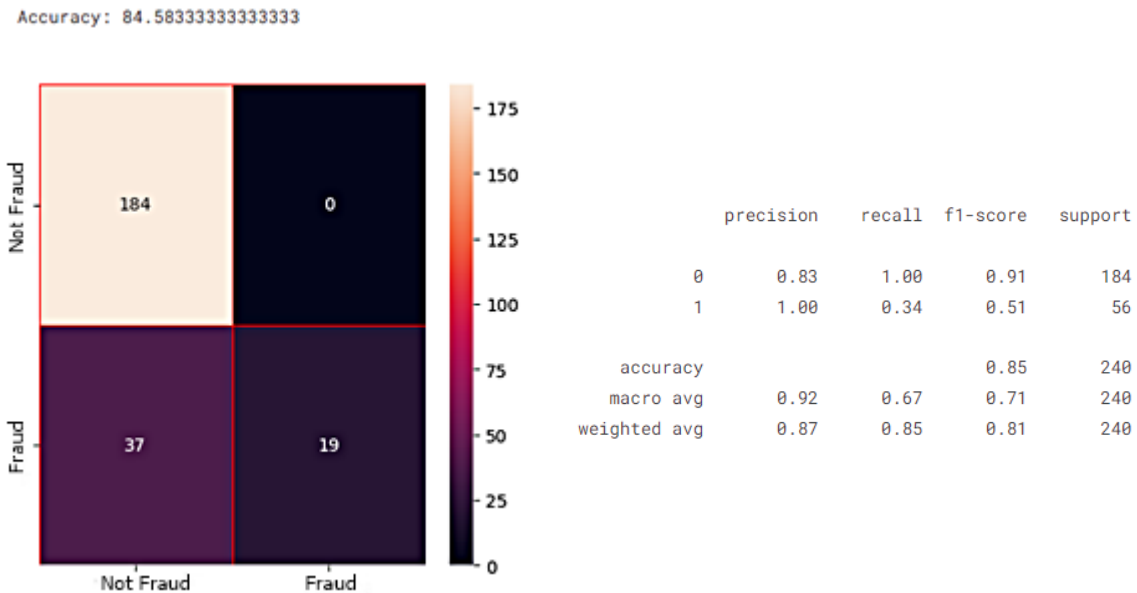


Figure 7.14: Confusion Matrix & Classification Report using Logistic Regression.

7.6.4.2 Linear Support Vector Classifier

Linear SVC is a linear classification algorithm that aims to find a hyperplane that best separates data points of different classes in feature space. It's a variant of the Support Vector Machine (SVM) algorithm that works well for binary and multi-class classification tasks. When considering its application to fraud detection using topics extracted from the LDA model, Linear SVC can offer advantages. Results are shown in Figure 7.15.

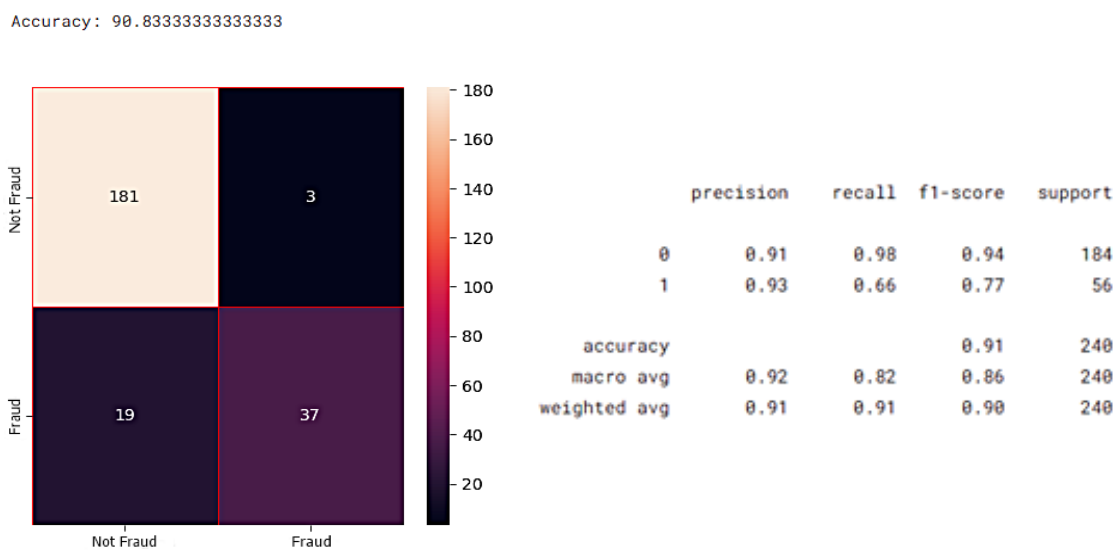


Figure 7.15: Confusion Matrix & Classification Report using Linear SVC.

7.6.4.3 Bernoulli Naive Bayes

It is a variant of Naive Bayes algorithm that is particularly well-suited for working with binary data, such as presence or absence of specific features. In context of fraud detection using topics extracted from LDA model, Bernoulli Naive Bayes can be a valuable choice due to simplicity and effectiveness in handling binary features. Results are shown in Figure 7.16.

Accuracy: 76.66666666666667

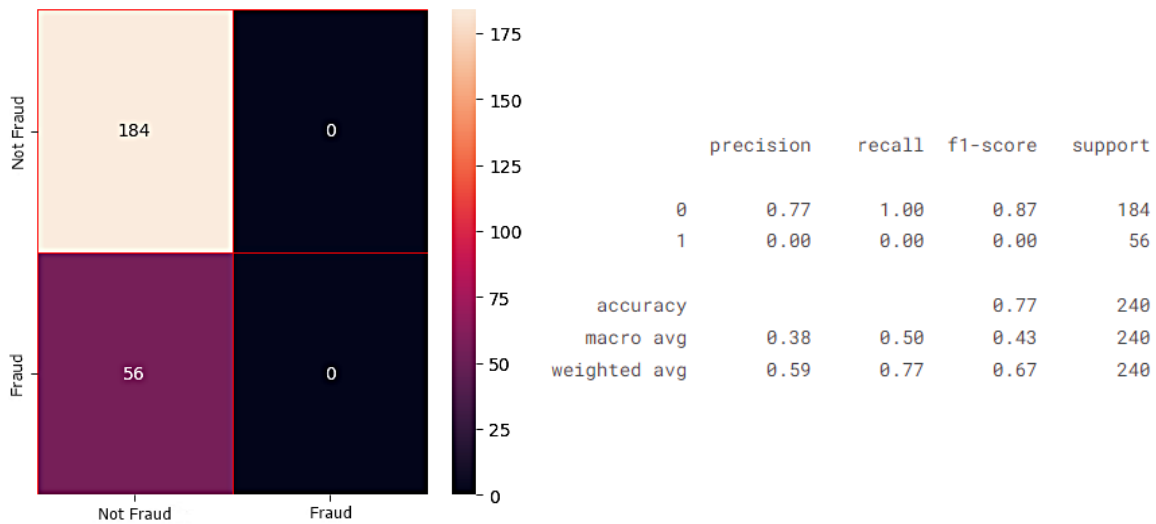


Figure 7.16: Confusion Matrix & Classification Report using Bernoulli Naive Bayes.

7.6.4.4 K-Nearest Neighbours

KNN is a simple instance-based learning algorithm used for smaller datasets. It classifies data point by considering class labels of its k-nearest neighbours. Results shown in Figure 7.17.

Accuracy: 86.66666666666667

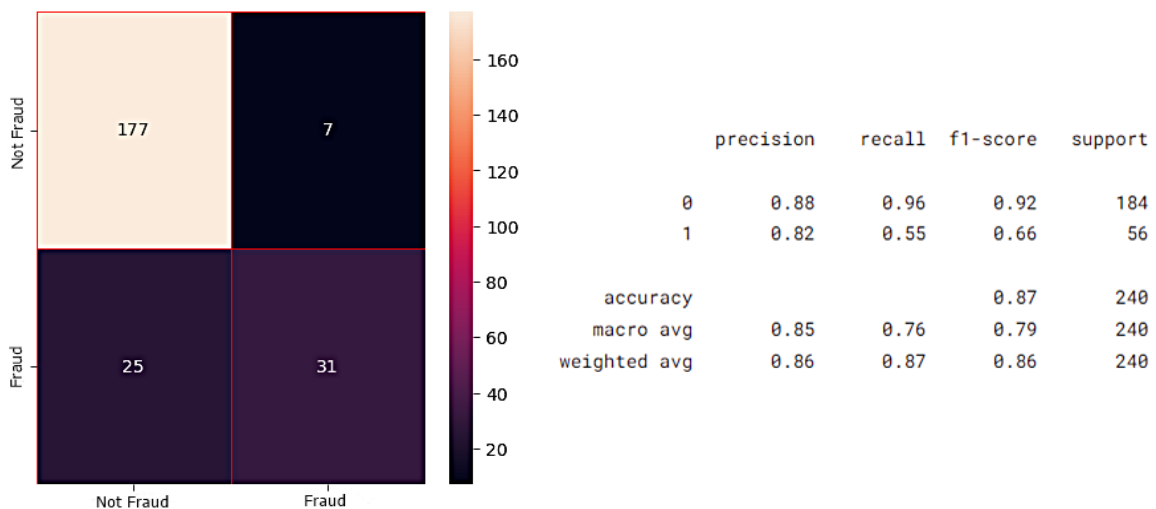


Figure 7.17: Confusion Matrix & Classification Report using K-Nearest Neighbours.

7.6.4.5 Random Forest Classifier

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve accuracy and reduce overfitting. It's effective for handling high-dimensional data & can work well with numerical & categorical features. Results are shown in Figure 7.18.

Accuracy: 84.58333333333333

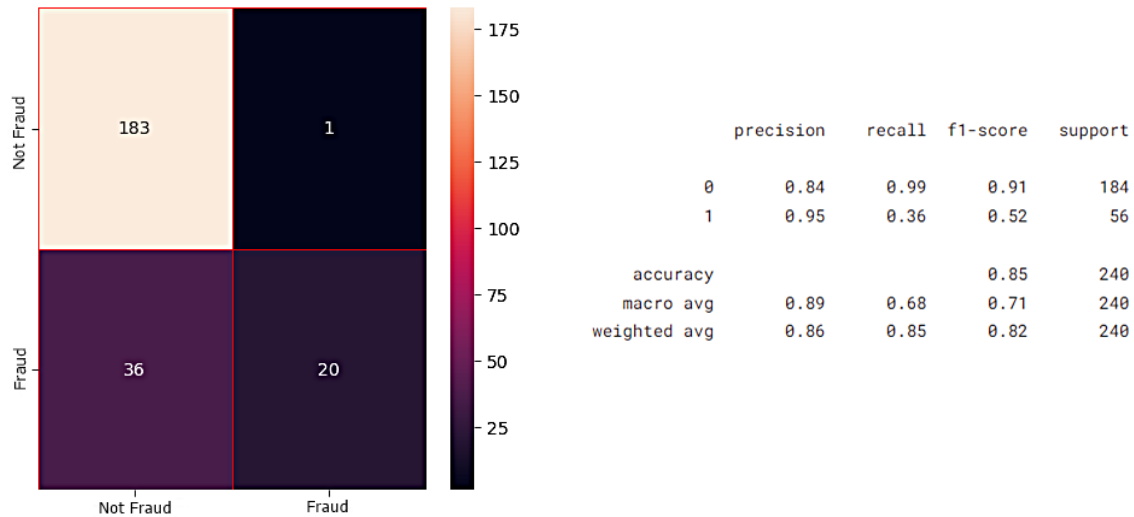


Figure 7.18: Confusion Matrix & Classification Report using Random Forest.

7.6.4.6 Gradient Boosting

Gradient Boosting is another ensemble technique that builds multiple models sequentially, each trying to correct the errors of the previous one. It's powerful for capturing complex relationships in the data & perform well in fraud detection. Results are shown in Figure 7.19.

Accuracy: 85.41666666666666

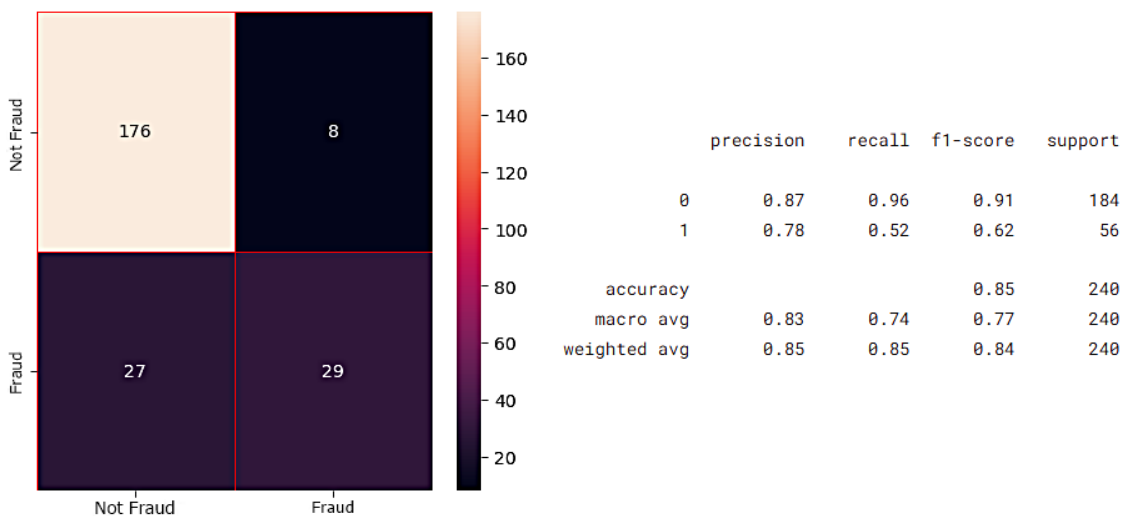


Figure 7.19: Confusion Matrix & Classification Report using Gradient Boosting.

7.6.4.7 Decision Tree Model

Decision trees can be used independently or as part of ensemble methods like Random Forest. They provide interpretable rules for classification. Results are shown in Figure 7.20.

Accuracy: 77.5

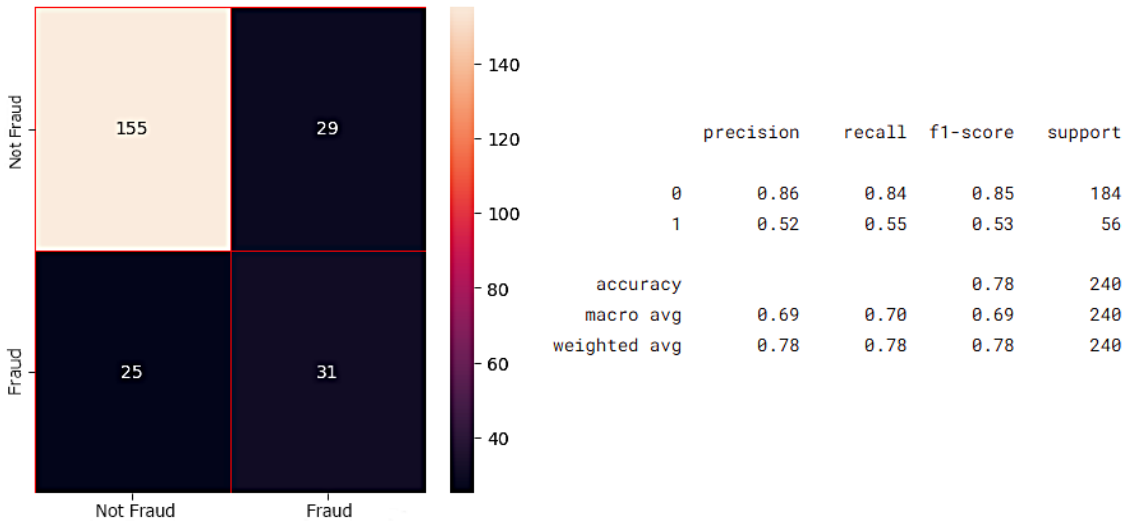


Figure 7.20: Confusion Matrix & Classification Report using Decision Tree.

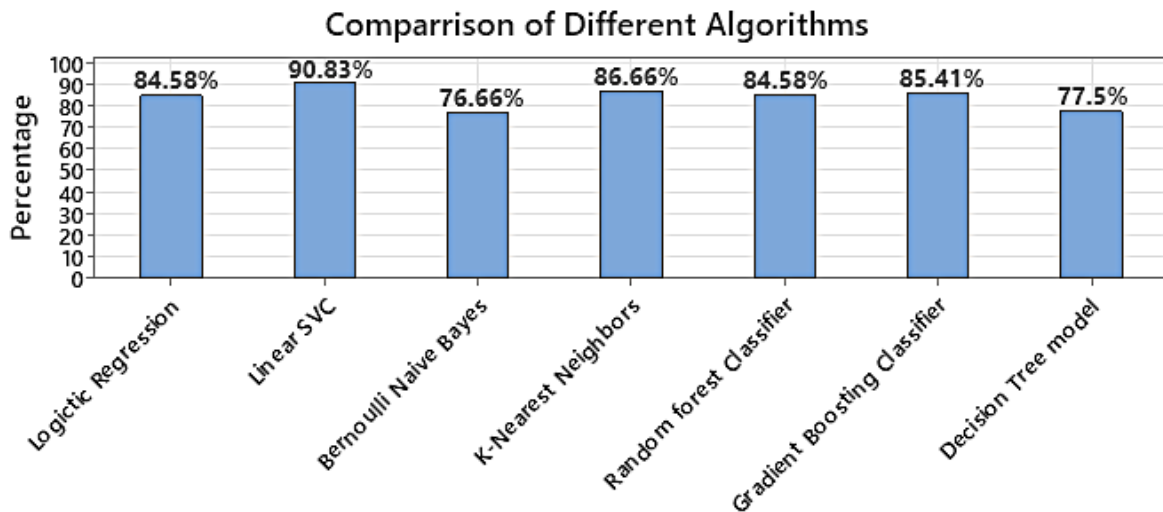


Figure 7.21: Comparison of Accuracy of Different ML Algorithms.

In Figure 7.21, an accuracy comparison of various machine learning algorithms is provided. Each algorithm's accuracy is listed as a percentage, indicating how well each algorithm performs in classifying fraud email data accurately. The accuracy value reflects the proportion of correctly classified instances out of the total instances in the dataset. In this context, higher accuracy percentages are generally indicative of better-performing models. Other metrics such as precision, recall, and F1-score were also found out for each algorithm respectively. Because they are important to comprehensively evaluate performance of these algorithms.

7.6.5 Model Evaluation and Improvement

The effectiveness of the fraud detection model can be assessed through various evaluation metrics. Iterative refinement of the model can involve adjusting the number of topics in the LDA model, tuning hyperparameters, and incorporating additional features for improved accuracy.

7.7 Prioritizing Fraudulent Email Data to Reduce Forensic Backlog

After flagging data that seems to be fraudulent, now we can easily identify which emails are of fraudulent nature and which are not. When dealing with a large volume of emails flagged as potentially fraudulent using the LDA model, it's important to efficiently manage the testing process to reduce the cloud forensic backlog. One effective approach is to focus testing efforts primarily on the emails that are predicted to be fraudulent by the LDA model. This strategy optimizes resources by prioritizing the most suspicious cases while discarding non-fraudulent emails. Following email data that is stored in original text are selected and other email data which is not flagged is discarded. The process is as follows:

	Dominant_Topic	% Score	Original text	fraud_lda_model
0	1.0	0.843944	[entity, regarding, following, list, make, fol...	1
1	1.0	0.689911	[california, department, water, resource, rece...	1

Figure 7.22: Prioritize Flagged Topic 1 as Fraudulent.

7.7.1 LDA Model and Fraud Detection

The LDA model has been utilized to identify patterns and topics associated with fraudulent emails. By analysing the topics extracted from emails, the LDA model assigns probabilities indicating the likelihood of an email being fraudulent.

7.7.2 Flagging Fraudulent Emails

Through the LDA model's predictions, emails are flagged as either likely fraudulent or non-fraudulent. This initial flagging is based on the model's assessment of the emails' content and the presence of topics linked to fraud.

7.7.3 Prioritization for Testing

Instead of testing every flagged email, a prioritization strategy is employed. In this strategy, the focus is primarily on testing the emails that the LDA model has identified as likely fraudulent as shown in Figure 7.22. By concentrating resources on these emails, investigation process becomes more efficient, and the most suspicious cases are dealt with promptly.

7.7.4 Resource Optimization

This prioritization approach optimizes the utilization of cloud forensic resources. Since testing, analysing, and investigating flagged emails can be resource-intensive, directing efforts towards those emails most likely to be fraudulent reduces the overall workload and accelerates the identification of actual cases of fraud.

7.7.5 Risk Mitigation and Efficiency

By emphasizing testing on the most suspicious cases, the risk of overlooking potentially critical fraudulent activities is minimized. At the same time, the strategy maximizes efficiency by reducing the need to invest significant resources in testing emails that the model has indicated are less likely to be fraudulent.

7.7.6 Post-Testing Steps and Continuous Monitoring and Adaptation

The results obtained can be used to refine and improve the LDA model or any other model used for fraud detection. This iterative process helps the model become more accurate over time, enhancing its ability to distinguish between fraudulent and non-fraudulent emails. As new data becomes available, the model can be adapted and trained to capture emerging patterns of email fraud. This ensures that the system remains effective in identifying fraudulent emails in an ever-evolving landscape.

7.8 Discussion

By extracting relevant topics from email data, the LDA model identifies patterns associated with fraudulent activities, enhancing the accuracy of detection. These identified topics serve as indicators, allowing for the categorization of emails as potentially fraudulent or non-fraudulent. Leveraging this information, a strategic prioritization process comes into play. Emails that the LDA model flags as having a higher likelihood of being fraudulent are given precedence during forensic analysis. This approach effectively minimizes cloud forensic backlog by focusing investigative efforts on the most suspicious cases. The combination of LDA model-based email data extraction and subsequent prioritization for cloud forensic backlog reduction is a proactive and efficient approach to combating fraudulent activities. It empowers organizations to effectively allocate resources, swiftly identify critical instances of fraud, and continually refine their fraud detection capabilities. By prioritization strategy discussed above, organizations can effectively manage the cloud forensic backlog, focus their efforts on the most suspicious cases, and ensure a more streamlined and efficient approach to fraud detection and prevention and to lower cloud forensic backlog.

Chapter 8: Name Entity Recognition Using BERT Model & Relation Extraction

8.1 Introduction

After the initial steps of data extraction and prioritization in fraud detection, advanced techniques like Named Entity Recognition (NER) and Relation Extraction (RE) further enhance the analysis process. NER identifies entities like names, locations, and dates in the extracted emails, while RE uncovers relationships between these entities. This enriches the understanding of fraudulent activities and aids in uncovering complex schemes. By integrating NER and RE, the flagged emails are subjected to a more comprehensive analysis, revealing hidden connections and patterns that might otherwise go unnoticed. This combined approach not only sharpens the accuracy of fraud detection but also provides deeper insights into the intricate web of fraudulent activities, thereby strengthening an organization's defence against financial threats.

The integration of NER and RE enhances information extraction's accuracy. Information extraction is an important process in natural language processing that involves identifying and extracting structured information from unstructured text. Named Entity Recognition (NER) and Relation Extraction (RE) are two fundamental techniques employed in this process, each playing a distinct role in uncovering valuable insights from text data.

NER is the task of identifying and classifying entities within a text, such as names of people, organizations, locations, dates, and more. By using linguistic patterns and context clues, NER algorithms automatically tag and categorize these entities, transforming unstructured text into structured data. For example, in a fraud detection context, NER could identify names of individuals, company names, transaction dates, and locations mentioned in emails or documents. RE focuses on identifying and extracting relationships between entities within text. It goes beyond the mere identification of entities and aims to discover how they are connected. For instance, in fraud detection, RE might reveal relationships between individuals and organizations, financial transactions, or specific activities mentioned in emails. RE algorithms analyse syntactic and semantic patterns to determine the nature of relationships, providing a deeper understanding of the context.

8.2 Name Entity Recognition & Relation Extraction

8.2.1 Name Entity Recognition (NER)

A natural language processing technique, NER helps in identification of named entities. And then further distribute it in categories/ labels like PER, ORG, LOC etc. These three are some

of the most important categories of NER. Name entity recognition is dependent on POS tagging. The goal of NER is to extract structured information from unstructured text and to assign appropriate labels to each recognized entity. NER plays a critical role in understanding and organizing textual data, enabling machines to comprehend the context and relationships between entities. By automatically identifying and categorizing named entities, NER contributes to various applications, including information retrieval, question answering, sentiment analysis, and, as mentioned earlier, fraud detection.

NER algorithms employ a variety of techniques, including rule-based methods, machine learning models (such as conditional random fields or deep learning approaches like LSTM and BERT), and combinations of these methods. These algorithms analyse linguistic patterns, context, and syntactic structures to accurately identify and categorize named entities within text. We will be using BERT.

8.2.2 Relation Extraction (RE)

In the context of NLP, entities are typically recognized and classified using techniques like Named Entity Recognition (NER). Once the entities are identified, RE goes further by determining the type of relationship that exists between them. These relationships can be diverse, including actions, affiliations, ownership, temporal relationships, and more.

Relation Extraction is a vital NLP task that goes beyond identifying named entities, focusing on understanding how these entities are interconnected. It's a cornerstone for creating structured knowledge representations and enabling machines to grasp intricate relationships within textual data.

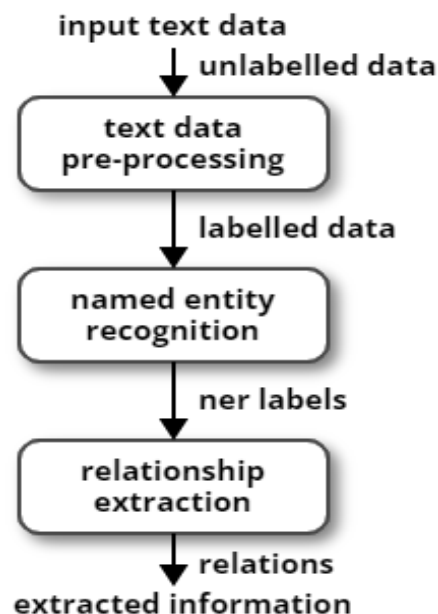


Figure 8.1: NER and RE After Topic Modelling.

8.2.3 NER and RE After Topic Modelling

After identifying fraudulent emails using Topic Modelling with the LDA model, the next step is to extract valuable information from the text, particularly through Named Entity Recognition (NER) and Relation Extraction as shown in Figure 8.1. These techniques add another layer of insight to enhance fraud detection and provide a more comprehensive understanding of the fraudulent activities described in the emails. After detecting fraudulent emails using the LDA model, NER can be applied to identify specific entities that are relevant to the email fraud context.

- **Persons:** Detecting names of potential fraudsters, accomplices, or victims mentioned in the email.
- **Organizations:** Identifying companies or groups involved in fraud.
- **Locations:** Recognizing locs where fraudulent transactions might take place.

RE is the process of identifying and classifying relationships between entities within text. After NER, the extracted entities can be analysed to understand the connections between them, aiding in the identification of complex fraud networks & schemes. For instance:

- **Person-Organization Relations:** Uncovering connections between individuals and organizations involved in fraudulent activities.
- **Person-Location Relations:** Finding locations where the involved personal performed fraudulent tasks.
- **Location-Organization Relations:** Identifying the locations where fraudulent operations are being carried out.

By combining NER and Relation Extraction with the insights gained from the LDA model, you create a more detailed and structured understanding of the fraudulent activities described in the emails. This enriched information can be used to:

- **Enhance Detection Accuracy:** Incorporating specific entity names and relationships as features in your fraud detection model can improve its ability to identify intricate fraud patterns.
- **Build Graph Representations:** Constructing graphs that represent relationships between entities can help visualize and analyse the fraud networks.

- **Provide Context:** Understand the who, what, where, and when of fraudulent activities, aiding in comprehensive investigation.

8.2.4 Name Entity Recognition & Relation Extraction Process

The process of Named Entity Recognition (NER) and Relation Extraction involves identifying specific entities and relationships within textual data. These techniques enhance the understanding of information in the text, allowing for the extraction of valuable insights as shown in Figure 8.2.

8.2.4.1 Preprocessing Before NER

Start by preprocessing the text data. This includes tokenization (breaking text into words or tokens), sentence segmentation, and removing unnecessary characters or formatting and converting them to lowercase. Preprocessing prepares the text for further analysis.

8.2.4.2 Named Entity Recognition (NER)

NER involves identifying and categorizing specific entities within the text. The entities can include names of people, organizations, locations, dates, monetary values, percentages, and more. The process involves the following steps:

- **Tokenization:** Break the text into individual words or subunits.
- **Part-of-Speech (POS) Tagging:** Assign POS tags to each word to determine their grammatical roles (e.g., noun, verb, adjective).
- **Entity Detection:** Using patterns, rules, or machine learning models, identify words or sequences of words that correspond to entities.
- **Entity Classification:** Classify the identified entities into predefined categories (person, organization, date, etc.). Popular NER libraries, such as spaCy, NLTK, and Stanford NER, offer pre-trained models for accurate entity recognition.

NER Techniques include rule-based approach and machine learning based approach. In Rule-Based Approach we use design rules and patterns to match entity names and structures. For instance, recognizing capitalized words as potential names or using regular expressions to identify date formats. While in machine learning Approach we train machine learning models using BERT on labelled training data to predict entity labels for new text.

8.2.4.3 Relation Extraction (RE)

Relation Extraction aims to identify and classify relationships between named entities within the text. This involves uncovering how entities mentioned in the text are related to each other. The process includes:

- **Dependency Parsing:** Analyse the grammatical structure of the sentences to determine the relationships between words (e.g., subject, object, verb).
- **Entity Pair Identification:** Identify pairs of entities within the same sentence that could potentially be related.
- **Feature Extraction:** Extract features from the text that provide contextual information about the entity pairs and their surroundings.
- **Classification or Clustering:** Use machine learning algorithms to classify relationships or group similar relationships together.

Relation extraction can be rule-based, supervised (using labelled data), or unsupervised (using clustering techniques). Deep learning models like transformers have shown promise in relation extraction tasks.

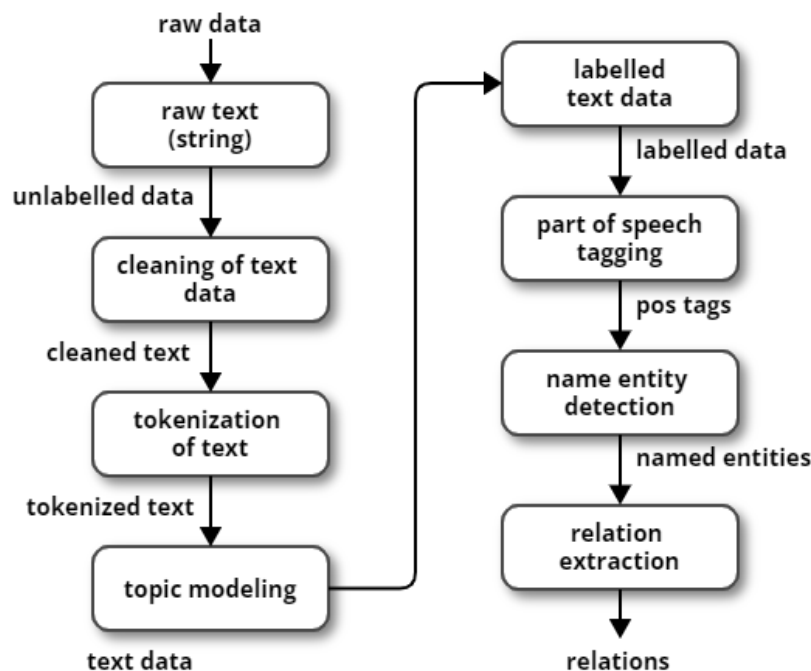


Figure 8.2: NER and RE Process.

8.2.4.4 Postprocessing and Analysis

After NER and Relation Extraction, postprocess the results to ensure accuracy and coherence. You can analyse the extracted entities and relationships to gain insights into the structured

information present in the text. First, we clean raw text by data cleaning, then we normalize it and tokenize it for topic modelling. After topic modelling we get labelled data by applying LDA model. Then we apply POS tagging on text data to extract POS tags by using the Natural Language Toolkit (NLTK) library. Then by using the spaCy library to perform NER to get Named tags. And at last relaxations are extracted from these named entities.

8.3 Name Entity Recognition Using BERT

BERT is a pre-trained language model developed by Google in 2018. It is based on the Transformer architecture and has revolutionized the field of NLP by achieving state-of-the-art results on a wide range of NLP tasks, including Named Entity Recognition (NER) shown in Figure 8.3. NER is a common NLP task that involves identifying and classifying named entities (such as names of people, organizations, locations, etc.) within a text, in our case we have fraudulent email text data. BERT can be used for NER by finetuning its pre-trained model on a labelled NER dataset. Here's how it works:

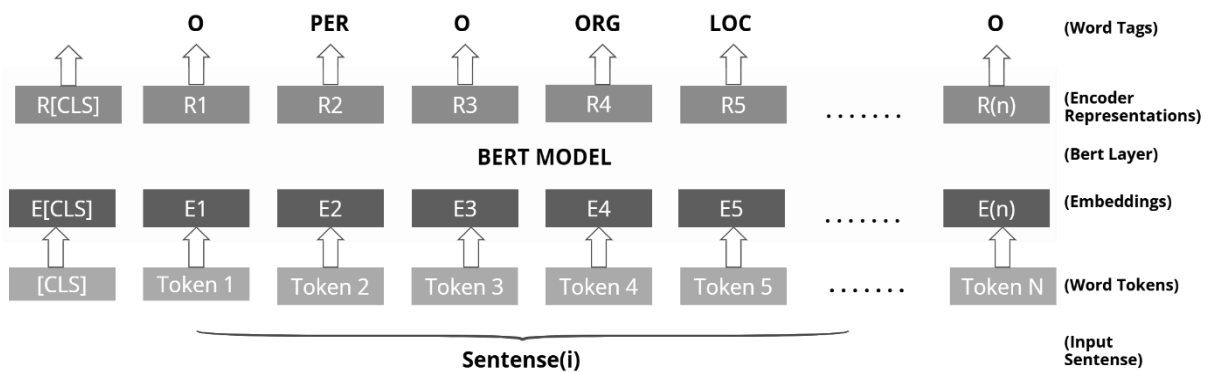


Figure 8.3: Bidirectional Encoder Representations from Transformers Model.

8.3.1 Preprocessing Data

BERT is first pre-trained on a large corpus of text data i.e., fraudulent email data using a masked language model objective. During pre-training, it learns to predict missing words in sentences (masked tokens) based on the surrounding context. The key innovation of BERT is its bidirectional nature, meaning it considers both left and right context words when making predictions. This enables it to capture deep contextual relationships between words. The email text data in “Original text” column as shown in Figure 8.4 is renamed to “body” and all other columns were dropped. The body column is converted into string & dataset is reshuffled as shown in figure 8.5. Only those email data is selected which were fraudulent and are prioritized.

	Dominant_Topic	% Score	Original text	fraud_lda_model
0	1.0	0.843944	[entity, regarding, following, list, make, fol...	1
1	1.0	0.689911	[california, department, water, resource, rece...	1

Figure 8.4: Fraudulent Email Text Data After Topic Modelling Using LDA Model.

	body
0	REVISED Weekly Cost Reports Attached is a revi...
1	Re Original Message From Storey Geoff mailtoGe...

Figure 8.5: Reshuffled body Column.

After that we split the body column in such a way that each email text data of a particular index was given a sentence number and after that body column is exploded to have each word in a separate row and renamed the column to word, this is called tokenization in which we break the input text into individual words as shown in figure 8.6.

	word	sentence_no
0	Enron	1
1	Mentions	1

Figure 8.6: New word Column with sentence_no.

By using the Natural Language Toolkit (NLTK) library POS tags are extracted. After that by using spaCy library, we get NER tags with respect to each word as shown in figure 8.7.

	word	sentence_no	pos	tag
0	Enron	1	NN	ORG
1	Mentions	1	NNS	O
2	Cover	1	NN	O
3	Story	1	NN	O
4	THE	1	DT	O

Figure 8.7: New word Column with POS & NER tags.

In the preparation of our data, which consists of sentences and corresponding entity tags. The ‘SentenceGetter’ class helps organize our data into sentences and tags. By extracting the words from each sentence in the DataFrame, and then storing the words in a list of lists (sentences) and then extracts the labels for each word in the sentences and stores them in a list of lists (labels).

Then we tokenize the sentences and create corresponding labels. We also added special tokens i.e., [PAD] tokens which are added to make all sequences of equal length. Later we converted these tokens to token IDs using the BERT tokenizer in the BERT vocabulary. The ‘tag2idx’ dictionary is used in converting tags to numerical indices i.e., IDs, which is often required when working with machine learning models.

The function `'tokenize_and_preserve_labels'` is used for tokenizing input sentences while preserving the corresponding labels for each token. This is essential when preparing data for sequence labelling tasks, as in NER. In the preprocessing stage we are tokenizing sentences, aligning labels, converting tokens to IDs, and then padding or truncating the sequences to a uniform length. Then this processed data will be used for training BERT.

8.3.2 Model Initialization & Training Setup

By initializing a `'BertForTokenClassification'` model, a BERT variant fine-tuned for sequence labelling tasks like NER. By training this model using our prepared data and defined dataloaders, to use it for making predictions on new NER sequences. The model has a token classification head that outputs the predicted label for each token.

By setting up the training process, including moving the model to the GPU if available. We can configure the optimizer, weight decay, and learning rate scheduler. Then we define the number of training epochs, which in our case is 30 and maximum gradient norm.

After setting up the optimizer, weight decay, and the learning rate scheduler for training our sequence labelling model. We are now ready to start the training loop, where we'll iterate through batches of data, will perform forward and backward passes, update the model's parameters, and adjust the learning rate according to the scheduler's plan.

8.3.3 Fine-tuning BERT Model

After pre-training, BERT is fine-tuned on specific downstream tasks like NER. For NER, the model is trained to predict the named entity labels for each token in a sentence. Tokens that are not part of a named entity are often assigned a special "O" label (for "Other"), while tokens that belong to named entities are labelled with specific entity types.

By utilizing the pre-trained BERT model as a feature extractor. We added a classification layer on top of BERT to predict entity labels for each token. We trained data labelled NER training data where each token is labelled with its corresponding entity type (e.g., PER for person, ORG for organization, etc.). We fine-tuned BERT model to minimize NER-specific loss, often using cross-entropy loss between predicted labels and ground-truth labels.

In our typical training loop for sequence labelling tasks using BERT. It trains the model on our training data and evaluates its performance on the validation data after each epoch. The last 2 epoch are shown in Figure 8.8. For each batch, we perform a forward pass to get logits without labels. we iterate over validation data in batches. The calculated training and validation losses are stored, which you can use for plotting the learning curve and tracking

the model's progress over each epoch. And then we store predictions and true labels for further analysis.

When training loop starts, it iterates over the training data in batches. For each batch, we move the batch to the GPU. We then perform a forward pass through the model and calculate the loss using provided labels. Then we perform a backpropagation to compute gradients. we clip gradients to prevent explosion. And we update the model's parameters using the optimizer and also updated the learning rate using the scheduler.

```
Average train loss: 0.007738838083155099

Epoch: 97%|██████████| 29/30 [12:06<00:25, 25.03s/it]

Average train loss: 0.007987956472095988

Epoch: 100%|██████████| 30/30 [12:31<00:00, 25.05s/it]
```

Figure 8.8: Last 2 Epoch with Average Train Loss.

8.3.4 Token-level Predictions

Fine-tuned BERT model with NER layer performs token-level predictions, meaning it assigns a label to each individual token in a sentence. This fine-tuned model can then be used to predict named entity labels for new, unseen sentences.

8.3.5 Post-processing

After making predictions, post-processing is often necessary to convert the token-level predictions into coherent named entities i.e., converting the predicted label IDs back to their corresponding entity labels. This involves merging consecutive tokens with the same entity type label and Group consecutive tokens with the same entity label into named entities. And handling cases where an entity spans multiple tokens. And then providing the final list of named entities along with their corresponding entity types.

8.3.6 Performance Evaluation

After training and validation loops, we evaluate the model's performance. We have calculated metrics such as validation loss, accuracy, and F1-score. By using these metrics, we will assess our model performance on the validation set. The classification report and confusion matrix provide detailed metrics for each NER label. The confusion matrix helps tell us which types of errors our model is making and provides information that can be used to calculate various metrics, such as accuracy, precision, recall, F1-score, and more. It summarizes the

performance of a classification model by showing the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions for each class.

- **True Positive (TP):** The model correctly predicted a positive class.
- **True Negative (TN):** The model correctly predicted a negative class.
- **False Positive (FP):** The model incorrectly predicted a positive class when the actual class was negative (Type I error).
- **False Negative (FN):** The model incorrectly predicted a negative class when the actual class was positive (Type II error).

Confusion matrix for tag 'PER':	Confusion matrix for tag 'ORG':
$\begin{bmatrix} 5503 & 1 \\ 5 & 14 \end{bmatrix}$	$\begin{bmatrix} 10409 & 48 \\ 5 & 15 \end{bmatrix}$
Confusion matrix for tag 'GPE':	Confusion matrix for tag 'O':
$\begin{bmatrix} 2430 & 5 \\ 11 & 9 \end{bmatrix}$	$\begin{bmatrix} 98797 & 5 \\ 0 & 10 \end{bmatrix}$

Figure 8.9: Confusion Matrix for NER using BERT.

A classification report is a summary of various evaluation metrics for each class in a classification task. It typically includes the following metrics for each class:

- **Precision:** The ratio of true positive predictions to the total number of positive predictions ($TP / (TP + FP)$). It measures the accuracy of positive predictions.
- **Recall (Sensitivity or True Positive Rate):** The ratio of true positive predictions to the total number of actual positives ($TP / (TP + FN)$). It measures the model's ability to identify all positive instances.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure between the two.
- **Support:** The number of actual occurrences of the class in the dataset.

```

Validation loss: 0.0034639021171079505
Validation Accuracy: 0.9993237433862434
Validation F1-Score: 0.9989116001613558

```

	precision	recall	f1-score	support
PER	1.00	1.00	1.00	5523
ORG	1.00	0.99	1.00	10477
GPE	1.00	0.99	0.99	2455
O	1.00	1.00	1.00	98812
accuracy			1.00	117267
macro avg	0.80	0.80	0.80	117267
weighted avg	1.00	1.00	1.00	117267

Figure 8.10: Classification Report for NER using BERT.

Confusion matrix for tags PER, ORG, GPE & O using BERT are shown in Figure 8.9. The results we've obtained from our BERT model are remarkably well on the validation set as shown in figure 8.10. The classification report provides more detailed information for each label. For GPE, Precision, Recall, and F1-score are all around 0.99, indicating strong performance for this label. For O, the "O" label represents non-entity tokens. Precision, Recall, and F1-score are all around 1.00, indicating excellent performance for these tokens. For ORG, Precision and F1-score are around 1.00, while Recall is slightly lower at 0.99, showing high performance for organization entities. For PER, Precision, Recall, and F1-score are all around 1.00, indicating strong performance for person entities.

The "accuracy" metric for all the classes combined is also very high, indicating that the model is correctly classifying a vast majority of tokens. The "macro avg" metrics for precision, recall, and F1-score are calculated as averages across all classes without considering class imbalance. In our case, since the classes are imbalanced, the macro avg values are somewhat lower than the "weighted avg" values. The "weighted avg" metrics take class imbalance into account, and these values are also around 1.00, indicating that our model is well-balanced in terms of its overall performance across different classes.

Overall, these results suggest that our model is performing exceptionally well on the validation set and is achieving high accuracy and F1-scores across different entity types. However, as always, it's important to carefully validate your model's performance on unseen data and consider potential sources of bias or error that might arise in real-world scenarios.

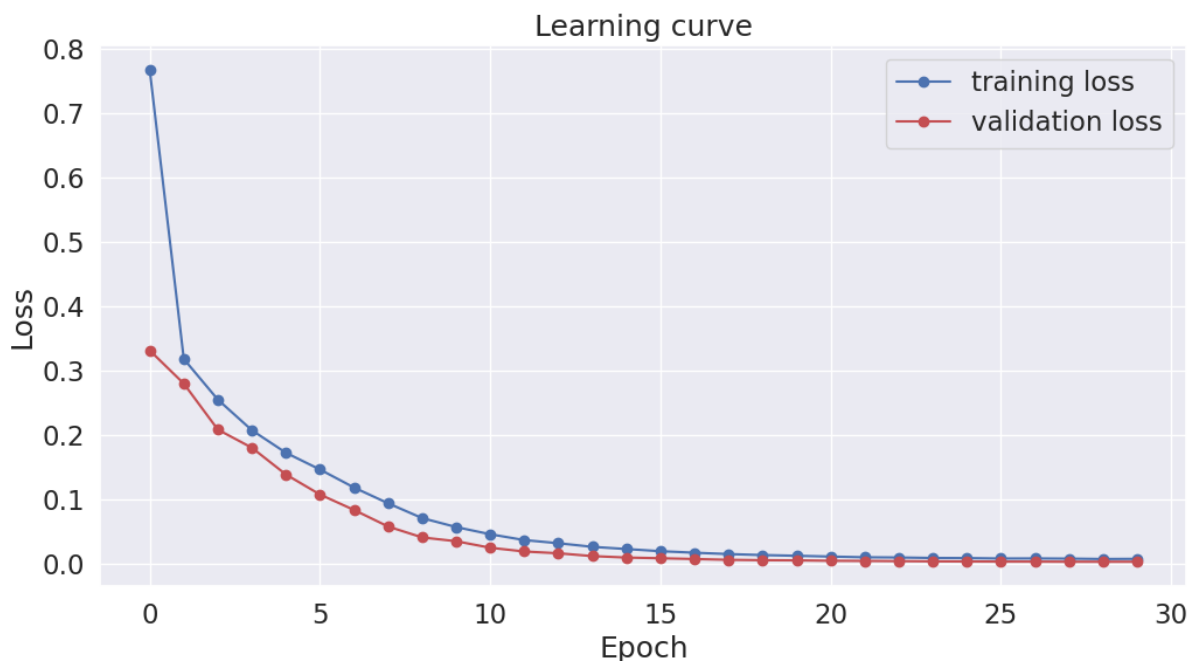


Figure 8.11: Learning Curve About Training and Validation Loss.

The learning curve of training and validation loss is a graphical representation as shown in figure 8.11, it shows how the loss values change as our model progresses through training epochs. It provides insights into how well our model is learning from the data and whether it's overfitting or underfitting. Initially, the training loss is high, and as the model iteratively updates its parameters through backpropagation, the loss gradually decreases.

The curve exhibits a steep initial decline, indicating rapid learning in the early epochs as evident in figure 8.10. The validation loss curve shows how the loss on the validation data changes as our model is trained. Validation loss is a measure of how well our model generalizes to unseen data. If the validation loss increases while training loss decreases, it indicates overfitting. Which is not the case, our curve shows a decreasing trend, reflecting that the model is improving its performance on unseen data. Our training and validation losses are converging which is showing model is generalizing well to new data.

8.4 Relation Extraction

The Apriori algorithm is a classic algorithm in data mining that is used for finding frequent itemsets in transactional databases. While it is more commonly associated with market basket analysis, it can also be adapted for certain types of relation extraction tasks where you have pairs of entities and you're interested in discovering common patterns between them.

8.4.1 Data Preprocessing

Start by preprocessing your text data, including tokenization, POS tagging, and named entity recognition. Identify the entities you're interested in and extract their corresponding labels.

8.4.2 Entity Pair Identification

Identify pairs of entities in email text that you want to extract relationships between.

8.4.3 Feature Extraction

Convert the entity pairs into feature vectors or transactional format. This could involve representing each entity pair as a set of attributes or features.

8.4.4 Apriori Algorithm

Apply the Apriori algorithm to discover frequent itemsets from your feature vectors. The algorithm will find entity pairs that frequently occur together in the dataset.

8.4.5 Association Rules Generation

From the frequent itemsets, generate association rules that express relationships between entities. These rules might have the form "entity A is associated with entity B."

8.4.6 Rule Filtering and Evaluation

Apply filters to the generated rules to remove trivial or uninteresting rules. Evaluate the generated rules using measures like support, confidence, and lift to assess their significance and quality.

e.g. (Tanya)->(Vince) it gives relation as (Vince is director of research, boss of Tanya)

It's important to note that while the Apriori algorithm can find co-occurring entity pairs, it doesn't capture more complex semantic relationships between entities. For example, it might identify that "Apple" and "CEO" often co-occur in a text corpus, but it won't necessarily understand the "works for" relationship between them. More advanced techniques like neural network-based models or knowledge graph-based methods are better suited for capturing and extracting more intricate relationships.

8.5 Comparison Between NER using BERT and NER using Combination of Rule-based Approach & CRF Model

Our model's NER performance using BERT appears to be very strong based on the provided metrics, achieving high precision, recall, and F1-scores across the mentioned classes. The reference paper's [] manually annotated NER tags in the detection phase and used union of rule based and CRF model for evaluation to increase NER performance, but with slightly lower scores compared to our BERT model. Following is a comparison based on the provided classification reports:

Our Classification Report Results

Validation loss: 0.0034639021171079505
 Validation Accuracy: 0.9993237433862434
 Validation F1-Score: 0.9989116001613558

	precision	recall	f1-score	support
PER	1.00	1.00	1.00	5523
ORG	1.00	0.99	1.00	10477
GPE	1.00	0.99	0.99	2455
0	1.00	1.00	1.00	98812
accuracy			1.00	117267
macro avg	0.80	0.80	0.80	117267
weighted avg	1.00	1.00	1.00	117267

Classification Report of Ref-paper

Named Entities	Precision	Recall	F1
Persons	0.84	0.92	0.88
Organizations	0.88	0.95	0.91
Locations	0.78	0.87	0.82

Min Yang, Kam-Pui Chow. AN INFORMATION EXTRACTION FRAMEWORK FOR DIGITAL FORENSIC INVESTIGATIONS. 11th IFIP International Conference on Digital Forensics (DF), Jan 2017, Orlando, FL, US. pp.61-76, ff10.1007/978-3-319-24123-4_4ff. fihal-01449071

Figure 8.12: Comparison of NER using BERT & NER using Combination of Rule-based Approach & CRF Model.

8.5.1 Metrics Results

Our model's NER using BERT seems to achieve very high F1-scores, indicating strong performance across classes. The F1-Score of 0.9989 is exceptionally high and suggests that the model is performing very well. On the other hand, the reference paper's F1-scores are generally lower, ranging from 0.82 to 0.91, which could be due to the NER using rule-based approach & Conditional Random Field Model & inherent challenges in manually labelling data as shown in Figure 8.13.

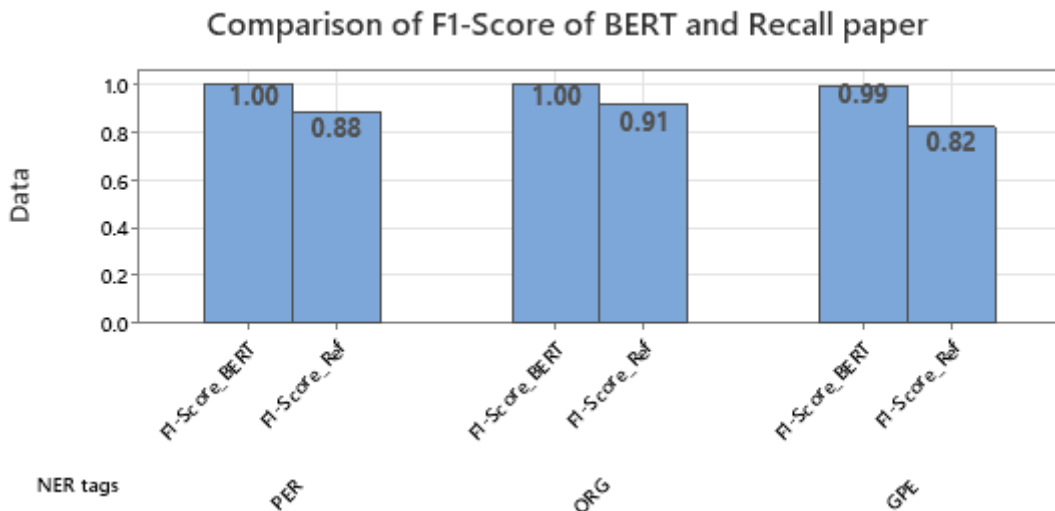


Figure 8.13: Comparison of F1-Score using BERT & Combination of Rule-based Approach & CRF.

Our model achieves high precision and recall values across the classes. This indicates that it can identify and classify named entities accurately, with very few false positives (low precision) or false negatives (low recall). The reference paper's precision and recall values are

also respectable, but slightly lower compared to our model's results as shown in Figure 8.14, and Figure 8.15.

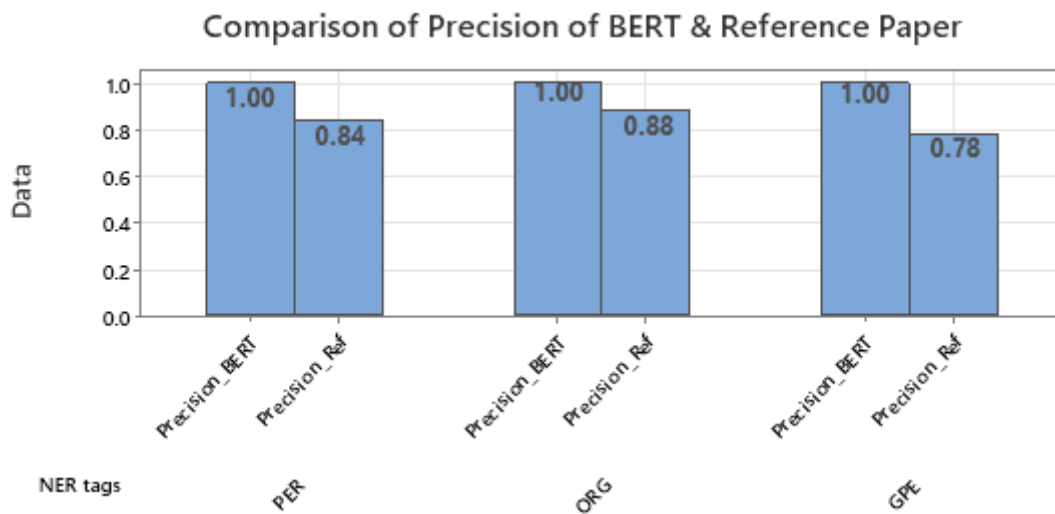


Figure 8.14: Comparison of Precision using BERT & Combination of Rule-based Approach & CRF.

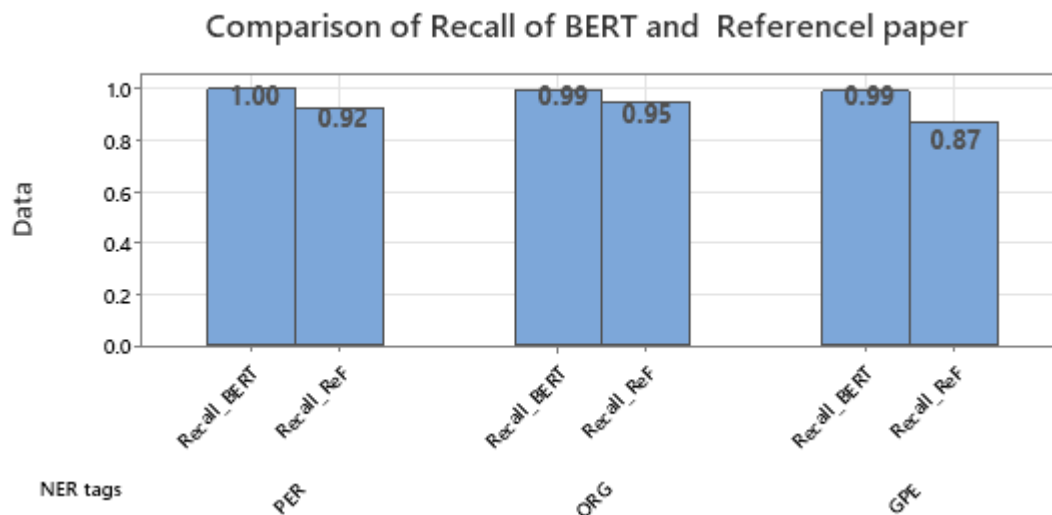


Figure 8.15: Comparison of Recall using BERT & Combination of Rule-based Approach & CRF.

8.5.2 Differences in Approach

BERT-based NER model appears to outperform the model described in the reference paper in terms of F1-Score. It achieves an extremely high F1-Score (0.9989) across multiple classes, which suggests that it is highly accurate and robust. BERT-based model has an advantage in recognizing additional entity types like GPE (Geopolitical Entity), which are not present in the reference paper. The reference paper's model is based on a combination of rule-based approaches and Conditional Random Fields (CRF), which is a different approach from the deep learning-based approach using BERT. Both methods have their strengths and weaknesses. It's important to note that in the reference paper, NER was done manually by three researchers. This manual annotation process may introduce human biases and

limitations. It should note that BERT's ability to capture contextual information and its bidirectional nature make it well-suited for NER tasks. However, it's worth noting that fine-tuning BERT for NER may require specialized datasets with labelled named entity information, and the model's performance can be influenced by factors like dataset quality, model size, and fine-tuning parameters.

8.6 Discussion

Fraudulent emails pose a significant threat to individuals and organizations, requiring advanced techniques for efficient detection and mitigation. Information extraction is a critical component of fraud detection, where identifying named entities and their relationships can unveil hidden patterns. We explored the synergistic approach of Named Entity Recognition (NER) using BERT and Relation Extraction after Topic Modelling for enhancing information extraction from fraudulent email data. Leveraging NER with BERT and Relation Extraction after Topic Modelling offers a powerful information extraction pipeline for fraud email data. This multi-step approach enables the identification of relevant entities, topics, and relationships, contributing to more accurate and targeted fraud detection. As the sophistication of fraudulent emails increases, combining advanced NLP techniques becomes crucial to stay one step ahead of malicious actors.

Chapter 9: Conclusion and Future Directions

9.1 Introduction

In our research journey, where we tried to understand cloud forensic backlog reduction through the innovative application of machine learning models. Our investigation into this critical domain has been a multi-faceted endeavour, encompassing various cutting-edge techniques and approaches.

One of the fundamental challenges we addressed in our research was the efficient management of vast volumes of digital evidence in the cloud. To tackle this issue, we employed hashing techniques to eliminate duplicate data, a crucial step in streamlining the forensic process and mitigating the backlog.

Furthermore, for email fraud detection, we used Latent Dirichlet Allocation (LDA) we have employed cutting-edge machine learning techniques, specifically the Latent Dirichlet Allocation (LDA) model. This methodology has empowered us to detect suspicious patterns and potentially malicious content with remarkable accuracy in the dataset that we used.

Another pivotal aspect of our research involved information extraction, where we used state-of-the-art BERT model for Named Entity Recognition (NER). This advanced approach not only enhances the precision and efficiency of identifying crucial entities within the labelled data, by using LDA model for fraudulent email detection, but also lays the foundation for more comprehensive data analysis. Next, we discuss relation extraction, offering a glimpse into its potential applications in cloud forensic analysis. As we wrap up our research, we contemplate the exciting prospects for this field, exploring future directions and emerging trends that could further refine and optimize forensic process in the cloud.

9.2 Data Deduplication of Cloud Forensic Evidence

The impact of deduplication methodology on reducing the cloud forensic backlog cannot be overstated. By systematically removing duplicates from the investigative queue, we have liberated valuable resources, enabling forensic analysts to focus their expertise and attention on truly unique and pertinent data. This streamlined approach translates into faster response times, more efficient case resolutions, and a substantial reduction in the backlog that once plagued the forensic workflow.

In essence, deduplication through hashing serves as the linchpin of our strategy, exemplifying how technology and innovation can be harnessed to enhance the effectiveness of cloud forensics. By unburdening investigators from the weight of redundant data, we have not only

expedited the process but also improved the overall quality and accuracy of our forensic analyses, setting the stage for a more agile and responsive approach to cloud-based investigations.

9.3 Relevant Data Extraction and Prioritization of Cloud Forensic Evidence

The detection and prioritization of fraudulent email data represent paramount challenges. Our research has leveraged the formidable Latent Dirichlet Allocation (LDA) model to address these challenges head-on. With its remarkable ability to uncover latent topics within a corpus of text, LDA has served as an indispensable tool in the identification of fraudulent email content. Through the application of LDA, we were able to find discern patterns, anomalies, and hidden connections within vast collections of email data. By examining the underlying themes and semantic structures of these communications, we've been able to flag suspicious content with a high degree of accuracy. This not only accelerates the identification of potential threats but also enables us to focus investigative efforts precisely where they are needed the most.

However, our research doesn't stop at detection alone. Equally critical is the prioritization of fraudulent data to alleviate the burden of cloud forensic backlog. With LDA's assistance, we've been able to assign priority levels to identified fraudulent emails based on the severity of their content and their potential impact on ongoing investigations. This intelligent prioritization system ensures that forensic experts can address the most critical cases first, significantly reducing the time and resources required to resolve cloud-related security incidents.

9.4 Data Analysis of Collected Evidence Using NER and RE

Our research has prominently featured Information Extraction as a pivotal component, and at its heart lies Named Entity Recognition (NER) powered by BERT. NER, a cornerstone of natural language processing, has revolutionized our ability to parse and identify crucial pieces of information within the vast sea of digital data. BERT, with its contextual understanding of language, has elevated the precision of NER to unprecedented levels. By deploying BERT-based NER models, we've been able to meticulously extract key entities such as names, dates, locations, and more from complex textual data, enhancing the granularity and organization of our forensic findings. By employing RE techniques, we're able to uncover the relationships and associations between entities, thereby reconstructing a more comprehensive and coherent narrative from fragmented digital evidence. This innovation holds immense promise for expediting forensic investigations, as it enables us to establish context and understand the flow of events more efficiently. For the BERT-based NER model, precision and recall for all

classes (GPE, ORG, PER) are very close to 1.00, indicating high accuracy and ability to capture most of the entities. In contrast, the rule-based and CRF-based model has lower precision and recall values, suggesting that it may have some false positives and false negatives, especially for the Persons and Locations classes.

9.5 Conclusion and Research Summary

In conclusion, for our extensive research which is aimed at alleviating the challenge of cloud forensic backlog through the strategic implementation of machine learning models. Our journey led us through various facets of this problem, and we employed cutting-edge techniques to address them. We leveraged hashing algorithms to effectively remove duplicate data, enhancing data deduplication processes. Furthermore, in our pursuit of detecting fraudulent emails, we harnessed the power of Latent Dirichlet Allocation (LDA) models to flag suspicious emails, subsequently prioritizing them to streamline the cloud forensic workflow in relevant data extraction and prioritization chapter. Then, Information extraction also played a pivotal role, where we harnessed BERT for Named Entity Recognition (NER), utilizing labelled data derived from the prioritization step. Lastly, we delved into the realm of relation extraction which we discussed in Name Entity Recognition Using BERT Model & Relation Extraction chapter, then setting the stage for future directions in our research. Our work not only offers a comprehensive approach to mitigating cloud forensic backlog but also opens doors to exciting prospects for further exploration in this critical domain.

9.6 Future Directions

The field of cloud forensic is continuously evolving, so staying updated with the latest developments and adapting your research accordingly will be essential for our success in the future. Some potential future directions for research based on our current work on reducing cloud forensic backlog using machine learning models:

- **Adaptive Prioritization:** Explore dynamic prioritization techniques that adapt to changing threat landscapes and forensic requirements. This could involve reinforcement learning or other adaptive algorithms for prioritization.
- **Streamlined Data Collection:** Consider ways to improve the efficiency of data collection and labelling processes. This could involve exploring semi-supervised or unsupervised learning techniques to reduce the reliance on labelled data.
- **Hybrid Models:** Investigate the feasibility of combining multiple machine learning models or techniques to create hybrid models. This could involve ensemble methods or integrating deep learning with traditional machine learning approaches.

- **Real-time Detection and Response:** Extend your research to focus on real-time detection and response mechanisms. Develop strategies for detecting and responding to fraudulent activities and security breaches as they occur, rather than waiting for a backlog.
- **Automation of Remediation:** Go beyond detection and focus on automating the remediation of identified issues. Develop methods for automatically mitigating threats and reducing the need for manual intervention.
- **Use Different Data:** To check the performance of current system, change data evidence to text it with same system to check compatibility.
- **Collaboration with Industry:** Collaborate with industry partners and organizations to validate and implement your research findings in real-world cloud forensic scenarios.

References

- [1] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, 'Systematic literature reviews in software engineering - A systematic literature review', *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.
- [2] S. Naaz and F. Ahmad, 'Comparitive Study of Cloud Forensics Tools', *Communications on Applied Electronics*, vol. 5, pp. 24–30, Jun. 2016, doi: 10.5120/cae2016652258.
- [3] B. Manral, G. Somani, K.-K. R. Choo, M. Conti, and M. S. Gaur, 'A Systematic Survey on Cloud Forensics Challenges, Solutions, and Future Directions', *ACM Comput. Surv.*, vol. 52, no. 6, p. 124:1-124:38, Nov. 2019, doi: 10.1145/3361216.
- [4] S. Simou, C. Kalloniatis, E. Kavakli, and S. Gritzalis, *Cloud Forensics: Identifying the Major Issues and Challenges*, vol. 8484. 2014, p. 284. doi: 10.1007/978-3-319-07881-6_19.
- [5] A. I. of McKemmish, 'What is forensic computing?', *Australian Institute of Criminology*, Jun. 30, 1999. <https://www.aic.gov.au/publications/tandi/tandi118> (accessed May 28, 2023).
- [6] K. Kent, S. Chevalier, T. Grance, and H. Dang, 'Guide to Integrating Forensic Techniques into Incident Response', National Institute of Standards and Technology, NIST Special Publication (SP) 800-86, Sep. 2006. doi: 10.6028/NIST.SP.800-86.
- [7] H. Guo, B. Jin, and T. Shang, 'Forensic investigations in Cloud environments', in *2012 International Conference on Computer Science and Information Processing (CSIP)*, Aug. 2012, pp. 248–251. doi: 10.1109/CSIP.2012.6308841.
- [8] G. Chen, Y. Du, P. Qin, and J. Du, 'Suggestions to digital forensics in Cloud computing ERA', in *2012 3rd IEEE International Conference on Network Infrastructure and Digital Content*, Sep. 2012, pp. 540–544. doi: 10.1109/ICNIDC.2012.6418812.
- [9] B. Martini and K.-K. R. Choo, 'An integrated conceptual digital forensic framework for cloud computing', *Digital Investigation*, vol. 9, no. 2, pp. 71–80, Nov. 2012, doi: 10.1016/j.diin.2012.07.001.
- [10] G. Sibiya, H. Venter, and T. Fogwill, 'Digital Forensic Framework for a Cloud Environment', May 2012. Accessed: May 31, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/Digital-Forensic-Framework-for-a-Cloud-Environment-Sibiya-Venter/104fcc8b4b9ddc4926d76bb85ce564adcaaf70b4>
- [11] K. Ruan and J. Carthy, 'Cloud Forensic Maturity Model', in *Digital Forensics and Cyber Crime*, M. Rogers and K. C. Seigfried-Spellar, Eds., in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Berlin, Heidelberg: Springer, 2013, pp. 22–41. doi: 10.1007/978-3-642-39891-9_2.
- [12] R. Adams, 'The Emergence of Cloud Storage and the Need for a New Digital Forensic Process Model', in *Cybercrime and Cloud Forensics: Applications for Investigation Processes*, 2012, p. pages 79-104. doi: 10.4018/978-1-4666-2662-1.
- [13] S. Saibharath and G. Geethakumari, 'Design and Implementation of a forensic framework for Cloud in OpenStack cloud platform', in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2014, pp. 645–650. doi: 10.1109/ICACCI.2014.6968451.
- [14] J. J. Shah and L. G. Malik, 'An approach towards digital forensic framework for cloud', in *2014 IEEE International Advance Computing Conference (IACC)*, Feb. 2014, pp. 798–801. doi: 10.1109/IAdCC.2014.6779425.
- [15] A. Pătrașcu and V.-V. Patriciu, 'Logging framework for cloud computing forensic environments', in *2014 10th International Conference on Communications (COMM)*, May 2014, pp. 1–4. doi: 10.1109/ICComm.2014.6866662.
- [16] S. Ahmad, N. L. Saad, Z. Zulkifli, and S. H. Nasaruddin, 'Proposed network forensic framework for analyzing IaaS cloud computing environment', in *2015 International Symposium on*

Mathematical Sciences and Computing Research (iSMSC), May 2015, pp. 144–149. doi: 10.1109/ISMSC.2015.7594043.

- [17] M. Banas, ‘Cloud Forensic Framework For IaaS With Support for Volatile Memory’, Sep. 2015. Accessed: May 31, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/Cloud-Forensic-Framework-For-IaaS-With-Support-for-Banas/5aa28a4c4ca23e58adaa347c02d21a278bd2739c>
- [18] S. Zawoad, R. Hasan, and A. Skjellum, ‘OCF: An Open Cloud Forensics Model for Reliable Digital Forensics’, in *2015 IEEE 8th International Conference on Cloud Computing*, Jun. 2015, pp. 437–444. doi: 10.1109/CLOUD.2015.65.
- [19] A. Faldu, ‘Authentication Framework in Forensic Science with Cloud Computing’, *International Journal on Advances in Engineering Technology and Science*, vol. 2, p. 7, Jan. 2016.
- [20] M. Faheem, D. Tahar, and D. An, ‘A Unified Forensic Framework for Data Identification and Collection in Mobile Cloud Social Network Applications’, *International Journal of Advanced Computer Science and Applications*, vol. 7, Jan. 2016, doi: 10.14569/IJACSA.2016.070103.
- [21] S. Datta, K. Majumder, and D. De, *DCF: A novel dynamic forensic framework towards cloud computing environment*. 2016, p. 764. doi: 10.1109/CCAA.2016.7813829.
- [22] M. Faheem, N.-A. Le-Khac, and T. Kechadi, ‘Toward a new mobile cloud forensic framework’, in *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, Aug. 2016, pp. 736–742. doi: 10.1109/INTECH.2016.7845142.
- [23] N. H. Ab Rahman, W. B. Glisson, Y. Yang, and K.-K. R. Choo, ‘Forensic-by-Design Framework for Cyber-Physical Cloud Systems’, *IEEE Cloud Computing*, vol. 3, no. 1, pp. 50–59, Jan. 2016, doi: 10.1109/MCC.2016.5.
- [24] S. Simou, C. Kalloniatis, S. Gritzalis, and H. Mouratidis, ‘A survey on cloud forensics challenges and solutions’, *Security and Communication Networks*, vol. 9, Nov. 2016, doi: 10.1002/sec.1688.
- [25] A. Alenezi, R. K. Hussein, R. J. Walters, and G. B. Wills, ‘A Framework for Cloud Forensic Readiness in Organizations’, in *2017 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud)*, Apr. 2017, pp. 199–204. doi: 10.1109/MobileCloud.2017.12.
- [26] V. R. KEBANDE, N. M. Karie, and H. S. Venter, ‘Cloud-Centric Framework for isolating Big data as forensic evidence from IoT infrastructures’, in *2017 1st International Conference on Next Generation Computing Applications (NextComp)*, Jul. 2017, pp. 54–60. doi: 10.1109/NEXTCOMP.2017.8016176.
- [27] M. N. Ahmed Khan and S. Ullah, ‘A log aggregation forensic analysis framework for cloud computing environments’, *Computer Fraud & Security*, vol. 2017, no. 7, pp. 11–16, Jul. 2017, doi: 10.1016/S1361-3723(17)30060-X.
- [28] P. Santra, P. Roy, D. Hazra, and P. Mahata, ‘Fuzzy Data Mining-Based Framework for Forensic Analysis and Evidence Generation in Cloud Environment’, in *Ambient Communications and Computer Systems*, G. M. Perez, S. Tiwari, M. C. Trivedi, and K. K. Mishra, Eds., in *Advances in Intelligent Systems and Computing*. Singapore: Springer, 2018, pp. 119–129. doi: 10.1007/978-981-10-7386-1_10.
- [29] S. S. Sampana, ‘FoRCE (Forensic Recovery of Cloud Evidence): A Digital Cloud Forensics Framework’, in *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)*, Jan. 2019, pp. 212–212. doi: 10.1109/ICGS3.2019.8688215.
- [30] Z. Umar and E. Emmanuel, ‘A Framework for Digital Forensic in Joint Heterogeneous Cloud Computing Environment’, *Journal of Future Internet*, vol. 3, pp. 1–11, Jun. 2019, doi: 10.18488/journal.102.2019.31.1.11.
- [31] D. Sudyana, N. Lizarti, and E. Erlin, ‘Forensic Investigation Framework on Server Side of Private Cloud Computing’, *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, p. 181, Dec. 2019, doi: 10.24843/LKJITI.2019.v10.i03.p06.
- [32] G. Pandi Jain and K. Wandra, ‘Secured Forensic Framework for Various Users in the Virtualized Environment of Cloud’, 2020, pp. 715–727. doi: 10.1007/978-981-13-7166-0_72.

- [33] S. Bhatia and J. Malhotra, 'CFRF: Cloud Forensic Readiness Framework – A Dependable Framework for Forensic Readiness in Cloud Computing Environment', in *Innovative Data Communication Technologies and Application*, J. S. Raj, A. Bashar, and S. R. J. Ramson, Eds., in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, 2020, pp. 765–775. doi: 10.1007/978-3-030-38040-3_88.
- [34] A. Razaque, M. Aloqaily, M. Almiani, Y. Jararweh, and G. Srivastava, 'Efficient and reliable forensics using intelligent edge computing', *Future Generation Computer Systems*, vol. 118, pp. 230–239, May 2021, doi: 10.1016/j.future.2021.01.012.
- [35] R. Rani and G. Geethakumari, 'A framework for the identification of suspicious packets to detect anti-forensic attacks in the cloud environment', *Peer-to-Peer Networking and Applications*, vol. 14, pp. 1–14, Jul. 2021, doi: 10.1007/s12083-020-00975-6.
- [36] M. Fadilla, B. Sugiantoro, and Y. Prayudi, 'Membangun Framework Konseptual Terintegrasi Menggunakan Metode Composite Logic untuk Cloud Forensic Readiness pada Organisasi', *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, pp. 144–153, Feb. 2022, doi: 10.30865/mib.v6i1.3427.
- [37] N. Kumari and A. K. Mohapatra, 'A Novel Framework For Multi Source Based Cloud Forensic', in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2022, pp. 1–7. doi: 10.1109/ICCMC53470.2022.9753849.
- [38] S. Simou, C. Kalloniatis, S. Gritzalis, V. Katos, and M. Psalidas, 'Revised forensic framework validation and cloud forensic readiness', *International Journal of Electronic Governance*, vol. 14, p. 236, Jan. 2022, doi: 10.1504/IJEG.2022.123254.
- [39] F. Ye, Y. Zheng, X. Fu, B. Luo, X. Du, and M. Guizani, 'TamForen : A tamper-proof cloud forensic framework', *Transactions on Emerging Telecommunications Technologies*, vol. 33, Apr. 2022, doi: 10.1002/ett.4178.
- [40] V. Prakash, A. Williams, L. Garg, P. Barik, and R. K. Dhanaraj, 'Cloud-Based Framework for Performing Digital Forensic Investigations', *Int J Wireless Inf Networks*, vol. 29, no. 4, pp. 419–441, Dec. 2022, doi: 10.1007/s10776-022-00560-z.
- [41] P. Mell and T. Grance, 'The NIST Definition of Cloud Computing', National Institute of Standards and Technology, NIST Special Publication (SP) 800-145, Sep. 2011. doi: 10.6028/NIST.SP.800-145.
- [42] S. Simou, C. Kalloniatis, H. Mouratidis, and S. Gritzalis, 'Towards the Development of a Cloud Forensics Methodology: A Conceptual Model', in *Advanced Information Systems Engineering Workshops*, A. Persson and J. Stirna, Eds., in Lecture Notes in Business Information Processing. Cham: Springer International Publishing, 2015, pp. 470–481. doi: 10.1007/978-3-319-19243-7_43.
- [43] S. Simou, C. Kalloniatis, H. Mouratidis, and S. Gritzalis, *A Meta-model for Assisting a Cloud Forensics Process*, vol. 9572. 2015. doi: 10.1007/978-3-319-31811-0_11.
- [44] D. Lillis, B. Becker, T. O'Sullivan, and M. Scanlon, *Current Challenges and Future Research Areas for Digital Forensic Investigation*. 2016. doi: 10.13140/RG.2.2.34898.76489.
- [45] G. Palmer, 'A road map for digital forensic research', in *First digital forensic research workshop, utica, new york*, 2001, pp. 27–30.
- [46] I. Orton, A. Alva, and B. Endicott-Popovsky, 'Legal process and requirements for cloud forensic investigations', in *Cybercrime and Cloud Forensics: Applications for Investigation Processes*, IGI Global, 2013, pp. 186–229.
- [47] G. Grispos, T. Storer, and W. B. Glisson, 'Calm before the storm: The challenges of cloud computing in digital forensics', *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 4, no. 2, pp. 28–48, 2012.
- [48] M. Rafique and M. N. A. Khan, 'Exploring static and live digital forensics: Methods, practices and tools', *International Journal of Scientific & Engineering Research*, vol. 4, no. 10, pp. 1048–1056, 2013.
- [49] S. Rahman and M. N. A. Khan, 'Review of live forensic analysis techniques', *International Journal of Hybrid Information Technology*, vol. 8, no. 2, pp. 379–88, 2015.

- [50] S. Almulla, Y. Iraqi, and A. Jones, 'A state-of-the-art review of cloud forensics', *Journal of Digital Forensics, Security and Law*, vol. 9, no. 4, p. 2, 2014.
- [51] 'Cloud Adoption and Risk Report 2016 Q4', Oct. 22, 2018. https://library.cyentia.com/report/report_002130.html (accessed Sep. 13, 2023).
- [52] U. S. G. A. Office, 'Cybersecurity | U.S. GAO', Dec. 21, 2021. <https://www.gao.gov/cybersecurity> (accessed Sep. 13, 2023).
- [53] K. Ruan, J. Carthy, T. Kechadi, and M. Crosbie, 'Cloud Forensics', in *Advances in Digital Forensics VII*, G. Peterson and S. Shenoi, Eds., in IFIP Advances in Information and Communication Technology. Berlin, Heidelberg: Springer, 2011, pp. 35–46. doi: 10.1007/978-3-642-24212-0_3.
- [54] K. Ruan, J. James, J. Carthy, and T. Kechadi, 'Key terms for service level agreements to support cloud forensics', in *Advances in Digital Forensics VIII: 8th IFIP WG 11.9 International Conference on Digital Forensics, Pretoria, South Africa, January 3-5, 2012, Revised Selected Papers 8*, Springer, 2012, pp. 201–212.
- [55] N. C. C. F. S. W. Group, 'Nist cloud computing forensic science challenges', National Institute of Standards and Technology, 2014.