DATA VISUALIZATION OVER ENCRYPTED DATABASES

By

Kalim Ullah

Submitted to the Faculty of Department of Information Security
Military College of Signals, National University of Sciences and Technology,
Islamabad in partial fulfillment of the requirements for the degree of MS in
Information Security

NOVEMBER 2022

Data Visualization over Encrypted Databases

Author
Kalim Ullah
Regn Number
00000359358

A thesis submitted in partial fulfillment of the requirements for the degree of
MS Information Security

Thesis Supervisor:
Shahzaib Tahir, PhD

Thesis Supervisor's Signature: _____

Department Of Information Security
Military College of Signals
National University of Sciences and Technology,
Islamabad
November 2022

CERTIFICATE OF CORRECTNESS AND APPROVAL

It is certified that work contained in this thesis “Data Visualization over Encrypted Databases”, was carried out by Kalim Ullah under the supervision of Dr. Shahzaib Tahir, for partial fulfilment of Degree of Master of Information Security, is correct and approved. This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Approved by

_____

(Shahzaib Tahir, PhD)

Thesis Supervisor

Military College of Signals (MCS)

Dated: _____Nov 2022

DECLARATION

I certify that this research work titled “Data Visualization over Encrypted Database” is my own work. No portion of this work presented in this dissertation has been submitted in support of another award or qualification either at this institution or elsewhere. The material that has been used from other sources has been properly acknowledged / referred.

_____

Signature of Student

Kalim Ullah

00000359358

This page is left intentionally blank.

ABSTRACT

With the progression of IoT, the concept of big data and artificial intelligence became a focus of great interest. However, big data security and analysis is becoming a challenging task for the researcher. Similarly maintaining big data confidentiality and privacy over unsecure media is also a daunting assignment especially where data privacy is of utmost importance like cellular companies customer phone call data and medical records of patients etc. “Data Visualization over encrypted data” is a possible solution for preserving the privacy & confidentiality of data and even data analysis through visualization tool can be performed at the same time. In this research we are going to suggest a framework followed by practical implementation of a case scenario where data is secured over untrusted media like public cloud and implement visualization tools for easy analyzation of data without shifting to trusted domain. In this solution computation and analysis for visualization would be carried out over untrusted cloud platform on hashed and encrypted database and data decryption be carried out in trusted domain for only required data. This solution is capable to preserve the confidentiality and privacy of data over the cloud.

Key Words: Data Visualization, Data Confidentiality, Searchable Encryption

COPYRIGHT STATEMENT

DEDICATION

“The example of those who take allies other than Allah is like that of the spider who takes a home. And indeed, the weakest of homes is the home of the spider; if they only knew. Indeed, Allah knows whatever thing they call upon; other than Him. And He is the Exalted in Might; the Wise.

And these examples, We present to the people, but none will understand them except those of knowledge.**”**

(Chapter 20: Surah Al-'Ankabut: Ayat 41-43)

Dedicated to my beloved country Pakistan.

ACKNOWLEDGEMENT

I am grateful to Allah Almighty for giving me strength to keep going on with this thesis, irrespective of many challenges and troubles. All praises for HIM and HIM alone.

Next, I am grateful to all my family and especially to my parents. Without their consistent support and prayers, this thesis would not have been possible.

I am very grateful to my Project Supervisor Dr. Shahzaib Tahir who supervised the thesis / research in a very encouraging and helpful manner. I am also grateful to my Co-Supervisor Brig Dr. Imran Rashid who took me under his wing. As supervisor and co-supervisor, their support and supervisions have always been a valuable resource for me.

I am also thankful to committee members who have always guided me with their profound and valuable support that has helped me in achieving my research aims.

I would also thank my friends Dr. Farhan Ullah and Mr. Hamaad Saeed who helped me during challenging task of coding and implementation.

Finally, I would like to express my appreciation to all the people who have provided valuable support to my study and whose names I couldn’t bring to memory.

This page is left intentionally blank.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

ACRONYMS

Advanced Encryption Standard	AES
Application Programing Interface	API
Amazon Web Services	AWS
Brakerski-Gentry-Vaikunathan	BGV
Confidentiality, Integrity and Availability	CIA
Computerized National Identity Card	CNIC
Central Processing Unit	CPU
Cloud Service Providers	CSPs
Call Data Records	CDR
Data Base	DB
Dijk-Gentry-Halevi-Vaikuntanatha	DGHV
Data Base Management System	DBMS
General Data Protection Regulation	GDPR
Google Compute Engine	GCE
Global System for Mobile communication	GSM
Gentry-Sahai-Waters, 2013	GSW13
Health Insurance Portability and Accountability Act	HIPPA
Hash Function	H
Infrastructure as a Service	IaaS
Internet of Thing	IoT
Identity	ID
Mobile Switching Center	MSC
Message Digest	MD
National Institute of Standards and Technology	NIST
Number Theory Research Unit	NTRU
Operating System	OS
Platform as a Service	PaaS
Personal Identification Number	PIN
RIPE Message Digest	RIPEMD
Personal Data Protection Bill	PDPB
Provable Data Possession	PDP
Pseudorandom function	PRF
Pseudo Random Number Generator	PRNG

Proof of Retrievability	PoR
Software as a Service	SaaS
Service Level Agreement	SLA
Secure Socket Layer	SSL
Storage as a Service	STaaS
Subscriber Identity Module	SIM
Structured Query Language	SQL
Secure Hashing Algorithm	SHA
Virtual Machine	VM
Virtual Private Network	VPN

Chapter 1

INTRODUCTION

"It's hard to beat a person who never gives up"

-Babe Ruth

1.1 Overview

Exponential growth and advancement in IoT have changed Human life. Wi-Fi Alliance [1] forecasts by 2022 over 400 M wifi devices be serving smart devices which will be carrying half the data traffic of globe which is a proof towards IoT growth rate. For management of big data produced from IoT devices are normally outsourced to third party like cloud which eliminates the problem of largescale data administration. Massive volume of data, as patients health records, contacts, e-mail communication records, IoT data and much more are outsourced.

The problem arises when sensitive and private nature data are transferred to curious / untrusted CSPs servers which is ultimate a privacy concern of customers. To address data privacy issues, sensitive data are normally encrypted before outsourcing to third party like cloud, which preserves confidentiality of the data and privacy of customers as well. However, for analytical purpose the data are normally downloaded to trusted domain to perform decryption and run analytical algorithm to extract required results. Similarly, big data analysis on encrypted data stored on untrusted domain is also a challenging task.

Pakistan's federal government has approved Cloud First Policy [2] and Personal Data Protection Bill in February 2022 [3]. The cloud policy has five classes of data. Open data, public data, Restricted data, and Sensitive / Confidential data includes information not to be published and can be accessed by certain peoples having authorization of access like phone number, registration numbers, passport details, etc. 5th category is Secret data, which are the "information requiring the highest level of protection from serious threats, whose breach will likely cause threats to life or public security, financial losses, serious damage to public interests". After the approval of the law, data centres of the federal government's departments and ministries' will be transferred to the central cloud. This change will be a great help to lower government spending, improve data security, and boost the effectiveness of the government department's online operations and services. However, these bills [2][3] also enforces us for safe and secure implementation mechanisms be adopted for data security on the cloud.

In this research we are going to propose an encryption scheme for preservation of confidentiality and privacy before outsourcing the data to third party like cloud. In second phase, data visualization technique shall be applied over the privacy preserved data and extract required data through polygenetic trees. In third phase only required data should be shifted to trusted domain for decryption. This solution will enable us to encrypt data for untrusted domain, search mechanism over encrypted database and provides easy analysis through visualization over encrypted data.

1.2 Motivation

The global end-user spending on public cloud is expected to expand at a rate of 20.4% in 2022 by $494.7 billion and by $600 billion in 2023 [4]. As per Cybersecurity Insiders [5] cloud security report 2022, today, 39% of respondents have more than half of their workloads on the cloud, and 58% intend to do so in the next 12 to 18 months. Despite increasing competition and demand, the cloud technology is becoming more complex, and security is becoming a top concern.

As per Thales Cloud Security Report 2022 [6], 45% of businesses have experienced a cloud-based data breach or failed audit in the past 12 months, which is increased by 5% from the previous year, Such issues prevent people from benefiting from resource sharing and prohibit them from externalizing their private and confidential data over the cloud, raising superior concerns about the protection of sensitive data. Enterprises and individuals are motivated to use cloud services due to multiple benefits associated with the usage of services provided by CSPs. There are some associated limitations as well. Like a semi-trusted or curious CSP can make use of users' personal data by sharing the identity, profile information, etc to third party vendors for their business analytics purposes. This confines the adoption and usage of cloud services for sensitive data.

To deals with such a problem, encryption of data locally and then outsources to cloud is a possible solution. However, if we required data for use or analysis then, encrypted data would be shifted to trusted domain every time for decryption and analysis. The problem necessities us to device a mechanism which resolve the challenge of search over encrypted data and possibility of implementing visualization technique for easy data analysis as well.

1.3 Problem Statement

In the present world, no one is sure about the fact that the data which is outsourced to public cloud is secured or not. Over the years many efforts have been devoted for securing the

data by encryption. However, searching and data visualization over encrypted databases without compromising confidentiality and privacy for data analytics is a demanding requirement. There is an important requirement of devising a mechanism where data can be kept secure and at the same time data can be searched and visualize over encrypted databases for analysis purpose.

1.4 Research Objectives

The main objectives of thesis are:

a. Suggest framework for data visualization over encrypted databases for preservation of data confidentiality and privacy.
b. Imply search methodology over encrypted data in untrusted domain.
c. Data visualization over encrypted data in untrusted domain for data analytics.
d. Implementation of concept on available sensitive dataset for proof of concept and result validation.

1.5 Scope

There are two main forms of data termed as structured and unstructured data. Structure data are normally stored in tabular format like excel sheet and SQL databases. Whereas unstructured data are stored as NoSQL database and media files etc. The scope of this study is confined to the structure data only.

1.6 Contribution

The need of such survey is multifaceted. This thesis will contribute in the following ways:

a. Due to weak security mechanisms of data security, private and sensitive data are available on social media especially in undeveloped /developing countries for sale thus, there is a dire need of implementation of strong data security mechanism.
b. Small organizations who deal with sensitive data but cannot afford private cloud can outsourced data to public cloud with more protection.
c. Big data analysis of sensitive data is not carried out due to the confidentiality and privacy concerns, however same data can be used for data analytics by researchers without compromising the privacy concerns.
d. A huge number of the naive and uneducated users fall prey to scammers every year, lose money as well as valuable data and are subjected to blackmailing and extortion due to weak data security mechanism over the cloud.

e. Public clouds are defiantly considered curious if not secure, and users are normally reluctant to use it for confidential and sensitive data. However, this solution will provide security to acceptable extent.

1.7 Thesis Outline

This research work is comprised of six chapters:

Chapter 1: Introduction is given, including the motivation of research, problem statement, research objectives, scope, and contributions.

Chapter 2: This outlines basic preliminaries of cryptography, databases, data visualization and literature review of the subject.

Chapter 3: This chapters cover the cloud data threat modelling which encompasses cloud deployment models, cloud security issues and cloud data security goals.

Chapter 4: This presents the proposed framework of data visualization over encrypted database.

Chapter 5: Physical implementation of proposed framework for case scenario of call data record security and visualization followed by leakage profiling.

Chapter 6: Conclusion and open areas for research are highlighted.

Chapter 2

PRELIMINARIES

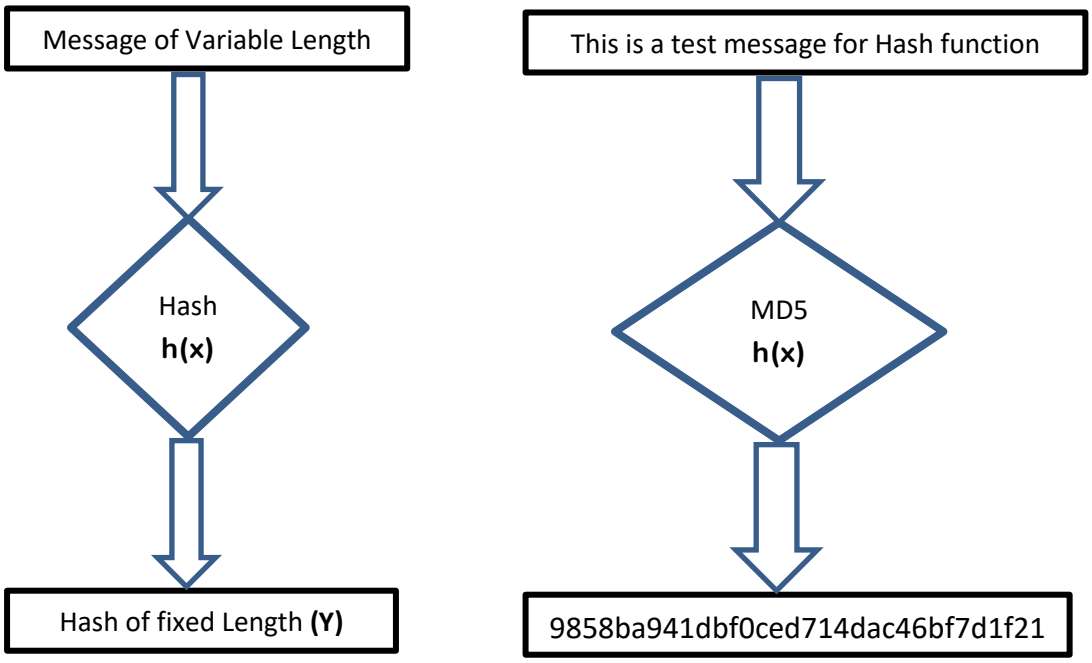
Before exploring the solution of structured data security, some basic concept should be revied for better understanding.

2.1 Cryptographic Background

2.1.1 Hash functions "H"

A variable number input of numeric data is provided to the hash function, which transforms it to a fixed length value of numeric data output. The values returned by these functions are called hashes or message digest [7]. Figure 2.1 represents a conceptual overview of the hash function.

Figure 2.1 Concept of Hash Function



2.1.1.1 Features of Hash functions

The following are some important features of the hash function: -

a. Variable length input of hash function produces fixed length output.
b. Hash functions are computationally much faster than a symmetric encryption function.

2.1.1.2 Properties of Hash functions

a. **Pre-Image Resistance.** This property implies that if a hash function H produces a digest **Y** from input **X,** then it should be computationally difficult to reverse a hash function. Equation 2.1 presents pre-image resistance property.

$$\left.\begin{aligned} H(X) &\longrightarrow Y \\ Y &\nrightarrow H(X) \end{aligned}\right\} \quad \text{(Equation 2.1)}$$

b. **Second Pre-Image Resistance**. This property implies that if an input X is passed into hash function H, it produces an output, then it should be hard to obtain another input Y, which is passed into same hash function H and produce same value. Hash function is collision resistant. Equation 2.2 presents second pre-image resistance property

$$\left.\begin{aligned} H(X) &\neq H(Y) \\ X &\neq Y \end{aligned}\right\} \text{(Equation 2.2)}$$

2.1.1.3 Popular Hash functions

a. **Message Digest (MD)**. The message-digest MD5 algorithm is among one of the most popular hash functions in message digest family. It produces 128-bit hash value which is cryptographically broken but still widely used. Its other variants are MD2, MD4, MD5 and MD6.

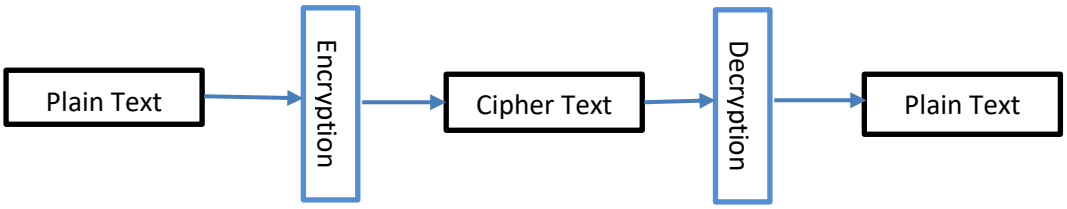
b. **Secure Hash Function (SHA)**. This algorithm is again a popular cryptographic function. The hash algorithm consists of modular additions, bitwise operations, and compression functions. Its variants are “SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512”.

c. Other Hash Functions are RIPEMD and its different variants, Whirlpool is derived from modified version of AES.

2.1.2 Cryptography

Cryptography is the knowledge of encryption and decryption used for securing data from unauthorized access. As obvious from the name, Crypto means hidden, and graph means writing which cumulatively means the security of data from fraud and unauthorized access [8]. Figure 2.2 gives pictural view of cryptography.

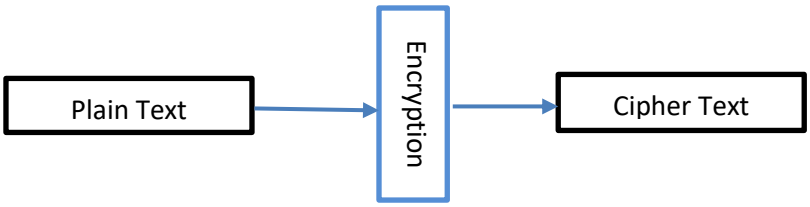
Figure 2.2 Cryptography



2.1.2.1 Encryption

Encryption is a process of securing data by converting it mathematically such that it cannot be translated by the unauthorized person. It can also be described as the conversion of data from readable (plaintext) format to a form which cannot be easily understood (ciphertext) [9] as shown in Figure 2.3.

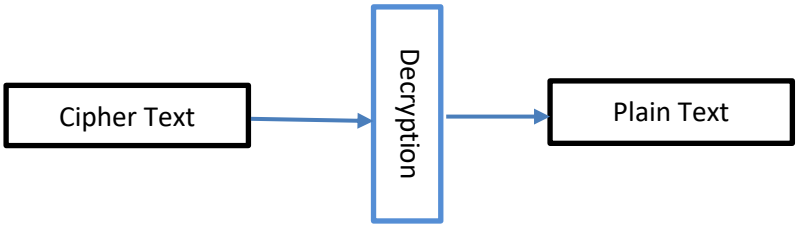
Figure 2.3 Encryption



2.1.2.2 Decryption

Decryption is a procedure of decoding secure data mathematically such that it can be brough in original readable format as shown in Figure 2.4. It can also be described as the conversion of encrypted data (ciphertext) back in original (plaintext) format [9].

Figure 2.4 Decryption



2.1.2.3 Types of Encryptions

Encryption can be divided into two main categories i.e symmetric and asymmetric encryption. The only difference between symmetric and asymmetric encryption is the use of keys. In symmetric encryption, one key is utilized for both encryption and decryption. However, in asymmetric encryption two distinct keys are employed, one for encryption and other for decryption then it is called asymmetric or public key encryption. Symmetric key encryption is computationally faster than Asymmetric encryption [10].

2.1.3 Random Number Generator

RNG is a function or hardware that produces sequence of numbers and, as the name suggests the generated number cannot be predicted. There are two distinct types of random number generators [11].

2.1.3.1 Pseudo-random Number Generator (PRNGs)

PRNGs are also called deterministic random bit generator (DRBG). It does not generate sequence of truly random number because it is completely depending on initial value to the function also known as seed value. Such generators are used by different applications like games, simulations and in cryptography as well.

2.1.3.2 True-random Number Generator (TRNGs)

TRNGs are device generators that produce true random numbers. These types of generators uses physical process, quantum phenomenon, bioelectrical and physical signals [12] such as thermal noise, the photoelectric effect, An electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG) and electrooculogram (EOG)" etc are not to be confused with Pseudo Random Generators. The later generates single random output for random input. The Pseudorandom function (PRF) generates random outputs regardless of the input provided.

2.2 Database

Database is an organized collection of data, with a sole purpose of allowing data to be easily accessed, manipulated, and updated in systematic way [13][14] with different requirements. The requirements may vary like security, availability, confidentiality, performance etc. The purpose of database is easy management of data as per user requirement. There are different types of databases used in market some are following.

a. Centralized database
b. Distributed database

c. Personal database
d. End-user database
e. NoSQL database
f. Relational database
g. Cloud database
h. Object-oriented database
i. Graph database

2.2.1 Database Security

Database security refers to various protective measures required to ensure the security of data from various internal and external threats [15]. There are three layers of database security and at each layer different security solution are applied. At Database level masking, tokenization and encryption are used, at Access level, permissions and access control list are use and at Perimeter level, firewall and VPNs are used for ensuring database security. Furthermore, database security includes followings.

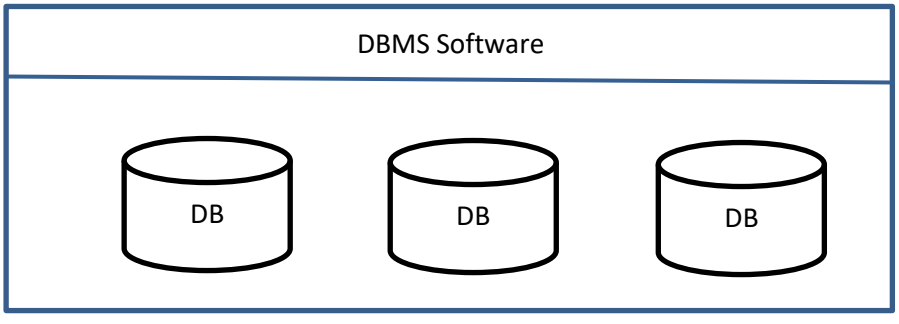
a. Protection of Database
b. Data contains in Database
c. Database Management System
d. Applications that access Data

2.2.2 Difference of DB and DBMS

Some time we confuse database (DB) with data base management system (DBMS). We normally describe SQL Server, DynamoDB, Oracle, MySQL as databases. But they are DBMS which is a software and used for management of different databases [16]. Figure 2.5 illustrates the difference of DB and DBMS. There are different DBMS and can be categorized as followings.

a. Relational Database Management Systems
b. Hierarchical Database Systems
c. Network Database Systems
d. Object-Oriented Database Systems
e. NoSQL Database Systems

Figure 2.5 Difference of DB and DBMS



2.3 Data visualization

Data visualization is a powerful technique of explaining data into a visual context, for example converting data into graph, charts, polygenetic trees etc. Visualization offers easy approach for human brain in summarize the data in minimal timeframe as compared to other methods [17]. The aim of data visualization techniques is to identify patterns, trends, and outliers in large data sets in minimum time. Visualization is a necessary part of every profession. It starts from the day a child opens his eyes, uses his/her five senses for visualizing his / her world experience for mind grooming. And this practice is the primary factor of human mind to easily understands the visualization instead of plain data. Modern visualization technique offers the ability to absorb information quickly, improve insights and helps in decisions making faster. Some of visualization techniques commonly used in data analytics are following.

a. Network Diagram
b. Polygenetic trees
c. Pie Chart
d. Bar Chart
e. Gantt Chart
f. Heat Map
g. Scatter Plot
h. Pictogram Chart
i. Timeline
j. Choropleth Map
k. Correlation Matrices

2.4 Cloud Computing

National Institute of Standards and Technology (NIST) refer cloud computing as "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" [18]. As per the NIST definition there the three main service models.

a. Software as a Service (SaaS)
b. Platform as a Service (PaaS)

c. Infrastructure as a Service (IaaS)

2.4.1 Software as a Service (SaaS)

In SaaS model the applications are delivered over the internet. SaaS provide the facilities to access the application over the internet, shifting the burden of installation and maintenance of applications from the users to cloud. In this model customer does not control or maintain the infrastructure where the application is installed. User / customers can access and use the services through thin client user interface like web browser or program interface. The maintenance of infrastructure, operating system (OS) and updates is the sole responsibility of the CSPs. Examples [19] of SaaS are Zoom, Netflix, Salesforce, Google Applications (G Suite), Adobe Creative Cloud, HubSpot.

2.4.2 Platform as a Service (PaaS)

In PaaS model the customers are offered a platform where the application can be installed and manged by the customer however, maintenance of the underlying infrastructure like server. OS, network storage etc are managed and maintained by the CSPs. PaaS provides clients /customers a complete development and deployment environment over the cloud. Some examples [20] of PaaS are AWS Elastic Beanstalk, Windows Azure, Heroku, Force.com, Google App Engine, Apache Stratos, OpenShift.

2.4.3 Infrastructure as a Service (IaaS)

Asper the NIST [18] IaaS provides the capabilities of storing, processing, networks and other fundamental computing resources. Hardware and computational power are provided to customer to use it as per the requirements. In this model the customer does not maintain and control physical infrastructure of cloud however, OS, storage, hosted application and some network resources are in user control. Some of the examples [21] of IaaS are Amazon Web Services (AWS) Elastic Compute Cloud (EC2), Amazon S3, Amazon VPC, Microsoft Azure, Google Compute Engine (GCE).

2.5 Cloud Deployment Model

To understand cloud threat spectrum, it is essential to appreciate and comprehend cloud models first. As per the NIST definition of cloud [18] there are four deployment models. Threat continuum of data varies with cloud deployment model.

2.5.1 Private Cloud

Single organization have a dedicated cloud environment for their customers. It is normally designed according to the needs of organization and can be set up on or off premises.

It can also be termed as corporate or internal cloud. Private cloud provides a greater security & privacy and have full control over the data.

2.5.2 Community Cloud

Community cloud can be described as a form of hybrid private cloud. Community cloud is shared by limited organization having common interest such as mission, policy etc.

2.5.3 Public Cloud

In public cloud, resources are owned and offered by CSPs through internet for usage. Resources like hardware, software and infrastructure are shared amongst the people. Customers adopt the model of paying as they use the resources.

2.5.4 Hybrid Cloud

Hybrid cloud is a combination of private, public or community cloud with same infrastructure through some standard technology.

2.6 Literature Review

Data security always remains challenging task in today's era where high reliance on internet communication increases the risk of data privacy. Novel mechanisms for data privacy are a vast field where tons of research have already been conducted. The researchers have identified several risks associated with data specifically structure data. There is numerous research work available on data encryption and data visualization in the literature, however there is a gap in implementation of both in one scenario.

Pal et al. [22] has presented the smart city concept like smart health care system, smart homes etc, produce big data. The author represented a 5-dimentional taxonomy of Big-data and proposed a framework for big data. However, security and secrecy, data integration and data analytics required to gain useful insights is the future challenges which required more attentions.

Zhang et al. [23] provides an overview of relevant visualization techniques and visual analysis systems regarding network anomaly data. Visualization addresses the problems arising from large-scale and heterogeneous data. however, real time data analysis is lacking as mostly existing tools provides visualization for offline data analysis. Visualization provides an overview but do not provide exact picture.

Raghav et al. [24] presented an overview of visualization tools with its efficacy for use. we achieve following from data visualization: -

a. Improvement in decision making.
b. Increase in return on investment.
c. Time Saving
d. Real-time Data Analysis

The author has focused only on static data visualization techniques for business and dynamic data visualization techniques are missing. Moreover, direct analysis without prepossessing for anomalies of data through visualization can miss lead the decision makers.

Yang et al. [25] explains live problem area with examples like Identify taxi communities waiting areas for passengers and Traffic congestion exploration. An important aspect facilitating exploration-driven trajectory analysis is interactive speed and amazing visualisation ideas. The Visualization-Assisted Interactive Big Urban Trajectory Data Exploration (Vaite) system has not yet been compared to any existing systems.

Important feature is interactive speed and excellent visualization concepts to support exploration-driven trajectory analysis. However, comparison of the Vaite system with existing system is lacking.

Shahzaib et al. [26] have taken the database of research journal as a use case and used it for the purpose of depicting data and their relationships in the form of phylogenetic tree and mesh. The idea can be used for the analysis of big data stored on unsecured third-party service provider like cloud etc.

Alhanjouri et al. [27] has proposed a new layer over the encrypted data for querying in database with the intention that on query over encrypted data will decrease performance. To overcome he used hashed map table for individual identity as a primary key and associated the data to hashed primary key. However, this method can lead towards confidentiality issues as anyone can calculate hashed for some name and search the database.

Matthew N. O. Sadiku [28] has presented that data visualization through graphical form creates the data simple in understanding being an innovative method for presenting large and complex data set. The author has discussed few Common visualization techniques like line graph, bar chats pie charts etc followed by its applications in real word use cases like health, fraud detection etc. However, data visualization has some challenges like large and time-varying datasets pose great challenge of users proactively respond issues at right time. Similarly, development challenge of big data visualization based on speed, size, and diversity of the data.

B. R. Kandukuri et al. [29], has discussed the security concerns from centralized (client-server based) to distributed systems like cloud. The author has the opinion that, Service Level Agreements (SLA) is the only way to keep check on CSPs, however there is no standard SLAs. Present SLAs provides waivers to customers but doesn't cover the original coast of lost data. For data security, customer required to trust CSPs and provide controlled physical and logical access to the sites and information. As per the author perspective the ultimate responsibility of data security and integrity lies on the customer even it is held with the CSPs. Investigation in cloud is a challenging task due to logging and recording of data for multiple users may be co-positioned or scattered across multiple hosts. The author has discussed the present SLAs contents and suggested a way forward to standardize the SLAs. Furthermore, the author has also suggested security at following access levels.

a. Program Security
b. Database security
c. Internet security
d. Server security
e. Data privacy security

Kacha et al. [30] considers data security in cloud is more challenging task as compared to the traditional infrastructure due to three main issues. Data is stored on CSPs environment, same physical infrastructure is shared by different users and accessibility of data by internet. The author has classified the cloud issues in three categories, according to the cloud types and characteristics, data security attributes and data life cycle. As compared to traditional infrastructure, cloud have numerous security concerns due to leased, open, shared, elastic, virtualized and distributed infrastructure. According to the cloud data life cycle the author has classified the security issues in data-in-rest, data-in-transit and data-in-use. The CIA trades have been debated according to the data security attributes. The writer has tabulated the common solutions for data security according to the mention characteristics of cloud at the end with conclusion that there isn't a complete technical solution for cloud security rather hinges on several attributes.

With the enhancement of applications for cloud data storage and computation has increased the tendency of improving the data security trends. Data availability and computational power are the drawing factors for migration to cloud, however the data is at extreme risk if not protected in a correct pattern. In this paper the authors have discussed the threats and risk related to the data in virtualization, especially threats associated with the hypervisors used for cloud environment. Similarly, security challenges, threats related to

multitenancy and public cloud have been discussed by the authors. In the end the writers have presented an overview of block cypher, stream cypher and hashing used for encryption to be used for security of data during rest and transit [31].

Soofi et al. [32] have carried out research on data security in cloud, based on two questions "What approaches have been introduced to ensure data security in cloud computing?" and "How the approaches have been validated?". The time-period of research was restricted from 2007 to 2014 and IEEE Xplore, Google scholar, science direct, portal digital library, Scopus, ACM, IJSI, IJERA were used with keywords Cloud computing, data security, security/data, concealment and data storage. Total 31 papers were selected out of which 14 (45%) used encryption, 6 (21%) applied guidelines and 5 (16%) utilized Framework for data security as an answer of first question. The outcomes shows that bulk of research are focused on encryption and out of encryption techniques, 71% results are proven out of which 67% use experimentation. The results also showed that 42% of papers gives no validation of outcomes, which mostly includes guidelines.

Cloud data security is forthcoming challenging task. Dependency on cloud services increases due to its obvious advantages of availability / accessibility, performance, low cost and numerous other benefits. The *Sood et al.* [33] has labelled the attackers as CSPs herself and external hackers. The proposed framework has been designed to fulfil the security of data during rest and transit. To protect the crucial data from unauthorized attackers, the author has suggested multiple techniques and divided the proposed framework into two phases. First phase covers the process of transmission of data and then securely storing over the cloud. The phase includes classification of data based on CIA trads using a proposed algorithm followed by index building and encryption before transmission. Before transmission Message Authentication Code (MAC) is generated and associated along with the encrypted data to ensure data is not altered and integrity remain intact. The second phase cover the retrieval of data from cloud based on the sensitivity and access rights of data which includes digital signature and key words in sensitive data cases. The proposed framework vanquishing many problems like tempering (integrity) of data, data leakage and unauthorized access from CSPs and achieves the availability and integrity of data during data transmission from owner to CSPs and CSPs to user.

Basu et al. [34] has presented a paper which has covered up the security gaps in the cloud environment in general. The authors have presented an overview of the cloud services and cloud deployment models followed by cloud security issues in details. The paper has

segregated the loopholes of cloud environment as per CIA trads to understand the security landscape of cloud virtualized environment. After a wholesome view of cloud computing security challenges, the paper has drawn a comparison of latest scheme in cloud data confidentiality, cloud virtualization confidentiality, cloud data integrity and cloud virtualization integrity. The author is in the opinion that ultimate responsibility of privacy and security of data maintained by CSPs, lies on the organization himself. Therefore, proper legislative protection be acquired in SLAs.

This paper has categories the homomorphic encryption algorithm in two categories, single homomorphic and fully homomorphic algorithms. In single homomorphic encryption algorithms *Geng et al.* [35] has drawn a comparison of Hill, RSA, ElGamal and Paillier algorithms. According to the mathematical problem Paillier algorithm has highest security and RSA & ElGamal algorithms have medium security whereas Hill algorithm has lowest security which is based on linear transformation matrix. According to encryption and decryption time, efficiency of Hill algorithm is highest whereas ElGamal, RSA and Paillier efficiency decreases respectively as per the output of experiment.

Geng et al. [35] has also reviewed four fully homomorphic encryption algorithms namely DGHV, BGV, GSW13 and NTRU. These algorithms performance comparison has been drawn according to mathematical problem, construction method, circuit calculation complexity and safety index. Vital applications of homomorphic encryption in present cloud computing are retravel of encrypted data, encrypted data processing, secure multi party computation and privacy of store data on cloud.

A summary of literature review discussed above is presented in a tabulated form in Table 2.1.

2.7 Summary

In this chapter basic idea of cryptography including properties of hash functions, encryption & decryption and random number generator have been discussed in context of data security. Database and visualization techniques are overviewed in context of cloud environment followed by cloud computing services. In the last a comprehensive literature review has been presented covering data security over the cloud environment and associated risks. Data visualization techniques are considered as powerful tools for data analytic.

Table 2.1 Summary - Literature review based on Visualization, Data, Security and Cloud

		Areas Covered	Pros	Cons

Research Paper	Topic discussed	Visualization	Data	Security Aspects	Cloud		
Pal et al. [22] "Big Data in Smart-Cities: Current Research and Challenges"	Big-Data characteristics, overview of analytical platform, framework, and applications of smart cities	Yes	Yes	No	No	• Overview of big data analytical platforms • Big data framework for smart cities • Big data taxonomy of smart cities according to following characteristics of data ➢ Computing infrastructure ➢ Storage infrastructure ➢ Data variety ➢ Data-analytics ➢ Data visualization	• Data integration is a challenging task due to wide variety of sensor devices. • Security and Privacy concerns direct to Internet connected smart devices. • Extracting precise information from the huge pool of data is the key challenge.
Zhang et al. [23] "A survey of network anomaly visualization"	Network anomaly data and its properties, visualization tasks process, applications of visualization and Network alert visualization	Yes	Yes	No	No	• Visualization addresses the problems arising from large-scale and heterogeneous data. • Paper provides an overview of relevant visualization techniques and visual analysis systems regarding network data anomalies. • Interactive visual tools integrate human perception and knowledge into the analysis process of network security.	• Real time data analysis is lacking as mostly existing tools provides visualization for offline data analysis. • In this paper the author introduced the pre-processing phase for reduction of anomalies which increase more accurateness. • Visualization gives us general

							overview but not the exact numerical pictures.
Raghav et al. [24] "A Survey of Data Visualization Tools for Analysing Large Volume of Data in Big Data Platform"	Data analytics techniques, usage of data visualization, features of big data visualization & associated errors, techniques of visualizations & presently used visualizations tools	Yes	Yes	No	No	• Improvement in decision making. • Increase in return on investment. • Time Saving • Real-time Data Analysis	• Author has focused only on static data visualization techniques and dynamic data visualization techniques are missing. • Direct analysis without prepossessing for anomalies of data through visualization can miss lead the decision makers
Cawthon et al. [60] "The Effect of Aesthetic on the Usability of Data Visualization "	The author has presented a study on 11 data visualization techniques based on standard data set. Aesthetic ranking was carried for 7 visualization technique selected for survey. The resulted data was scrutinized to achieve refined results.	Yes	Yes	No	No	• Error removing from survey for achieving best result. • It is not compulsory that fastest and most accurate technique be also aesthetically best. • 11 visualization techniques were used for survey out of which SunBurst resulted to be highest level of beauty	• Survey is confined to 11 techniques only which does not represent the complete picture of Aesthetic view. • The survey focus is mainly on Aesthetic view • We cannot project the results of small data set as deciding factor • It is not compulsory that fastest and most accurate

							technique be also aesthetically best
Shahzaib et al. [26] "A novel phylogenetic tree data visualization application for researchers"	The author has taken database of research journal as a use case and used it for the purpose of depicting data and their relationships in the form of phylogenetic tree and mesh.	Yes	Yes	No	No	Data visualization for science journal data through polygenetic trees Idea can be utilized for easy understanding of relationship	• Security aspect not reviewed
Alhanjouri et al. [27] "A New Method of Query over Encrypted Data in Database using Hash Map"	The writer has proposed a new layer on top of the encrypted data layer for querying in database using hashed map table as a primary key for searching	No	Yes	Yes	No	• Increase the performance of querying in databases • Possible searching over encrypted data	• Radom hash search possibility in database • Privacy issue if hash compromise d
Matthew N. O. Sadiku [28] "Data Visualization "	Common visualization techniques (line graph, scatter plot, bar & pie charts) and its applications in real word use cases	Yes	No	No	No	• Powerful and widely applicable tool for analysing and interpreting large and complex data • Communicate complex ideas with clarity, accuracy, and efficiently	• Large and time-varying datasets pose great challenge of users proactively respond issues at right time. • Complicated and time-consuming process to generate a big data set visualization

B.R. Kandukueri et al. [29] "Cloud Security Issues"	Security concerns from centralized to distributed systems, typical SLAs and how to standardize SLAs	No	Yes	Yes	Yes	• SLA is a legal approach to keep check on CSPs • Suggestion of standardizing the SLAs • Provide controlled physical and logical access to CSPs	• No standard SLAs among CSPs • SLA does not cover the original coast of data if lost • Investigation in cloud is a challenging task
Kacha et al. [30] "An Overview on Data Security in Cloud Computing"	Data Security Issues and classification, Data Security Attributes (CIA) and Data Security Solutions comparison accordance to cloud	No	Yes	Yes	Yes	• Cloud security is more challenging due to shared infrastructure and accessibility of data by internet • CIA trades have been debated according to the data security attributes	• Numerous security concerns due to leased, open, shared, elastic, virtualized and distributed infrastructure • There isn't a complete technical solution for cloud security
Albugmi et al. [31] "Data security in cloud computing"	Threats & risk related to the data in virtualization (hypervisor), security challenges related to multitenancy.	No	Yes	Yes	Yes	• Threats & risk awareness in cloud environment • Security issues of virtualization and multitenancy in cloud • Cryptography is best way out for security	• No security solution suggested • Only presented risk and threat hypothetically
Soofi et al. [32] "A Review on Data Security in Cloud Computing"	Cloud computing and its classifications , security approaches of cloud computing	No	No	Yes	Yes	• Presented a survey of security approaches • Encryption is resulted as the best approach for cloud security	• 42% of the research papers have no proof of the end results out of which 67% are guidelines.
Sood et al. [33] "A combined approach to ensure data security in	Proposed framework for security of data at rest and transit includes classification	No	Yes	Yes	Yes	• Proposed framework protects from data tempering, leakage & unauthorized	

cloud computing"	of data, index builder, SSL encryption, MAC, and digital signatures					access from CSPs • Achieves availability and integrity of data during data transmission	
Basu et al. [34] "Cloud computing security challenges & solutions-A survey"	Covered security gaps of cloud environment in general, security issues in detail as per CIA traits, comparison of latest schemes in cloud	No	Yes	Yes	Yes	• Comprehensive overview of cloud security flaws • Comparison of cloud security schemes in detail	• Ultimate responsibility of privacy and security lies on owner of data
Geng et al. [35] "Homomorphic encryption technology for cloud computing"	Single and fully homomorphic algorithms, comparison of Hill, RSA, ElGamal and Paillier algorithms, overview of fully homomorphic algorithms (DGHV, BGV, GSW13 and NTRU)	No	No	Yes	Yes	• Comparison of single homomorphic algorithms • Comparison of fully homomorphic algorithms •	• Single homomorphic algorithms are fast but less security efficient • Fully homomorphic algorithms are security efficient but low in performance
This research	Data threat analysis in cloud perspective, framework for data visualization over the encrypted data for security, practical implementation based on proposed framework	**Yes**	**Yes**	**Yes**	**Yes**	• Dual security mechanism • Suited for cloud and traditional infrastructure • Search capability over encrypted data • Capability of visualization over encrypted data	• Based on structured data only • Limited to confidentiality • Security dependency on primary database

Chapter 3

CLOUD DATA THREAT MODELLING

This chapter explores the threat landscape of data storage at cloud starting from the production of data at trusted domain followed by the data transfer to untrusted or curious domain like cloud. The third phase encompasses data at rest on cloud environment and then retrieval of data from the cloud to trusted domain for use and analysis purpose.

3.1 Introduction

The evolution of telecommunication has opened a new approach to the digital world. Huge manufacturing and sale of digital devices like smart phones and tablets are the result of this phenomenon. These digital devices not only provide a medium of communication but also provide access to the remote data. However, traditional infrastructure is no longer compatible to the modern requirements of the huge data management to some extent. Today, the fast speed of internet and computational power has emerged a new approach to access, manage and compute the data instantly, which is possible through cloud computing. Cloud computing has elevated the burden of ownership and management of the massive infrastructure from the customers. At a same time, shifting of infrastructure control responsibility from the user to the CSPs does not offers security responsibilities of data. Which remains a genuine concern of customer where data is no longer in their physical control.

Globally, cloud adoption for routine application remains on rise. As per the Thales [6] survey, organizations worldwide were using an average eight SaaS application in 2015, whereas the number increased to 110 in 2021. However, with the increase usage of the cloud, thread landscape has also been increased significantly. As per the Thales report -2022, “In comparison to 35% in 2021, 45% of organisations have had data breaches or have failed audits involving data and cloud-based applications.”

In last few years cloud data breaches has drastically increased due to cloud security misconfiguration or lack of security awareness. In 2017, UpGuard Cyber Risk Team [36] has uncovered the fact that "At least four AWS S3 storage buckets were left insecure by Acenture in 2017". UpGuard team has also claimed that 137GB data was available for public access which included user data, decryption keys, API data, metadata and digital certificates. Accenture again became victim of LockBit ransomware attack in 2021 and reportedly 6 TB of data was captured for which a ransom of $50 million was charged. 14 M Verizon Customer Records [37] was exposed due to AWS S3 misconfiguration. The S3 repository were utilized

for storing customer data which resulted customer data i.e., account details, names, addresses and Personal Identification Numbers (PINs) for gaining access to Verizon call centres agents. 29207 incidents [36] were again reported by the Verizon in 2020 out of which 5,200 were verified breaches.

ZDNet [38] reports that 44 million Pakistani mobile customers' data was offered for sale on the dark web for 2.1 million dollars in bitcoin. On analysis by ZDNet, the data comprises of customers personal identification details and telephonic records from late 2013. In April 2021, Facebook data was breached [39], and the attacker also succeeded to affect Facebook founder Mark Zuckerberg account. The volume of data was expected up to hundreds of millions of records. The records were exposed on Amazon's cloud computing service, the problem was immediately resolved by disconnected the affected server through Amazon.

Security of cloud domain is a vast subject and nearly impossible to cover its all aspects. However, in this chapter our focus would be limited to security of data hosted on a cloud infrastructure only.

3.2 Cloud Data Security Issues

Cloud data security can be best described as "the technologies, policies, services and security controls that protect any type of data in the cloud from loss, leakage or misuse through breaches, exfiltration and unauthorized access" [40]. Data security in cloud is somehow different from data security in traditional infrastructure due to following factors; -

a. Data is stored outside the physical control of the owner/customers.
b. Same physical infrastructure is shared for different user data.
c. Only medium to access the data is internet.
d. Geographically distributed nature of infrastructure.
e. Classical associated security risks of virtualization.

Cloud data security issues can be classified in several domains, however in this research we are classify them according to data life cycle and security attributes.

3.2.1 Cloud Security Issues corresponding to Data Life Cycle

CSPs provides two main services, storage facilities and computational power [41]. Being distributed nature of cloud infrastructure [42], these resources are served from different datacentres located around the world. We can distribute data life cycle in three phases data at rest, data at move and data in use. Data threat spectrum of each phase is different from other. An overview of data security characteristics and common solutions according to the data life cycle is mentioned in Table 3.1.

Table 3.1 Cloud characteristic and solutions according to data states and data life cycle [42]

Characteristics	Data states			Data security attributes			Main common solutions to data characteristics
	At-rest	In-use	In-transit	Confidentiality/privacy	Integrity	Availability	
Leased infrastructure	Impact	Impact	Impact	Impact	Impact	Impact	Encryption, access control, better transparency of service provider
Open infrastructure	Impact	Impact	Impact	Impact	Impact	Impact	Encryption, Access control
Shared infrastructure	Impact	Impact	Impact	Impact	Impact	Impact	Encryption, Access control
Elastic infrastructure	Impact	Impact	Impact	Impact	Impact	Impact	Encryption
Virtualization	Impact	Impact	Impact	Impact	Impact	Impact	VM securing, hypervisor securing, VPN
Distributed infrastructure	Impact	Impact	Impact	Impact	Impact	Impact	Better transparency of service provider, anonymization techniques
Main common solutions to data state/security attributes	Encryption, access control	Encryption, anonymization	Encryption, network security equipment, security protocols	Encryption, Access control, anonymization, data concealment	Integrity technique verification, Encryption	Resources/data duplication, backup	

3.2.1.1 Data at Rest

There are numerous associated risks available in literature related to the data stored on cloud. These risks can be categorized as per data location, shared storage media or according to the CSPs as following: -

a. **Risks related to Data location.**

1. Cloud infrastructures are distributed in different geographically dispersed location around the world. Cloud customer don't know the physical location of the data unless bounded through SLA. These geographical locations abide regional laws. As per section 215 of US Patriot Act [43], intelligence agencies of US can spy on any data held by third party. Mostly CSPs datacentres are in USA which can influence data security, especially confidentiality and privacy.
2. CSPs ensures data availability through backups and replication of data in different datacentres which increase the spectrum of data exposure.
3. In cloud environment large data is distributed in parts at multiple location with multiple copies for redundancy and availability [44]. However, in traditional data systems system sensitive data is protected in multiple security layers which is not possible due to the dispersed data.

b. **Risks related to Shared Storage Media**

1. In shared environment [45] multi-tenant users are separated at virtual level, but physical resources are shared. Corrupted data or malicious code of one tenant can affect the data of other tenants if CSPs have improper configured infrastructure (co-tenant and external attack).
2. The concept of shared resources amongst user have increased the risk of data compromise. CSPs allocate and deallocate shared resource according to the customers use unless dedicated resources are acquired through SLA. Unauthorized access to customers resources (shared) cannot be ruled. This drawback of shared resources can be overcome with adoption of appropriate deletion method before allocation to the next user.
3. Due to weak isolation in multitenant environment, access to hypervisor by malicious code cannot be ruled out which can be curtailed by robust authentication and access control [46].
4. CSPs provide multitenant and shared environment for data storage. Since multitenancy permits multiple users to store data. This increases the possibility of data intrusion by injection of malicious code. This problem can be mitigated by strong data segregation mechanism. *Rao et al.* [47] considers data segregation as first major concern (92%) of cloud data security followed by data leak prevention (88%).

c. **Risk related to CSPs**

1. As per the Google Cloud [48] data is encrypted over the cloud before it is written on the disk. Google Cloud also provide additional feature of server-side encryption however, question still arises on encryption keys which is owned and managed by CSPs. Similarly, trapdoors in security algorithms cannot be ruled out.
2. Customer has no control over the data once transmitted to CSPs storage. CSPs are considered as reliable but curious. There is a possibility that CSPs can use the customers data for own analysis at least.
3. According to Google Cloud Platform Terms [49], "Google and its suppliers are not responsible or liable for the deletion of or failure to store any customer data and other communications maintained or transmitted through use of the services. customer is solely responsible for securing and backing up its application, project, and customer data". Thus, arises serious concerns about cloud data and applications.

4. Cloud is a diverse combination of technologies and mixed nature of hardware with different software versions [44]. Maintaining security for such a diverse system is a challenging task where a minor flaw of hardware or software or even version exploits can be misused.
5. API acts as gateway between cloud and customer which is responsible for all services including management, provisioning and monitoring [46]. If security of API is compromised by malicious code or accidently can affect the cloud services.
6. Due to heavy workload, CSPs outsource hardware / software resources for maintenance and management. This model is good but defiantly it increases the spectrum of insider threat. Malicious insider threat is a worst threat where data of customers can be threatened or compromised by employees. It happens when a legitimate employee exercises his authorized role or get/acquired unauthorized access through privilege escalation or social engineering for illegal activity.

3.2.1.2 Data in Move

a. When data is transmitted to cloud storage resources man-in-the-middle attack is possible due to absence or lack of SSL configurations like SSL, VPN, authentication protocols.
b. Phishing attack is a possible method to manipulate CSPs link and capture customers credentials for conceding the information in transit.
c. Service / Account hijacking is a method of acquiring credential of the legitimate customer through phishing, fraud, man in the middle or social engineering attack. Once the credential is acquired by hacker then he/she can do any activity like legitimate users.

3.2.1.3 Data in Use

a. Cloud provides elasticity in resource allocation as per the requirements of the users. However, when resources are released by a user is reallocated to another user can recover the previous user data through recovery tools. Which is a serious threat to loss of confidentiality.
b. Cloud is based on distributed system where multiple nodes are physically dispersed but communicate with each other for different processes. Data is processed by those nodes which have computational resources [44] so computation can take place anywhere in the cluster. It is difficult to find the exact node where the computation took place and ensure security for sensitive data.

c. Risk associated with shared media and distributed nature of cloud is fully applicable in data in use as CPU and RAM resources are always used during accessing the data.
d. Cloud data recycling [46] is another vulnerability which can be exploitable if data sanitization is not carried. Data sanitization is a process of disposing of data sent to garbage before allocating same resources to the other customer.
e. Due to dynamic nature of cloud, it provides the liberty to create, modify and copy Virtual Machines (VM) image which is kept in DB repository. It can be turn on and off easily. Cloud also provides the freedom to create her own VM image or copy a previously created image. This phenomenon can be misused by creating an image of malicious code VM. Same infected VM can be used for malicious activities like compromising user privacy and data theft. Improper management of VMs can resulted a situation of VM sprawl [46] where new VMs are created and previous VMs remain idle.

3.2.2 Cloud Data Security Goals

According to the requirements of the customer, security goals varies as per the data type. The security requirements also depend on the state of data (rest, transit and use). However, confidentiality, integrity, and availability are the three cornerstones of data security (CIA). These attributes are applicable on cloud main characteristics like virtualization, dynamicity and distributed architecture of cloud.

3.2.2.1 Confidentiality

Confidentiality can be referred as protection of data from unauthorized access or leakage of information without owner permission. The risk spectrum of unauthorized access to data increased when transferring data to cloud environment as compared to traditional storage system. There are three phases where data is exposed to attacker once data is stored on cloud or when data is accessed for use.

a. Data in transit (risks and vulnerabilities associated with shifting of data from traditional storage to cloud storage)
b. Data in rest (risks and vulnerabilities associated with the cloud environment)
c. Data in use (risks and vulnerabilities associated with cloud computation environment)

In private and community cloud risk spectrum is less as compared to public and hybrid cloud model. Multiple solution exists in literature like encryption, access control, anonymization and data concealment for preservation of confidentiality. Cloud is a heterogeneous nature environment where a single solution for confidentiality is not possible.

In multiple research papers [32] [41] [42] [46] [47] [50], encryption is considered as best solution for data confidentiality however strong authentication and authorization [51] cannot be ruled out as a solution for confidentiality issues. Encryption solves the issue of confidentiality to some extant however, there are challenges of key management, searching over encrypted data and efficiency of system. To enhance the efficiency of the encryption methodology we should also consider some following factors: -

a. Data in transit have minimum exposure time thus provide less time opportunity for attacker to decrypt the data. To increase efficiency short length of encryption key is recommended instead of long keys.
b. According to sensitivity of data, classification of data is recommended before encryption instead of encrypting the whole data.
c. Homomorphic encryption is recommended for sensitive data where frequent accesses is expected instead of whole data keeping in mind the computational complexity and cost.
d. Encryption and decryption of data should be done on client side to curtail the associated challenges of untrusted / curious cloud.

3.2.2.2 Integrity

Data integrity can be defined as unauthorized change in the data. The change includes modification and deletion. In cloud data integrity is considered as important facet of cloud security. The aim of integrity is to shield the data from change by illicit user (internal and external). If we consider the life cycle of data, integrity can be compromised during transit, rest or during computation. Keeping in view the cloud distributed nature, maintaining data integrity is a multifaceted task as compared to the traditional and centralized system. To check the integrity, present data on cloud is compared with the original data which involves the process of downloading [41]. Other methods of data integrity check are integrity technique verification, encryption and hashing which does not involved the downloading of data.

Keeping the sensitivity of data, cloud user may be reluctant to trust the cloud for data integrity. Reason being cloud services are offered by the third parities who does not fall under the same trust sphere [50]. However, SLA can be done with CSPs for data integrity services being considering the CSPs under the trusted domain. Another factor is timely data integrity services should be ensured. It is important for long term storage because it is difficult to recover data on a disk which is multiple time over written.

There are multiple strategies of integrity check without downloading data from cloud. These strategies can be divided in three main groups

a. Provable Data Possession (PDP)

b. Proof of Retrievability (PoR)

c. Third Party Auditing Technique

Walker et al. [52] defines provable data possession (PDP) is a technique in which tenants are verifying integrity of data stored on untrusted domain like cloud. In this strategy tenants pre-process the data and send the data to untrusted domain keeping a small amount of metadata. To check the integrity of outsourced data, the owner challenges the server and response of server is verified by the owner of data. There are multiple variants of PDP like Scalable PDP, dynamic PDP etc.

Proof of Retrievability (PoR) is somehow similar scheme like PDP. The difference is that PoR guarantee that tenant can recover the file even with minor corruption of data [52]. In PoR methodology the customer can retrieve the file and fix the corrupted data through error correction code. In PDP and PoR method integrity check is carried out by the data owner whereas in Third Party Auditing Technique the integrity is checked by third party inside the system and effectively check the integrity without making any local copy. A comparison of all the above-mentioned techniques is given in the Table 3.2.

Table 3.2 Cloud Data Integrity Checking Techniques [41]

Employed Technique	Pros	Cons
Provable Data Possession (PDP)	Minimal computational cost Minimal storage overhead	Doesn't suit dynamic data
Scalable PDP	Supports dynamic data	Advance fixation of challenges and responses limited count of updates
Dynamic PDP Supports	Dynamic data without any limits for updates	Time complexity
Proof of Retrievability (PoR)	Supports dynamic data operations Resilient to reset attacks	-
Third Party Auditing (Multiple versions)	Data audit takes place without local copy of data	Trust related issues

3.2.2.3 Availability

Data availability means access to required resources 24/7, which can be attained through redundancy of resources in cloud. Main driving factor of business migration to cloud is availability. There are some scenarios where data availability is can be compromised due to

network failure, disasters like flood, earth quick, fire etc. In such situation, the data owner is concern about data availability. He must know about the response time of CSPs to recover the data which can be assured through SLAs.

Data availability can be ensured by disaster recovery strategy which includes redundancy and spread of resources across geographical location. Data availability is generally good and more reliable in clouds like Google, Amazon and Microsoft due to large infrastructure as compared to traditional systems. Cloud availability also encompasses the quality parameters as promised by the cloud. This means that data is accessible but experience time delay in operation as promised in SLA or in service quality parameters will be considered as nonavailability.

Availability and security in cloud environment are somehow considered as inversely proportional. Security of cloud is increasing with complexity of operation and complexity required more computational power for speedy operations sometime affect performance and impose delay resulted to nonviability.

3.3 Deductions

After detail analysis of the data threat spectrum in cloud, following points are deduce:

a. Attack spectrum of data is different as per the state of the data.
b. One security mechanism is not applicable in all states of data security.
c. Data security risk is different according to the cloud deployment model.
d. A strong security system depends upon a balance in CIA traits.
e. For confidentiality of data, encryption is considered as most popular methodology [32].
f. Encryption schemes are used for data security in complete data life cycle in multiple forms like hashing, encryption and anonymization.
g. Strong authentication and authorization mechanism in cloud enhances data security in general.

3.4 Summary

In this chapter a comprehensive threat modelling of data security in cloud is presented. Initially deployment model has been discussed in context of cloud security in general. Cloud data security issues have been discussed in data life cycle. Security issues corresponding to data at rest, transit and use are deeply reviewed with possible solutions available in literature. In the last common security goals (CIA) related to data and cloud environment have been summarised with few deductions.

Chapter 4

FRAMEWORK FOR DATA VISUALIZATION OVER ENCRYPTED DATABASES

This chapter explains the proposed framework of data security over an untrusted source through mix scheme of encryption & hashing and design a mechanism of data extraction and data analytics over encrypted database through data visualization.

4.1 Introduction

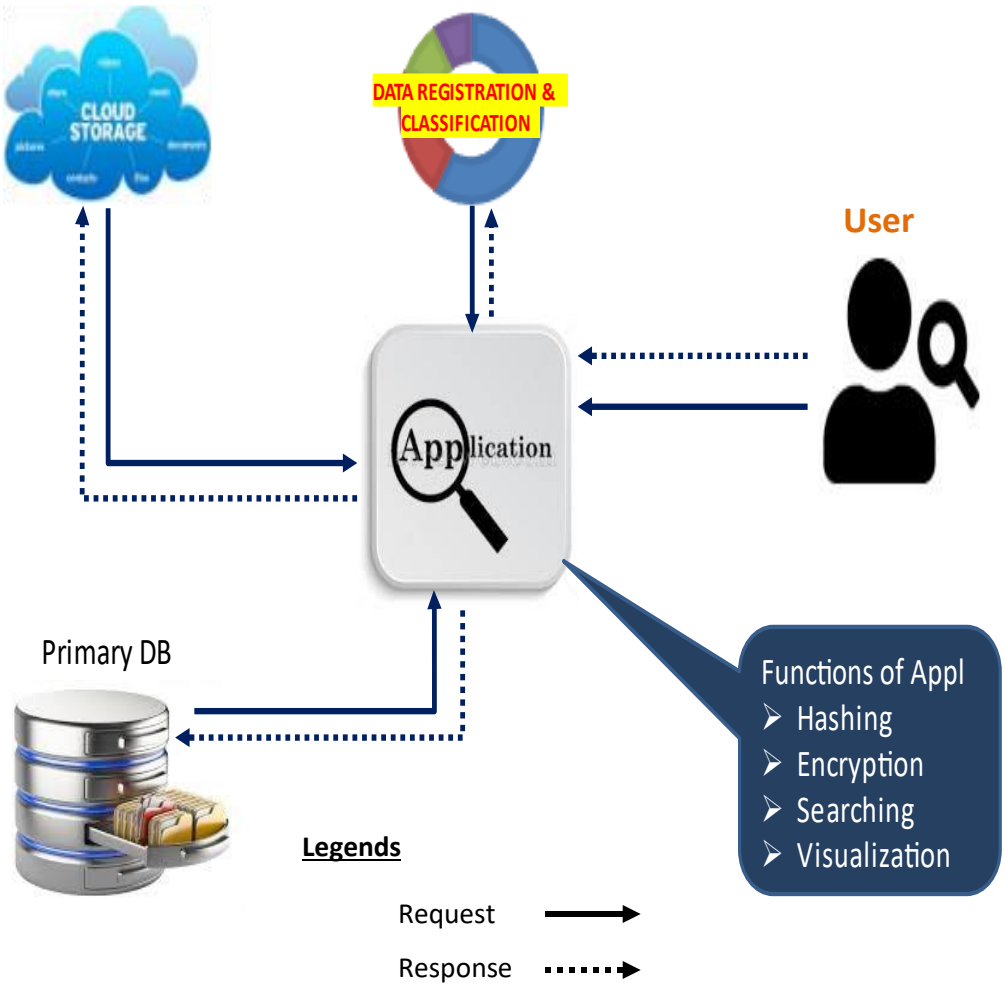
The exponential growth and advancement in IoT have changed human life. Wi-Fi Alliance [1] forecasts by 2022 over 400 M wifi devices be serving smart devices which will be carrying half the data traffic of globe which is a proof towards IoT growth rate. For management of big data produced from IoT devices are normally outsourced to third party like cloud which elevate the problem of largescale data administration. Massive volume of data, as patients health records, contacts, e-mail communication records, IoT data and much more are outsourced.

The problem arises when sensitive and private nature data are transferred to curious / untrusted CSPs servers which is ultimate a privacy concern of customers. To address data privacy issues, sensitive data are normally encrypted before outsourcing to third party like cloud, which preserves confidentiality of the data and privacy of customers as well. However, for analytical purpose the data are normally downloaded to trusted domain to perform decryption and run analytical algorithm to extract required results. Similarly, big data analysis is also a challenging task.

In research we are proposing a framework for protection of confidentiality and privacy of data before outsourcing the data to third party like cloud. Followed by data visualization technique be applied over the stored data on cloud and extract only required data. In the end only required data should be shifted to trusted domain for decryption.

Figure 4.1 represents an overview of proposed framework. Application acts as a central entity isolating cloud and local database from each other. The application performs

Figure 4.1 Framework of Data Visualization over Encrypted Database



hashing, encryption searching and visualization function. Before outsourcing, data is passed through classification and registration process. After segregation process, data is outsourced in cloud data base and registration data in primary database. User can use application services (, registration, searching, visualization) through application only.

4.2 Assumptions

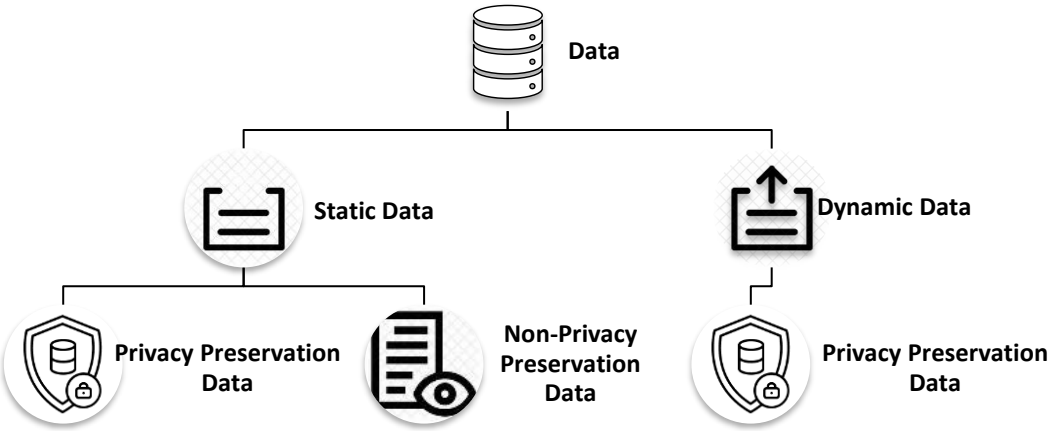
Before proceeding to the suggest framework for data visualization over encrypted data bases, following points are assumed for better understanding of the system model.

a. Cloud environment has multiple security threats and risks. However, in this research we implicitly selected data security.
b. Data life cycle is comprised of data at rest, move and in-use. This framework is restricted to data security at rest only.
c. Data can be divided in two main types, structure and unstructured. We assumed structure data for our research and system model.
d. We have assumed that data is outsourced to public cloud by enterprise for sensitive data management and access is only available to data administrators of same organization. However, this concept can be extended for enterprise-customers model as well.
e. The focus of this framework is data confidentiality rather than availability and integrity.

4.3 Data Classification

Data classification is the first step. We can define data classification as the process of organizing data according to the relevant categories. The reason of classification may be according to the domain knowledge, regulatory compliance, ease of access or various other personal and business objectives. As per this research scenario we can categorize the data according to privacy as following (Figure 4.2).

Figure 4.2 Data Classification



4.3.1 Static Data

The data that does not changed once it's being recorded. Moreover, the data represent fixed data which is seldom changed. This type of data gets more importance in financial and privacy preserved data related sectors. Examples of static data are country name, transection ID, biometric data, DNA etc.

4.3.2 Dynamic Data

As the name signifies, the data which is changing frequently once it has been recorded or data which is not fixed and increases or decreases frequently. Sensor data, logs of different systems, cellular phone call data records etc.

4.3.3 Non-Privacy Preserved Data

Data, which is publicly available for use, reuse and can be distributed freely without any legal restriction (local, national or international). These data also include personal Data, which is publicly available for use like name, contact number, name etc.

4.3.4 Privacy Preserved Data

The data which is private in nature and be protected from public. The data includes health records, system logs, financial records, private files and documents etc.

4.4 Data Storage

After classification of data, second phase is the storage of data in database. Storage is classified as outsourced data base and primary storage database as per the framework.

4.4.1 Primary Storage Database

This storage will keep the primary data required to access outsourced database. The database is recommended to be in full access of the owner. Primary data includes the access ID of outsourced database. The ID would be obtained by taking hash of the data of the customer plus random number. The ID and basic data would act as lookup table in the outsourced database. Hash function would be used as per the user requirement and satisfaction. In Table 4.1, Fields (1…. n) represents the primary data like Name, CNIC, address, Mobile number etc. Random number is the instant random number generated by the operating system for adding anonymity. Ref ID (Primary key) value is the hash of all the fields (1…. n) plus, random number.

Table 4.1 Primary Database

Ref ID (Primary key)	Field 1	Field 2	Field ...	Field n	Random Number	Encryption key (optional)
Hash (field (1‖2‖…… n) ‖ Random number)						
11ceb13b3489ebd39855ce1aef8ada62eeeb5756d0fedb94698ff0f34e598158						

4.4.2 Outsourced Database

The outsourced data base keeps all the data which is required to be stored and of privacy concern with reference ID (of primary Database). This data would be stored on untrusted domain. Therefore, the data would be encrypted before transferring to outsourced database. The encryption mechanism would be used as per the user satisfaction / requirements. Table 4.2 represents a general scheme of outsourced database. Hash primary ID is associated with all the data in the outsourced database and all the data fields (1…n) are encrypted with a secret key before transferring the data to cloud/outsourced database.

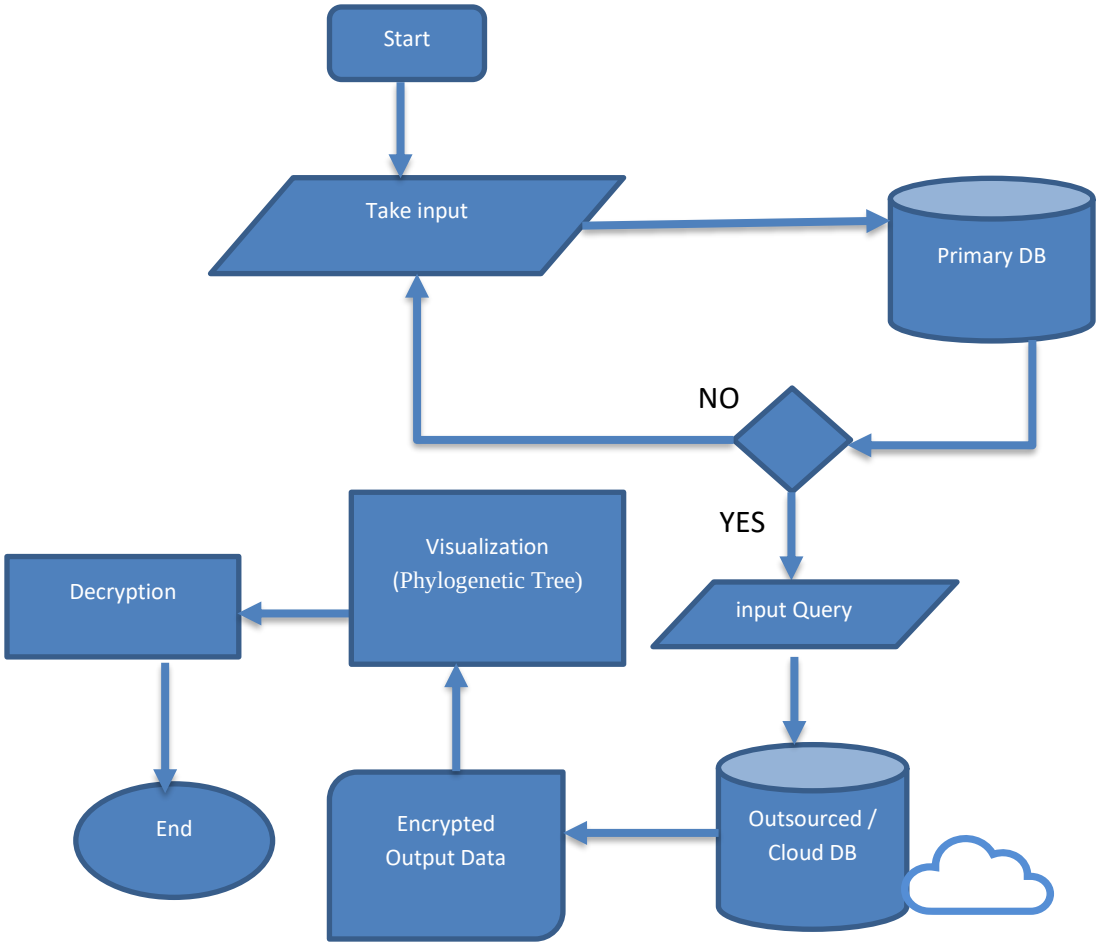
Table 4.2 Outsourced Database

Ref ID (Primary key)	Data 1	Data 2	Data 3	Data ....
11ceb13b3489ebd39855ce1aef8ada62eeeb5756d0fedb94698ff0f34e598158				

4.5 Data Search

Search over encrypted database (outsourced database) is illustrated in flow chart Figure 4.3. In case information of a customer is needed to be searched, the user having administrator rights will search the individual ID as per the basic information available in the primary database illustrated as field 1…n. once the ID is searched in the primary database, then the user would be able to search the encrypted data in the outsourced database based on the hashed ID (Primary Key). However, at this stage the data is encrypted format thus preserving the confidentiality of data. Based on the search data matched with the ID would be brought to trusted domain and then decrypted with decryption key.

Figure 4.3 Data Search Flow



4.6 Data Visualization

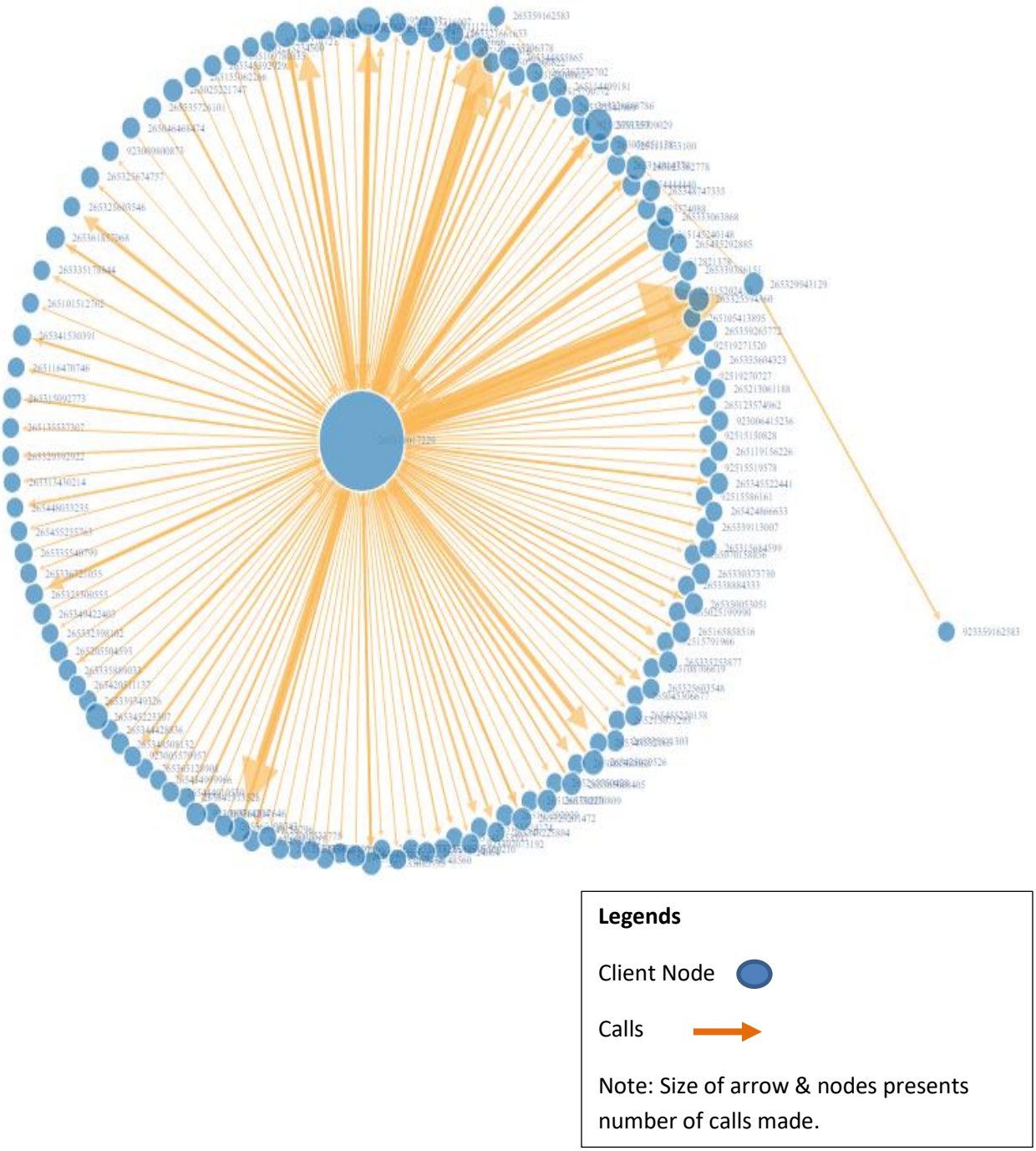
The data is searched in the outsourced database, based on the hashed ID. The output data would be passed through visualization function. The function would draw visual tree over encrypted data. The visualization technique presents a good analytical tool for analysis purpose. Figure 4.4 represents a centrally located node connected with many other nodes. In this polygenetic tree structure, size of arrow represents amount of communication took place between the nodes. Size of every node also depends on the amount of communication took place with all other nodes in the polygenetic tree.

4.7 Summary

In this chapter we have presented a conceptual framework for data visualisation over encrypted databases. The system model can be formulated in four phases based on this framework. In first phase data classification is performed followed by data storage in two databases. The primary database holds basic data which is provided by the customer for

registration including system generated primary key. In outsource (cloud DB) database dynamic nature data is outsourced to third party cloud after encryption. Third and last phase comprises of data search over the cloud (encrypted database) and generation of visual trees and graphs respectively.

Figure 4.4 Data Analysis through Visualization



Chapter 5

CASE SCENARIO – CALL DATA RECORD SECURITY AND VISUALIZATION

This chapter explain sensitive CDR protection through proposed framework followed by practical implementation of concept through coding and visualization of CDR for pragmatic analysis.

5.1 Introduction

According to the General Data Protection Regulation (GDPR), "Personal data are any information relating to a recognised or identifiable natural person" (i.e., name, identification number, location data, etc.) [53]. Additionally, the GDPR defines "sensitive personal data" as information that requires a higher level of protection, such as genetic, biometric, and health information as well as information about racial and ethnic origin, political views, religious or ideological beliefs, or membership in a trade union. The Personal Data Protection Bill 2021 (PDPB) of Pakistan [26] states that "best international standards be applied to protect personal data from any loss, abuse, modification, unauthorised or accidental access or disclosure, alteration, and destruction."

As per GDPR and PDPB of Pakistan [53][54], it is mandatory for all the custodians of sensitive data to adopt international standards for the protection of sensitive data. However, it is some time financially cumbersome for small companies to build own infrastructure keeping the frequent technology changes over time. Public cloud provides easy access of services and compliance of standards, however public cloud is considered as curious and cannot be fully trusted. Similarly, Health Insurance Portability and Accountability Act (HIPPA) also contends to protect patients data electronically stored by adopting suitable administrative, physical and technical protections to make sure confidentiality, integrity and availability of information remain intact.

In this chapter we are taking CDR sensitive data as a use case and secure the data by adopting our proposed framework, then deploying the data on untrusted cloud platform followed by searching, visualization and analysis of data over the cloud.

5.2 Call Data Record

Call Data Record is sensitive data and generally not available even for educational purpose. However, as per section 24-a of PDPB [54] offers the right to subject for acquiring their personal data. Law enforcement agencies can also acquire the information required in investigation or to prevent further crimes. Pakistani cellular companies like Ufone, Zong and Telenor provide limited data record facility through their official application and websites after necessary security verification. Similarly, some mobile application can also extract call records of mobile phones, but the data is limited to fields as available in Table 5.1.

Table 5.1 Call Data Record

Called Number	Call Type	Call Time	Call Duration	Call Charges
923076360822	Outgoing	01/01/2022 12:30	29s	0
923123574962	Outgoing	02/01/2022 21:13	16m, 2s	0
923145240148	Incoming	02/01/2022 17:57	1m, 6s	0
Internet	GPRS	24/12/2021 21:02	11h, 58m, 21s	0
923341530391	SMS	27/12/2021 20:45		0

The information available in logs maintained by Network Service Provider (NSP) depends upon the equipment's used [55]. This information is generally archived by billing departments. A general example of a CDR from a GSM MSC [56] is presented in Table 5.2 to offer a good understanding of information these logs can include.

5.3 CDR Data Set

For practical implementation of concept, legally CDR data of multiple people was a big challenge. However, the problem was resolved by acquiring CDR of five Ufone user from the author family. The selected members were frequently in contact during last three months. Ufone official application were used by each member personally for acquiring the CDR Data. The consent of all the members were obtained for use of CDR for educational purpose only. Data duplication was removed manually, and actual mobile numbers were altered with specific format to mask the actual identity of each number.

The actual CDR are composed of around 20-25 different details. As already highlighted in Table 5.2 and depends upon the technology as well [55]. For practical implementation we made few assumptions where data was not available.

Table 5.2 Generic Example - Call Data Record

Generic CDR Collected from a GSM MSC (Gibbs and Clark, 2001)
Example: Mobile originated call (MOC)
CDR HEADER
CALL REFERENCE
NUMBER OF SUPPLEMENTARY SERVICE RECORDS
CALLING IMSI
CALLING IMEI
CALLING NUMBER
CALLING CATEGORY
CALLED IMSI
CALLED IMEI
CALLED NUMBER
DIALED DIGITS
CALLING SUBSCRIBER FIRST LOCATION AREA CODE
CALLING SUBSCRIBER FIRST CELL ID
CALLING SUBSCRIBER LAST LOCATION AREA CODE
CALLING SUBSCRIBER LAST CELL ID
OUT CIRCUIT GROUP
OUT CIRCUIT
BASIC SERVICE TYPE
CHARGING START TIME
CHARGING END TIME
CAUSE FOR TERMINATION
ORIGINATING CALL CHARGE TYPE
ORIGINATING CALL TARIFF CLASS
CONNECTED TO NUMBER
CHARGE NUMBER
CHARGE NATURE
CARRIER SELECTION
SPEECH VERSION
INTERMEDIATE CHARGE CAUSE
CLOSED USER GROUP INFORMATION

5.3.1 Static Data

In this case scenario the static data are the information required for initial registration of mobile Subscriber Identity Module (SIM). This data can be privacy preserved or non-privacy preserved data which is static in nature. This category of data may vary according to country / region and laws. In Pakistan CNIC information and biometric are used for SIM registration which are static data.

5.3.2 Dynamic Data

In this case scenario the dynamic data is the information logs continuously generated during the use of the mobile SIM. Dynamic data are of privacy concern data as discussed in Table 5.2. For practical purpose we will considers the logs in Table 5.1 as Dynamic data.

5.3.3 Privacy Preserved Data

Privacy preserved data in this case scenario are Table 5.2 data, which is only available to law enforcement agencies. For practical purpose we considered Call Time and Call Duration in Table 5.1 as privacy preserved data set. All the security mechanism would be applied on call time and call duration and would be consider applicable for the data [56] as mentioned in Table 5.2.

Table 5.3 Privacy Preserved Data

From	TO	Call Time	Call Duration
265330788508	265319017229	01/10/2021 8:43	1m, 14s
265330788508	265319017229	01/10/2021 9:41	43s
265319017229	265088309068	01/10/2021 12:20	21s
265319017229	92515150828	01/10/2021 19:34	24s
265319017229	265116470746	01/10/2021 20:08	8m, 12s

5.3.4 Non-Privacy Preserved Data

Data, which is publicly available and used for identification of individual or items in public domain. In this case scenario Name, Mobile number, CNIC are considered as non-privacy preserved data.

5.4 Data Storage

After segregation of data, the next step is data storage. We can identify two main problems as per our case scenario, which should be cater for before storage of data.

d. Security of data through encryption.

e. Search mechanism over encrypted data (searchable encryption)

To achieve above two challenges, data are segregated in two databases. Primary storage database and outsourced database. For practical purpose pgAdmin is used for depiction of primary storage database and outsourced database. A web-based GUI management tool pgAdmin is used to interact with Postgres and related relational databases on both local and distant servers [57]. The best feature of this application is the availability of an intuitive data administration interface to manage SQL queries, maintenance, and other essential tasks without the need for command line prompts. [58]. After installation of the pgAdmin we created to two databases as following.

a. Primary Storage Database as "primary_dataset"
b. Outsourced Database as "cloud"

5.4.1 Primary Storage Database

Data in primary storage database are normally static in nature and limited in size. Due to sensitivity and limited size of the data in this database, it is recommended to be in direct access of the application owner only and should not be outsourced to cloud domain. This primary database will act as a lookup table to the outsource database, where the primary data of customer along with the unique hash ID of each customer is preserved in the database.

For obtaining the unique hash ID of each customer, primary data along with random number are passed into a MD-5 hash function as visible in Figure 5.1. This primary data of customers is normally obtained, once a customer visits mobile service provider company franchise (Ufone, Telenor, Zong etc) for registration / obtaining of new SIM. The function generates a unique hash ID for the customer, which remains as an identification of the customer in primary and outsourced database. Purpose of adding random umber in hash function is to introduce randomization for non-generation of hash ID if primary data is acquired and provide opportunity for registration more mobile number on a single customer data. The best part of hash function is non reversibility. Extraction of primary data from fixed size hash ID is not easily possible. Non reversibility of hash ID phenomenon is used for maintaining confidentiality of data.

Figure 5.1 MD-5 Hash Function

```
UPDATE public.user_info SET id = md5(mobile || cnic || address || random_number)
WHERE id = md5($1)
```

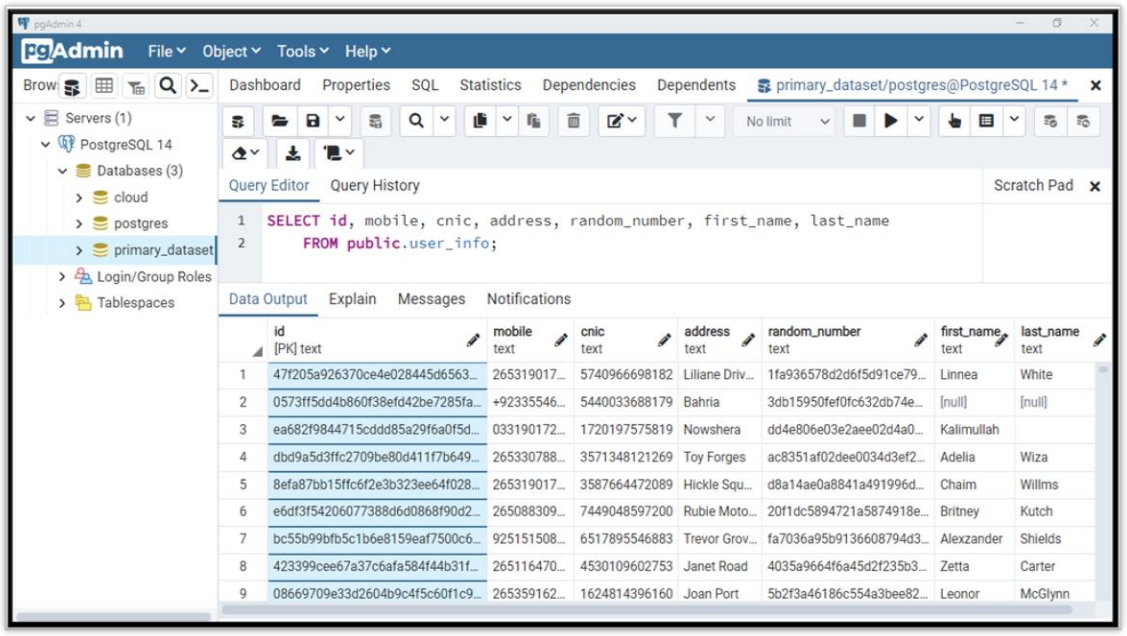
For available set of call data record, the primary data of most of the mobile number are assumed for practical purpose. In build library function “faker” is used for filling the dummy data in the primary data where data was not available. Figure 5.2 presents faker function used in application.

Figure 5.2 Faker Function

```
connection.query(`INSERT INTO user_info(id, mobile, address,cnic,  first_name, last_name)
                                        VALUES (md5($1), $1, $2, $3, $4, $5)`,
                                        [item[0].replace(/\s/g, ""),
                                        faker.address.streetName(),
                                        faker.random.numeric(13), faker.name.firstName(),
                                        faker.name.lastName()],
            (error, result) => {
                                    if (error) {
                                    console.warn(error)
                                        }
                                    else {
                                        connection.query(`UPDATE public.user_info SET id
                                        = md5(mobile || cnic || address ||
                                        random_number)
                                        WHERE id = md5($1)`, [item[0].replace(/\s/g, "")],
                                    (error1, result1) => {
                                        if (error1) {
                                        console.warn(error1)
                                                }
                                                })
                                            }
                                    }
```

After performing the basic function, primary data is stored in the primary database. An overview of the primary database is shown in the Figure 5.3.

Figure 5.3 Primary Storage Database



5.4.2 Outsourced Database

The outsourced database maintains logs of each customer with unique hash ID along with the data produced from the customer call as shown in the Table 5.2. This database is depicted as cloud database in pgAdmin. Data in this database would be outsourced to third party like cloud to reduce the administrative burden of the mobile service providing company and ease of access. Same framework is fully applicable even in private cloud / traditional infrastructure and can acquire full benefits.

Before outsourcing the data in cloud database, two functions are applied on the data to meet the requirement of confidentiality and search. First function is the encryption function and sectioned is the search function.

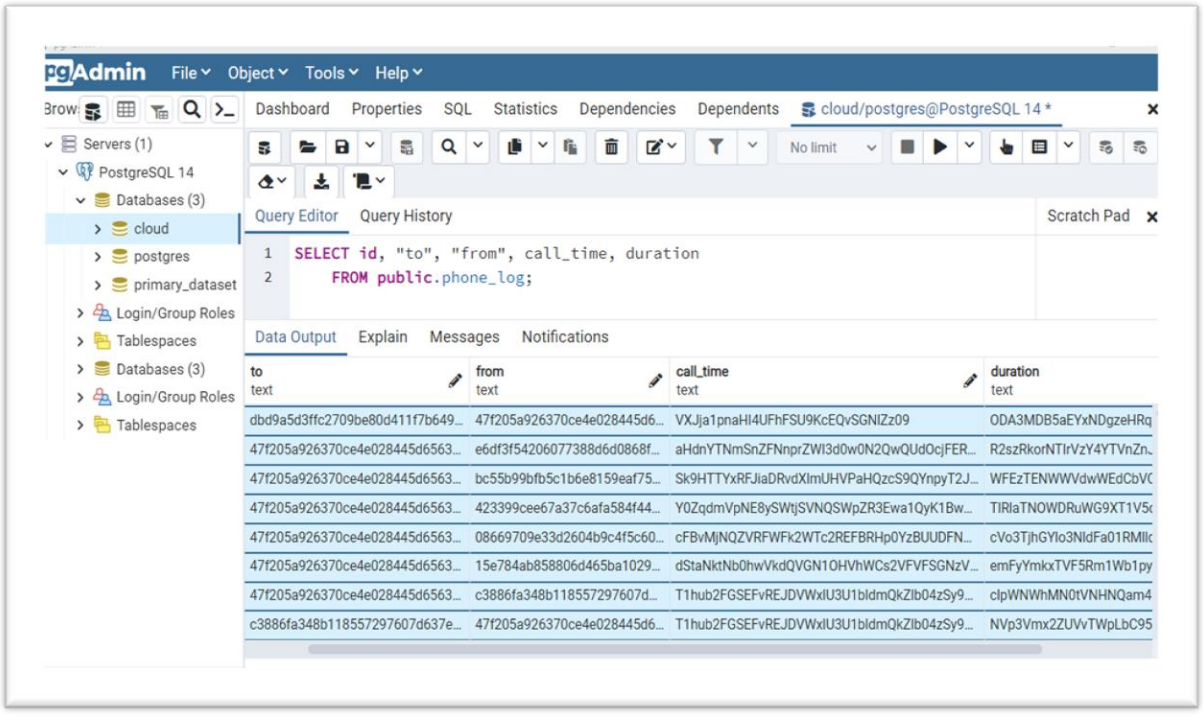
In this case scenario, we have assumed Call Duration and Call Timing as sensitive data where confidentiality is important. As per the framework the sensitive data is required to be encrypted by adopting any encryption algorithm as per the satisfaction of the MSPC. In our case scenario the data is encrypted with AES-256-CBC function as given in Figure 5.4.

Figure 5.4 AES-256-CBC

```
function encrypt_string(plain_text) {
   var encryptor = crypto.createCipheriv('AES-256-CBC', key, iv);
   var aes_encrypted = encryptor.update(plain_text, 'utf8', 'base64') +
encryptor.final('base64');
   return Buffer.from(aes_encrypted).toString('base64');
};
function decrypt_string(encryptedMessage) {
   const buff = Buffer.from(encryptedMessage, 'base64');
   encryptedMessage = buff.toString('utf-8');
   var decryptor = crypto.createDecipheriv('AES-256-CBC', key, iv);
   return decryptor.update(encryptedMessage, 'base64', 'utf8') +
decryptor.final('utf8');
};
```

After encryption, the encrypted data is stored with hashed ID on cloud database. The data in cloud database are a combination of hashed and encrypted data, thus make a reasonable strong confidentiality mechanism. The best part of this framework is provision of dual security mechanism. If encryption key is compromised and data is decrypted, even then data cannot be mapped on the mobile numbers until unique hashed ID is also compromised. An overview of hash and encrypted data is visible in the Figure 5.5.

Figure 5.5 Cloud Database after Encryption



5.5 Data Search

Data on cloud database are in encrypted format as shown in Figure 5.5. Search over the encrypted database(cloud) for customer data is possible as shown in Figure 5.6. In present case scenario if call record data of some customer is required then, the customer hash ID is searched in primary database through his basic data which he provided during SIM registration. Search function of primary database is shown in Table 5.4 and result of the function in Figure 5.7. After returning hash ID by the function, next step is searching of data in cloud database which resides in untrusted domain (curious cloud). Search flow between trusted and untrusted domain (cloud database) is shown in Figure 5.6. Data on cloud database is a combination of hash IDs and encrypted data. The basic concept of hashed IDs and encryption is to ensure confidentiality of data over untrusted / curious cloud.

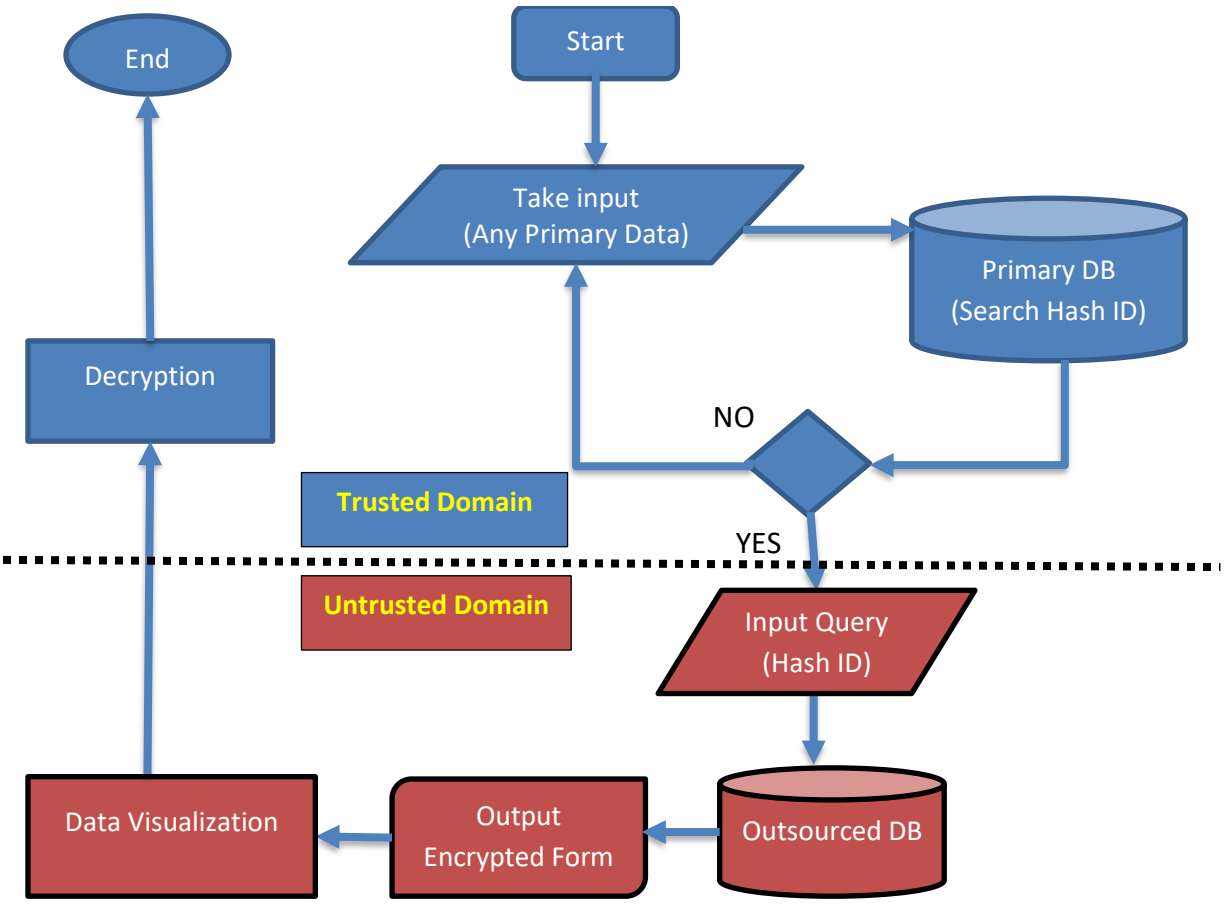
Table 5.4 Search Function

```
connection1.query("SELECT * from phone_log WHERE \"to\" = $1 OR \"from\" = $1",
[id], (err1, result1) => {
        if (err1) {
            res.send({
                statusCode: 300,
                message: err1.message
            })
        }
        else {

            let calls = result1.rows.map((item) => {
                return ({
                    ...item,
                    call_time: decrypt_string(item.call_time),
                    duration: decrypt_string(item.duration)
                })
            })

            res.send({
                statusCode: 200,
                user: result.rows[0],
                calls: calls
            })
        }
    })
```

Figure 5.6 Data Search Flow between Trusted and Untrusted Domain



The next search is performed over the cloud database through hash ID which is extracted from the primary database against the customer data. The result of the search is populated in a visual format. Application provides the facility of performing multiple searches of customers associated with the primary customers just by clicking on the customer. After searching over encrypted data, the result is produced over untrusted domain is in encrypted format. The encrypted result can be visualized for analysis purpose, but the actual contents can only be achieved through decryption. The Final step is data decryption but before decryption the data is brought to trusted domain to avoid loss of confidentiality of data.

5.6 Data Visualization

After data search over encrypted cloud database, the result is also in encrypted form and untrusted domain. This encrypted data is populated through visualization with associated details like total numbers of call made, associated customers etc as shown in Figure 5.7. The best part of this tool is provision of data visualization over encrypted data for easy data analysis on cloud database which is considered curious for sensitive data. Month wise call record is populated in graphical form using same mechanism over the encrypted data (Figure 5.8).

Figure 5.7 Data Visualization and Search over Encrypted Data

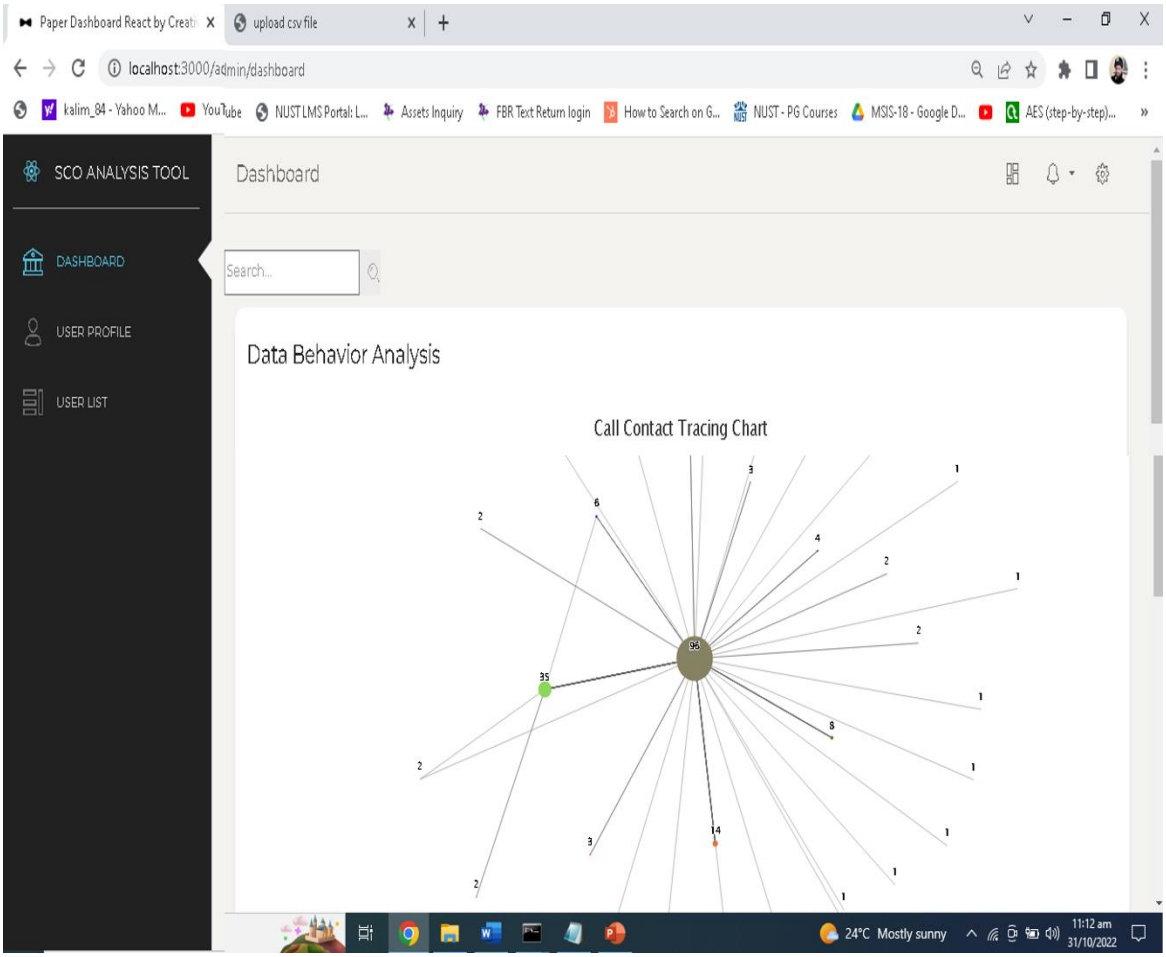
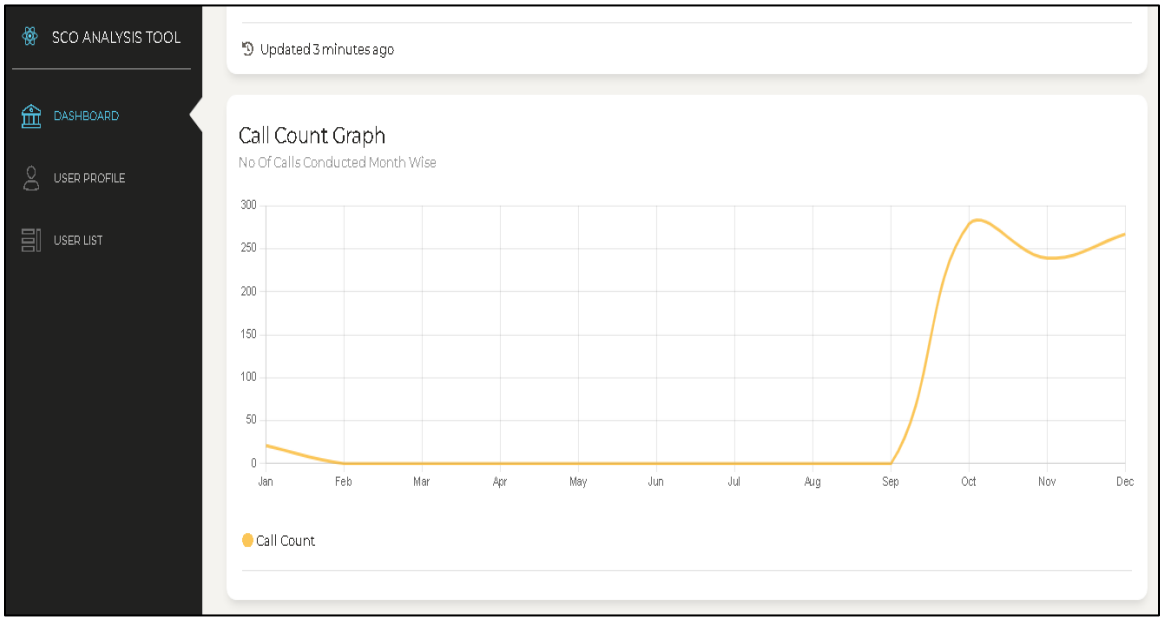


Figure 5.8 Monthly Call Record Graph



5.7 Application Features

Following are some related features of application developed by adopting same framework.

a. New customers can register directly through user profile dashboard or new user data can be uploaded directly through csv file (Figure 5.9).
b. Already register user data can be viewed in user list tab (Figure 5.10).
c. Main dashboard provides three features search, tree of connected calling nodes (Figure 5.7) and call record graph (Figure 5.8).
d. Search features provide the liberty to search customer data with any attribute which is provided during registration process like Name, CNIC, Address, mobile number (Figure 5.11).
e. Click on any node decrypt all relevant data associated with the clicked node like call time, duration or any other information encrypted by the application (Figure 5.12).
f. Index based searching over encrypted data is more efficient as compared to other searching mechanism over the encrypted data. Figure 5.13 shows application load time, Figure 5.14 shows call data load time & Figure 5.15 shows application and data load time in graphical format against different records.

Figure 5.9 Customer Registration

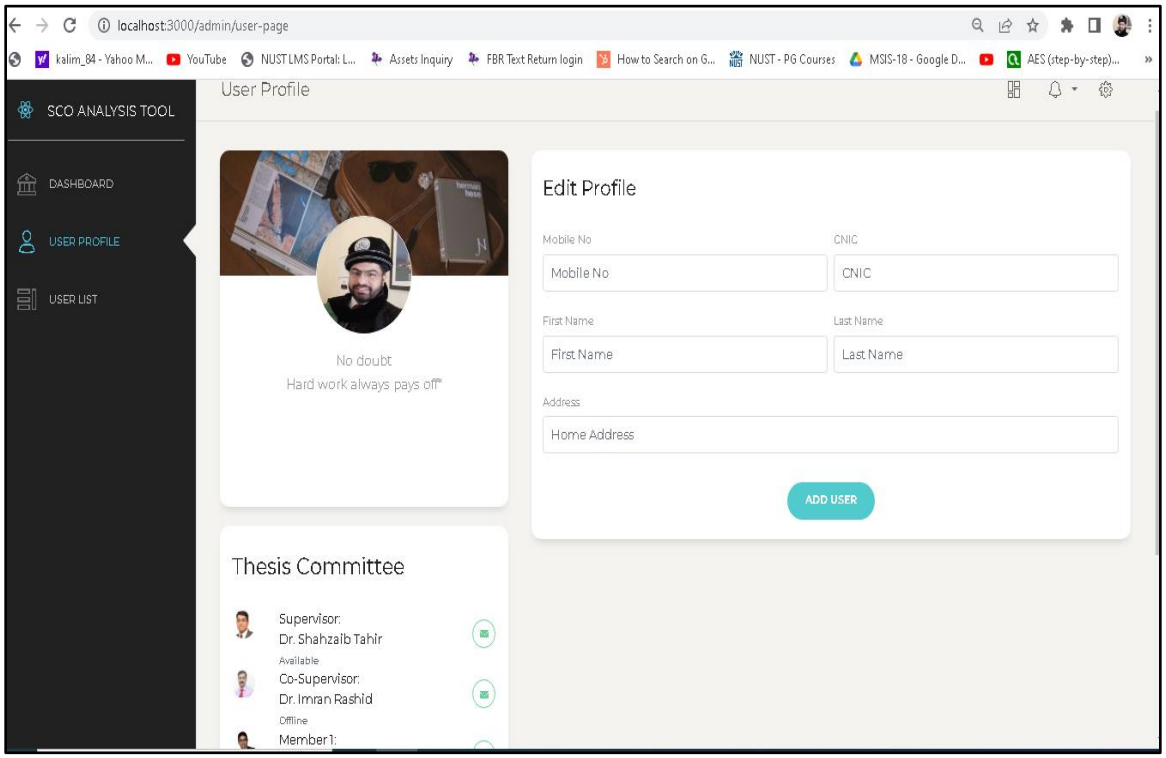
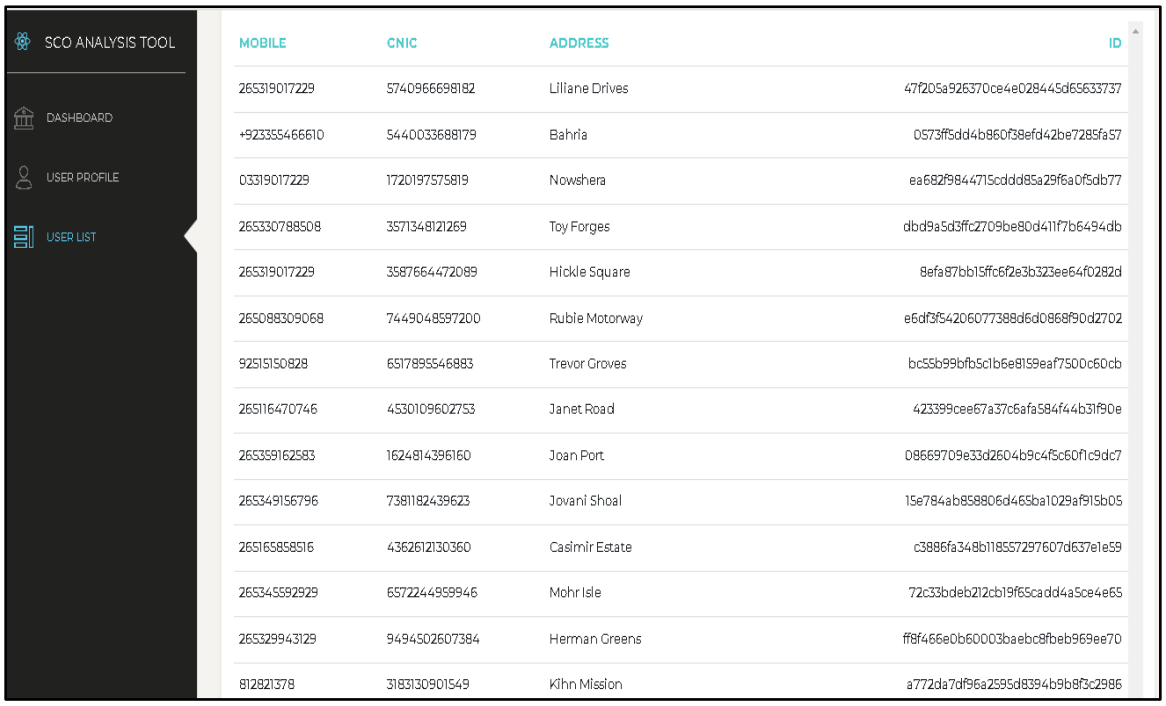


Figure 5.10 Registered User Data



MOBILE	CNIC	ADDRESS	ID
265319017229	57409666698182	Liliane Drives	47f205a926370ce4e028445d65633737
+923355466610	5440033688179	Bahria	0573ff5dd4b860f38efd42be7285fa57
03319017229	1720197575819	Nowshera	ea682f9844715cddd85a29f6a0f5db77
265330788508	3571348121269	Toy Forges	dbd9a5d3ffc2709be80d411f7b6494db
265319017229	3587664472089	Hickle Square	8efa87bb15ffc6f2e3b323ee64f0282d
265088309068	7449048597200	Rubie Motorway	e6df3f54206077388d6d0868f90d2702
92515150828	6517895546883	Trevor Groves	bc55b99bfb5c1b6e8159eaf7500c60cb
265116470746	4530109602753	Janet Road	423399cee67a37c6afa584f44b31f90e
265359162583	1624814396160	Joan Port	08669709e33d2604b9c4f5c60f1c9dc7
265349156796	7381182439623	Jovani Shoal	15e784ab858806d465ba1029af915b05
265165858516	4362612130360	Casimir Estate	c3886fa348b1185572976O7d637e1e59
265345592929	6572244959946	Mohr Isle	72c33bdeb212cb19f65cadd4a5ce4e65
265329943129	9494502607384	Herman Greens	ff8f466e0b60003baebc8fbeb969ee70
812821378	3183130901549	Kihn Mission	a772da7df96a2595d8394b9b8f3c2986

Figure 5.11 User Data Search

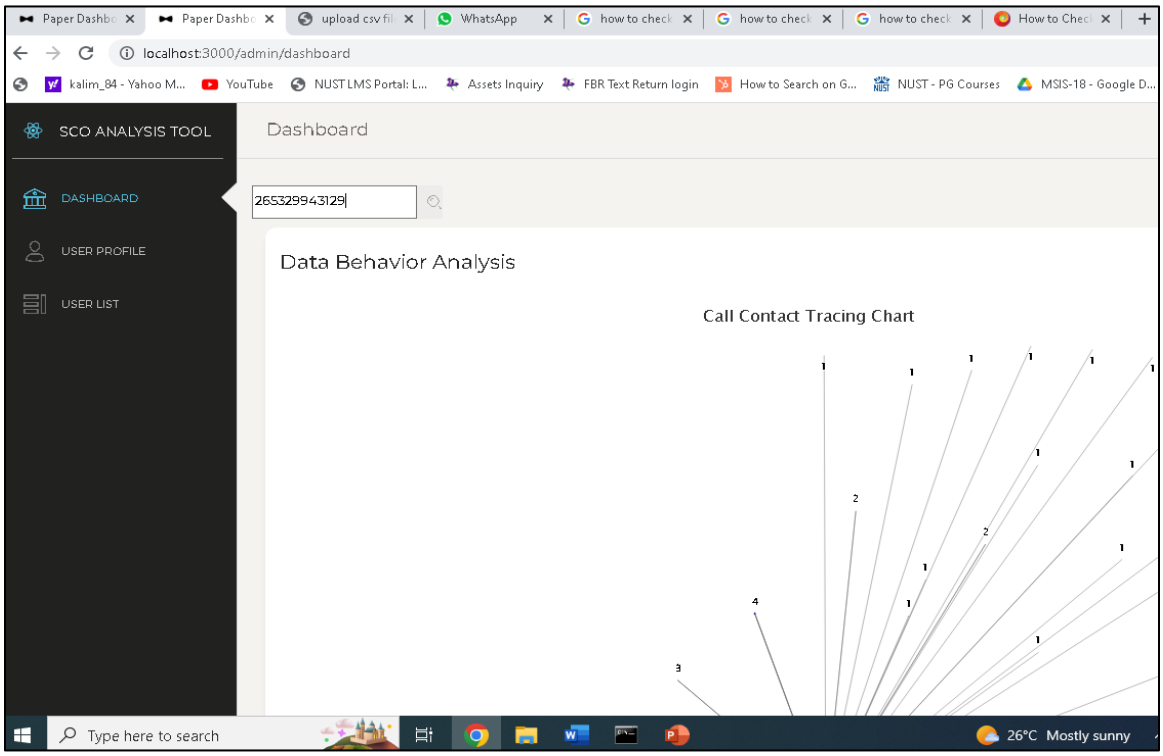


Figure 5.12 On Click Data Decryption

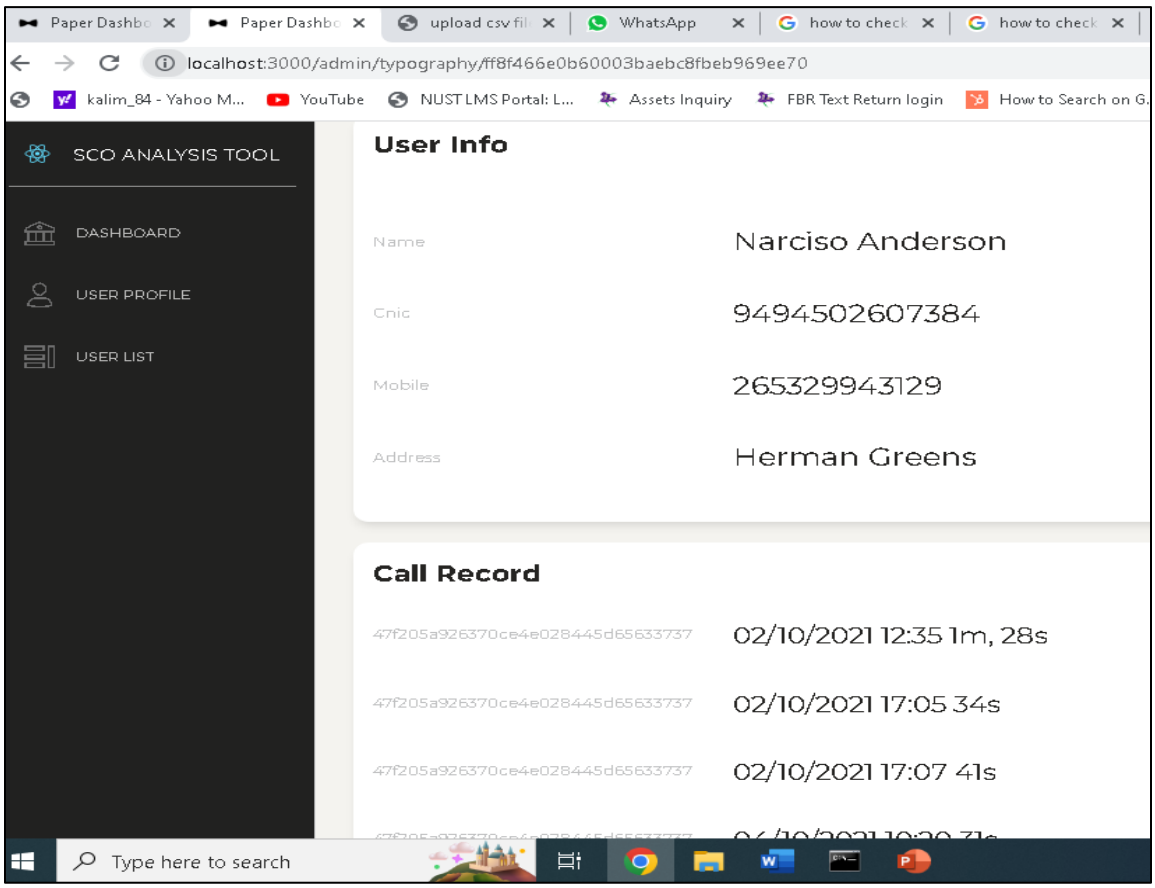


Figure 5.13 Application Load Time

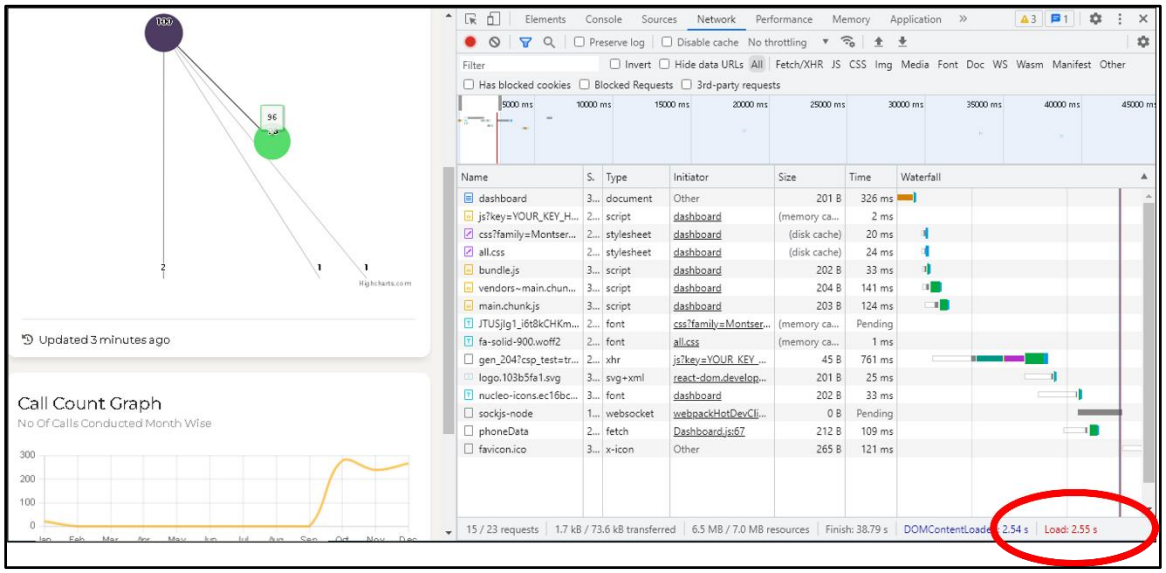


Figure 5.14 Search Time Efficiency

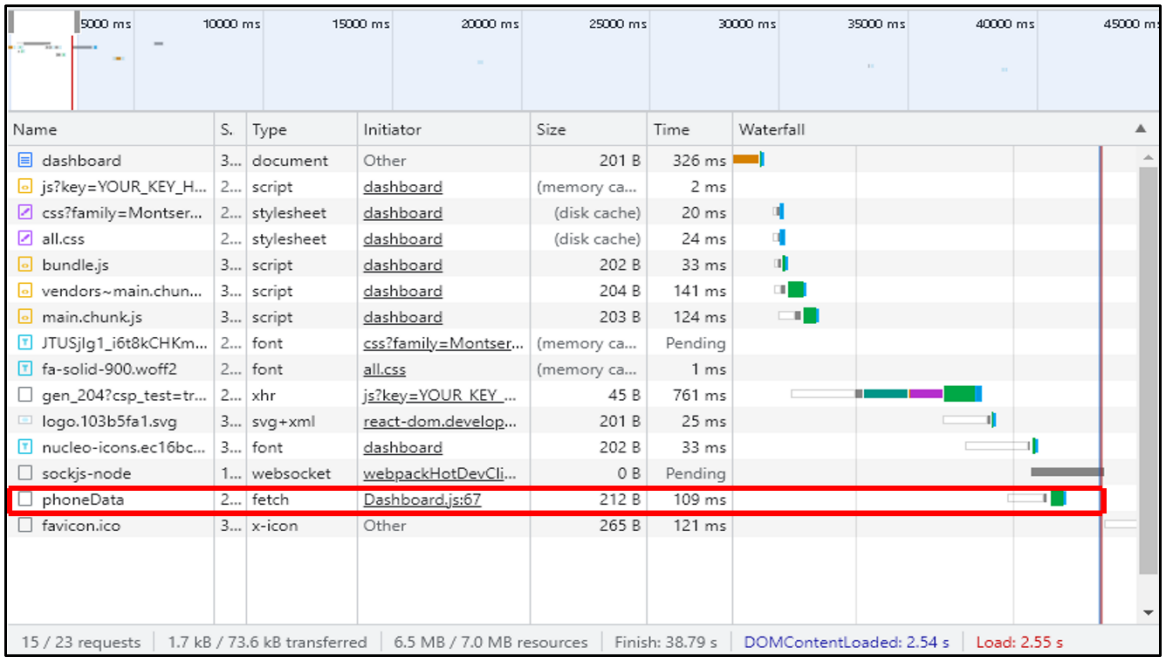
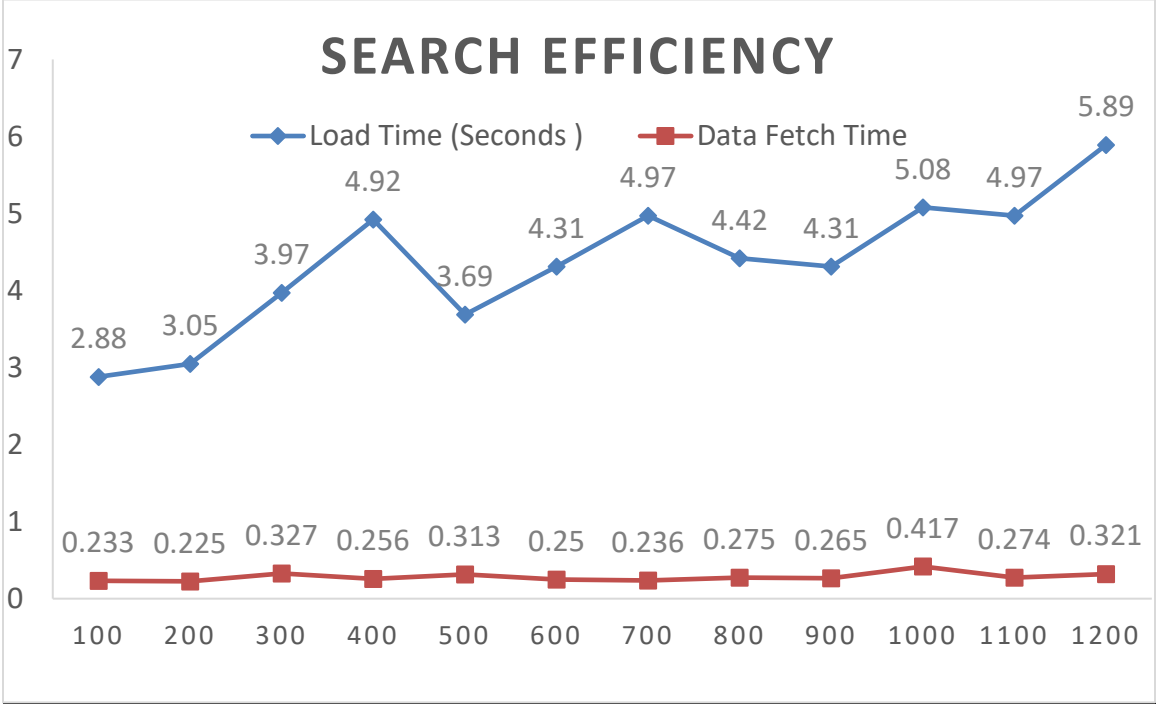


Figure 5.15 Search Efficiency Graph



5.8 Leakage Profiling

After implementation of concept and detail analysis of case scenario, following are some of the limitations observed which needs to be reviewed before implementing in actual security scenarios.

5.8.1 Structure Database Constrain

For proof of concept, the framework was adopted for a structured data of call records however, for unstructured or NoSQL data, a filtering mechanism is required to be adopted for converting unstructured data into a structured data.

5.8.2 Primary ID Dependency

Hashed primary ID is only possible access mechanism to search over encrypted data and even for creation of graphical structure for assessment of data. Thus, creation of graphical structure of encrypted data and its analysis is possible by beholder of data, however actual data extraction through reverse engineering of hashed ID and decryption of data is difficult due to dual security mechanism.

5.8.3 One Point Failure

Security of entire system depends upon the primary data base which is in direct access of the owner. If the primary database compromised, then entire data can be

compromised which necessitated the owner to keep effective security mechanism for primary Database. This drawback can be overcome by adopting a suitable backups and redundancy mechanism.

5.8.4 Integrity of Data

This framework doesn't cover the integrity of outsourced / cloud database to some extent. The hashed ID is created from primary data of customer with combination of random number and data is encrypted through encryption function, however once data is outsourced to cloud DB, then for integrity check we have to rely on the cloud traditional integrity mechanisms for checking whether data is manipulated or not. The hacker / curious CSPs can delete, recopy or association of fake primary ID with fake data possibility cannot be ruled out.

5.9 Summary

The chapter explains case scenario of call data record privacy through proposed framework. Data of call records were obtained, which was classified as per the framework guidelines. After classification of data, sensitive call record data was secured on untrusted domain. Search and analysis mechanism over encrypted database were devised through Data Behaviour Analysis application.

Chapter 6

CONCLUSION AND FUTURE WORK

This chapter provides a conclusive summary of the research work completed in the thesis. We will also see how this research can further be extended in the future.

6.1 Overview of Research

With the advancements in cloud computing, individual users and organizations are interested to outsource their personal and business-related data to the CSP. The organizations and businesses are sometime involved in handling of sensitive data of customers. The most alluring advantage cloud offers to clients and organizations are storage as a service facility. This service elevates the burden of large data management and computational cost from the users with minimum financial cost. However, data confidentiality risks are a major concern of customers till date. Customers are more concerned about privacy of sensitive nature data. Once data is outsourced to third party like cloud then it is in direct administrative control of cloud and indirect control of customers.

To address this issue, data is encrypted first in trusted domain and then outsourced to cloud as a possible solution to address the security concerns of customers. However, when outsourced data is require for some process, the entire data is brought to trusted domain first, then decrypted for final use. This process is time consuming and required a huge computational power as well. This phenomenon leaded the researcher to adopt searchable encryption. Searchable encryption is an effective solution, but resource constrain and time-consuming process.

In this research (as discussed in chapter 4), we have presented a novel solution of index based searchable encryption scheme that enables the user to search over the encrypted data stored on untrusted domain like cloud. Moreover, the scheme also provides visualization tool of polygenetic tree over the encrypted data for analysis purpose. This framework does not require to bring the data to trusted domain. Through this framework we achieved following: -

6.1.1 Data Confidentiality

Data confidentiality is achieved by adopting encryption (used AES 256 CBC Mode in application for testing purpose). The encryption key is in direct control of the user. Cloud is unaware of adopted encryption scheme.

6.1.2 Data Search

Data search over the cloud in encrypted form is possible through hashed primary key. The hash primary keys act as a lookup table in the cloud database.

6.1.3 Dual Security Mechanism

This framework provides dual security mechanism in the form of encryption and hashing. To compromise the data, hackers / curious cloud required encryption key and reversal of hashed ID which is not possible according to the pre-image resistance property of hash function.

6.1.4 Search Efficiency

The data is stored with hashed ID, which is a form of index-based search. *Song et al.* describes the fast hash lookup table as more efficient as compared to other methods.

6.1.5 Data Visualization

According to this framework data is stored on cloud based on hashed primary ID. Primary ID relationship with other IDs can easily be mapped, searched and plotted in graphical format in the form of visualization.

6.2 Conclusion

Fast progress in technology related to cloud computing, machine learning, and big data analysis, clients are relying on outsourcing their data to cloud. The cloud services providers, provides storage as a service (STaaS). The individual users and enterprises are motivated to outsource their personal and business-related data on to the servers. While on the other hand, outsourcing restricts the users as the outsourced data is out of control from the users. The confidentiality of data is no longer exist in most of the cases being consider the cloud as curious / untrusted for sensitive data. The user needs to perform normal operations over the data keeping the confidentiality of data intact. This leads us to the searchable encryption schemes for sensitive and confidential data.

In this thesis, we have designed a method of searching over encrypted data, based on index-based search through unique hashed ID. The method provides dual mechanism of confidentiality over encrypted data. This methodology also provides the advantage of creating polygenetic trees / visualization pattern of the data for analysis purpose over the encrypted data present on untrusted / curious cloud.

6.3 Future Work

Open areas for future research are

a. Data visualization over encrypted database for unstructured/ No SQL data through proposed framework.
b. Comparative analysis of the encryption efficiency for different types of data (voice, text messages, pictures etc) adopting this framework.
c. Hardware implementation of the concept (embedded device hardware designing).
d. Search efficiency analysis of different encrypted databases.

BIBLIOGRPAHY

[1] ZTE. (2022, September 29). Retrieved September 20, 2022, from https://res-www.zte.com.cn/mediares/zte/Files/PDF/white_book/Wi-Fi_6_Technology_and_Evolution_White_Paper-20200923.pdf?la=en

[2] "Pakistan Cloud First Policy" accessed on 23 Aug 2022, available on: https://moitt.gov.pk/SiteImage/Misc/files/Pakistan%20Cloud%20First%20Policy-Final-25-02-2022.pdf

[3] "Personal Data Protection Bill 2021", accessed on 23 Aug 2022, available on: https://moitt.gov.pk/

[4] Conn., S. (2022, April 19). Gartner Forecasts Worldwide Public Cloud End-user spending to reach nearly $500 billion in 2022. Retrieved August 23, 2022, from https://www.gartner.com/en/newsroom/press-releases/2022-04-19-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-nearly-500-billion-in-2022

[5] INSIDERS, C. (2022, September 10). 2022 cloud security report [(ISC)2]. Retrieved September 23, 2022, from https://www.cybersecurity-insiders.com/portfolio/2022-cloud-security-report-isc2/

[6] Group, T. (2022, June 7). Cloud data breaches and cloud complexity on the rise, reveals Thales. Retrieved September 23, 2022, from https://cpl.thalesgroup.com/about-us/newsroom/thales-cloud-data-breaches-2022-trends-challenges

[7] T. (n.d.). Cryptography hash functions. Retrieved July 27, 2022, from https://www.tutorialspoint.com/cryptography/cryptography_hash_functions.htm

[8] Says:, B. (2022, August 3). What is cryptography? types of algorithms & how does it work? Retrieved August 3, 2022, from https://intellipaat.com/blog/what-is-cryptography/

[9] Ries, D., & Simek, J. (n.d.). Encryption overview " aba TECHSHOW. Retrieved August 3, 2022, from https://www.techshow.com/2015/04/encryption-overview/

[10] Fontaine, C., & Galand, F. (2007). A survey of homomorphic encryption for nonspecialists. EURASIP Journal on Information Security, 2007, 1-10.

[11] Sönnerup, J. (2019, April 12). On the difficulty of generating random numbers. Retrieved November 20, 2022, from https://debricked.com/blog/difficulty-of-generating-random-numbers/

[12] Arslan Tuncer, S., & Kaya, T. (2018). True random number generation from bioelectrical and physical signals. Computational and mathematical methods in medicine, 2018.

[13] Sharma, O. (2019, October 11). Introduction to databases. Retrieved August 4, 2022, from https://www.c-sharpcorner.com/article/introduction-to-databases/

[14] Peterson, R. (2022, October 01). What is a database? definition, meaning, types with example. Retrieved August 4, 2022, from https://www.guru99.com/introduction-to-database-sql.html

[15] Logic, S. (2019, June 06). What is database security: Sumo Logic. Retrieved November 20, 2022, from https://www.sumologic.com/blog/what-is-database-security/

[16] Elgabry, O. (2016, September 14). Database - introduction (part 1). Retrieved November 20, 2022, from https://medium.com/omarelgabrys-blog/database-introduction-part-1-4844fada1fb0

[17] Brush, K., & Burns, E. (2020, February 20). What is data visualization and why is it important? Retrieved November 20, 2022, from https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization

[18] Mell P.M. and Grance.T. 2011. “The NIST Definition of Cloud Computing.” In Computer Security Publications from the National Institute of Standards and Technology (NIST) SP 800145.Gaithersburg: National Institute of Standards & Technology.

[19] Technolabs, C. (2022, March 24). 10 popular examples of SAAS applications [proven strategies]. Retrieved October 23, 2022, from https://citrusbug.com/blog/saas-application-example

[20] Watts, S and Raza, M. (2022, March 24). SAAS vs paas vs iaas: What's The Difference & How to choose. Retrieved October 23, 2022, from https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/

[21] AWS. (2022, October 23). What is Infrastructure as a Service (IaaS)?. Retrieved October 23, 2022, from https://aws.amazon.com/what-is/iaas/

[22] Pal, D., Triyason, T., & Padungweang, P. (2018). Big data in smart-cities: Current research and challenges. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 6(4), 351-360.

[23] Zhang, T., Wang, X., Li, Z., Guo, F., Ma, Y., & Chen, W. (2017). A survey of network anomaly visualization. Science China Information Sciences, 60(12), 1-17.

[24] Raghav, R. S., Pothula, S., Vengattaraman, T., & Ponnurangam, D. (2016, October). A survey of data visualization tools for analyzing large volume of data in big data platform. In 2016 International Conference on Communication and Electronics Systems (ICCES) (pp. 1-6). IEEE.

[25] Yang, C., Zhang, Y., Tang, B., & Zhu, M. (2019, April). Vaite: A visualization-assisted interactive big urban trajectory data exploration system. In 2019 IEEE 35th International Conference on Data Engineering (ICDE) (pp. 2036-2039). IEEE.

[26] Tahir, S., & Afzal, M. T. (2014, August). A novel phylogenetic tree data visualization application for researchers. In 2014 Science and Information Conference (pp. 93-99). IEEE.

[27] Alhanjouri, M. A., & Al Derawi, A. M. (2012). A New method of query over encrypted data in database using hash map. International Journal of Computer Applications, 41(4).

[28] Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., & Perry, R. (2016). Data visualization. International Journal of Engineering Research And Advanced Technology (IJERAT), 2(12), 11-16.

[29] Kandukuri, B. R., & Rakshit, A. (2009, September). Cloud security issues. In 2009 IEEE International Conference on Services Computing (pp. 517-520). IEEE.

[30] Kacha, L., & Zitouni, A. (2017). An overview on data security in cloud computing. Proceedings of the Computational Methods in Systems and Software, 250-261.

[31] Albugmi, A., Alassafi, M. O., Walters, R., & Wills, G. (2016, August). Data security in cloud computing. In 2016 Fifth international conference on future generation communication technologies (FGCT) (pp. 55-59). IEEE.

[32] Soofi, A. A., & Khan, M. I. (2014). A review on data security in cloud computing. International Journal of Computer Applications, 94(5).

[33] Sood, S. K. (2012). A combined approach to ensure data security in cloud computing. Journal of Network and Computer Applications, 35(6), 1831-1838.

[34] Basu, S., Bardhan, A., Gupta, K., Saha, P., Pal, M., Bose, M., ... & Sarkar, P. (2018, January). Cloud computing security challenges & solutions-A survey. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 347-356). IEEE

[35] Geng, Y. (2019). Homomorphic encryption technology for cloud computing. Procedia Computer Science, 154, 73-83.

[36] "Top 5 Cloud Security Data Breaches in Recent Years" written by GAURAV SIYAL available on: https://www.makeuseof.com/top-recent-cloud-security-breaches/ (Accessed on 25 Oct 2022).

[37] "Third Party Exposes 14 Million Verizon Customer Records" by Michael Mimoso available on: https://threatpost.com/third-party-exposes-14-million-verizon-customer-records/126798/ (Accessed on 25 Oct 2022).

[38] "Details of 44m Pakistani mobile users leaked online, part of bigger 115m cache" by Catalin Cimpanu, available on: https://www.zdnet.com/article/details-of-44m-pakistani-mobile-users-leaked-online-part-of-bigger-115m-cache/ (Accessed on 25 Oct 2022).

[39] "Top 5 cloud security breaches (and lessons)" by cybertalk.org, available on: https://www.cybertalk.org/2022/04/26/top-5-cloud-security-breaches-and-lessons/ (Accessed on 25 Oct 2022).

[40] "What is Cloud Data Security" by Gui Alvarenga, available on: https://www.crowdstrike.com/cybersecurity-101/cloud-security/cloud-data-security/ (Accessed on 26 Oct 2022)

[41] Arjun, U., & Vinay, S. (2016, March). A short review on data security and privacy issues in cloud computing. In 2016 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC) (pp. 1-5). IEEE.

[42] Kacha, L., & Zitouni, A. (2017). An overview on data security in cloud computing. Proceedings of the Computational Methods in Systems and Software, 250-261.

[43] "SURVEILLANCE UNDER THE USA/PATRIOT ACT" written by American Civil Liberties Union, available on: https://www.aclu.org/other/surveillance-under-usapatriot-act (Accessed on 26 Oct 2022)

[44] Inukollu, V. N., Arsi, S., & Ravuri, S. R. (2014). Security issues associated with big data in cloud computing. International Journal of Network Security & Its Applications, 6(3), 45.

[45] "Multi-tenant Security in the Cloud What You Need to Know" by Cloudreach, available on: https://www.cloudreach.com/en/blog/multi-tenant-security-in-the-cloud-what-you-need-to-know/ (Accessed on 26 Oct 2022)

[46] Singh, A., & Chatterjee, K. (2017). Cloud security issues and challenges: A survey. Journal of Network and Computer Applications, 79, 88-115.

[47] Rao, R. V., & Selvamani, K. (2015). Data security challenges and its solutions in cloud computing. Procedia Computer Science, 48, 204-209.

[48] "Data encryption options" by Google cloud available on: https://cloud.google.com/storage/docs/encryption (Accessed on 26 Oct 2022)

[49] "Google Cloud Platform Terms of Service" by google cloud, available on: https://cloud.google.com/terms/index-20180605 (Accessed on 26 Oct 2022)

[50] Yu, S., Lou, W., & Ren, K. (2012). Data Security in Cloud. Handbook on Securing Cyber-Physical Critical Infrastructure, 389.

[51] Joaquim, J. L. M., & dos Santos Mello, R. (2020, November). An analysis of confidentiality issues in data lakes. In Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services (pp. 168-177).

[52] Walker, I., Hewage, C., & Jayal, A. (2022). Provable Data Possession (PDP) and Proofs of Retrievability (POR) of Current Big User Data. SN Computer Science, 3(1), 1-9.

[53] Consulting, I. (2018, March 29). Art. 4 GDPR – definitions. Retrieved August 23, 2022, from https://gdpr-info.eu/art-4-gdpr/

[54] Pakistan, M. (2021, August 28). Ministry of Information Technology & Telecommunication. Retrieved August 23, 2022, from https://moitt.gov.pk/SiteImage/Misc/files/25821%20DPA%20Bill%20Consultation%20Draft_docx.pdfForte, Dario (2010). Handbook of Digital Forensics and

Investigation || Mobile Network Investigations, (), 517–557. doi:10.1016/b978-0-12-374267-4.00010-0

[55] Gibbs, K. E., & Clark, D. F. (2001). In E. Casey (Ed.), Handbook of computer crime investigation. Academic Press.

[56] “pgAdmin PostgreSQL Tools”, available on: https://www.pgadmin.org/ [Accessed on 2nd September 2022].

[57] “What-is-pgadmin”, available on: https://www.adservio.fr/post/what-is-pgadmin [Accessed on 2nd September 2022].

[58] Cloud Customer Architecture for e-Commerce., https://www.omg.org/cloud/deliverables/cloud-customer-architecture-for-ecommerce.htm (Accessed: 24 June 2021)

[59] Cawthon, N., & Moere, A. V. (2007, July). The effect of aesthetic on the usability of data visualization. In 2007 11th International Conference Information Visualization (IV'07) (pp. 637-648). IEEE.

[60] Song, H., Dharmapurikar, S., Turner, J., & Lockwood, J. (2005). Fast hash table lookup using extended bloom filter: an aid to network processing. ACM SIGCOMM Computer Communication Review, 35(4), 181-192.