CHAPTER 1

DATABASE AND SECURITY

1. INTRODUCTION

In the last decades, phenomenal growth is noticed in the amount of data available electronically, which leads to various legal and confidentiality concerns. Due to increase in the usage of cyberspace and its applications in e-Commerce and social networking it is essential to preserve the data properly. To preserve the confidentiality, in this scenario, various algorithms have been developed and accounted in literature. There are new challenges in various internet applications, related to individual confidentiality and security.

Security and confidentiality are always linked together, but actually they are different. Basically, security means that data access is controlled, and safe from attacks. It helps to ascertain confidentiality. Without security one cannot assure confidentiality. Confidentiality concern, deals with legitimate usage of personal sensitive data by authorised personnel and prevention from malicious participants during the data access and computation.

Due to increase in digitization huge amount information is available online. It can be accessed for personal and organizational benefit; this is a critical problem, which needs concern. In the collaborative computation main issues is, information and identity disclosure related to confidentiality and anonymity. Once the information is shared the control is lost, and it can be misused or harm the party by disclosing it to unauthorized party. An outcome can be secure or confidential, if no participant can learn anything other than, what can be interpreted from the result and his own input.

In various applications, a group of parties contribute their input for collaborative computation. To preserve individual sensitive data, the participants need to accomplish confidential computing. It can be implemented using trusted third party (TTP), or without TTP. In first method, the level of trust that must be employed in such TTP is undesirable, illegal, or unadvisable. So to protect application, either TTP should be removed or identity of party should be hidden. This work uses concept of hiding the individuals' identity.

1.1 NEED OF CONFIDENTIALITY AND SECURITY

Due to digitization the data inflow-outflow is increased beyond a limit. These data are kept in repositories for unspecified time durations. This data could be related to health, wealth or business. Here, the problem is relevant as, when personal data is considered individual and organizations have various reasons to protect it. For ex. an individual may wish to keep his genetic health record confidential as revealing it may lead to social negligence.

If this information is available digitally in the form of plain text, it may lead to emotional or personal outages. So it is necessary to keep the data confidential and secure. For existing confidentiality conservation techniques [*See Section 1.3*]

1.2 BACKGROUND

Now a day everything around the world is becoming smarter, as internet is omnipresent. Smart technology is going beyond imagination (medical, security, fitness, traffic management *etc*.). This technology trends leads to various concerns such as data confidentiality, security, and anonymity. There are few technology trends in data base which are widely accepted and need to be explored for organizational and individuals' growth.

1.2.1.1 DATA MINING

Data mining (DM) is a well-known process by which useful knowledge can be extracted from the raw dataset. Data mining is defined as "Knowledge Discovery from Data" or in other words, "Data Mining is a practice or a process which is used to draw useful understanding from the huge amount of dataset collected by human". Dataset can be collected from different organizations, which is stored by them for different purposes.

For example, the business organization such as amazon, flipkart etc. maintain customer data and buying pattern record for future marketing; the federal agencies for countrywide security keep records of people with felonious intent and criminal's for security reasons; and the health organization collect medical records of the patients for better treatment and medical research. Some of the major domains are Business, Education, Government and Medical.

1.2.1.2 SCOPE OF DATA MINING

When voluminous data is accessible, data mining can avail new opportunity by providing:

- Computerized Behaviour and Trend Discovery: It helps to predict large data sets automatically, which needs hand-on analysis traditionally.
- Unknown patterns discovery automatically: Tools used in data mining traverse the database to identify hidden patterns in a step. For example, Customers buying patterns, to identify the product purchased together. Other pattern discovery problems include hacker's browsing pattern etc.

1.2.1.3 ARCHITECTURE OF DATA MINING

Modelling is fundamentally the action of building a prototype in a circumstance where the outcome is known, then linking it to another circumstance that is not known. *Figure 1.1* illustrates architecture for analysis in a large data warehouse.

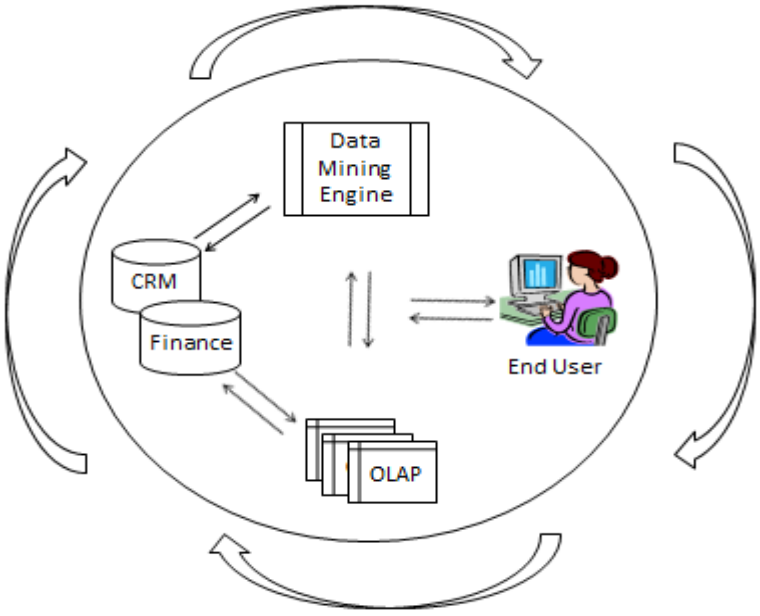

Figure 1.1: Data Mining Architecture

This design signifies fundamental change from traditional decision support systems. It delivers practically analyzed relevant information instead of just providing data to the end user via query processing and reporting package.

1.2.1.4 APPLICATION AND CHALLENGES

Earlier main focus was simply query processing over tabular data; data mining has enormous set of techniques to mine tabular data to get the desired result. In this section few application of data mining has been briefly reviewed.

Data mining techniques are relevant to any establishment, willing to process a large data warehouse for betterment of customer relationships. Essential facets for success in DM are: a huge, well- structured data warehouse with the well-defined understanding of the business process where DM is to be applied.

Some effective application areas comprise:

a) A pharmaceutical organization can study the current sales records. To meliorate targeting of general practitioners and identify the next promotion strategy which will have great impact in due course. This will include information about local health care providers and other organizations market activity. Mining the result on the basis of these two characteristics will give an idea to sales team to improve sales strategies.
b) Data Mining can be used to identify emerging trends in education system in the world. It can bridge the knowledge gap in higher education systems. For this various data mining tools could be used to extract the knowledge and this can be used to identify emerging trends in education systems, and to provide quality education.
c) It can be used to identify and schedule sports event worldwide using previous sports record.
d) A bank can process its huge depository of customer operation to recognize prospective customers to be fascinated to a newly launched credit policy.

Application of data mining is not limited to these; it can be applied in intelligence agency, e-commerce, digital library etc. Although data mining is quite beneficial for organizations and individuals, there are few

challenges in data mining such as confidentiality, data integrity, and security.

1.2.2 MULTI-PARTY COMPUTATION ENVIRONMENT

In multi-party computations, various parties in a network wish to perform operation on a common function. These parties could be a PDA, a laptop, desktop or any other input device connected within the network. Here, Network could be a system taking input from various files for computation, an intranet connecting systems or an internet where systems are connected globally.

This computation could be performed in presence of TTP [Ideal Model, *See Section 2.3*] or without TTP [Real Model, *See Section 2.3*]. This raises the concern of confidentiality and leads to Secure Multi-Party Computations.

1.2.3 BIG DATA

Big data as the name implies "Voluminous Data". It could be from various traditional and digital sources signifying source for on-going discovery and study. It is combination of unstructured data and multi-structured data. Unstructured data such as unorganized data from traditional data models, any social networking site's post or twitters tweets, and multi-structured data is combination of images and texts along with structured data such as customer transactions information and forms.

1.2.3.1 SCOPE OF BIG DATA

Big data analysis has wide scope in various real life domain such departmental stores, Bank transactions, e-commerce, social networking, health care, where pica bytes of information is generated daily, this need to be analysed for better research and education purpose.

a) Big data can benefit organization by calculating high risk region for the business, targeted high-value marketing campaigns, and to detect and prevent financial frauds.
b) Big data can be used to make smarter city, like city police can use predictive software for reducing serious crimes, or it can be used for better traffic management by optimizing traffic signals to reduce CO2 emission.

1.2.3.2 ARCHITECTURE OF BIG DATA

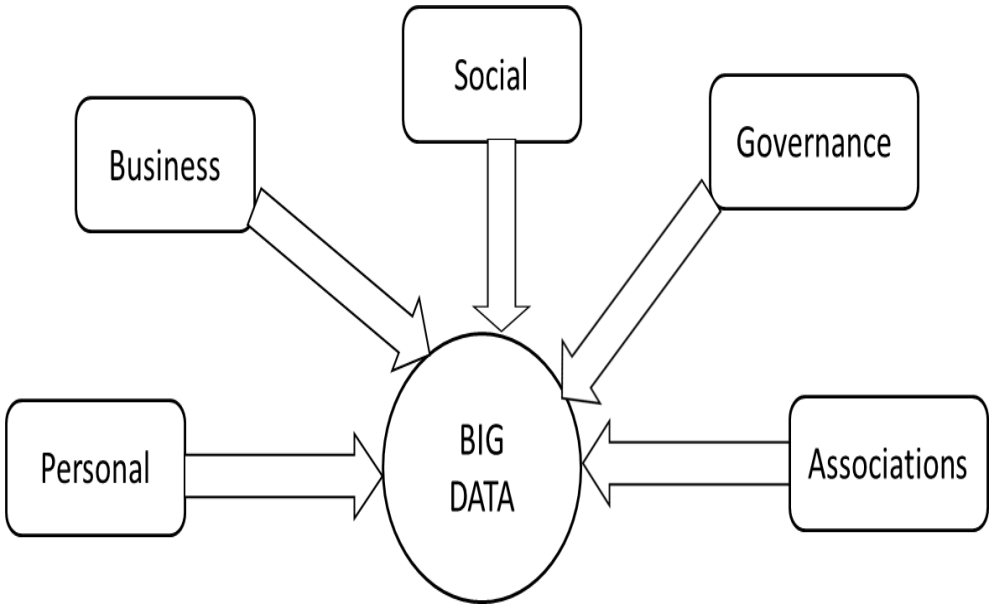

Figure 1.2: Architecture of Big Data

This design in *figure 1.2* indicates that big data has wide scope from personal, business, social, government and joint association (personal-business, social-personal). In this pica bytes of unstructured data is generated daily and it need to be analyzed for advancement in all the aspect. It provides analyzed and significant information rather than data to the end user.

1.2.3.3 APPLICATIONS AND CHALLENGES

Now a day's users are flooded with high volume of data. In many Big Data applications, data is gathered at higher rate. Earlier results were based on

estimation, or meticulously build model of the world, now it could be established by data itself. The big data analytics is touching nearly all facet of modern society constituting daily needs, e-commerce, finance, health care, life sciences etc.

a) Quality Education is one of the applications of big data as using big data different approaches adopted by institutions can be monitored thoroughly to design most effective academic approaches.
b) Quality health-care, it is broadly believed that the application of "IT" in health-care can minimize the cost of quality health-care. It can make patient care more protective by providing personalized continuous monitoring instrument.
c) Market analysis is quite easy with big data, as it adds pica bytes of customer transaction data daily to the warehouse, thereby it becomes easy for the vendor to analyse customer buying habit.

Though big data has wider application area, it does have few challenges.

a) Size- The first thing anyone observes in big data is its size. Organization of bulky and quickly growing dimensions of data has been a provocative problem for years. Earlier this challenge was satisfied by increasing processor speed. But now the scenario is changed, dimensions of data are growing much faster than computer resources.
b) Heterogeneity and Incompleteness- As data is collected from various sources there may be possibility of different database at different sites, or few fields are left NULL by some users. It makes data incomplete. This type of heterogeneous environment is difficult for data analysis. This heterogeneity and incompleteness demands,

error correction and data cleaning during analysis. This is a challenging task.

c) Timeliness- As the size of data increases; it takes longer time for analysis. Sometime query need to give quick response, for example if fraudulent behaviour in a credit card transaction is suspected than it should be identified before completion of transaction. Design of such system becomes quite challenging when volume of data is increasing at faster and response time is restricted to a limit.

d) Confidentiality- It is one of the major concern which increases with Big Data. For example data collected from location based service, in this user has to share his location with the service provider, could be a confidentiality concern for the user, such as health issue like presence of user in a cardiac centre.

1.2.4 CLOUD COMPUTING

Cloud is on-demand, location independent, online service. It provides hardware and software as service to enterprises, general public, and business markets. Clouds provide pay as you go service. It reduces overall cost of ideal resources. It is not only cost effective rather it is faster, and flexible. Security is basic issue in cloud, hybrid clouds are at high security risk.

1.2.4.1 SCOPE OF CLOUD COMPUTING

Cloud computing has wider scope from Business, Personal and Engineering perspective as it deals with services, such as Platform as a Service (PaaS), Software as a Service (SaaS), and Infrastructures as a Service (IaaS).

a) Business- Cloud computing is cost effective as in cloud computing organization need not to purchase and maintain the resources. It

provides pay-as-you-go service for organizations so overall costs incur can be reduced. Backup and recovery is much faster and easier in cloud environment. Quick deployment and automatic integration of software makes it more appealing than any other available technology.

b) Personal- It increases accessibility of client as they can access data from anywhere. It reduces licencing cost. Minimize expenditure on technological infrastructures'.

c) Society-Cloud computing helps society by providing green computing, as it decreases electricity usage, dropping carbon emission; reduce overall IT cost by use of infrastructure and software's.

1.2.4.2 CLOUD COMPUTING ARCHITECTURE

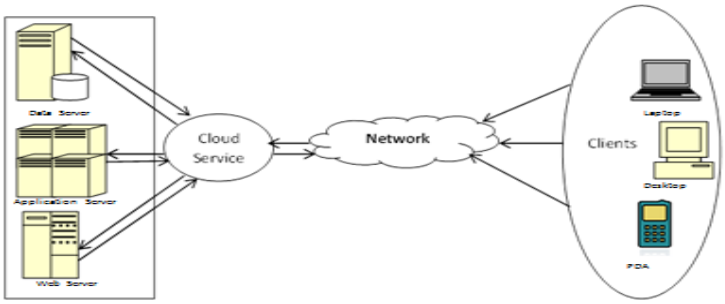

Figure 1.3: Cloud Computing Architecture

The cloud computing architecture is represented in *Figure 1.3*, it is basically divided into front-end and back-end usually connected via internet. Front-end side is for the customers or the users. Cloud is at the back-end. The front-end includes the clients, the applications and software desired to access the cloud services. For example, in-order to access web service browser such as Firefox, Internet explorer. On the back-end various computers, servers, devices, data storage systems forms "cloud". Central

server is used to administer all the activity, and fulfil clients request to run the system smoothly. Cloud keeps copy of data, to overcome failure.

1.2.4.3 APPLICATIONS AND CHALLENGES

Cloud has wide range of application from web service, health care, backup to name a few.

a) Web service- Analytical tools and web server are moving to cloud. It provide low cost infrastructure.
b) Cloud backup- Organizations are moving business and disaster recovery data to cloud server.
c) Cloud provides complete business solution within a single, fully integrated system, covering CRM, finance, marketing, e-commerce, inventory etc.
d) Meetings and conference management is one of the widely used application in IT infrastructure.

Although cloud is touching heights in IT market there are still few challenges in the technology which keep customers away from adopting at wider scale.

a) Connection Speed- Cloud computing is nothing without internet connection. If internet connection is down than you cannot access your own data, means no internet no work. If connection speed is low than it's very difficult to use cloud computing. For example if you are working with web-based application lot of bandwidth is required for download.
b) Security- In cloud computing the users give control of application, data and service to the cloud there are various security hazards, as user need to trust a third party for all of its operations. Multi-tenancy also stem to security challenges.

c) Lack of Control- Resources, application and data are placed with the cloud provider. User's identity, security policies, access control rules are managed by cloud provider. Consumers should have trust on cloud service provider for data confidentiality and security, resources availability and management.

1.3 COMPUTATIONAL TECHNIQUES

1.3.1 CRYPTOGRAPHIC TECHNIQUES

These approaches to solve SMC problems are based on basic framework of cryptography approach such as symmetric and asymmetric encryption. The solution proposed earlier using cryptographic approach is centred on various cryptographic tools such as oblivious transfer (Rabin, 1981), Yao's millionaire protocol (Yao, 1982), and secret sharing (Dorothy, 1979).

The concept of SMC was initially introduced by Yao (1982) and then carried by many researchers. Initially the computation problem was presented as combinatorial circuit; secondly the participants run a protocol for every gate to ensure security. This approach depends on the circuit size. This result has not addressed the efficiency.

1.3.1.1 SYMMETRIC ENCRYPTIONS

If the encryption key and decryption key are computationally deducible from each other, then it is called as symmetric encryption.

In this type of encryption the key for encryption and decryption is same. It means there is a need to share in advance the secret key '*k*' between sender and receiver. The essential feature of this encryption mechanism is that, the communicating parties of a secured communication use the same key for data preservation. Here the question is how to firmly accept, or exchange a suitable key to establish a secure communication link.

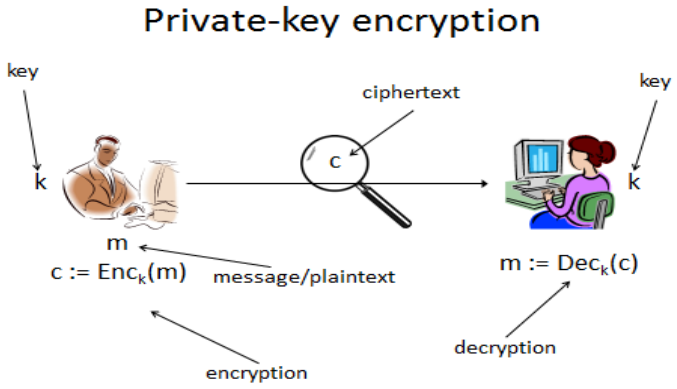

Figure 1.4: Private Key encryption for two parties

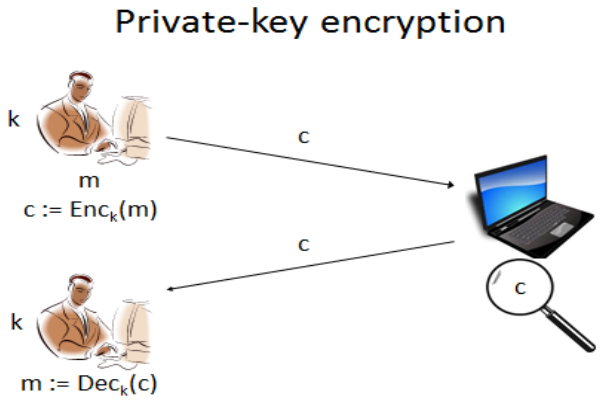

Figure 1.5: Private Key encryption for one party

The symmetric key encryption mechanism in this *'c'* is sent over the public communication channel is shown in *Figure 1.4 and 1.5*. No eavesdropper who observes the cipher text sent across the channel can figure out anything about the underlying message.

A private key encryption is defined by a message space *'M'* and three probabilistic polynomial time (PPT) algorithms (Gen, Enc, Dec)

Gen: This represents a key generation algorithm, generates the private key *'k'*.

Enc: It is an encryption algorithm, accepts the private key 'k' and a message $m \epsilon M$ as input and outputs cipher text $c \leftarrow Enc_k(m)$.

Dec: It is a decryption algorithm, accepts the private key *'k'* and cipher text *'c'* as input and outputs the message '*m'* or error §

$$m = Dec_k(c)$$

Although the symmetric key scheme doesn't tell anything about actual message to a eavesdropper there are few known limitation of the scheme: (i) The key distribution problem: As the sender and receiver share the same key, key need to be communicated to/shared with the other party using a secure channel, for this sender and receiver should be in physical proximity or there should be some trusted courier. It will be expensive and slow, here there is a probability that the courier is compromised. (ii) The key management problem: imagine the organization with *'n'* employee where each pair of employee may need to communicate securely in this case each user must store and manage *'n-1'* secret keys.(iii) lack of support to "open system" for example two users who have no prior relationship want to communicate securely.

While the private key cryptography has few drawbacks; it is still used in certain application like disk encryption. Public key cryptography is 2-3 times less efficient than private key cryptography, so if in a setting private key cryptography is an option then it is preferred to use private key cryptography. Indeed a private key cryptography is used for efficiency in public key setting.

1.3.1.2 ASYMMETRIC ENCRYPTIONS

If the encryption key is distinct from the decryption key, then it is called as public-key encryption or asymmetric encryption.

In the public key settings, a party generates a pair of key (P_k, S_k); P_k is supposed to be widely disseminated, S_k is private key, it is kept secret by

party and shared with no one. Anyone who wants to communicate with the party is able to access the public key. In this type of encryption it is assumed that the attacker is passive at least during key distribution.

In this type of encryption, the sender has publicly available keys which can be accessed by all the participants. When sender (parties) wishes to send the message to receiver (TTP), he uses the public key of receiver to encrypt the message. Receiver has the private key through which he will decrypt the message. This provides security, in a sense that only the authentic receiver can decrypt the data.

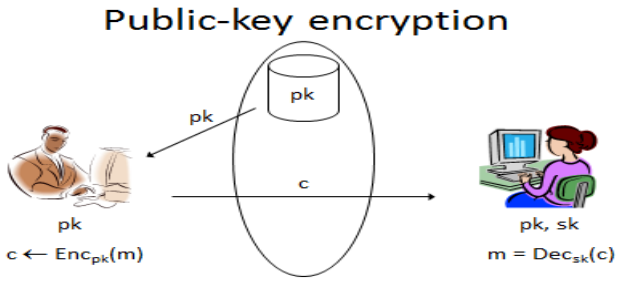

Figure 1.6: Public key encryption

A public key encryption algorithm is defined by three probabilistic polynomial time (PPT) algorithms (Gen, Enc, Dec)

Gen: It is a public key generation algorithm, generates the key pair (P_k,S_k) on input 1^n.

Enc: It is an encryption algorithm, accepts the public key 'P_k' and a message $m \epsilon M$ as input and outputs cipher text $c \leftarrow Enc_{pk}(m)$.

Dec: It is a decryption algorithm, on the key *'S_k'* and *'c'* as input and outputs the message *'m'* or error §

$$m = Dec_{sk}(c)$$

With the help of Asymmetric encryption scheme one can overcome the drawbacks of Symmetric key encryption scheme, (i) Key distribution in public key is over the public (authenticated) channel. (ii) In this scheme the user's need to manage only one key as the remaining '*n-1'* public key could be stored at a public repository or even if they are storing it locally they need not to bother about the secrecy. (iii) Applicability in "Open Systems" i.e. in this scheme party who doesn't have any prior relationship can find other's public key and use it to communicate.

Drawback

Speed: The main drawback of public-key cryptography in SMC is the time taken for encryption and decryption of packets. As the time increases the throughput for a computation takes longer and it decreases the speed of computation. The aim of this research is to protect the confidentiality and security and this mechanism assures the same but with the reduced computation speed.

1.3.2 ANONYMIZATION TECHNIQUES

Anonymization is the technique of hiding the identity of sender (party in this case). Data anonymization is the first step of providing anonymity in network environment; in this the communication is performed in stateless manner.

For example, in the web services, the users' identity or locations are sometime used to send advertisements on the basis of browsing pattern, so there is a requirement here to hide the identity to get rid of unwanted promoters. Even in sensitive cases, where multiple participants are involved for mining or computation, there is a need to ensure, that their identity is preserved then only they will share fair inputs to get the desired result.

It is very good approach for data and identity protection so that snooper cannot get the electronic trail to attack the sender.

1.3.3 PSEUDO-RANDOMIZATION TECHNIQUES

The pseudo-randomization method play a key role in cryptography, as it can be used for hiding data, digital signature and key generation etc. It is used to enable security and confidentiality in multi-party environment depending on the inconstancy of the pseudo-random numbers. Basically random number are generated in two ways: true random number generators (non-deterministic random number generators (NDRG)) and pseudo random number generators (deterministic random number generators (DRG))

An NDRG derives a sequence or number which can be generated from unordered physical practices like noise of free running oscillator or a diode, nuclear decay. NDRGs are usually relatively *incompetent* than DRGs, it takes long time to generate sequence. These are non-deterministic; means that sequence of numbers cannot be repeated, or it can be repeated several numbers of times. This technique is used when multiple outcomes are allowed each with different path, irrespective of which outcome will be selected at run time.

A deterministic pseudo-random number generator uses mathematical formulae to produce number or sequence of numbers. These are *competent*; to generate many number in short time interval, and *deterministic*; means that a sequence can be repeated if beginning of sequence is known. These are periodic means sequence repeats itself after a time interval.

1.3.4 SECURE MULTI-PARTY COMPUTATIONS

In SMC, a number of parties can cooperatively perform some global function on their private data, without any loss of data confidentiality. SMC is a new dimension of collaborative computation for joint benefit of participants. In SMC (Dorothy, 1979; Sheikh, 2010a) multiple parties carry collaborative computation on the confidential data inputs, preserving the confidentiality of participant's personal inputs. Due to socialization, higher use of Internet and huge number of wired transactions, the idea of individual data confidentiality and secure collaborative computation are matter of concern. Organizations and individuals frequently demand for collaborative computations for mutual profits, but they are at the same time worried about confidentiality of their individual input. It is because of less trust among participating and computing entities.

For example, Meeting Scheduling Problem (MSP), in this, a manager wants to schedule a new meeting for number of employees of an organization, in a programmed time. To do schedule it successfully, he want feedback from employees' personal schedule information, to evade clash with other individual programmes, and find an accessible time period in the stated time and individual preferences for employee satisfaction. Here the requirement is to find meeting date/time slot that satisfies all the constraints and approved by all the participants, at the same time without disclosing individual schedule.

SMC can be referred as, the problem when two or more parties wish to conduct a collaborative computation on the basis of their private inputs, but without revealing their personal data. SMC problem (Yao, 1982) is the problem when '*n*' parties wish to jointly compute a function on their private inputs in a secure manner, where requirement is to compute the correct result by maintaining the confidentiality of the parties as some of the parties may misuse the other party's data. It is assumed that inputs are

$x_1, x_2, \ldots, x_n$, where x_i is the data of party P_i, and the TTP will compute a function $f(x_1, x_2, \ldots, x_n) = (y_1, y_2, \ldots, y_n)$ and send the results to respective parties so that party P_i will receive only y_i and not the result of other parties. This implies the data of all parties must be secure.

Let us consider, a group of organizations working in the same domain. All of them have their own customer set, and thus its own vision of a market segment. On the other hand, in order to formulate enhanced business choices, these organizations would be willing to have a good market outline. For example, let us take the mobile phone service provider company (Sheikh, 2011), these companies wish to know total number of subscribers in a predefined region, but no one wants to reveal its number of customer to other. Various mobile phone service providers are operating in the market. Everyone has its own customer loyalty plan, which they use to collect data about the customers' calling pattern. They use that data to perform various analyses, to provide better plan which suit customer requirements. (E.g. frequent item-set mining). In this manner each company can only investigate and predict the call patterns of its own customer set and they do not have an outline of the whole mobile phone user market.

1.4 CONTRIBUTION OF RESEARCH

Initially, the existing protocols have been studied, and gap is identified to chalk out the factors, which cause the loss of confidentiality, that give generic parameters to design of protocol which reduces the confidentiality loss. Firstly, the thesis proposes a novel protocol for Secure Sum using pseudo-randomization techniques called, Joint Computation with Randomization and Anonymization (*JCRA*), in this data confidentiality is prevented using pseudo-randomization, and probabilistic mechanism is used to identify malicious behaviour of anonymizers and parties. Secondly, an asymmetric encryption based protocol (*JCAE*), has been proposed which could be applicable to generic SMC problem like finding total wealth of a family without revealing individual's wealth, total number of patient having a specific disease in a state without disclosing individual patients identity. Through intuitive analysis it is proved that proposed protocol will perform very well in certain set of situations. Thirdly, a distributed randomization based technique (*DRSS*), has been proposed in which data confidentiality has been improved by distributed randomization. Thesis mainly focuses on confidentiality, security and anonymization. Whenever users deal with confidentiality, they need to compromise with complexity issue.

1.5 THESIS ORGANIZATION

The thesis consists of 7 chapters. Chapter 1 describes about the recent technology trends in first section. Section 2, describes the terminology used in research. Different computation techniques, in this first sub-section describe the various cryptographic techniques used in SMC. The 2 and 3 sub-sections describe the anonymization and pseudo-randomization approaches for SMC. Sub-section 4, discusses SMC, and the problem

which arises in collaborative computations. Section 4 presents thesis overview. Last section discusses the contribution of research.

The chapter 2 reviews the work that has been done on SMC and includes different entities related to SMC. This chapter is divided in 7 sections. Section 1 gives introduction. Section 2, describes major concept of research. Related work and various application domains are discussed in section 3. The 4 section discusses various problems which occur during SMC. Types of adversaries are discussed in section 5. Section 6, describes scope of SMC. Section 7 gives summary of the work done so far in SMC.

Chapter 4 describes research problem and methodology used. In this section 1 and 2 defines motivation of the research, and problem definition. Objectives of research and probable design issue are discussed in section 3 and 4. In section 5 research methodology used for the research work is explained. Section 6 and 7 discusses system domain and application domain respectively. Section 8 gives the outcome of research. Section 9 gives chapter summary.

Chapter 4 describes two protocols of the research work, namely *JCRA and DRSS*. This chapter is divided in 6 sections. Section 1 is about the introduction of the chapter. Section 2, discusses secure sum problem. Section 3 gives formal definition of secure sum problem. The next section is about the proposed layered architecture (*JCRA*), its formal and informal description, and characteristics. Section 5 gives second layered architecture (*DRSS*) with its formal and informal description, and characteristics proposed. Section 6 gives probabilistic performance analysis of both the proposed protocol is presented in presence of adversaries. In last section summary of proposed protocols is presented.

In chapter 5, the proposed protocol (*JCAE*) is explained. This chapter is divided into 6 sections. Section 1 gives chapter introduction. Section 2 discusses the security mechanism. 3 and 4 section explains the architecture and characteristics of the proposed protocol. In Section 5 probabilistic performance analysis of proposed protocol is presented. Section 6 gives overall summary for the proposed protocol.

In chapter 6, simulation of protocol and results are presented. Section 1, discusses test environment and requirements. Section 2, discusses software development and algorithms. Section 3, describes simulation setup. In section 4 simulation results and graphs are shown. Section 5, describes the simulation results, graphs and snapshots. In section 6 sample codes for algorithms are given. At last summary of the chapter is given.

In Chapter 7, Conclusions and future work is discussed in section 1 and 2 respectively. Section 3 shows the Scope for future work. Section 4 gives final remarks.