INCEPTION OF DATA CREATION PHASE AND PREVENTION OF DATA LEAKAGE IN BIG DATA LIFE CYCLE

Thesis submitted in fulfillment of the requirements for
the Degree of

DOCTOR OF PHILOSOPHY

BABASAHEB
BHIMRAO
AMBEDKAR
UNIVERSITY

LUCKNOW

प्रज्ञा शील करुणा

ESTABLISHED 1996

in

INFORMATION TECHNOLOGY

by

KANIKA

Supervised by

PROF. R. A. KHAN
Department of Information Technology
Babasaheb Bhimrao Ambedkar University, Lucknow

Co-Supervised by

Dr. ALKA
Department of Information Technology
Babasaheb Bhimrao Ambedkar University, Lucknow

Submitted to

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW

FEBRUARY-2019

ABSTRACT

Big data has almost become lifeblood for most of the business to gain profit. It is the combination of popular 3V's, which defines its nature with the three basic characteristics Volume, Velocity and Variety. Big data deals with the creation, collection, storage and transformation of different formats from various sources. Google, using online search to mining large customer's data or predicting big data can be used to support medical and health related tasks including other clinical decision support, disease monitoring, and population health. Amazon may know about each and every book, a user viewed or bought by analyzing huge amount of data collected over the years. The National Security Agency (NSA) can know all phone numbers a person dialed. Facebook may and will examine big data and can tell the birthdays of persons that you did not know. With the introduction of various digital methods all this data has become big data and is still growing.

Eventually, big data technologies can improve decision-making and can provide more insights with faster results but with the negative side of data privacy loss. With the development of more advanced analytical tools for big data, increasing availability of large data sets from different sources makes it more difficult to ensure security. From the last few years, big data research has been spread worldwide. Currently, user's data is one of the most essential assets for the organizations. The constant growth in volume of data has raised a crucial and sensitive problem which cannot be managed by the traditional techniques. Big data has created new challenges linked not only to data's volume, variety or velocity, but also to security and privacy of data. However, this big data's proliferation is not without its risks. The gathered data contains personal information about users or corporate secrets which causes great harm if caught by wrong hands.

The attackers create underground markets where someone can purchase and sell stolen sensitive or personal information. This imposes the need for improved security techniques are to secure big data stored at scattered systems from such ruinous attacks. There have been great efforts to employ a wide range of mechanisms to enhance the privacy of data and thus to make environments more secure. The techniques that have been used for securing data include encryption, trusted platform

module, tokens, access control etc. However, building usable privacy-preserving systems to handle sensitive data securely is still an open problem. Existing privacy and data protection legislation demand strong privacy policy, security and transparency of data usage. In addition, prevention from data leakage with a broad range of emerging or existing security solutions to build efficient secure environments is strongly required.

Security is the basic need for the user's sensitive and personal data. The big data's enhancement is approaching to provide the secure environment. Although there are several cryptography techniques (that can secure data) available, yet due to existence issues or problems there is need of more work in this field. First contribution of this thesis is that the researcher has proposed a unique big data lifecycle which introduces various security issues and their possible solutions at each phase. The researcher is sure that the thesis will provide better understanding to the two main big data issues including misuse of personal data and unauthorized access. This study explores the relation between privacy and security of big data. The objective the research has been to collect knowledge on how adoption of big data affects the security and privacy in this connected world as well as their solution. In the context of big data, sensitive information includes data from an extensive range of various domains and areas. Data related to health or even basic information, both can be the example of sensitive information and it is clear that most of the users do not want to disclose their sensitive information. Therefore, in recent days, with the growth of big data, data privacy and security requirements are in the air to secure users against monitoring and data disclosure.

This thesis focuses on the design and development of methodologies for handling sensitive data appropriately in the area of big data. The idea behind the proposed solutions is enforcing the privacy requirements mandated by existing legislation that aims to secure user's privacy in big data environment. The researcher begins with a description of background material in addition to reviewing existing security and privacy solutions that are being used in the area of big data. It then continued to develop an improved lifecycle for big data and phase wise security threats, followed by identifying the problems on very phase of big data life cycle that are essential to be solved.

This lifecycle also addresses the security attacks on the data creation phase as well as their remedies. This is the complete lifecycle. It has five phases: data creation, data collection, data mining, data analytics and decision making phase. Every phase has its own attacks and data is travelling from one phase to another phase. To provide security at every phase is a difficult task. The researcher has tried to secure the two initial phases i.e. data creation and data collection phase. It is observed that if these two phases can be secured less effort will be required to secure further phases. Every phase is minimizing the flow of vulnerable information throughout the lifecycle.

Further, the research aims to tackle the privacy issue. This issue appears when user's personal information is shared or sold to the unauthorized party or to the third party. Therefore, the organizations collecting user's data should provide the customized privacy policies for the user. Hence, the researcher has proposed the privacy policies to address privacy and security of user's data with the aim to minimize security risks and privacy as the second contribution. In the recent years, security of user's data concerns are largely rooted in the rapidly expanding big data ecosystem conceptualized as the privileges of persons whose data is shared with others. Information security and privacy of private data have long been viewed as primary human rights. It is expected that these policies will secure as well as aware individuals about their security. These state that while asking for personal data, organizations must provide a choice to the user where they wish to share their data or not. Policies are clear and concise that clearly explains expectation from users. In the same row confusion matrix is developed to calculate the accuracy of classification. This approach has been implemented in Hadoop 2.5.2 (pre-built i386-Linux native and java version SDK Oracle Java 1.8.0 25 (8u25).

The third main contribution is a novel approach for securing Sensitive Health Information (SHI) in big data heterogeneous environments. Health data is very sensitive to any patient (data owner). The goal has been design an approach to prevent the SHI from unauthorized access. With the help of this scheme, an unauthorized user cannot access the patient's SHI without user's consent. A novel Information Leakage Prevention Scheme (ILPS) has been developed for SHI. The proposed approach aims to secure patient's SHI by improving RSA encryption technique. This approach works on data collection phase. It achieves better security in comparison with existing methods available in the area. The results of the

experiments reflect improvement in the existing RSA. Comparative analysis shows that the proposed approach is taking less time in comparison to the other existing approaches. According to the statistical results the proposed approach is acceptable. This proposed approach has been implemented in Hadoop 2.5.2 (pre-built i386-Linux native and java version SDK Oracle Java 1.8.0 25 (8u25) with the system specifications CPU Intel i7-7700, 4.2 GHz CPU and Microsoft Windows 7 as the operating system.