PROTECTION OF BIG DATA PRIVACY

Vishwanath Burkpalli[1],
[1]Department of Information Science and Engineering, P.D.A College of Engineering College,Kalburagi

Abstract

Despite big data could be effectively utilized for us to better understand the world and innovate in various aspects of human endeavors, the exploding amount of data has increased potential privacy breach. For example, Amazon and Google can learn our shopping preferences and browsing habits. Social networking sites such as Facebook store all the information about our personal life and social relationships. Popular video sharing websites such as YouTube recommends us videos based on our search history. With all the power driven by big data, gathering, storing and reusing our personal information for the purpose of gaining commercial ports, have put a threat to our privacy and security. In 2006, AOL released 20 million search queries for 650 users by removing theAOL id and IP address for research purposes. However, it took researchers only couple of days to re-identify the users. Users' privacy may be breached under the following circumstances:1.Personal information when combined with external datasets may lead to the inference of new facts about the users. Those facts may be secretive and not supposed to be revealed to others.2.Personal information is sometimes collected and used to add value to business. For example, individual's shopping habits may reveal a lot of personal information.The sensitive data are stored and processed in a location not secured properly and data leakage may occur during storage and processing phases.

Keywords: *Big data, data privacy, encryption techniques.*

1. Introduction

Big data have become a hot research topic. The increasing amount of big data also increases the chance of breaching the privacy of individuals. Since big data require high computational power and large storage, distributed systems are used. As multiple parties are involved in these systems, the risk of privacy violation is increased. There have been a number of privacy-preserving mechanisms developed for privacy protection at different stages (e.g., data generation, data storage, and data processing) of a big data life cycle. The goal is to provide a comprehensive overview of the privacy preservation mechanisms in big data and present the challenges for existing mechanisms.

Due to recent technological development, the amount of data generated by social networking sites, sensor networks, Internet, healthcare applications, and many other companies, is drastically increasing day by day. All the huge amount of data generated from different sources in multiple formats with very high speed is referred as big data. Big data has become a very active research area for last couple of years. The data generation rate is growing so rapidly that it is becoming extremely difficult to handle it using traditional methods or systems. Meanwhile, big data could be structured, semi-structured, or unstructured, which adds more challenges when performing data storage and processing tasks. Therefore, to this end, we need new ways to store and analysis the data in real time. Big data, if captured and analyzed in a timely manner, can be converted into actionable insights which can be of significant value. It can help businesses and organizations to improve the internal decision making power and can create new opportunities through data analysis. It can also help to promote the scientific research and economy by transforming traditional business models and scientific values.

1.1. Literature Survey

Big Data Security and Privacy

While Big Data gradually become a hot topic of research and business, and has been everywhere used in many industries, Big Data security and privacy has been increasingly concerned. However, there is an obvious contradiction between Big Data security and privacy and the widespread use of Big Data. In this paper, author firstly reviewed the enormous benefits and challenges of security and privacy in Big Data.

Big data: Issues challenges tools and good practices

Big data is defined as large amount of data which requires new technologies and architectures so that it becomes possible to extract value from it by capturing and analysis process. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Big data due to its various properties like volume, velocity, variety, variability, value and complexity put

forward many challenges. Here author introduces the Big data technology along with its importance in the modern world and existing projects which are effective and important in changing the concept of science into big science and society too.

Toward scalable systems for big data analytics:

Recent technological advancements have led to a deluge of data from distinctive domains over the past two decades. The term big data was coined to capture the meaning of this emerging trend. In addition to its sheer volume, big data also exhibits other unique characteristics as compared with traditional data. For instance, big data is commonly unstructured and require more real-time analysis. This development calls for new system architectures for data acquisition, transmission, storage, and large-scale data processing mechanisms. In this paper, the author present a literature survey and system tutorial for big data analytics platforms, aiming to provide an overall picture for non-expert readers and instill a do-it-yourself spirit for advanced audiences to customize their own big-data solutions, the author presented a systematic framework to decompose big data systems into four sequential modules, namely data generation, data acquisition, data storage, and data analytics. These four modules form a big data value chain In addition, we present the prevalent Hadoop framework for addressing big data challenges.

Security and privacy in cloud computing

Recent advances have given rise to the popularity and success of cloud computing. However, when outsourcing the data and business application to a third party causes the security and privacy issues to become a critical concern. Throughout the study at hand, the authors obtain a common goal to provide a comprehensive review of the existing security and privacy issues in cloud environments. They have identified five most representative security and privacy attributes (i.e., confidentiality, integrity, availability, accountability, and privacy preservability)

1.2. Problem statement

Despite big data could be effectively utilized for us to better understand the world and innovate in various aspects of human endeavors, the exploding amount of data has increased potential privacy breach. For example, Amazon and Google can learn our shopping preferences and browsing habits. Social networking sites such as Facebook store all the information about our personal life and social relationships. Popular video sharing websites such as YouTube recommends us videos based on our search history. With all the power driven by big data, gathering, storing and reusing our personal information for the purpose of gaining commercial ports, have put a threat to our privacy and security. In 2006, AOL released 20 million search queries for 650 users by removing theAOL id and IP address for research purposes. However, it took researchers only couple of days to re-identify the users. Users' privacy may be breached under the following circumstances:

1. Personal information when combined with external datasets may lead to the inference of new facts about the users. Those facts may be secretive and not supposed to be revealed to others.
2. Personal information is sometimes collected and used to add value to business. For example, individual's shopping habits may reveal a lot of personal information.

The sensitive data are stored and processed in a location not secured properly and data leakage may occur during storage and processing phases.

1.3. Proposed System

In order to ensure big data privacy, several mechanisms have been developed in recent years. These mechanisms can be grouped based on the stages of big data life cycle, i.e, data generation, storage, and processing. In data generation phase, for the protection of privacy, access restriction and falsifying data techniques are used. While access restriction techniques try to limit the access to individuals' private data, falsifying data techniques alter the original data before they are released to a non-trusted party. The approaches to privacy protection in data storage phase are mainly based on encryption techniques. Encryption based techniques can be further divided into attribute based encryption (ABE), Identity based encryption (IBE), and storage path encryption. In addition, to protect the sensitive information, hybrid clouds are used where sensitive data are stored in private cloud. The data processing phase includes privacy preserving data publishing (PPDP) and knowledge extraction from the data. In PPDP, annonymization techniques such as generalization and suppression are used to protect the privacy of data. Ensuring the utility of the data while preserving the privacy is a great challenge in PPDP. In the knowledge extracting process, there exist several mechanisms to extract useful information from large-scale and complex data. These mechanisms can be further divided into clustering, classification and association rule mining based techniques. While clustering and classification split the input data into different groups, association rule mining based techniques and the useful relationships and trends in the input data.Privacy protection in data processing part can be divided into two phases. In the first phase, the goal is to safeguard information from unsolicited disclosure because the collected data may contain sensitive information about the data owner. In the second phase, the goal is to extract meaningful information from the data without violating the privacy. We will discuss the two phases in this section.

A PPDP during PPDP, the collected data may contain sensitive information about the data owner. Directly releasing the information for further processing may violate the privacy of the data owner, hence data modification is needed in such a way that it does not disclose any personal information about the owner. On the other hand, the modified data should still be useful, not to violate the original purpose of data publishing. The privacy and utility of data are inversely related to each other and will be discussed in detail later in this section. Many studies have been conducted to modify the data before publishing or storing them for further processing. To preserve the privacy of a user, PPDP mainly uses annonymization techniques. The original data are assumed to be sensitive and private and consist of multiple records. Each record may consist of the following four attributes.

• Identifier (ID): The attributes which can be used to uniquely identify a person e.g., name, driving license number, and mobile number etc.

• Quasi-identifier (QID): The attributes that cannot uniquely identify a record by themselves but if linked with some external dataset may be able to re-identify the records.

• Sensitive attribute (SA:) The attributes that a person may want to conceal e.g., salary and disease

• Non-sensitive attribute (NSA): Non-sensitive attributes are attributes which if disclosed will not violate the privacy of the user. All attributes other than identifier, quasi-identifier and sensitive attributes are classified as non-sensitive attributes.

2. Design approach

2.1. Setup and data upload

In order to verify the data without retrieving the actual data, the client needs to prepare verification metadata. Metadata are computed from the original data and is stored alongside the original data. For practical use, the metadata should be smaller in size compared to the original dataset. The metadata are computed with the help of homeomorphic linear authenticator (HLA) or Homeomorphic verifiable tag (HVT). HLA or HVA have evolved from digital signatures like RSA and BLS. Each block stored on cloud is accompanied with an HVT or HLA tag. Current integrity verification methods also utilizes authenticated data structure like Merkel Hash Tree (MHT) . MHT is similar to binary tree, each node will have maximum of two child nodes. MHT is a tree of hashes in which leaves are hashes of data blocks.

2.2.Authorization for TPA

The TPA who can verify data from cloud server on data owner's behalf needs to be authorized by the data owner. There is also a security risk if the third party can ask for integrity proofs over certain dataset. This step is only required when client wants some third party to verify data.

2.2. Challenges and verification of data

To verify the integrity of the data, a challenge message is sent to the server by TPA on client's behalf. The server will compute a response based on the challenge message and send it to TPA. The TPA can then verify the response to and whether the data are intact. The scheme has public verifiability if this verification can be done without the client's secret key. Most of the schemes, such as provable data processing (PDP) and proofs of irretrievability (POR), support public data verification. The major issue with public verification schemes is that it can enable malicious practices. For instance, the challenge message is very simple and everyone can send a challenge message to CSS for a proof of certain _le block. A malicious user can launch a distributed denial of service (DDOS) attacks by sending multiple challenges from multiple clients by causing additional overhead and congestion in network traffic.

2.3. Data update

Data update occurs when some operations are performed on the data. The client needs to perform updates to some of the cloud data storage. Common could data update includes insert, delete, and modify operations

3. Implementation

3.1. Minimum support

Minimum support plays an important role in mining frequent itemsets. We increase minimum support thresholds from 0.0001% to 0.0003% with an increment of 0.00005%, thereby evaluating the impact of minimum support on Pfp and our proposed algorithms containing three Map Reduce jobs using both celestial spectral and synthetic datasets respectively. In this set of experiment, we increase the minsupport from $1\times10-4$ to $3\times10-4$ with an increment of $0.5\times10-4$. A small minimum support slows down the performance of the evaluated algorithms. This is because an increasing number of items satisfy the small minimum support when the minsupport is decreased; it takes an increased amount of time to process the large number of items

3.2. Load Balancing

In this group of experiments, we measure BDCaM's workload balance metric on a low- and high-dimensional datasets. In the case of high-dimensional dataset, we test our algorithms using the celestial spectral dataset. Recall that our analysis shows that the load balancing mechanism of the third Map Reduce job substantially improves the performance of BDCaM. Section V-A formally introduces the workload balance metric. We measure BDCaM's workload balance metric when the celestial spectral dataset is used as an input. We partition the initial input using the default settings. The workload balance metric defined in Section V-A does not incorporate the skewers of initial input; rather, the balance metric is measured based on the load of decomposing itemsets in the third Map Reduce job

3.3. Speedup

We evaluate the speedup performance of Pfp, BDCaM, and BDCaM-HD by increasing the number of data nodes in the.. test Hadoop cluster from 4 to 16 with an increment of 2.The celestial spectral dataset is applied to drive the speedup analysis of the three algorithms.

3.4. Scalability

In this group of experiments, we evaluate the scalability of BDCaM when the size of input dataset grows dramatically. HD when we scale up and process the dimensionality of the series of D1000W The parallel mining process is slowed down by the excessive data amount that has to be scanned twice. The increased dataset leads to a long scanning time. Interestingly, BDCaM scales better than BDCaM-HD. For example, when data size is large, the cost of reading and writing files from and to
HDFS in the BDCaM-HD case grows much faster than that ofthe BDCaM case.

4. Screen shots

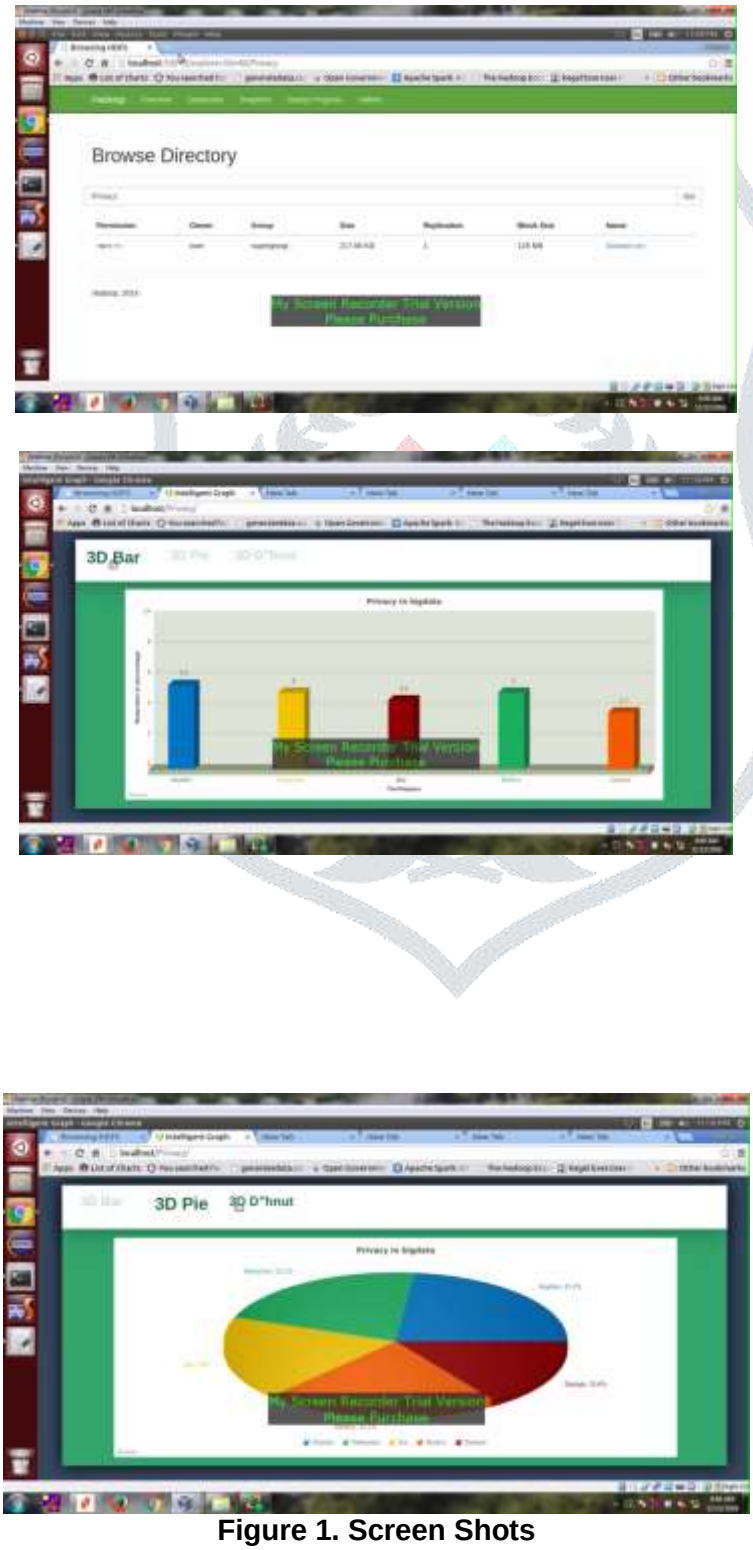

Figure 1. Screen Shots

5. Conclusions

To ensure that the data are only accessible by authorized users and for end to end secure transfer of data access control methods and different techniques are used. Data is anonymized by removing the personal details to preserve the privacy of users. It indicates that it would not be possible to identify an individual only from the anonymized data. As our personal data are gradually collected and stored on centralized cloud server over the time, we need to understand the associated risk regarding privacy. So by anonymization technique we are providing security to the big data due to that, the privacy of users data may be secured and there will be no fear of privacy violations.

References

[1] J. Manyika *et al.*, *Big data: The Next Frontier for Innovation, Competition,and Productivity*. Zürich, Switzerland: McKinsey Global Inst., Jun. 2011, pp. 1_137.

[2] B. Matturdi, X. Zhou, S. Li, and F. Lin, ``Big data security and privacy: A review," *China Commun.*, vol. 11, no. 14, pp. 135_145, Apr. 2014.

[3] J. Gantz and D. Reinsel, ``Extracting value from chaos," in *Proc. IDC IView*, Jun. 2011, pp. 1_12.

[4] A. Katal, M. Wazid, and R. H. Goudar, ``Big data: Issues, challenges, tools and good practices," in *Proc. IEEE Int. Conf. Contemp. Comput.*, Aug. 2013, pp. 404_409.

[5] L. Xu, C. Jiang, J.Wang, J. Yuan, and Y. Ren, ``Information security in big data: Privacy and data mining," in *IEEE Access*, vol. 2, pp. 1149_1176, Oct. 2014.

[6] H. Hu, Y. Wen, T.-S. Chua, and X. Li, ``Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652_687, Jul. 2014.

[7] Z. Xiao and Y. Xiao, ``Security and privacy in cloud computing," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 843_859, May 2013.

[8] C. Hongbing, R. Chunming, H. Kai,W.Weihong, and L. Yanyan, ``Secure big data storage and sharing scheme for cloud tenants," *China Commun.*, vol. 12, no. 6, pp. 106_115, Jun. 2015.

[9] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, ``Privacy-preserving multikeyword ranked search over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*,vol. 25, no. 1, pp. 222_233, Jan. 2014.

[10] O. M. Soundararajan, Y. Jenifer, S. Dhivya, and T. K. P. Rajagopal, ``Data security and privacy in cloud using RC6 and SHA algorithms," *Netw. Commun. Eng.*, vol. 6, no. 5, pp. 202_205, Jun. 2014.

[11] S. Singla and J. Singh, ``Cloud data security using authentication and encryption technique," *Global J. Comput. Sci. Technol.*, vol. 13, no. 3, pp. 2232_2235, Jul. 2013.

[12] U. Troppens, R. Erkens, W. Muller-Friedt, R. Wolafka, and N. Haustein, *Storage Networks Explained: Basics and Application of Fibre Channel SAN, NAS, iSCSI, In_niBand and FCoE*. New York, NY, USA: Wiley, 2011.