A survey on Scalable Big Data Challenges and Privacy Preservation in Cloud

D.Priyadarshini
Assistant Professor, Department Of Computer Science, Sree Narayana Guru College, K.G.Chavady, Coimbatore, India

Jamsheena Nellisseri
M.Phil Scholar, Department Of Computer Science, Sree Narayana Guru College, K.G.Chavady, Coimbatore, India

Abstract – Big Data commences under the unstable rise of overall data as an automation which is capable of storing up and technique to handle huge and diverse volumes of data, offering both activities and science with deep approaching over its clients/experiments. Cloud computing endows a trustworthy, error bearable, accessible and flexible atmosphere to handle big data dispersed management systems. Within the framework of this paper an overview of big data applications which has been applied to the cloud environment has been given. A BigData application access Cloud computing environments offers scalable infrastructure and elevated storage space, which practices enormous quantity of unstructured data in Big data applications. A survey has been taken in order to review the Big data applications over Cloud and the consequences faced. Methodologies such as proximity-aware, privacy-aware, image data, anonymization, etc. have been discussed.

Index Terms – Big Data, Cloud environment, Proximity-aware, Anonymization, MapReduce.

1. INTRODUCTION

Developing cloud services are fetching certain components of recent information and dissemination systems and move into our daily lives. Few cloud services such as Amazon's Simple Storage Service, Box.net, CloudSafe etc. access user identity, individual data and/or the locality of clients. Therefore, these cloud computing services unlock an amount of security and privacy worry. The present anxiety in cloud services is the secure and privacy-preserving confirmation of users. Users, who accumulate their perceptive information like monetary information, physical condition records, etc., have an essential right of confidentiality. There are few encoding tools and method like anonymous validation schemes, grouping signatures, zero acquaintance protocols that can both screen user individuality and offer authentication. The bringers of cloud services necessitate organizing the validation process to allow the admittance of only applicable clients to their services. Additionally, they must be able to invalidate malicious clients and expose their identities. It is observed that, hundreds of users can use cloud services at the same time. CLOUD computing and BigData, two troublesome tendencies at current, create an important bounce on existing industry and examine community. Today, a huge number of big data services are organized or transfer to cloud for data mining, processing or sharing. The prominent features of cloud computing such as high scalability and pay-as-you-go manner make Big Data necessarily available by several associations to the entire public cloud communications. Data sets in Big Data functionalities frequently enclose individual confidential perceptive data like electronic health records and monetary transaction records.

The examination of these data sets offers thoughtful approaches into a quantity of key areas of society namely healthcare, medical, etc. The data sets are frequently mutual or free to third party associates or the public. So it is necessary for sturdy protection of data confidentiality. Data anonymization operates as a chief role in privacy protection in non-interactive data distribution and discharge process. Data anonymization specifies to hiding individuality of perceptive data so that the confidentiality of a person is efficiently conserved even positive collective information can be still uncovered to data users for different examination and mining tasks. A multiplicity of privacy representations and data anonymization advances have been planned and broadly considered.

2. RELATED WORKS

2.1 Preservation in Big Data

In big data circumstances [1], numerous perceptive characteristics are habitually enclosed in data sets, while offered proximity-aware privacy models suppose only one solitary perceptive quality, either definite or arithmetic. Hence, it is believed that several insightful attributes in the privacy model, including both categorical and numerical attributes. As proximity privacy attacks twigs from arithmetical features, present proximity-aware privacy models imagine that definite featured values have no intelligence of allowable closeness.

That is, definite values are only observed whether they are accurately alike or unlike. Also, privacy models for definite features only intends at hiding accurate reform of perceptive values via preventing the sum or allocation of sensitive values without taking into account of semantic proximity. On the other hand, perceptive definite values habitually have the intelligence of semantic proximity in real-life applications since the values are regularly organized in a classification tree in terms of domain particular knowledge.

In paper [2], an improved History record-based Service optimization method, named HireSome-II based on the earlier basic one of HireSome-I, has been urbanized for privacy-aware cross-cloud service composition for dealing out big data applications. It can efficiently encourage cross cloud service work in the circumstances where a cloud declines to release all particulars of its service operation records for industry confidentiality issues in cross-cloud situation. This work estimation advent accomplishes two advantages.

Firstly, this technique extensively decreases the time complication as only some illustrative history records are employed, which is extremely required for big data applications. Secondly, this technique defends cloud privacy as a cloud is not mandatory to uncover all of its operation records, which consequently care for privacy in big data. Replication and logical consequences have established the strength of the method evaluated to a benchmark.

2.2. Big Data Privacy

To encourage the cloud computing as a resolution for big data, this paper [3] planned a proficient method to concentrate on the rising apprehension of data privacy in cloud for image data. This method segregates an image into chunks and mixes up the chunks with hit and miss origin point and casual step. This system manages at the chunk level as an alternative of the pixel level, which significantly rapids up the calculation. The image privacy problem has been converted into the jigsaw puzzle problem. To make the jigsaw puzzle problem NP-complete, the tailored each pixel of the image data is take into account by accessing a random one-to-one mapping function. These procedures make the pair wise similarity un-reliable and make the shuffled image un-recognizable. The implementation of this method is done onto the real networks (including the Amazon EC2) and experienced the security and effectiveness.

In paper [4], the quality of anonymization used k-anonymity based metrics has been intended. While it is felt that this is an essential initiate, other hits survive than is an origin for privacy failure even with practical k-anonymity values. Additional procedures of privacy, like l-diversity can be used to enhance privacy and it are necessary to explore them. Furthermore, the existing approach does not appear at data longitudinally and does not cover chronological incidences that can expose users. Implementation of Hadoop has been used to evaluate the anonymized data and attain positive results for the Human Factors analysts. A number of disclosures have been found out. First, it is not imagined that prepared activity data like a web server log file would have modifiers dispersed all over a record, opposed so much from the canonical case. The reference pointer technique has been used to translate entries into the canonical case without loss of information. Second, the browser and user agent information would be so personally attached to persons. Third, it is found that anonymization tools intended for the activities usually did not appear to believe the excellence of anonymization and whether an anonymized data set was weak to connection attacks.

2.3 Multi-Objective Approach

In paper [5], the development of privacy-preserving multifactor validation system has been incorporated without beginning of any additional objective mechanism for cloud systems operating big data characteristics. In the system a technique called MACA (Multi-Factor Cloud Authentication) which includes the first feature a password and the second feature a hybrid user profile has been used that recapitulates the user behavior.

MACA spots on the privacy protection of the second feature, which has two advantages over formerly projected systems. First, user confidentiality is not disclosed to everywhere cloud computing environment with FHE and fuzzy hashing. Second, the hybrid user profiling representation is extremely utilizable and integrates a lot of characteristics and related data, which facilitates simple privacy-preserving Multi-Factor Authentication (MFA) operations with Fully Homomorphic Encryption (FHE) and fuzzy-hashing computations. One can always alter the attribute list for user profiling in MACA corresponding to the actual circumstances.

In paper [6], a new approach has been proposed in cryptography by association between evolutionary cryptography and homomorphic cryptography. The mixing of these two criterions can improve the protection level of the other. Hypothetically, it seems that the encryption of a datum numerous times will generate more security but still need more investigation in terms of complexity of computing and complexity against attacks.

2.4 Privacy Challenges in Big Data

In paper [7], an investigation done on the extent nervousness of multidimensional anonymization over big data on cloud, and intended a flexible MapReduce based method. It has been determined that the flexibility problems of discovering the median appropriate to its core role in multidimensional division, and have planned an extremely flexible MapReduce based algorithm of discovering the median by means of utilizing the scheme of the median of medians and the scatter graph system. Collaborative of divergences has been commenced as a search measure to direct the collection of dividing the measurement for generating fine division occupancy.

It has also analyzed extremely flexible MapReduce work to compute collection of dissimilarity and to divide datasets. Recursion granularity of multidimensional partitioning has been studied to gain good balance among cost-effectiveness and efficiency.

In paper [8], a consideration regarding the disputes mounted when constructing systems which need at the same time a wide level of user-profiling and an elevated level of user privacy. Structuring and releasing fine-grained user profiles can be extremely efficient in affording excellence suggestions. This is specifically accurate when suggesting long-tail data. Big data methodologies play a vital role in assembling this summary even more exact. On the other hand, this elevates the problem regarding a specified user's privacy. Big data may even enhance this risk by affording attackers the means of avoiding privacy protective actions. Author had demonstrated these issues by establishing the disputes mounted by EEXCESS, aiming both to afford high quality suggestions and to value user privacy.

2.5 Issues in Big Data

In paper [9], the region of privacy in big data framework which comprises of a group of key problems has been addressed by research. Many of these problems do not branch from technical issues, but purely are based on regulation and organizational affairs. However, it can be predictable that it is possible to gather each of the issues regarding with correct technical measures. For example, keeping track of a specific individual's data all over big data analytics contexts is purely an organizational necessity that can be meeting up by means of logfiles. Association of displace datasets can frequently be executed without hoping on linkage via user's identities, but depending on other types of data. In the similar way, many types of data can be preprocessed with proper *anonymization* or *pseudonymization* previous to distribution, such that linkage of datasets remains possible, but linkage to an individual's identity becomes tough.

The paper [10] examines the characteristics of the Cloud and presents the opportunity to take care it as a latest type of Public Utility, specifically Information Utility. This theory should be discarded, because there are vital variations in the organization, in spite of obvious comparisons in service characteristics. Instead, this paper highlights the requirement of caring privacy as an industrial norm.

Taking into account the long convention of free market for computing trade, self-regulation is fundamentally desirable to government guidelines. But from an unusual viewpoint of "nudge", a hybrid mixture of libertarianism and paternalism, this paper concludes by offering short recommendations with fair contract terms as well as individual confidentiality from privacy.

Table 1.0: Summary of Big Data application in Cloud environment

Paper Number	Technique	Advantages	Disadvantages
1	MapReduce	More powerful, elastic and cost-effective	Lots of data are shuffled, Unsuitable for short online transactions, computation of a value depends on previously computed values MapReduce is not applicable.
2	K-means clustering	Faster than hierarchical, produce tighter clusters	Difficult to predict K-value, with group cluster it didn't work well, Different initial partitions can result in different final clusters
3	AES encryption and decryption	Good performance , secured, easily implemented and ubiquitous.	Uses too simple algebraic structure, every block is encrypted in same way, AES in counter mode is complex.
4	Anonymization	Effectiveness, accuracy, active data and published	Hard to prevent attackers, API's can be abused by

		data, no key needed.	malicious third parties.
5	Fuzzy Hashing,	user privacy is not leaked to ubiquitous Cloud computing environment , highly usable and configurable .	Weighted performance must be improved
6	Hommomorphic Encryption	High protection	complexity of computing and complexity against attacks has to be improved
7	Multi-dimensional anonymization	Cost-effectiveness and efficiency	Scalable privacy preservation of data has to be concentrated
8	Network analysis, sentiment analysis	Gains information from user friends, co-workers, etc., determines opinion and subjectivity of users	Reliability has to be increased, More complicated
9	Pseudonymization	Can be managed at earlier stage, Owner is responsible for Data source, Data privacy challenge can be solved for data stored.	Individual solution per data/source is required.
10	Demarcation	Self-regulation is promising, effective.	Protection has to be concerned.

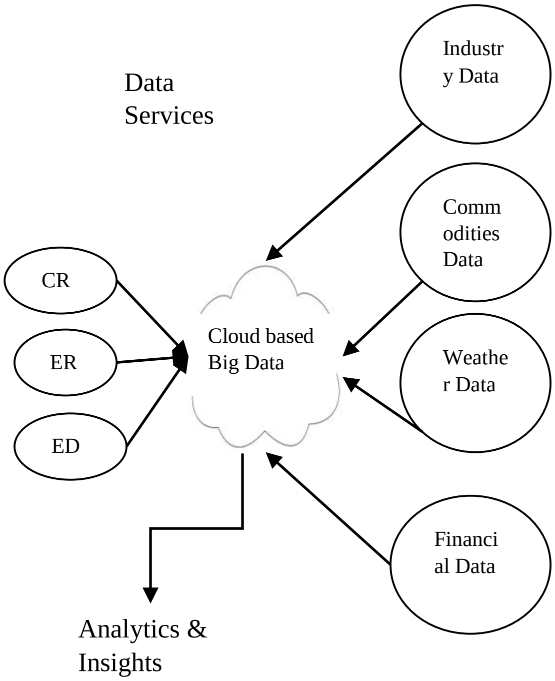

Figure 1.0: Representation of Big Data in Cloud environment

3. CONCLUSION

Data processing on a cloud based cluster provides additional features such as error bearable, varied, ease of use, unlock, proficient, endow with performance and “tool plug-ability” which most DBMS do not provide. Big Data applications such as MapReduce and Hadoop seem to be an efficient one. Big Data applications focus on the privacy preservation mechanisms when it has been applied to the Cloud environment. By keeping all the above concluding remarks of this paper, design and development of, MapReduce techniques work well with the time-cycle reduction and handles multi-variety data. Methodology which involves encryption and decryption could be enhanced. The outcome of Big Data application on Cloud involves lot of privacy preserving methodologies and security concerns. Cloud Computing and big data accepts massive consideration globally due to diverse business-driven assurances and outlooks such as lower upfront

IT costs, a faster time to market, and opportunities for creating value-add business.

REFERENCES

[1]. Zhang, Xuyun, et al. "Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud." *IEEE transactions on computers* 64.8 (2015): 2293-2307.

[2]. Dou, Wanchun, et al. "HireSome-II: Towards privacy-aware cross-cloud service composition for big data applications." *IEEE Transactions on Parallel and Distributed Systems* 26.2 (2015): 455-466.

[3]. Huang, Xueli, and Xiaojiang Du. "Achieving big data privacy via hybrid cloud." *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*. IEEE, 2014.

[4]. Sedayao, Jeff, Rahul Bhardwaj, and Nakul Gorade. "Making big data, privacy, and anonymization work together in the enterprise: experiences and issues." *Big Data (BigData Congress), 2014 IEEE International Congress on*. IEEE, 2014.

[5]. Liu, Wenyi, A. Selcuk Uluagac, and Raheem Beyah. "MACA: A privacy-preserving multi-factor cloud authentication system utilizing big data." *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*. IEEE, 2014.

[6]. Rahmani, Amine, Abdelmalek Amine, and Reda Hamou Mohamed. "A multilayer evolutionary homomorphic encryption approach for privacy preserving over big data." *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on*. IEEE, 2014.

[7]. Zhang, Xuyun, et al. "A mapreduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud." *Cloud and Green Computing (CGC), 2013 Third International Conference on*. IEEE, 2013.

[8]. Hasan, Omar, et al. "A discussion of privacy challenges in user profiling with big data techniques: The EEXCESS use case." *Big Data (BigData Congress), 2013 IEEE International Congress on*. IEEE, 2013.

[9]. Jensen, Meiko. "Challenges of privacy protection in big data analytics." *Big Data (BigData Congress), 2013 IEEE International Congress on*. IEEE, 2013.

[10]. Hayashi, Koichiro. "Social issues of big data and Cloud: privacy, confidentiality, and public utility." *Availability, Reliability and Security (ARES), 2013 Eighth International Conference on*. IEEE, 2013.