J. Glob. Innov. Agric. Soc. Sci., 2015, x(x): xxx.
ISSN (Online): 2311-3839; ISSN (Print): 2312-5225
http://www.jgiass.com

CHALLENGES OF MAPREDUCE IN PRIVACY AND SECURITY IN BIG DATA

Sara Masood, Rabia Zafar, Samia Akbar,Zahid Javed
Department of Computer Sciences, Agriculture University, Pakistan

***Corresponding Author's email: saramasood@gmail.com**

Abstract:- MapReduce paradigm is highly scalable which allow it to process large volume of dataset by distributing it over large number of nodes, its scalability make it key approach to solve Big Data problems. The exposure of sensitive data is a big threat to personal and organizational security. So protection of sensitive data is most prominent issue to handle large scale security sensitive datasets. In this paper map reduce challenges in handling big data security issues are identified for better planning to cope with security issues. Moreover, current efforts and solutions to address security challenges by extending and improving map reduce is presented. Consequently, by identifying challenges of MapReduce in handling big data security issues, open horizon for new research in this field.

Keywords :-Big Data, Map Reduce ,Accountability , Access Control, privacy preservation; Big Data, Homomorphic

INTRODUCTION

Traditional approaches to store and process data are inadequate to address big data challenges, Growing trend in developing web, sensors, social media and mobile devices results in the explosion of data set sizes, for example Facebook have more then 1 billion users, with 618 million users adding more than 500 terabyte of new data each day[1],Consequently adding more complexity to handle data set by traditional approaches leads to many security issues.

Big data is characterised by 4,vs [10] volume, velocity, variety, and veracity. Volume refers to large and complex data sets, velocity refers to fast generation and processing of data, Varity refer to diversity of data set, generally these data sets are from different sources and are of different types such as half structured medical records and business transactions,[6] and varsity is ability ensure that data s reliable when making critical decision on data.[10]. The Map Reduce is widely adapted by business and organization to process large volume of dataset.[11].

MapReduce privacy concerns are aggravated, because the sensitive information are spread in various data sets, [6] and exposure of personal and origination, confidential data is a big threat. This paper identify challenges that MapReduce faces handling privacy issues of Big Data.[6].Security challenges comprises Auditing, Access control, privacy , Authoriser Access. Recent research on privacy issues of MapReduce is originated, Different mechanism such as access control, differential privacy, and auditing is subjugated to achieve data security in MapReduce .[6] These mechanism are four pillars of data security and privacy in MapReduce frame work, and still have open question for security challenges of MapReduce in big data.[10]

Table 1 describe main challenges of map reduce in terms of security and privacy. While detail of each challenge is discuss in section sections III to V. Moreover, current efforts and solutions to address security challenges by extending and improving map reduce is presented. Consequently, by identifying challenges of MapReduce in handling big data security issues, open scope of further improvements.

This paper is organized as follows: Section II introduces the MapReduce (MR) paradigm. Section III identifies Auditing challenges while Section IV discusses Big

Access control issues. Section V discusses and privacy and security challenges in Section VI Unauthorised access .Finally, Section VII concludes the paper.

Table 1: Over view of Challenges of Map Reduce

Challenges	Solution
Accountability and Auditing	Trusted third party monitoring, security Analytics [9]
Access control	Mandatory access control approach with semantic understanding [9]
privacy	Privacy policy enforcement with security to prevent information leakage[9]
Data protection and Encryption	Encrypting algorithm schemas, homomorphic schema

MAP REDUCE

The major advantage of the MapReduce Paradigm is its high scalability which allow parallelized and distributed execution on large number of nodes. In MapReduce model map or reduce task is divided in to high number of jobs, these jobs are assigned to nodes in the network. MapReduce is fault tolerance and reliable, if a job failed on any node , job is re assigned to other node in this way reliability is achieved. Popular implementation of MapReduce is Hadoop which implement it on top of Hadoop Distributed File System (HDFS).[9]
MapReduce is a processing technique and programming paradigm designed for processing large data sets in distributed environments [3]. Map Reduce algorithm comprise two important tasks, namely Map and Reduce.

2.1 Map :Map takes a dataset and convert it in to (key, value) pairs .In other words it performs filtering and sorting,

2.2 Reduce: Reduce function take input from Map and carries aggregation an grouping operations.[9].

2.3: MapReduce flow.

In map reduce one node is assigned as master, which is responsible to assign work to different workers, Input data is divided in to small chunks or splits and master is responsible for assigning splits to Map workers. Worker generates key/value pairs and writes them to disk or in memory. Once data is on intermediate file(disk or memory) Master notify *Reducer workers* about location of intermediate file, reducer read data from intermediate storage , perform processing according to reducer functions and finally output data to output file[9].

Accountability and Auditing

Accountability and auditing are security one of the prominent issues that Present in both MapReduce and Big Data. Accountability means to track when someone perform the action and who is responsible for that action, and it is generally traced through auditing.[1] accountability is MapReduce is provided when mappers and reducer are held responsible for the task they have completed.[1] One solution is purposed by creating Accountable MapReduce[1], which utilize set of auditors and perform real-time accountability tests as an audit trial on mappers and reducers , as a result malicious mappers and reducers can be identified and on that ground accountability is provided [9]

ACCESS CONTROL

User access control is very important on access sensitive and personal data.[13].Providing access control to large volume of access sensitive data is one of the big challenge face by MapReduce and big data , it is explain by “Big Data” 3v prosperities: volume, variety and velocity[2]. When dealing with large dataset information’s are scattered on multiple storage locations, and to perform work on that information require access to multiple storage locations and devices.[9].Therefore, multiple access requirements are required to perform a task. When handling large diverse data set semantic understanding is required to provide access control efficiently.[2].finally velocity require that whenever access control approach is used, it must be optimized to determine access control rights in a reasonable amount of time.[9]

Basically access control fail to maintain data privacy, because if data user access unencrypted data, they can understand privacy sensitive information's.Roy et al. [9] examine this problem caused by MapReduce and present a system *Airavat,* that provide access control by incorporating a heuristic approach called differential privacy. To ensure privacy preservation , mandatory access control is triggered when privacy leakage increase a threshold, though noise is mixed in the result produced by the system, which is not suitable in many applications, e.g. medical experiment data mining and analysis.[6]

Data protection and Encryption

Another major challenge faced by map reduce and big data as well as cloud computation, is security of intermediate data, and does not allow operation on ciphertext. [15]
Map reduce take set of input of key/value pair and create set of intermediate key/value pair, then map reduce assemble the same key and pass data sets to reducers, which then aggregate data.in this process intermediate files are not protected[14] .To get trust worthy map Reduce there is need to not only protects intermediate data but also allow specific computations on encrypted intermediate data.[14].if we encrypt large datasets, it is a challenge to carry out processing on encrypted datasets efficiently, because most of the applications run on unencrypted data sets[15],However a novel approach namely homomorphic encryption (FHE),allows performing operations on encrypted data as well as gain protection of data against malicious users, but implementing these schemas are expensive due to their efficiency. [14, 15],Another Solutions is instead of encrypting dataset Puttaswamy et al. [16] present a set of tools called *Silverline ,that* works by separation cloud based data from all encrypted data for privacy preservation ,and unencrypted data for applications functions.

PRIVACY

When dealing with large data set privacy measurement is major topic of concern. Data linkages can be deduced and discover in processes of predictive analytics and data Mining . Data linkages are very beneficial to organization, it allows them to provide better understanding of target and provide to their client and users. Though, on individual level this discovery of information causes data provider's identities to be exposed.
A level of control is required to maintain over the personal information's. This control can be provided via transparency and allowing inputs from the data provider. Taking inputs from user allow them to define level of privacy to their information on their wishes. Transparency provided to an individual can be described by following .

- How private information is collected?
- What private information is collected?
- How the private information is being used? And
- Who has access to information?

When there is large number of mappers and reducer it is difficult to provide transparency to individual users. It is feasible that the ability to provide transparency and control is stated in legislation that must strictly be followed to avoid penalties.[9] Example issues are stated below that lead to penalties ,using the example of Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada [3], and Data Protection Directive of the European Union [4]:

- Those individuals who have data collected on them are Able to understand. How data is being used, and by whom, and for what purpose. Such legislations is difficult to apply on large data environment.
- In some circumstances permission must be given before data can be used. During map reduce analytics on large data it is very difficult to inform individual users about what happening to their data and to request access permission is a big challenge.
- If individual withdraw permission to use information's, it must be deleted from data repository. though, in Big if information is put into the system it is hard to remove .[9]

To provide privacy and protection to map reduce some efforts are being made. Airavat [5] is designed to provide protection to MapReduce. It enables the execution of trusted and untrusted MapReduce computations on access sensitive dat. While enforce privacy policies belonging to data providers[5].Airavat divide the MapReduce process into two parts, the un trusted mapped code, and the trusted reducer code. Disadvantage of the Airavat solution is the mandatory use of an Airavat provided Reducer, which reduces its ability to operate in any domain. While this initial approach has shown some promise, there is still room for improvement.

CONCLUSION

In big data community, volume of data is rapidly increasing ,traditional data processing and storage approaches are facing many challenges in meeting the continuously Increasing computing demands of Big

Data. This paper focused on MapReduce, one of the prominent approaches for meeting Big Data demands by providing highly parallel processing on a large number of commodity devices, also we identified most of challenges face by MapReduce in terms of security. It is categorized as following: Accountability and Auditing, access control, Authentatication and privacy. Moreover, efforts to improving and extending MapReduce to identified and address challenges are presented. By identifying security and privacy MapReduce challenges in Big Data, this paper facilitates for better planning of Big Data projects and encorsge for future research.

REFERENCES

[1] Z. Xiao and Y. Xiao, "Achieving accountable MapReduce in cloud computing," Future Generation Computer Systems, 30,pp. 1-13,2014.

[2] W. Zeng, Y. Yang and B. Luo, "Access control for Big Data using data content," IEEE International Conference on Big Data, 2013.

[3] The Personal Information Protection and Electronic Documents Act (PIPEDA), http://www.priv.gc.ca/leg_c/r_o_p_e.asp.

[4] Protection of Personal Data, http://ec.europa.eu/justice/dataprotection.

[5] I. Roy, S. T. Setty, A. Kilzer, V. Shmatikov and E. Witchel, "Airavat: Security and privacy for MapReduce." Proc. of the 7th Usenix Symposium on Networked Systems Design and Implementation,2010.

[6] X. Zhang, C. Liu, S. Nepal, W. Dou, and J. Chen, "Privacy-preserving layer over MapReduce on cloud," *Proc. - 2nd Int. Conf. Cloud Green Comput. 2nd Int. Conf. Soc. Comput. Its Appl. CGC/SCA 2012*, pp. 304–310, 2012.

[7] C. Gentry, "Fully Homomorphism Encryption Using Ideal Lattices," in *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC'09)*, pp. 169-178, 2009.

[8] K. Grolinger, M. Hayes, W. a. Higashino, A. L'Heureux, D. S. Allison, and M. a. M. Capretz, "Challenges for MapReduce in Big Data," *Proc. IEEE 10th 2014 World Congr. Serv. (SERVICES 2014)*, 2014.

[9] F. Ohlhorst, Big Data Analytics: Turning Big Data into Big Money, Hoboken, N.J, USA: Wiley, 2013.

[10] J. Dean and S. Ghemawat, "Mapreduce: A Flexible Data Processing Tool," *Communications of the ACM*, vol. 53, no. 1, pp. 72-77, 2010.

[11] S. Chaudhuri, "What Next?: A Half-Dozen Data Management Research Goals for Big Data and the Cloud," in *Proceedings of the 31st Symposium on Principles of Database Systems (PODS'12)*, pp.1-4, 2012.

[12] Y. B. Reddy, "Access Control for Sensitive Data in Hadoop Distributed File Systems," no. c, pp. 72–78, 2013.

[13] X. Chen and Q. Huang, "The data protection of mapreduce using homomorphic encryption," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, pp. 419–421, 2013.

[14] N. Cao, C. Wang, M. Li, K. Ren and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," in Proceedings of the 31st Annual IEEE International Conference on Computer Communications (INFOCOM'11), pp. 829-837, 2011.

[15] K.P.N. Puttaswamy, C. Kruegel and B.Y. Zhao, "Silverline: Toward Data Confidentiality in Storage-Intensive Cloud Applications," in *the* 2nd ACM Symposium on Cloud Computing (SoCC'11), Cascais, Portugal, October 27-28, 2011.