The Prototype for Implementation of Security Issue in Big Data Application using Hadoop Server

Shalini Singh
PG-Scholar
IET DAVV, Indore

Meena Sharma
PhD, Computer Department
IET DAVV, Indore

ABSTRACT
A large amount of data can be referred as BigData. A vast size of data requires special kind of methodology to process and store. BigData research consortium team developed a distributed server known as Hadoop Server, to divide and partition large data into multiple pieces for fast and efficient processing. Hadoop is an open source solution developed by Google Corporation for large data processing. It supports variety of components and distributed file system. MapReduce, Pig, Hive are the components used for efficient development of software, together with Hadoop Distributed File System which is responsible for storing and processing large data with multiple nodes. The complete study observes that advance level of processing is required for large data scale, thereby to accomplish level of concert. In order to circumvent problem of privacy leakage and access maintenance, an elucidated security model has been developed for BigData application. This paper describes the security issue along with its solution. The proposed solution is implemented with Hadoop server in single node and multinode environment.

Keywords
Big Data, Hadoop, Hive, Sqoop, MapReduce, RSA Cryptographic.

1. INTRODUCTION
A very large amount of data is being created and processed on the internet and this amount is increasing exponentially day by day. To store and retrieve such an amount of data efficiently we have to acquire faster techniques, comfort by BigData [1]. Therefore the provision of data security and data confidentiality is still required on the user level. Privacy mean a user who is not required to see the information, should not be able to access it. Security is an outsider should not be able to see the information stored. The core of Hadoop [2] is MapReduce framework, created by Google to solve the problem of web search indexes. The nonprofit organization Apache Software Foundation (ASF) maintains and manages Hadoop framework and Hadoop environment technology. The framework such as Hive, Pig, NoSql, MongoDB [3], and several other are announce in BigData environment to manage massive amount of sensitive data at any given time. Several technologies related to Hadoop includes, the HDFS which is used for distributed file system. The Hive is a data warehouse implementation for Hadoop. The MapReduce is a programming model of Hadoop. The Pig is used for querying language in Hadoop, which is similar to SQL language but SQL is used for relational database. The Sqoop, provide connectivity to upload data to HDFS [4] and to Hive from MySQL. There are several other technologies developed in the Hadoop environment to play with BigData and expert one's own skills. The MapReduce [5] framework has been widely adopted by various companies and organization to process huge volume of datasets. Thus, it solves the problem of data that is being too large. The Hadoop is integrated in Linux environment lessens cost-effective for computing array. To support distributed file system design, Hadoop Distributed File System (HDFS) was developed. It is java based file system. It is reliable and scalable to data storage. HDFS is designed using low cost hardware and is highly fault tolerant. HDFS replicate the data across the cluster. It continues computation without halting the process in case of single server failure. HDFS has no restriction on dataset storage. It support both structured and schema-less data.

The challenge of big data is its distributed environment and thus it is more complicated and vulnerable to attack. Without right security and encryption BigData means big problem. BigData environment may include dataset with personal identifiable information such as account number, mobile number, social security number etc. Therefore it is important to address the information ownership and classify the data according to its criticality.

2. RELATED WORK
Brief overviews of research work done for privacy issue in MapReduce framework are discussed in following section.

The introduction of Hadoop, MapReduce and its distributed environment in technological world give rise to privacy and security issues. The research has been done to address security threat on this framework to achieve integrity and availability of data. Tran, Q. et. al. [6] examines the data privacy problem triggered by MapReduce framework and presented a system named Airavat which incorporate mandatory access control with differential privacy. Since, the result produced in this system are mixed with certain noise, it is unsuitable to many application that need dataset without noise, such as data mining and analysis operation containing medical experiment.

A practical and widely-adopted technique for data privacy preservation is to anonymize data via generalization to satisfy a given privacy model. For example, km-anonymity [7], lm-diversity [8], complete k-anonymity [9], ρ-uncertainty [10], (h,k,p)Coherence [11] and PS-rules [12] have been proposed to protect transaction dataset through data transformation. These models differ in their assumptions about how data may be attacked by an adversary. Zhang et al [13] leveraged Map-Reduce to automatically partition a computing job in terms of data security levels which will help to maintain data privacy in hybrid cloud. Iwuchukwu et al. [14] proposed bulk-loading techniques which use an R+tree index to enhance anonymization performance. Lefevre et al. [15] and Loukides et al. [16] proposed sampling based methods to anonymize large datasets. These approaches are however designed to work on a single machine, and thus their scalability is limited. Federated MapReduce proposed by Wang.et.al [17], provide capability to run MapReduce across geographical distributed

cluster without specifying global reduce function. It ensures that sensitive data do not violate during analysis process. It includes automatic proxynation synthesizing to deploy application on multicluster environment. But it has a limitation to aggregate it with fine grained resources and run Fed-MR on higher version of Hadoop.

In contrast to above work, proposed work in these paper ensures security and privacy preservation through access control and encryption. The encrypted data is put on the Hadoop Distributed File System in Hadoop. According to role based access control right user can access corresponding data, thereby, right operational services can be enforced on role. Irrespective of the increasing number of user and services, matrix table is designed to provide scalability for mapping role and services.

3. PROBLEM DOMAIN

The development of technology becomes the backbone of daily routine. It is changing the habit of not only business but daily life too. Internet has become large source for processing and service provider. The rapid growth of such solutions not only increases the expectation of users but also generates the huge amount of data for storage and processes. The rate at which huge data produce and processes, certain restriction and shortcomings may occur. Security is one of the big challenges to maintain the originality and preserve the privacy of information throughout the processing.

The privacy is a primary requirement of growing technology. To maintain isolation over sensitive data (such as transactional data, medical diagnosis, and customer personal information in market dataset) is a big challenging task. The MapReduce Hadoop framework come with several advantage suffer from privacy issue as it discloses private and sensitive information. To prevent the information leaks, and to balance the goal of permissive model, the entrusted code should be confined. The traditional approach to data privacy is based on cryptography [18], which alone cannot enforce privacy demanded by BigData services. The conventional system proposed for privacy problem in MapReduce is also unsuitable to much application that needs data sets without noise, e.g., medical experiment data mining and analysis. Therefore, the complete study concludes that one of the restrictions in BigData storage and processes is its security features. The confidentiality, authentication and access control are the three major security principles should be involved in BigData elucidation. The study also perceives that most of researchers consider security as the supplementary requirement and only emphasis about the encryption and decryption of the information. They do not concern about the leakage of information from unauthorized access or fabrication in process module or information communication.

For that reason problem observation address that there is a need to develop a solution which should not only increase the performance for large data processing but also help to maintain security during transaction, storage, and operation.

4. SOLUTION DOMAIN

The BigData as a large amount of data reflect a new technology for consortium and become one of the trending researches today. There are two fundamental different approaches to control the visibility of data to different entities such as individual, organization and systems. The first approach is to limit the data access within the underlying system to control the visibility of data. The second approach encapsulates the data itself in a proactive shell using encryption technique. Both approaches have their benefits and detriments.

The complete solution has implemented into Hadoop framework using MapReduce component and Hadoop Distributed File System. The following steps are proposed to achieve security policy in large database based super market application.

1. The proposed solution considers a dataset from foodmart cooperation as the sample database for proposed application. The foodmart cooperation was founded in 1995 and it is an authorized consortium to import and export for food and beverages in VIETNAN. Huge amount of transactions make there working, complex and create trauma environment for processing and storage.
2. The Hadoop framework is used with Hadoop Distributed File System and MapReduce framework for parallel processing of BigData.
3. The Access Control [19] is one of the crucial modules of the proposed solution. It has implemented with Role Based Access Control theory. The RBAC is the method of regulating access to resources, rights, and roles of individual with proper planning.
4. The definition of confidentiality states that it is one of the moral and core rehearse of security to retain data safe, secure and isolated from unauthorized, unwanted, undesirable access and interrupt. It is one of the key principles of security and requires encryption and decryption process. Encryption is an approach which converts a plain text into non-human readable cipher text. It uses key attribute to make every iteration process unique. The confidentiality can achieve through cryptographic algorithm. The proposed solution follows the policy of asymmetric key cryptography algorithm and implements RSA algorithm [20] to make data safe and secure. It uses public key for encryption and private key for decryption to avoid unauthorized access.
5. In order to avoid unauthorized access, propose solution implements access control authentication scheme to give proper user access with specified user roles. The username and password authentication scheme has been implemented with RBAC model. A RBAC defines individual role like administrator, manager, and customer for users. It also defines the level of access and rights of operation. So, when user attempts to get access into supermarket application, a specific role definition with right permission classifies user operations and database access.
6. The MapReduce component has been implemented with mapper and reducer class, to process and execute search and information fetching request in large dataset. It has implemented to partition large dataset into multiple parts to distribute the searching load and to increase system performance.
7. Performance of proposed system has measured in terms of computation time. The complete proposed solution is tested with one node Hadoop framework

and compare with three node based distributed environment.

The complete solution concludes that proposed solution will not only be able to maintain user classification and right distribution but will also help to maintain data safety in Hadoop Distributed File System. It implements Hive and MapReduce component to make work easy and efficient. A block representation of proposed solution is drawn in Figure1.

5. IMPLEMENTATION VIEW

This section explores the implementation view of proposed solution and also gives a description of packages, classes and methods used during implementation. In summary, this section gives the details about:

1. User Interface View
2. Logical View
3. Background View

5.1 User interface view

User Interface view describes about the graphical interface of proposed supermarket application along with source code module. It specifies the action and description with the details of used packages and classes. The front end of application contain login page, registration page and the operational interface which user are privileged to operate according to their credentials with the help of access control mechanism. It includes java file and its associated packages.

5.2 Logical view

The proposed solution is design to maintain privacy and transparency of credentials during data distribution, so, a logical interface is developed.

It implements confidentiality in HDFS and secure large data distribution. The Hive and MapReduce components are used with Hadoop Distributed File System in Hadoop Server. A Sqoop component is used to import data from MySQL database to Hive Component. MapReduce helps to encrypt imported data and forward to HDFS. Afterwards, HDFS distribute all the large data file into multiple nodes. The complete scenario concludes that HDFS will always consist of encrypted data and null possibility of information leakage is possible.

5.3 Background view

Furthermore, the complete execution also consists of the background process to achieve the desire requirement. It implements the system role of user with rights and responsibility classification. It also helps to perform user operation of supermarket dataset.

There are different types of database table maintained here. First database table contain actual supermarket dataset of foodmart. It contains details of food warehouse description, cost, and customers purchase and personal details. Second database table play system management role. It classify user based on different role and manage access control mechanism.

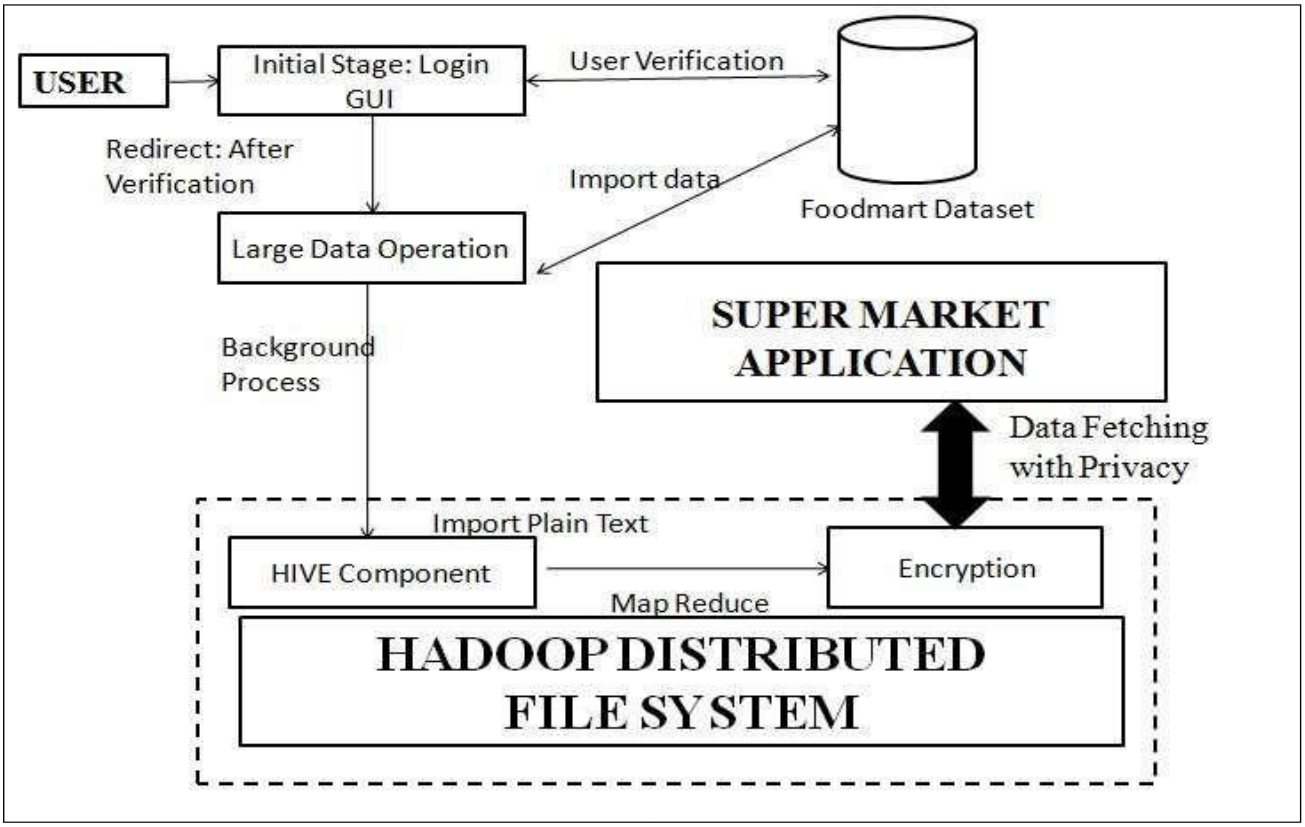

Figure 1: Block Representation of Proposed Solution

Analysis of system performance is done in two manners. The first way is observation of performance in terms of computation time with respect to single node and multinode (e.g. 3nodes). And the second way is execution of testing and verification of developed application with respect to proposed solution.

In the Hadoop distributed processing application, involvement of other machine partition the complete process and large data into multiple part and decrease the computation overhead. To achieve this, proposed solution develops a network with three computers where one is master and another two computers are slave machine. All machines have adequate hardware and software specification which are shown in implementation phase. The observation computation time describe that three node configuration gives low computation time then single machine in case of large data. But result observations also conclude that, it may create burden in case of small data size due to extra overhead of Hadoop Server. Two different memory size tables are processed on single node and Multi Node. The Multi Node (Three Node) configuration take 15 seconds average map time for small size data(approx 60,000 records) and single node consumed 14 second for average map time. This results conclude that single node consume low timing for small data than Multi Node. So, traditional technique with single node is suggested for small dataset. Subsequently, 9mins 37sec is recorded with single node for large dataset (approx 1,000,000 records) and 8mins 14sec is recorded with large data set on multinode. It concludes that multi node Hadoop server perform better in case of large data but perform poor for small data size due to multi node overhead. Thus complete work strongly suggests that multi node help to process large data with efficient manner. Other diagnostics of proposed solution with respect to single node and multi nodes for small data and large data is shown in Table1. The Average Map Time computation on Hadoop Server is depicted in graph as shown in Figure 2 and Figure 3.

7. CONCLUSION

The implementation of privacy preservation and access control mechanism, configured with multiple nodes justify the improved performance with desire level of security. The purpose behind observation of problem in HDFS is to enhance the level of security during distributed storage and processing. The development and deployment of complete proposed system is evaluated on three steps. The first step is verification of encrypted data access in unauthorized manner in HDFS. Here, proposed solution hides the plain text into encrypted format, so, third party will not be able to view the table data. In the second step, unauthorized user access is also attempted to perform operation but role based operation classification doesn't gives any opportunity to perform any unauthorized operation. Third step implements the performance evaluation of proposed solution in terms of computation time. The computation time of proposed solution is observed with single node and multinode network. The Hadoop server overhead is more for smaller sets of data. Therefore, computation time is more on multinode compared to single node in case of smaller sets of data and give low performance. But, for larger sets of data Hadoop works intelligently and give low computation time with multinode environment compared to single node.

The complete work concludes that proposed solution not only implements confidentiality, authentication and access control for different user roles but also perform better with Hadoop server. So, proposed solution can be used with Hadoop server to maintain security in HDFS for large data based supermarket application.

Table 1: Table for Computational Measurement

	Single Node		Multi Node (Approx 3 Node)	
	Small Data	**Large Data**	**Small Data**	**Large Data**
Average Map Time	14sec	9min37sec	15sec	8min14sec
Average Shuffle Time	15sec	1min13sec	8sec	30sec
Average Reduce Time	15sec	27sec	2sec	48sec

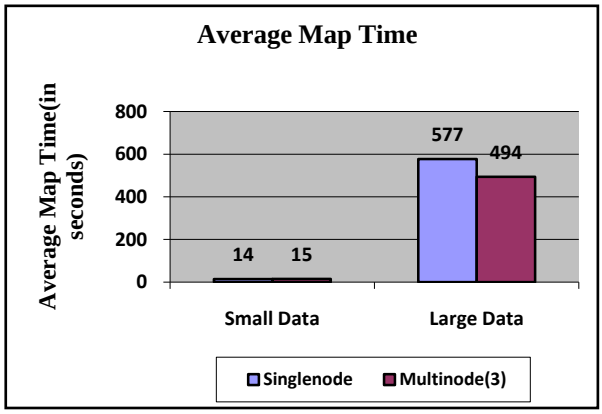

Figure 2: Average Map Time for Small Datasets Vs Large Datasets

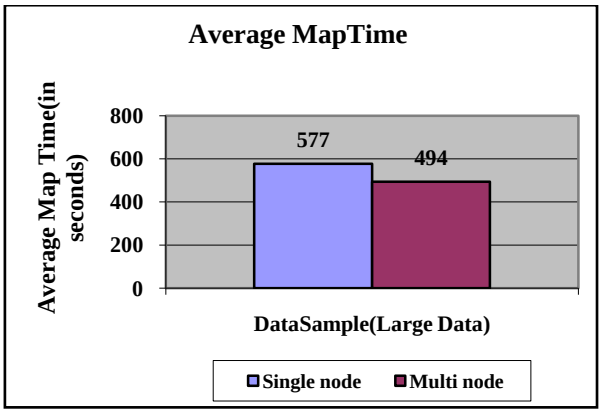

Figure 3: Average Map Time for Large Datasets

8. REFERENCES

[1] Interactions with Big Data Analytics, Danyel Fisher Microsoft Research | danyelf@microsoft. com.

[2] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#Overview

[3] MongoDB: The Definitive Guide,2013, by Kristina Chodorow

[4] The Hadoop Distributed File System Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler Yahoo! Sunnyvale, California USA {Shv, Hairong, SRadia, Chansler}@Yahoo-Inc.com

[5] Introduction to MapReduce and Hadoop Matei Zaharia UCBerkeley RAD Lab matei@eecs.berkeley.edu

[6] A Solution For Privacy Protection In MapReduce Quang Tran,Hiroyuki Sato Graduate School of Engineering, The University of Tokyo.

[7] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving anonymization of set-valued data," Proc. VLDB Endow., vol. 1, no. 1, pp. 115–125, Aug. 2008.

[8] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Local and global recoding methods for anonymizing set-valued data," The VLDB Journal, vol. 20, no. 1, pp. 83–106, Feb. 2011.

[9] Y. He and J. F. Naughton, "Anonymization of set-valued data via topdown, local generalization," Proc. VLDB Endow., vol. 2, no. 1, pp. 934– 945, Aug. 2009.

[10] J. Cao, P. Karras, C. Raïssi, and K.-L. Tan, "ρ-uncertainty: inferenceprooftransactionanonymization," Proceedings of the VLDB Endowment, vol. 3, no. 1-2, pp. 1033–1044, 2010.

[11] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu, "Anonymizing transaction databases for publication," in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '08, 2008, pp. 767–775.

[12] G. Loukides, A. Gkoulalas-Divanis, and J. Shao, "Anonymizing transaction data to eliminate sensitive inferences," in Proceedings of the 21st International Conference on Database and Expert Systems Applications: Part I, ser. DEXA'10, 2010, pp. 400–415.

[13] X. Zhang, L. Yang, C. Liu, and J. Chen, "A scalable two-phase topdown specialization approach for data anonymization using MapReduce on cloud," Parallel and Distributed Systems, IEEE Transactions on, vol. 25, no. 2, pp. 363–373, Feb 2014.

[14] T. Iwuchukwu and J. F. Naughton, "K-anonymization as spatial indexing: Toward scalable and incremental anonymization," in Proceedings of the 33rd International Conference on Very Large Data Bases, ser. VLDB '07, 2007, pp. 746–757.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workloadaware anonymization techniques for large-scale datasets," ACM Trans. Database Syst., vol. 33, no. 3, pp. 17:1–17:47, Sep. 2008.

[16] G. Loukides, A. Gkoulalas-Divanis, and J. Shao, "Efficient and flexible anonymization of transaction data," Knowledge and information systems, vol. 36, no. 1, pp. 153–210, 2013.

[17] Wang.et.al, Federated MapReduce to Transparently Run Applications on Multicluster Environment,2014 IEEE International Congress on Big Data

[18] Travis Mayberry, Erik-Oliver Blass, Agnes Hui Chan, "PIRMAP: Efficient Private Information Retrieval for MapReduce", Proceedings of Financial Cryptography and Data Security (FC'13), pp. 371—385, Okinawa, Japan

[19] Access Control for Sensitive Data in Hadoop Yenumula B. Reddy Department of Computer Science Grambling State University, USA

[20] Ron Rivest, Adi Shamir, Leonard Adlema. "RSA algorithm,"and approached by Avi Kak (kak@purdue.edu) February,2016 Avinash Kak, Purdue University.

[21] Marko Grobelnik "Introduction to bigdata" marko.grobelnik@ijs.si Jozef Stefan Institute Ljubljana, Slovenia.

[22] Yenumula B Reddy "Access Control Mechanisms in Big Data Processing" Department of Computer Science Grambling State University, Grambling, LA 71245, USA.

[23] C. Dwork, "Differential privacy" in Encyclopedia of Cryptography and Security Springer 2011.

[24] Foodmart Dataset, https://technet.microsoft.com/en-us/library .

[25] Jim Kurose, Keith Ross Addison-Wesley, "Security Principle", March 2012.

[26] Huseyin Ulusoy, Murat Kantarcioglu, ErmanPattuk, Kevin Hamlen, et al. , 2014"Fine grained Access Control in MapReduce ".

[27] Neelam Memon,Grigorious Loukides, Jianhua Shao, et. al., 2014, "A Parallel Method for Scalable Anonymization of Transaction Data" School of Computer Science & Informatics Cardiff University, UK.

[28] Weidong Shi, TaeweonSuh, et. al., IEEE 2014,"A FPGA cloud for Privacy Preservation computation "

[29] Xianfeng Yang and Liming Lian,et. al., 2014, "A New Data Mining Algorithm based on MapReduce and Hadoop," Xinxiang University, Xinxiang Henan, P.R.CHINA

[30] B.C.M. Fung, K. Wang and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 5, pp. 711-725, 2007.

[31] M. V. Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, Fully homomorphic encryption over the integers, presented at the 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Riviera, French, 2010.

[32] Privacy-preserving Anonymization of Setvalued Data Manolis Terrovitis Dept. of Computer Science University of Hong Kong rrovitis@cs.hku.hk Nikos Mamoulis Dept. of Computer Science University of Hong Kong nikos@cs.hku.hk Panos Kalnis Dept. of Computer Science National University of Singapore kalnis@comp.nus.edu.sg