# A Keyword Recommendation of Policies to Achieve Security Using Mapreduce

**Sushma M S[1], Mallikarjuna S B[2]**

*[1]Student, Dept. of Computer Science and Engineering, BIET college, Karnataka, India.*
*[2]Associate Professor, Dept. of Computer Science and Engineering, BIET college, Karnataka, India.*

---

***Abstract -*** A lot of data proprietors are fundamental while in transit to free the data as an alternate form of unique globe importance. Subsequently, most it is crucial principally to decide an extensive variety of re-unmistakable evidence threats through recommending convincing de-acknowledgment methodologies to guarantee both assurance and utility of the data. De-recognizing confirmation approaches is one of the models it could be used for the most part to achieve such essentials, regardless, the amount of de-acknowledgment techniques is exponentially colossal due to the wide zone of selectrifier characteristics. Predominantly to manage the exchange among data ease of use and data assurance, skyline estimation it could be used to pick those methodologies, despite the fact that it for the most part striving for beneficial skyline taking care of over significant number of courses of action. It likewise prescribe its estimation called security protecting gathering k-implies computation for tremendous data batching, which relies upon MapReduce to crush the test. Wide tests above including honest to goodness and gathering of data demonstrate that the suggested SKY-FILTER-MR estimation, which shows extraordinary adaptability over broad course of action gatherings.

*Keywords*- De-acknowledgment policy, anonymization, skyline computation, data privacy.

## 1. INTRODUCTION

At present, individualized calculative innovation and social media channels, for example, Facebook and Twitter, turn out to be progressively famous, bigdata technology is in the massive development. Bigdata consist of variety of heterogeneous, means individual object in bigdata posses variety of modules. Extraordinarily, enormous bigdata collections incorporate different relative sorts of articles, let's consider an example, writings, pictures and sounds, bringing about huge variety related to structured format of data, including organized information and unstructured format of information. In addition, distinctive sorts of items convey diverse data until it is connected with other format of data. For instance, a small bit of game frame video with data about data utilizes a substantial number of resulting pictures to show the activity procedure and utilizes a small amount of meta data, for example, comment and encompassing writings, mainly to demonstrate extra data it is not shown in the current frame of video, for example the labels of competitors. Despite the fact that the ensuing pictures travel through various data from the encompassing writings, it mainly depict similar articles from alternate points of view.

Bunching is intended mainly to isolate article in the direction of a few unique gatherings as indicated by uncommon measurements, making the articles with comparative highlights in a similar gathering. Grouping system is an effectively connected to learning revelation of knowledge and information designing associated along with the expanding prevalence of big data. The big data technique of bunching is drawing is very much consideration in distinction to information architects and analysts. Let's consider an instance, GAO et al. outlined a chart positioned bunching calculation for the technique that is bigdata by summing up its past picture content grouping strategy. Present positive grid tri factorization calculation mainly to perform cluster of vast datasets by catching the relationship all over the numerous process. Zhang et al. suggested a large arrange bunching calculation for the technique of big data mainly by utilizing the area to show the connections over the numerous processes. Be that as it may, if it is troublesome for such to group big data viably, particularly variety data, because of the accompanying the main two cause. To start with, it connect the most of highlights from various methodologies straight and overlook the most tedious relationships covered up in the variety of data sets, that is why it is outstanding to deliver wanted outcomes.

Further, it is frequently posses large pace unpredictability, to make them just applicable to little data sets. In this manner, they can't bunch a lot of diversified data proficiently. The handle the mentioned issues, a security safeguarding large arrange PCM plot (PPHOPCM) for the bigdata technique of grouping has been utilized. PCM is one of the critical plans of fuzzy bunching. PCM it can mirror the normality of each and individual article to distinctive groups adequately and it can keep away from the defilement of clamor in the bunching methodology. In any case, PCM can't be connected to big data bunching straightforwardly because it is at first intended for the little organized dataset. Extraordinarily, it can't catch the most tedious relationship above the different process of the diversified data articles. To build the productivity for grouping

the collection of information resides at bigdata, plan a clustering k-Means calculation in light of MapReduce is utilized. Be that as it may, the most secrete data has a tendency to be in exposure in case of performing HOPCM on the specified cloud. Let's consider the bank application financial data it is a custom kind of big data for instance: It consist of most of the confidential information of each and individual account information, for example, individual email, phone number, amount details, loan details, deposits details etc.,

The motivation behind the effort is to avoid the manual work of the bank data to perform programming knowledge to be utilized effortlessly, basic and financially savvy. It deals with the client's data and focus on the considered data set which has been utilized. The defined limit of the architecture is mainly to select and collect the individual data and to receive data as and when needed and furthermore ready to regulate it effectively and furthermore give data protection.

### 1.1 RELATED WORK

System requirement specifications (SRS) which gives information about the system behaviour to be created. SRS includes both functional and non-functional requirements.

**FUNCTIONAL REQUIREMENTS:** The user should have the ability to a viable treatment of systems and data should be taken care of by using MapReduce. This technology will be used to provide security to bank user's data.

**NON-FUNCTIONAL REQUIREMENTS** is a requirement that explains the criteria requested to analyze the operation of a system. It describes how the system works and what standard should be provided. It describes some system attributes like accessibility, availability, security, reliability etc. Service level requirements are the measures of the quality of service required and are crucial to capacity planning and physical design. For each service level, we need to identify the realistic, measures target values. These target values are like service hours, throughput etc. Access restrictions must specify what data should be protected, which data should be restricted to a particular user.

### 1.2 SYSTEM DESIGN

The different stages of project design has been described in the system design which includes description of the overall project, algorithms used to implement and high level diagrams like class diagram, data flow diagram, sequence diagram.

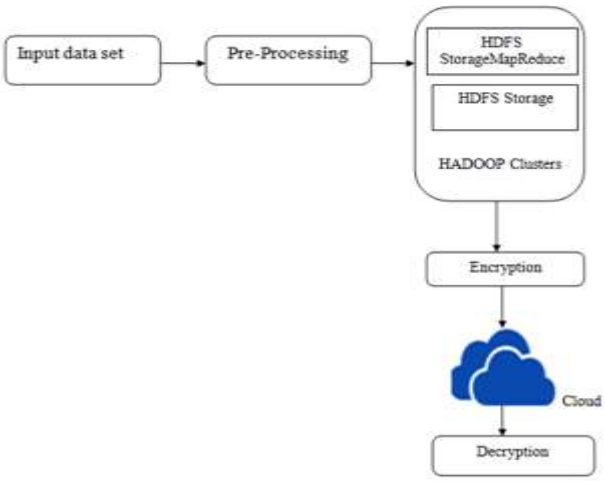


Fig 1: Architecture Diagram

**Description:**

**The following steps describes in detail:**

Step 1: Here input is raw dataset. It means, dataset might consist of unwanted data, unnecessary data, and redundant data. This kind of raw dataset will be provided as an input. System uses the bank related dataset as the input dataset.

Step 2: Preprocess means the removal of unnecessary data from the raw dataset. Redundant data must be eradicated. And unwanted dataset must be eradicated. This step leads the dataset it consist of the required attributes and properties.

Step 3: Clustering process has been done after preprocess step. In this procedure, dataset will be splitting into few properties relied on few properties relative values. Clustering will follow the technique that is dividing and conquer policy. It means it will split based on some attributes and joins them based on some values.

Step 4: The modified dataset will be uploaded into the hadoop distributed file system. It will be used as a database for storing the dataset which has been resulted from the clustering process.

Step 5: Fifth step is using the mapreduce. It is a process of programming procedure and corresponding associative and relative implementation for further processing and producing the big dataset. Here hadoop platform and hadoop distributed file system components have been utilized.

Step 6: The modified dataset which has went through all the mentioned above steps will be either uploaded or saved on cloud. Here cloud can be any sort of freely available to the user cloud.

## 2 IMPLEMENTATION DETAILS

### 2.1 MODULES

**ECLIPSE**

It is a freely available tool to mainly use to develop and implement the java projects. Important thing to be noticed is many varieties of java development tools are available o the market. Based on the developer need the right one must be chosen. There multiple versions of the java integrated development environments are available. For example eclipse, Jupiter etc., It provides the developer friendly environment. Intelliscience will give the hints where the developer went wrong. Many built in functions are available. Also the pluggable applications are well available here. Developer can import from the library of the plugins.

**SQL**

Structured query language (SQL) is an inquiry programming dialect. It deals or handles the relational database. Here data is arranged by a well structured manner where table is called as relation, rows are termed as tuples and columns are termed as attributes and values which have been used in each row and column are called as domains. So relational database mainly involve the important operations such insert, modify, delete, update etc., To perform all these fundamental operations on database, sql provides the rich set of queries.

To perform proper insertion, deletion and modification sql also provides the normal forms. Where each normal forms deals with an anomalies' that could happen while performing the fundamental operations of the sql.

**WAMP Server**

It is a collection of packages that performs on the windows operating system. It utilises the apache as a web server. And it also provides the MYSQL as an open source to handle the database queries. PHP is pre processor hypertext. It uses the localhost server to execute the relative php files. It also provides the PhpMyAdmin to handle the database tables in a convenient manner.

The user interface provided by the PhpMyAdmin says about the general illustration for accessing the sql tables. PhpMyAdmin provides the all the features that has been provided by MYSQL. It makes user to perform insertion, deletion and modification in a more convenient way. And PHP is considered as a scripting language. It also provides the access to python and perl languages nothing but scripting languages. It is a collection of package.

### 2.2 EXPERIMENTAL RESULTS

This section shows snapshots of the results.


Fig 2: Data Fetching

In above figure the path of the selected file and name of the selected file is displayed and the data will be retrieved and the content will be displayed in an unstructured format which will be converted to structured format.


Fig 3: Data Pre-processing

In above figure the data which is retrieved from the selected file is preprocessed and it is converted into structured data. The unstructured data should be divided in the form of rows and columns.


Fig 4: Loading Data to Big Cluster

Above figure shows the loading data to big cluster, once the preprocessing is done and the data is converted to structured format, the formatted data is loaded to big cluster here we are selecting only specified attributes from the data and the sorted data will be uploaded to cluster.


Fig 5: Level I Clustering

In above figure, after loading sorted data to the big cluster, the data is divided into different levels of clusters here we are dividing on the basis of marital status, level I clustering based on marital status i.e. MARRIED.


Fig 6: Level II Clustering

In above figure, after loading sorted data to the big cluster, the data is divided into different
levels of clusters here we are dividing on the basis of marital status, level I clustering based on marital status i.e. SINGLE.


Fig 7: MapReduce Page

In above figure, we have to enter the no. of clusters then the list of clusters is displayed then we have to select any file out of those, then upload selected file to cloud and then perform MapReduce function.


Fig 8: Data encryption

In above figure we have to select the type of cloud to perform data encryption i.e. public or private type then the selected file is opened. Then encrypt the sorted data using secret key.


Fig 9: Secret key generation

In above figure the data is encrypted using this secret key. Once the secret is generated the above popup message is displayed.

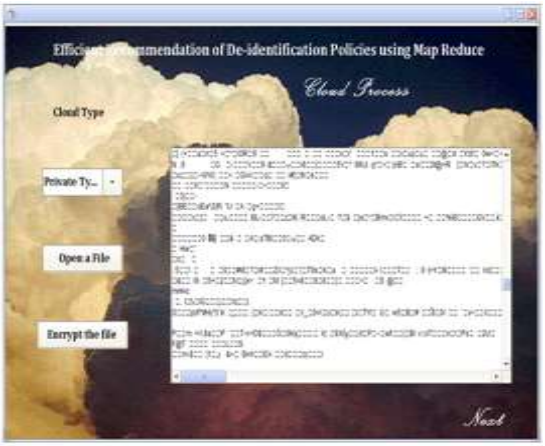


Fig 10: Encrypted file

In above figure, after the selection of cloud type i.e. public or private type the selected file is opened. Then the encryption is done privacy will be given to the data, in order to provide privacy data is encrypted by using secret key.


Fig 11: Decryption of data

In above fig download the encrypted file from cloud and then decrypts encrypted data by using the secret key. The decrypted data is displayed after entering of secret key.

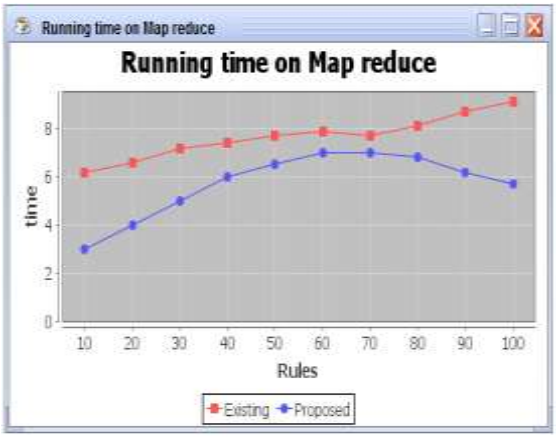


Fig 12: Graph

The above screenshot explains the running time on MapReduce.

**3. CONCLUSION**

The projected system is build up with a reason on condition that suggestion of regulations mainly to accomplish information security by means of Map reduce programming model. Initially a successful path for the regulations and rules which has been applies for security reason. The novel projected system characterization, it will reduce the amount of instance of producing the regulations and the quantity of another regulations group artificially. Subsequently projected system illustrates about the usage of SKY FILTER MR. It is one of the form of using MapReduce programming model. It mainly used to provide solution for the regulations proficiently. It executes the most composite and tedious environment of dataset to provide the high security for the files which has been uploaded to the cloud. Whereas the proposed system effectives uses the asymmetric encryption algorithm to achieve the high level security for the dataset. Results of the proposed system illustrates that caliber of the system is high.

**REFERENCES**

[1] Xindi MA, Hui LI, Jianfeng MA, Qi JIANG, Sheng GAO, Ning XI & Di LU, "APPLET: a privacy-preserving framework for location-aware recommender system", 2016.

[2] Weiyi Xia, Raymond Heatherly, Xiaofeng Ding, Jiuyong Li, Bradley Malin "Efficient discovery of de-identification policies through a risk-utility frontier," in CODASPY, 2016, pp. 59–70.

[3] Simson L. Garfinkel conducted research on De-Identification of Personal Information 2015.

[4] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," in SIGMOD, 2015, pp. 1423–1434.

[5] Wahbeh Qardaji, Weining Yang Ninghui Li "Priview: Practical differentially private release of marginal contingency tables," in SIGMOD, 2014, pp. 1435–1446.

[6] Theodoros Rekatsinas, Amol Deshpande and Ashwin Machanavajjhala SPARSI: Partitioning Sensitive Data amongst Multiple Adversaries 2013.

[7] B. C. M. Fung, K.Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, no. 4, pp. 14:1–14:53, 2012

[8] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 555–570, 2002.

[9] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," Proc. VLDB Endow., vol. 5, no. 11, pp. 1388–1399, 2012.

[10] Q. Zhang and Z. Chen, "A Weighted Kernel Possibilistic c-Means Algorithm Based on Cloud Computing For Clustering Big Data," International Journal of Communication Systems, vol. 27, no. 9, pp. 1378-1391, 2014.

[11] J Hoffstein, "NTRU Public Key Cryptosystem – Methodology", shodhganga.inflibnet.ac.in/ bitstream/1060 3/103254/10/10_chapter-iii.pdf.

[12] The hadoop offical website. http://hadoop.apache.org.