Overcoming the lack of Numerical Confidential on Big Data by Introducing Shuffling in MapReduce Algorithm

B Thirunavukarasu,K. Sangeetha, Dr T. Kalaikumaran, Dr. S.Karthik

Abstract – In any engineering field the data associated with knowledge is important one for taking decisions for solving problems in the current system development. The numerical confidential on certain big data gets vulnerable. The algorithm used in big data is MapReduce algorithm. This algorithm does not provide with the effective numerical confidential. So Shuffling of data can be done in order to overcome the numerical confidential on big data. Shuffling on data could be made on certain algorithm including Minimal MapReduce algorithm.

Key Terms – Shuffling, MapReduce, algorithm, Big Data, Numerical Confidential, Analysis, Privacy.

1 INTRODUCTION

Organizations of all kind can gather, store, and efficiently process large quantities of data. The ultimate goal for gathering such data is to gain necessary information from the data to improve business processes of the respective organization using statistical and data mining tools. The tremendous needs of data mining have been sequentially updated in the market-basket analysis, fraud detection, consumer profiling, medicine, agriculture, and many other domains. Well before data mining became popular, commercial organizations and government agencies were using statistical methods to analyse data to benefit consumers and society.

Today, data mining draws from and adds to the many statistical analysis techniques. One of the key objectives of data mining is the discovery of new and useful relationships and patterns in the data. Some of these discoveries occur when data is mined specifically for the purposes of discovering such relationships. These unplanned discoveries are facilitated when users are provided access to the stored data. Unfortunately, privacy and confidentiality issues are increasingly creating strong barriers that prevent us from realizing the full benefits of data. In many instances the data that was collected explicitly for analytical purposes sits in a secure facility where only a few authorized individuals are provided access to the data.

Obviously this limits the usefulness of the data and defeats the very purpose for which they were gathered. Numerical data are of particular importance in this regard. They pose the greatest threat yet offer the greater benefits. They pose the greatest threat since they tend to be almost unique and an intruder with numerical data can easily compromise the privacy and confidentiality of sensitive records. They offer the greatest benefit since much of the business intelligence comes from numerical data. Hence, it is important

1.1Privacy on large data

Companies have sought for decades to make the best use of information to improve their business capabilities. However, it's the structure (or lack thereof) and size of Big Data that makes it so unique. Effort Estimation

Companies that collect and analyze specific consumer behavioral data and then sell the results to other companies looking to improve their consumer marketing and sales efforts. However, it is important to acknowledge that growing privacy concerns about the use of big data are not limited to these conventional data brokers. The Economist Intelligence Unit, an independent business within the Economist Group, has published a study of leaders in the use of big data that spanned 19 industry sectors including manufacturing, IT and technology, financial services, professional services, healthcare, pharmaceuticals and biotechnology, and consumer goods.

1.2 Analysis on Big Data

Big data analytics is the application of advanced analytic techniques to very large, diverse data sets that often include varied data types and streaming data.

Big data analytics explores the granular details of business operations and customer interactions that seldom find their way into a data warehouse or standard report, including unstructured data coming from sensors, devices, third parties, Web applications, and social media - much of it sourced in real time on a large scale. Using advanced analytics techniques such as predictive analytics, data mining, statistics, and natural language processing, businesses can study big data to understand the current state of the business and track evolving aspects such as customer behaviour. New methods of

- *Author Thirunavukarasu B is currently pursuing Bachelor's degree program in Computer Science and Engineering in SNS College of Technology, India, PH-+917402184561. E-mail: droptothiru@yahoo.com*

working with big data, such as Hadoop and MapReduce, also offer alternatives to traditional data warehousing.

Analytics, providing deep insights on Big Data to optimize every customer touch point. Using personalized workspaces and self-service templates, analytics are rapidly assembled, customized and shared across business teams.

1.3 Introduction to Numerical Confidential

If your research involves human subjects, you will need to consider both legal and ethical obligations regarding sharing your data. The 1998 Data Protection Act affects the processing of personal or sensitive data and the circumstances under which you can share them with others. The University's Records Management Section has online guidance and offers direct support for decisions about information disclosure. The role of the University's various research ethics committees is to develop policy and general guidance for Colleges and Schools on ethical issues arising from non-medical research involving human participants. Your School may have its own research ethics committee or guidance.

Avoidance of disclosure of personal or sensitive data can be accomplished in a number of ways, including anonymisation techniques or data aggregation for numeric data, editing of video or sound recordings, use of pseudonyms in qualitative data.

Different methods have different consequences for data quality, and should be considered in tandem with the consent process, for example, what sort of informed consent you seek from your subjects.

1.4 Algorithms for MapReduce

- Sorting
- Searching
- TF-IDF
- BFS
- PageRank
- More advanced algorithms

2. EXISTING TECHNIQUE

MapReduce (M/R) is a technique for dividing work across a distributed system. This takes advantage of the parallel processing power of distributed systems, and also reduces network bandwidth as the algorithm is passed around to where the data lives, rather than a potentially huge dataset transferred to a client algorithm. Developers can use MapReduce for things like filtering documents by tags, counting words in documents, and extracting links to related data.

MapReduce is one method for non-key-based querying. MapReduce jobs can be submitted through the HTTP API or the Protocol Buffers API. Also, note that MapReduce is intended for batch processing, not real-time querying.

2.1 MapReduce Algorithm by Google

MapReduce Jobs Tend to be very short, code-wise
Identity Reducer is very common "Utility" jobs can be composed Represent a data flow, more so than a procedure

Sort Algorithm Takes advantage of reducer properties: (key, value) pairs are processed in order by key; reducers are themselves ordered. Mapper: Identity function for value
(k, v) _ (v, _)
_ Reducer:
Identity function (k', _) -> (k', "")
Sort: The Trick
_ (key, value) pairs from mappers are sent to a particular reducer based on hash(key)
_ Must pick the hash function for your data such that k1 < k2 => hash (k1) < hash(k2)

2.2 Local Aggregation

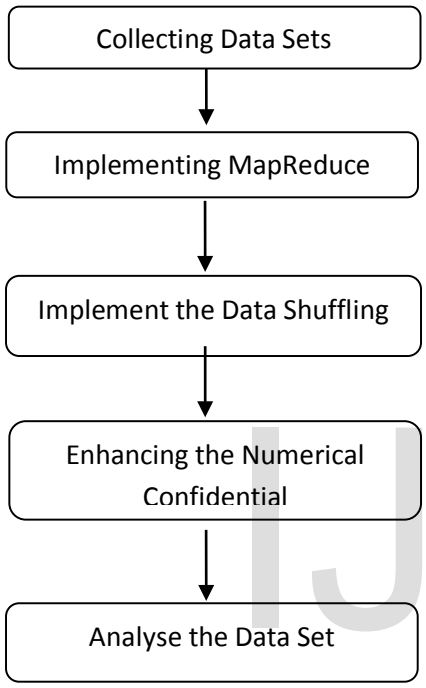
In the context of data-intensive distributed processing, the single most important aspect of synchronization is the exchange of intermediate results, from the processes that produced them to the processes that will ultimately consume them. In a cluster environment, with the exception of embarrassingly-parallel problems, this necessarily involves transferring data over the network. Furthermore, in Hadoop, intermediate results are written to local disk before being sent over the network. Since network and disk latencies are relatively expensive compared to other operations, reductions in the amount of intermediate data translate into increases in algorithmic efficiency. In MapReduce, local aggregation of intermediate results is one of the keys to efficient algorithms. Through use of the combiner and by taking advantage of the ability to preserve state across multiple inputs, it is often possible to substantially reduce both the number and size of key-value pairs that need to be shuffled from the mappers to the reducers.

Algorithm Word count (repeated from Algorithm 2.1)
The mapper emits an intermediate key-value pair for each word in a document.
The reducer sums up all counts for each word.
1: class Mapper
2: method Map (docid a; doc d)
3: for all term t 2 doc d do
4: Emit (term t; count 1)
1: class Reducer

2: method Reduce (term t; counts [c1; c2; : : :])
3: sum 0
4: for all count c 2 counts [c1; c2; : : :] do
5: sum sum + c
6: Emit (term t; count sum)

3. PROPOSED METHODOLOGY

In my proposed methodology, after the analysis of the data is done, the shuffling of the data is made. This will increase the efficiency. The data mart could be used to store the historical data that are often used by the user for the purpose of analysis.



3. Fig. 1. Flow of Proposed Methodology

The analysis tools are used to perform the analysis. MapReduce will be effective one but the main demerit of MapReduce is that it could not able to produce the Numerical Confidential privacy. Initially the data sets are to be collected from various resources. Then the algorithm of MapReduce was implemented. With this MapReduce algorithm, the data shuffling was done. This data shuffling on the unstructured data will increase the efficiency on the privacy of numerical data of the system. Thus finally the analysis on the data with numerical confidential can be done in effective manner.

3.1. Introduction to Shuffling

Data offers an excellent solution to this dilemma. Data shuffling is a masking technique where sensitive numerical values in sense the attributes are shuffled among the records. Based on high end statistical modelling, Data shuffling is performed in such a discipline so as to provide the highest possible level of protection from disclosure of sensitive information while preserving the analytical value of the data by tracking most of the relationships between the attributes. Specifically, Data shuffling maintains all linear and non-linear relationships among the masked variables to be same as that between the original variables. This allows the client end to analyse the data with the assurance that, for most techniques, analysing the current masked data will provide similar results as analysing the available original data. Among masking techniques for numerical data, data shuffling provides with the highest level of protection while also providing the highest analytical value. In short, Data shuffling thwarts the attempts of any individual intent on compromising the data while rewarding the attempts of any individual interested in performing valuable analysis.

3.2 A Description of Data Shuffling

Data Shuffling can be briefly described as follows. Data set consisting of a set of numerical confidential variables X and a set of numerical and categorical non-confidential variables S are considered. Data shuffling is implemented on these attributes as follows.

- Initially the rank order correlation of entire data set of the system is computed.
- Using the normal copula multivariate, the variables X and S are converted into X* and S* respectively, such that they have a joint multivariate normal distribution. The product moment correlation of the new variables is derived effectively by the rank order correlation.
- The transformed values Y* are then calculated by the help and implementation of general additive data perturbation or transformation method.
- Let y*ij represent the transformed value for the ith record and jth variable. In the original data set, replace y*(i),j with x(i),j ((i) = 1, 2, …, n; j = 1, 2, …,m) whereas y*(i),j and x(i),j represent the rank order observations of Y* and X respectively.

Comprehensive theoretical and empirical evaluation of the data utility and security of the data shuffling approach. The key characteristics of data shuffling can be summarized as follows. The shuffled values are actually the original values of the confidential variables assigned to a different observation. Hence, the univariate marginal distribution of the masked data is identical to that of the original data. The use of the copula-based perturbation approach enables data shuffling to maintain the rank order correlation of the masked data to be the same as that of the original data. This implies that data shuffling results in minimal information loss in linear and monotonic non-linear relationships among variables. In this paper we demonstrate three modes of delivery based on the specific needs of the organization: (1)
Excel based solution for small applications, (2) Web based solution for larger applications where the organizations

wish to perform Data shuffling by themselves, and (3) Third party solution (installable java based application) for organizations and agencies for large complex data sets. We believe that these modes of delivery will be useful for any organization that intends to analyse, share, or disseminate numerical confidential data without risk of disclosure.

4. RESULTS AND DISCUSSIONS

The approach in this proposed methodology is collecting data associated with Data Sets implemented in the developed Big Data project and applying shuffling over them. Here we have taken some sample data set for describing the function and output of the system.

4.1 Modes of Delivery

4.1.1 Excel Add-in Based Solution

For small to medium data sets with up to 20 variables and 10,000 records, we have developed a Microsoft Excel Add-in. The Add-in is easy to install and implement.

4.1.2 Web-Based Solution

Organizations that use medium-sized data sets (up to 100,000 records) and that only require masked data and do not want to install additional software, may take advantage of our web-based solution. The screen shots below show the web-based interface. As can be seen from the figure, the interface permits the uploading of a data set, the selection of appropriate variables to mask, and performing shuffling using a single button. One of the features of the web based solution is that it permits a user to store a history of their previous masking attempts, so that multiple runs (with different variables possibly masked each time) can be tracked easily.

5. MERITS AND DEMERITS

The main advantage of Shuffling – which is one of the most critical issues around data masking software – is that it quickly and efficiently deals with large tables, and it leaves the look and feel of the data intact. When used in small amounts of data, Shuffling is rarely effective. Another disadvantage is that, since the original data is still present, there is a chance the data may be "unshuffled," especially if the algorithm used was not too sophisticated. Number and Date Variance involves the use of an algorithm to modify each number or date value in a column by some random percentage of its real value.

The advantage of Number and Date Variance is its ability to reasonably obfuscate numeric data while still keeping the range and distribution of values within existing limits. Its main limitation is that Number and Date Variance can only be applied to numeric data. Encryption involves the algorithmic scrambling of data whereby only those with the appropriate key can view the encrypted data. Other than the actual obfuscation of the data, Encryption has no other real advantage; and even this seeming advantage is tenuous. It is easy to see when data has been encrypted as it destroys the formatting, and look and feel of the data. Similarly, it is almost always possible to break the encryption. Also, when using test or development databases, anyone with the key has access and this key can be easily revealed, rendering the encryption futile. Nulling Out/Truncating/Deletion involves the removal of the sensitive data.

Masking out Data is a very effective technique when the data is in a specific, invariable format. When the format varies, then masking can be very cumbersome and complex to administer, as well as leaving some data not well protected.

6. CONCLUSION

In this paper, we have described a new shuffling procedure for masking confidential data. The advantages of this approach can be summarized as follows:

- The released data consists of the original values of the confidential variables(i.e., the marginal distribution is maintained exactly),
- All pair-wise monotonic relationships among the variables in the released data are the same as those in the original data, and
- Providing access to the masked micro data does not increase the risk of disclosure.

REFERENCES

[1] Sarathy R., K. Muralidhar, R. Parsa. 2002. Perturbing non-normal confidential variables: The copula approach. *Management Science* **48** 1613-1627.

[2] K. Muralidhar and R. Sarathy, "A Theoretical Basis for Perturbation Methods," Statistics and Computing, vol. 13, pp. 329-335, 2003

[3] K. Muralidhar and R. Sarathy, "Data shuffling - A new masking approach for numerical data," Management Science, vol. 52, pp. 658-670, 2006.

[4] L. T. Willenborg and T. D. Waal, Elements of statistical disclosure control. New York: Springer, 2001.

[5] R. Nelsen, "An introduction to Copulas," New York: Springer, 2007

[6] K. Muralidhar, R. Parsa, and R. Sarathy, "A general additive data perturbation method for database security," Management Science, vol. 45, pp. 1399-1415, 1999.

[7] K. Muralidhar, R. Sarathy, and R. Parsa., "An improved security requirement for data perturbation with implications for e-commerce," Decision Sciences, vol. 32, pp. 683-698, 2001.

Mr. B.Thirunavukarasu is presently pursuing B.E Computer Science & Engineering at SNS College of Technology, affiliated to Anna University-Chennai, Tamilnadu, India. His research interests includes BigData , Data Mining and Business Analytics. He has published 2 papers in National conference and 2 in journal. He is an active entrepreneur involving web services and mobile application development.

Mrs. K.Sangeetha is presently Assistant Professor at Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University-Chennai, Tamilnadu, India.She received the B.E degree from Sasurie College of Engineering and M.E degree from Nandha Engineering College. Currently she is pursuing her doctoral degree in AnnaUniversity Chennai. Her research interests include Datamining, and Networking. She has published many papers in international journals, national and International conference.

Dr.T.Kalaikumaran is presently Professor & HoD in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University, Chennai Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Ann University, Chennai.. He is interested in the research areas of data mining, spatial data mining, machine learning, uncertain data classification and clustering, pattern recognition, database management system and informational retrieval system. He is a member of CSI and IEEE

Dr.S.Karthik is presently Professor & Dean in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University-Coimbatore, Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Anna University of Technology, Coimbatore. His research interests include network security, web services and wireless systems. In particular, he is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published more than 35 papers in refereed international journals and 25 papers in conferences and has been involved many international conferences as Technical Chair and tutorial presenter. He is an active member of IEEE, ISTE, IAENG, IACSIT and Indian Computer Society.