International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 18 (2015) pp 39789-39794

Privacy Preserving Clustering process with Cost minimization for Big Data Processing

V. S. Thiyagarajan

Research Scholar, Annamalai University, TamilNadu. Mail Id: thiyagu_cse86@yahoo. com

Dr. K. Venkatachalapathy

Professor, AnnamalaiUiversity, TamilNadu.

ABSTRACT

Clustering plays a major role in handling large data sets to do their relevant process in various domains. The main issue is the identification of clusters in multi-dimensional data sets. It also has the open problem of privacy and security issues. To overcome these challenges, we propose the clustering algorithm of Apriori enhancement algorithm for directly mines frequent closed itemsets. Multi dimensional data sets produce the best quality of clustering for the user needs because the datasets are gathering from various domains. The slicing can be used to handle the high dimensional data in the difficult situation. Slicing can support both the horizontal and vertical processes of the large data set; for that we will use both of the process. Initially the input data are divided into subsets; then the data are given as the inputs for another process. For that the horizontal process is used to reduce the time complexity and cost of the overall process. Vertical grouping of attributes is used to help the privacy preservation. Finally divided subsets are combined with the help of group heads.

KEYWORDS: Big data, vertical partitioning, clustering, Apriori enhancement algorithm, vertical grouping attributes, multi datasets.

Design and development of a data dividing and integration approach for parallel privacy preserving clustering process

1. Introduction:

The term “Big Data” appeared for the first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress". Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is datamining. The origin of the term Big Data is due to the fact that we are creating a huge amount of data every day. Big data has been used to convey all sorts of concepts: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more [1]. In data mining privacy is an important issue. It is very challenging to preserve the privacy information of the users. The importance of privacy preserving is to hide sensitive information so that they cannot be discovered through data mining techniques. Privacy preserving publishing of slicing has been studied expansively in recent years [3]. The new aspect of Big Data lies within the economic cost of storing and processing such datasets; the unit cost of storage has decreased by many orders of magnitude, amplified by the Cloud business model, significantly lowering the upfront IT investment costs for all businesses. As a consequence, the “Big Data concerns” have moved from big businesses and state research centers, to a mainstream status [20]. Big data doesn’t always fit into neat tables of columns and rows. There are many new data types, both structured and unstructured, that can be processed to yield insight into a business or condition [21].

1. 1 Big data:

Big Data is a collection of large-volume, complex, growing multiple data sets, autonomous sources in data mining. This can be mainly used in the applications of enterprises, government, social media and industries. The environment of a big data is suitable for any data structured, semi/ multi/ unstructured data from multiple sources.

The term big data was coined to address this massive volume of data storage and processing. It is increasingly becoming imperative for organizations to mine the data to stay competitive. Analyzing data can provide significant competitive advantage for an enterprise. The data when analyzed properly lead to a wealth of information which helps the businesses to redefine their strategies. However, the current volumes of the data sets are too complicated to be managed and processed by conventional relational database and data warehousing technologies.

The data sets are beyond the capability of humans to analyze manually. Big data tools have the ability to run ad-hoc queries against the large data sets in less time with a reasonable performance. Big data analysis enables the executives to get the relevant data in less time for making decisions. Big data can pave way for fraudulent analysis, customer segmentation based on the store behavior analysis and loyalty program that identifies and targets the customers. This enables us to perform innovative analysis which indeed changes the way we think about data [27].

1. 2 Big Data is characterized by the following 4 Vs:

- **Volume:** The vast amount of data generated every second that is larger than what the conventional relational database infrastructures can cope with.
- **Velocity:** The frequency at which new data is generated, captured, and shared.

- **Variety:** The increasingly different types of data (from financial data to social media feeds, from photos to sensor data, from video capture to voice recordings) that no longer fit into neat, easy to consume structures.
- **Veracity :** The disarrayed data (Face book posts with hash tags, abbreviations, typos, and colloquial speech)

1. 3 Types of Big Data and Sources:
There are two types of big data [1]:

➢ **Structured**
Structured data are words and numbers that can be easily categorized and analyzed. These data are generated by things like electronic devices, smartphones, and global positioning system (GPS) devices.

➢ **Unstructured**
Unstructured data are more complex information, such as customer's reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically. These types of datasets are difficult to handle in data mining. The multi data is under a unstructured type only.

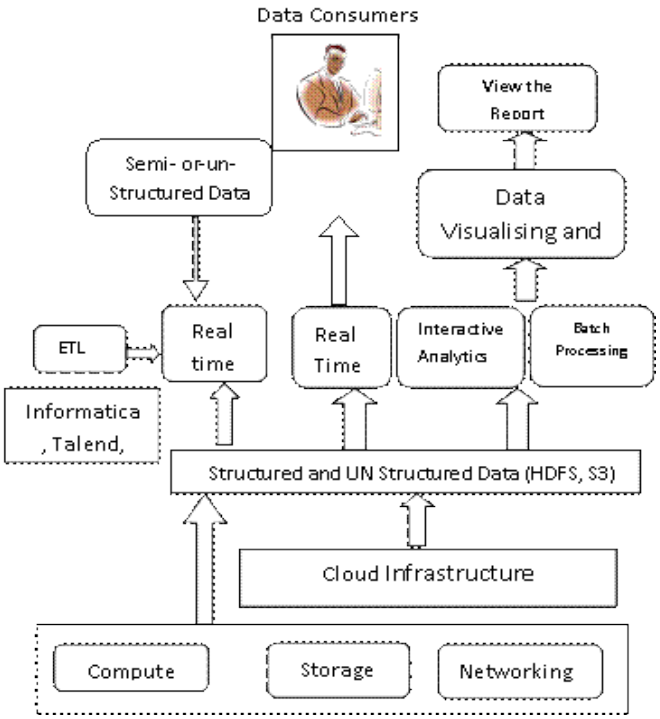

Fig 1: integrated view of big data process

2. Multi Datasets:
The multi datasets are basically defined as the collection of multiple data sets in a common pool. This is mainly used for segregation based on the types of an item sets in the database.

For example the fig below shows the types of datasets present in the database. In that different types of datasets are used:

- Log data
- Audio
- Video
- Images
- E-mails
- Social Media
- Events
- Transaction

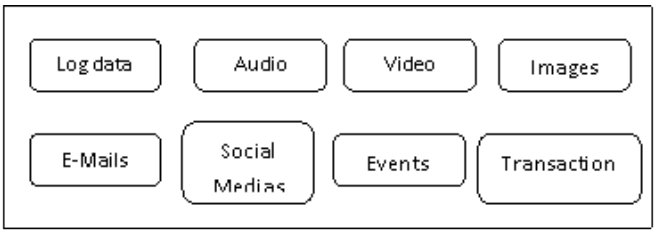

Fig 2: Sources of Big Data

Our paper mainly focuses on video data type. In that the large stream data are handled with the help of this way. The data items are processed based on the capacity of the data sets. For that the data items have to be selected segregated based on their preference, which can be processed first, and then sent after that send to requested one.

2. 1 Data mining in big data:
The data mining is a process of analyzing data from different views and recapitulates it into meaningful or useful information. Technically the data mining is a correlation between the large datasets and retrieves requested information from the data base. In that, six tasks or activities are used to define big data:

➢ **Classification:**
Classification means that based on the properties of existing data, we have made groups i. e. we have made classification.

➢ **Prediction**
It's a statement about the way things will happen in the future, often but not always based on experience or knowledge. Prediction may be a statement in which some the outcome is expected.

➢ **Estimation**
Estimation deals with continuously valued outcomes.

➢ **Extraction**
It is a process of extracting structured information from the unstructured datasets.

➢ **Feature selection**
Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction [5].

3. Literature survey:

Big Data is "a collection of complex data sets. Data mining is an analytic process designed to explore data in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. Big data is an emerging trend and the need for Big data mining is rising in all science and engineering domains. With Big data technologies most relevant and most accurate social sensing feedback can be had to better understand our society at real time [1]. Transitioning from Relational Databases to Big Data "Big Data will lead to a new wave of fundamental technological advances that will be embodied in the next generations of Big Data management and analyse platforms, products, and systems. A major investment in Big Data, properly directed, can result not only in major scientific advances, but also lay the foundation for the next generation of advances in science, medicine, and business [2]. Privacy and Data Utility for High- Dimensional Data can be achieved by using Anonymization Technique, "A new approach called slicing for privacy preserving in micro data publishing. By using Pearson and chi-squared based correlation coefficient the correlations for each pair of attributed are computed. To increase the data utility, an overlapping slicing method is used which duplicates an attribute in more than one column. The horizontal partitioning analyse the result by adding noise data to original data table and finally performs the vertical partitioning for different cases for both Pearson based and chi square based slicing [3].

Enhanced Slicing for Privacy Preserving Data Publishing: "Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. Before anonymizing the data, one can analyze the data characteristics and use these characteristics in data anonymization, Attribute partitioning enables slicing to handle the high dimensional data. Horizontal partitioning is done by grouping the tuples into buckets and within each bucket; values in each column are randomly permuted to break the linking between different columns. Slicing breaks the association of cross columns, but preserves the associations within each column [4]. Data Privacy Preserving using Random Grouping: "In this Anonymizati on countermeasures are still vulnerable as they still can expose protected information in released information. Sensitive attribute disclosures are based on random grouping, which is not very effective as randomly generating the associations between column values of a bucket significantly lowers data utility. In our quest for efficiency, we propose to replace random grouping with an optimized tuple grouping algorithms such as Tuple Space Search algorithm that are driven by hashing techniques [5].

Scalable Semantic Web Data Management Using Vertical Partitioning: "The previously proposed"property table" optimization has not been adopted in most RDFdatabases, perhaps due to its complexity and inability to handlemulti-valued attributes. To overcome that the proposed vertically partitioning tablesdemonstrated that they achieve similar performanceas property tables in a row-oriented database [6]. Integrating Vertical and Horizontal Partitioning intoAutomated Physical Database Design: "This proposed techniques enable a scalable solutionto the integrated physical design problem of indexes, materializedviews, vertical and horizontal partitioning for both performanceand manageability. Our work on horizontal partitioning focuseson single-node partitioning [7].

4. Slicing

To preserve the sensitive data, a slicing anonymization method is used. The dataset is partitioned horizontally and vertically. By grouping the attributes based on the associations among the attributes, the vertical partitioning is accomplished. The grouping of attributes is accomplished with the help of partitioning. At the end, within each and every group, the values in each column are arbitrarily sorted to split the linking between dissimilar columns. To improve the data utility, the modified slicing method is used [3].

4. 1 Attribute partitioning:

To handle high-dimensional data, slicing is an effective method. By partitioning the attributes into columns, this method diminishes the dimensionality of the data. It enables the slicing to handle high dimensional data. So highly correlated attributes are in the same column and the uncorrelated items are discarded. The highly correlated attributes are conserved so as to preserve the correlation among the attributes for high data utility. To evaluate correlations between two continuous attributes, where mean-square contingency coefficient is

$$\Phi 2(A_1, A_2) = [1/\min\{d_1, d_2\}-1] \sum_{i=1} d1 \sum_{j=1} d2$$

$$\{(fij - (fi, fj))^2/(fi, fj)\} \quad (1)$$

4. 2 Attribute Clustering:

After the computation of correlations of each pair of attributes, clustering is used to partition the attributes into columns. After finding the correlation between attributes, it is dissimilar from clustering, the vertical partitioning is done based on attribute correlation coefficient and the clustering can be applied based on the attribute selection of a database.

Step 1: To evaluate the association between two continuous attributes first to define correlation coefficient.

Step 2: Horizontal partitioning can be applied.

Step 3: Attribute clustering for vertical partitioning is based on attribute correlation coefficient to evaluate correlations between the sensitive attribute SA and each attribute, the identifier is defined below:

$$d(A_1, A_2) = 1 - \phi^2(A_1, A_2) \quad (2)$$

5. Vertical Partitioning of the Data:

Vertical partitioning of the dataset consists of subset of attributes made as columns. In that table T contains d attributes A= {A_1, A_2, ….. A_d} and c columns c_1, c_2, …cn. In our algorithm, grouping the highly correlated attributes are in the same column. Grouping the high dimensional correlated attributes preserves the data utility that means data should maintain privacy and be handled securely. The maintained data are useful for analysis in data mining. The remaining

uncorrelated attributes provide more identification risk in multi dataset. The identification risk of highly correlated attribute is less when compared withuncorrelated attributes. In the table frequent values of the attributes are mined with the help of closed Apriori enhancement algorithm [4].

5.1 Closed Apriori Enhancement:

In this enhancement the domain of each attribute is the itemset and implements the enhancement as an Apriori like algorithm. It directly mines frequent closed itemsets from the table. There are two main steps:

The first is to use Top- down search to identify generators, the smallest frequent itemset that determines a closed itemset. All generators are found using a simple modification of Apriori. After finding the frequent sets at level k, Close compares the support of each set with its subsets at the previous level. If the support of an itemset matches with the support of any of its subsets, the itemset cannot be a generator and is thus pruned.

The second step in Close is to compute the closure of all the generators found in the first step. To compute the closure of an itemset, we have to perform an intersection of all transactions, where it occurs as a subset. The closures for all generators can be computed in just one database scan, provided all generators fit in memory. Nevertheless computing closures this way is an expensive operation.

5. 1. 1 Closed Apriori algorithm Pseudo code:

Join Step: C_k is generated by joining Lk-1with itself

- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code: C_k: Candidate itemset of size k

L_k : frequent itemset of size k
L_1 = {frequent items};
For(k = 1; L_k !=∅; k++) do begin
C_{k+1} = candidates generated from L_k;
for each transaction t in database do increment the count of all candidates in C_{k+1} that are contained in t
L_{k+1} = candidates in C_{k+1} with min_support
End Return∪kL_k;

5.2 Vertical Overlapping Slicing

In slicing each attribute is exactly in one column. In this approach the sensitive attributes are combined at different sets to provide enhanced anonymity. Duplicating the attribute in more than one column provides better data utility, but this releases more attribute correlations. But the privacy implications must be carefully maintained.

5.2.1 Grouping of attributes:

The vertical grouping of attributes plays a major role in privacy preservation. In that the data items are considered as sets and named set D. From that we derived identification elements like ID and it is processed with the help of slicing methods**.** Given a data set D with elements e_i (i = 1, 2, ···, n) where |D|=n is the size of the data set. D can be made hierarchical by k where k = max (frequency (e_i)), frequency (e_i) is the occur times of $e_i$within D. Pair wise elements {e_i, e_j } stand for the co-occurrence within D where i ≠ j. The first order of data set identification is

$Id'_i = Id_i – Id_{i+1} = (slice_i – slice_{i+1}) – (slice_{i+1} – slice_{i+2})$ (3)

where $1 \le i \le k - 1$, and let. Id'_i= constant $_i$

In order to collect the second level below, we identify the repetition of data items in the collected datasets.

$Id''_i = (d'_i)'$

$= (Id_i – Id_{i+1})'$

$= ((slice_i – slice_{i+1}) – (slice_{i+1} – slice_{i+2})) – ((slice_{i+1} – slice_{i+2}) – (slice_{i+2} – slice_{i+3}))$ (4)

By Definition 1, we have$slice_i$ = {e_i|frequency (e_i) = 1, $1 \le i \le |D|$}. By algorithm we know that e_i i = 1, 2, ···, n are tagged with frequency frequency(e_i). Then, $slice_i$ is the collection of co-occurrent Descriptor Name pairs with frequency(e_i) = 1. Thus, we know that co-occurrent Descriptor Name pairs of different frequency value are also distinct. That is, $ld_i \cap ld_j = \emptyset$ with $i \ne j$.

6. Cost Minimization:

The cost is normally set to 1 for each of the operations. The diagonal jump can cost either one, if the two characters in the row and column do not match or 0, if they do. Each cell always minimizes the cost locally.

Consider the average cost of a single-output data

$$AC(w, y) = \frac{C(w, y)}{y} \quad . \qquad (5)$$

Economies of scale are present at any given output level if *AC(w, y)* falls as *y* increases. Similarly, when *AC(w, y)* rises with *y*, diseconomies of scale are present. In the multi-output case, average cost is not defined in the usual sense. We may, however, define the ray average cost for a given output bundle y^0 as

$$RAC(w, t; y^0) = \frac{C(w, ty^0)}{t}. \qquad (6)$$

As in the single output case, scale economies (diseconomies) are present when the ray average cost declines (increases) with an increase in the output scale. In processing economics the output level (scale) where the average cost (ray average cost) reaches a minimum is called the efficient scale of processing. The dual or economic scale efficiency of a data is measured by the ratio of the minimum (ray) average cost attained at this efficient scale and the average cost at its actual processing scale. This measure shows by what factor a data can reduce its average cost (ray average cost) by altering its output scale to fully exploit economies of scale.

The minimum average cost can be obtained by exploiting the following two useful propositions:

(P1) Locally constant returns to scale holds at the output where the average cost (ray average cost) is minimized; and

(P2). When constant returns to scale holds everywhere, the average cost (ray average cost) remains constant.

The costs are obtained from each of the aboveexperiments against the cost incurred when using datatransfersfrom the nearest (with respect to the computed resourcewhere the task is assigned) Cloud storage resource. Wemeasure the total cost

incurred for transferring data fromthe nearest location by making compute-resource cost: zero (relatingto publicly available resources) and non-zero (relatingto commercial Cloud resources), consecutively. Finally the privacy preservation successfully handles and reduces the overall processing cost of the data sets.

7. Result Evaluation:

The below figure shows the performance level between the existing and proposed algorithms of our concept. In that the classification and feature selection are achieved in a better manner. Figure 3 shows the classification between the AEA and SLS. Figure 4 shows the feature selection between the AEA and SLS. This can be explained as below:

Fig 3: This image shows the classification performance between the AEA and SLS. Through that the data items are segregated and classified based on the itemsets. The big data levels are mentioned in Y axis and the data items are mentioned in an X axis.

Fig 4: This image shows the feature selection between the AEA and SLS. Through that the data items attributes are mined and grouped under the cluster head. The cluster heads are allocated based on the type of data. The big data levels are mentioned in Y axis and the data items are mentioned in X axis.

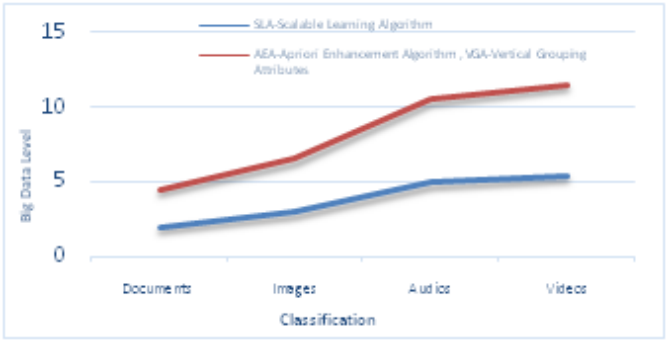

Fig3: the classification between the AEA and SLS

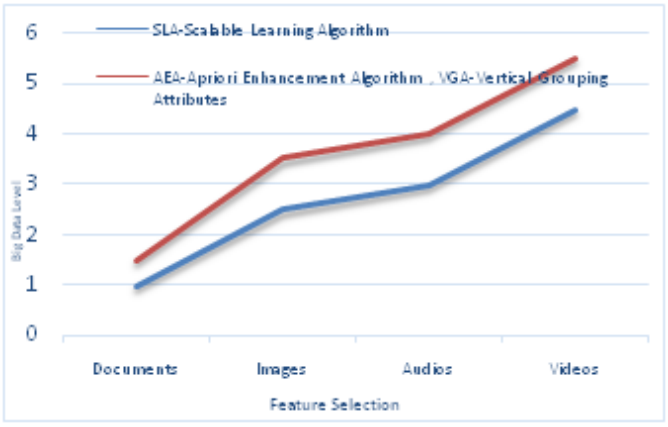

Fig4: the feature selection between the AEA and SLA

7. Conclusion:

In thispaper the problem of identifying the clusters in multi-dimensional data sets is handled by the proposed algorithm of an Apriori enhancement algorithm. This algorithm directly mines a frequent closed itemsets in a database. The mining can be processed based on the attribute of an itemsets. Based on this the related datasets are grouped. The open problem of privacy preserving can be achieved through vertical partitioning approach in our paper. And also the multiple datasets are analyzed and grouped under the concept of cluster heads in a dig data.

References:

[1] Bharti Thakur, Manish Mann, "Data Mining for Big Data: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 4, Issue 5, May 2014.

[2] SangeetaBansal, Dr. Ajay Rana, "Transitioning from Relational Databases to Big Data", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 4, Issue 1, January 2014.

[3] P. Nithya, V. Karpagam, "Improving Privacy And Data Utility For High- Dimensional Data By Using Anonymization Technique", International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801, Vol. 2, Special Issue 1, March 2014.

[4] K. Vani, B. Srinivas, "Enhanced Slicing For Privacy Preserving Data Publishing", The International Journal Of Engineering And Science (IJES), ISSN(e): 2319 – 1813, Volume 2 Issue 10 Pages 01-04-2013.

[5] http: //en. wikipedia. org/wiki/Feature_selection.

[6] EdaraRudrani Devi, MrK. RajaSekhar, "Data Privacy Preserving using RandomGrouping", International Journal For Development Of Computer Science & Technology, ISSN-2320-7884 (ONLINE), VOLUME-1, ISSUE-III, 2013.

[7] Daniel J. Abadi, Adam Marcus, " Scalable Semantic Web Data Management Using VerticalPartitioning", VLDB '07, September 23 - 28, 2007.

[8] Sanjay Agrawal, Beverly Yang, "Integrating Vertical and Horizontal Partitioning intoAutomated Physical Database Design", ACM 1-58113-859-8/04/06, SIGMOD 2004.

[9] Wei Fan and Albert Bifet "Mining Big Data: Current Status and Forecast to the Future", Vol. 14, Issue 2, 2013.

[10] Big Data Analytics by Sachidan and Singh – Paper published in 2012 International Conference on Communication Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India.

[11] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-Down Specialization for Information and Privacy Preservation, " Proc. Int'l Conf. Data Eng. (ICDE), pp. 205-216, 2005.

[12] G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data, " Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.
[13] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-Preserving Anonymization of Set-Valued Data, " Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 115-125, 2008.
[14] J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing, " Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.
[15] L. Sweeney, "k-Anonymity: A Model forProtecting Privacy, " Int' lJ. Uncertainty Fuzziness andKnowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.
[16] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation, " Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.
[17] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate Query Answering on Anonymized Tables, " Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 116-125, 2007.
[18] Cornell D. W., Yu P. S. An Effective Approach to Vertical Partitioning for Physical Design of Relational Databases. IEEE Transactions on Software Engg, Vol 16, No 2, 1990.
[19] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-Down Specialization for Information and Privacy Preservation, " Proc. Int'l Conf. Data Eng. (ICDE), pp. 205-216, 2005.
[20] "Big DataA New World of Opportunities", NESSI White Paper, December 2012.
[21] "Big Data Analytics Advanced Analytics in Oracle Database", An Oracle White Paper March 2013.
[22] Sankita Patel, Viren Patel, Devesh Jinwala, "Privacy Preserving Distributed K-Means Clustering in Malicious Model Using Zero Knowledge Proof", Distributed Computing and Internet Technology, vol. 7753, pp 420-431, 2013.
[23] Jinfei Liu, Li Xiong, Jun Luo, Joshua Zhexue Huang, "Privacy Preserving Distributed DBSCAN Clustering", transactions on data privacy, vol. 6, pp. 69–85, 2013.
[24] Bipul Roy, "Performance Analysis of Clustering in Privacy Preserving Data Mining", International Journal of Computer Applications & Information Technology, Vol. 5, no. II, May 2014.
[25] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin and Edward Y. Chang, "Parallel Spectral Clustering in Distributed Systems", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 33, No. 3, pp. 568 – 586, 2011.
[26] Eshref Januzaj, Hans-Peter Kriegel and Martin Pfeifle, "Scalable Density-Based Distributed Clustering", Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 231-244, 2004.
[27] Chandrashekhar Hegde, Srinivasan Gopalakrishnan "big data spectrum" Infosys powered by intellect driven by values, 2011.

V. S. THIYAGARAJAN obtained his Bachelor's degree in Computer Science Engineering from Department of Computer Science, Annamalai University. Then he obtained his Master's degree in Computer Science and Master's degree in Business Administration from the same university. He has also obtained Post Graduate Diploma in Data Mining from Directorate of Distance Education, Annamalai University. Currently he is doing his research in Big Data Analytics. His specializations include Big Data, Ontology, Data Security, Web Technologies, Database Management System and Computer networks.

Dr. K. VENKATACHALAPATHY received his Master's degree in Computer Applications from Pondicherry University in 1990. He completed his Ph. D in Computer Science & Engineering from Annamalai University, Tamilnadu, India in 2008. He is
He is currently working as Professor in the Department of Computer Science & Engineering, Faculty of Engineering & Technology, Annamalai University. He is having 22 years of experience in teaching. He has published more than 20 research papers in international conferences and journals. His field of interest includes Image Processing, Data Mining and Computer networks. He is currently guiding 9 research scholars towards Ph. D. He is a life member in various professional bodies like ISTE, CSI. Etc.,