ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

Two Layered Privacy Architecture for Big Data Framework

Namavaram Vijay[1], Ajay Babu Sriramoju[2], Ramesh Gadde[3]

IT Product Manager, Practicepa Ltd, UK[1]

Programmer Analyst , Randstad Technologies, *USA* [2]

Assistant Professor, Mekelle Institute of Technology, Mekelle University, Ethiopia [3]

ABSTRACT: Big data is used for gathering and analyzing a huge volume of real-time produced data efficiently and effectively, but it has also a great volume of sensitive data arising invasion of privacy. Big data analysis will give us very customized and effective analysis results, but this technology can be abused for privacy invasion of personal users. This paper introduces sensitive information which can be integrated at the data collection stage, data analysis stage, and presentation service stage. The proposed system will provide the steps for hiding protective information from the big data processing databases, and these steps are presented in detail with some examples in this paper.

KEYWORDS**:** Big Data, Privacy, Security, De-identification

I. INTRODUCTION

Mobile Ad Hoc Networks (MANETs) consists of a collection of mobile nodes which are not bounded in any infrastructure. Nodes in MANET can communicate with each other and can move anywhere without restriction. This non-restricted mobility and easy deployment characteristics of MANETs make them very popular and highly suitable for emergencies, natural disaster and military operations.

Nodes in MANET have limited battery power and these batteries cannot be replaced or recharged in complex scenarios. To prolong or maximize the network lifetime these batteries should be used efficiently. The energy consumption of each node varies according to its communication state: transmitting, receiving, listening or sleeping modes. Researchers and industries both are working on the mechanism to prolong the lifetime of the node's battery. But routing algorithms plays an important role in energy efficiency because routing algorithm will decide which node has to be selected for communication.

The main purpose of energy efficient algorithm is to maximize the network lifetime. These algorithms are not just related to maximize the total energy consumption of the route but also to maximize the life time of each node in the network to increase the network lifetime. Energy efficient algorithms can be based on the two metrics: i) Minimizing total transmission energy ii) maximizing network lifetime. The first metric focuses on the total transmission energy used to send the packets from source to destination by selecting the large number of hops criteria. Second metric focuses on the residual batter energy level of entire network or individual battery energy of a node [1].

Recent smart phones, real-time social networking services, and IT infrastructures make our daily lives very comfortable and effective, but those things also make huge data in real time. Frameworks of big data provide data storage architecture, data collecting functions, data analysis tools, and data presentation facilities for such huge and fast-producing data [1]. The big data framework can be utilized for high quality decision making system and more various value-aided business services [2].

Over the last few decades, most of data analyses had been focused on the specified or defined data collection. In the present days, they have changed into the big data analysis, which collect as many as possible from target domain or systems for various analysis. Big data would change the future business trend, and more it would make the industrial structures transformed [3]. Through big data analyses, many companies are trying to recognize various changes in consumer tastes and behavior in real time, in order to make new business models [4]. Different kinds of information are gathered through various channels in big data environment, and that information include demographic variables and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

personal sensitive data such as preferences, assets, health status, residence, contacts, content browsing, purchase history, and other personal data. Therefore such information can be seen in all procedures of big data analysis and of course invasion of privacy happens. Moreover, sometimes private sensitive information could be used intentionally for some commercial purposes [5-6].

Actually the entire data in big data environment could be disclosed and shared widely in the near future due to many technological and business reasons, so the privacy protection should be securely considered and expanded in big data processing environment [7]. Sometimes progress in technology gives us convenient and efficient environment, but also gives us possibility for invasion of privacy.

The pilot study of this paper has been presented at PlatCon-14 and this paper is an extended version of that paper [8].

This paper introduces the general concept of big data and its technological points, and proposes a security architecture for privacy protection in big data, and then presents its security and performance analyses. Section 2 surveys various categorized technologies of big data frameworks. Section 3 shows the proposed privacy protection scheme. Section 4 gives practical examples of the privacy invasion and the proposed scheme's role for making them up and the performance and security enhancement of the proposed scheme. Section 5 concludes the paper and discusses the open questions.

II. RELATED WORK

Big Data Processing Technology

Generally the big data domain can be categorized into the following four parts; data collection, data storing, data analysis, and data presentation. This classification and their flow are illustrated in Figure 1. This section introduces these four parts and their current technologies, security issues, and considerations [9-10].

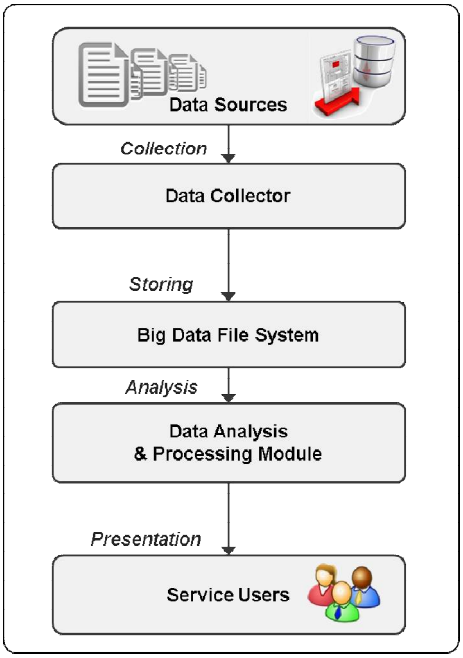

Figure 1. Conceptual Framework of Big Data Technology

Big Data Collection

This technology includes all data from internal and/or external sources in any forms. For example, Chukwa, Scribe and Flume can be used for collecting huge volume of log data, and web robots/crawlers collect all of web contents. Specially the Chukwa can collect monitoring logs, application logs, system logs, Hadoop logs and any kinds of log data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

using Chukwa Agent and Chukwa Collector, and finally stores all that collected log data in HDFS. Scribe is an application system which collects real-time streaming data log from large number of servers. Scribe server can provide collection service based on a layered and distributed architecture. Flume can provide reliable collection service on a distributed environment and it can extend its capability easily according to the service domain size. Flume is used for the collection service for being used in HDFS [11-12].

Big Data Storing

In the framework of big data, the amount of the data is too huge to be processed on a single computing device, but it should be dealt on the distributed computing environment. All kinds of distributed computing technologies should be prepared for the big data framework. HDFS, Hadoop Distributed File System which is an official file system of the Apache Hadoop project can store the collected huge amount data into a distributed storage system. And HDFS can monitor and control all distributed data on its hierarchy using HTTP protocol for moving, replications, and rebalancing of contents. No SQL is highly adopted for manipulating unstructured data in a different manner from the conventional relation based database architecture.

Big Data Analysis

All of recent emerging data analysis technologies include the function to analyze unstructured and informal text mining and they can be used easily for SNS analyses. Such analysis technologies are based on NLP, Natural Language Processing and they can find meaningful keywords, relations, or patterns from the flat and complicated texts. MapReduce is a software framework for processing big data in parallel computing resources and its Map phase and Reduce phase provide an effective text analysis results. The Map phase produces relations from the unrelated and huge data, and the Reduce phase removes redundant relations from the Mapped data, and then finally meaningful relations are popped-up after these two phases. By the way, Big Query, one of online OLAP systems, can provide real-time analyzing facility for Tera-Bytes level data using the Google search engine infrastructure.

Big Data Presentation

Representative analysis tool R provides modelling, statistical computation, cutting edge data mining methods, and finally visualization interpreter and integrated development environment. In Vis, an Interactive Visualization Framework for Massive Data supporting Multiple Users , is also well-known Info-Graphic tool set and its strong points include real-time visualization efficiency and user convenient interface.

Privacy Requirements

In these 4 processing states of the big data framework, the incursion of privacy can happen, so this paper investigates the privacy requirements on each phase. It is privacy invasions to collect personal information that is not permitted by the users on the basis of privacy protection laws [13]. And also there is a big privacy risk that sensitive personal information could be collected through the monitoring of services such as log data on users. Therefore, a notice must be made on the use of personal information through a legal consent process in the collection phase of user data service.

The big data storage phase has the need for secure storage and management on the collected data. When the stored data is illegally leaked to the outside of the system by the intrusion may secure the integrity of the data. In addition, if the collected data contains unnecessarily sensitive personal information, they should be de-identified. Personal information de-identification is generally to remove or mask a part or all of the personal data in order to make difficult to identify a specific person associated to other sensitive services or information.

The collected protective data should not be processed and analyzed for the purpose of use violating data acquisition consent. At this phase, non-sensitive data could be more privacy data by a combination of other data results or other data analysis, and it may cause an invasion of privacy.

Ultimately in the big data presentation phase, it should be controlled and blocked to abuse the analyzed data. The results obtained by the big data analyses should not be used or provided in addition to the purpose agreed upon the data collection phase. Personal sensitive information should be provided to comply with the protection policy, the use and service of the stored sensitive information must be carried out with access only by the legal representative.

Privacy requirements must be carried out in each phase to protect user privacy in big data processing framework. In addition, the development of technology and its corresponding privacy policy for the collection of sensitive personal information security classification and safe handling of data considering the process is required.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

All of the collected data should be classified according to its privacy level, and they must be processed through a secure private information processing according to its privacy level. For this, it is very important to develop privacy policy and its corresponding technology.

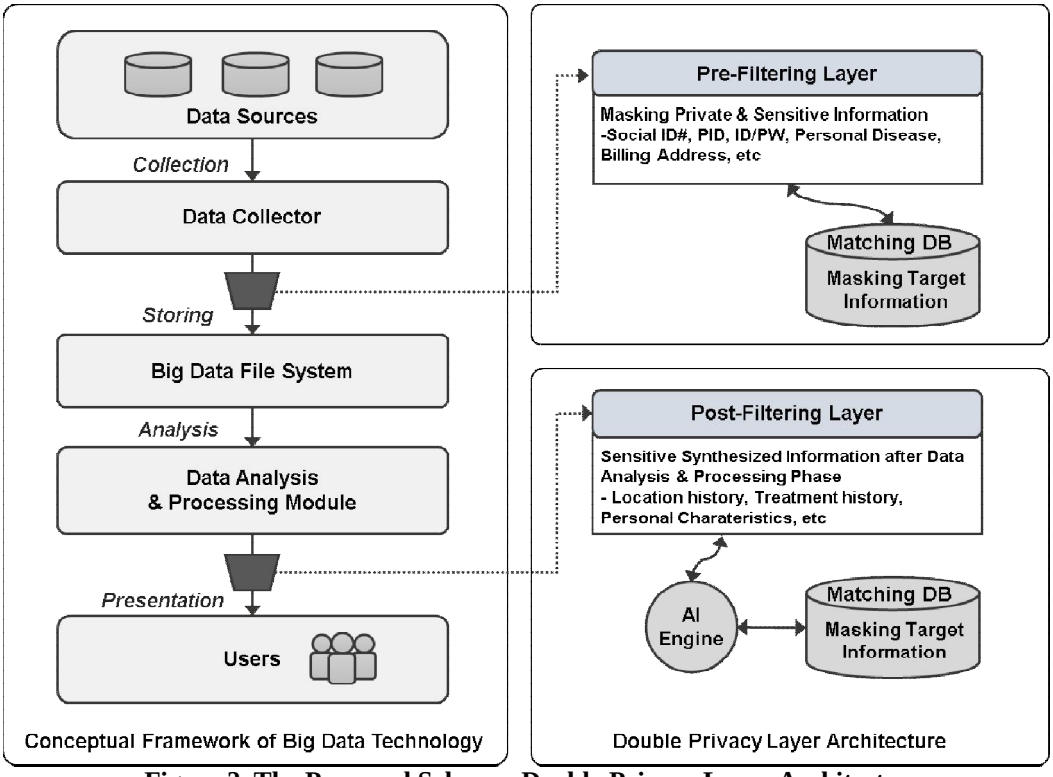

Figure 2. The Proposed Scheme: Double Privacy Layer Architecture

2. Double Privacy Layer Architecture

This paper proposed a security design which adds double layered privacy filters to the existing big data technology framework. This security architecture is divided into the pre-filtering layer and the post-filtering layer. The above figure shows the double privacy layer architecture which offers privacy enhancing in the big data environment.

The Pre-Filtering Layer

The pre-filtering layer is the first privacy layer of the proposed architecture and it works in the data collection phase, and it finds and deletes personal sensitive information form the collected data. This data reduction is called the de-identification process, it removes some or all individual sensitive information to be combined with other information, so that the system make difficult to identify a particular person.

For an example, if the personal ID information such as SSN(social security number) of the collected data has been detected by the pre-filter, the privacy layer will remove other number fields only except the information corresponding to the age and gender. In Korea, the social security number has some sensitive information includes the birth date, gender, nationality, the birth location, and even the secret parity number of the person. In this case, only birth date and gender can be stored in the big data system in this privacy layer architecture. All patterns should be removed and filtered are stored in the matching database system for continuous pattern updating. The pre-filter must deactivate all sensitive information depending on the prescribed guidelines in the matching DB. Examples of personal information to be filtered are as follows.

- ***Personally identifiable information****: Name, Telephone number, Address, Birth date, Face photo, etc.*
- ***A unique identification information****: Social Security Number, Driver's license numbers, Biometric information, Member ID, Email, etc.*

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

- ***De-identification Example***: *[Name, SID, Date of Birth, Gender, Job Title, Address, Email] → [Age, Gender, Area]*

Post-Filtering Layer

The post-filtering layer is masking sensitive information synthesized after the big data analysis. Sometimes non-identifiable data and non-sensitive personal information are aggregated and associated to be a sensitive personal information in the analysis phase. Therefore the post-filter should be located before the presentation phase and also the verification filtering and periodic monitoring should be performed for masking the re- identifying potentials. This is an essential mechanism for the privacy protection in the big data analyzing and presenting process of the big data framework. The system should always check the users' opt-in consent and finds re-identification cases before the presentation services.

In the presentation phase, within the scope permitted by the authority of the requesting authorities and service purpose, a dynamic privacy masking process can be performed. The followings are example of re-identifying information synthesized during the analysis phase.

- ***Re-identifying information***: *Behaviour information, Preference information on personal health, disease, religion, hobby and foods, Location history information, Political activities, Property information, Life style, Living pattern and so on.*

III. SECURITY ANALYSIS

There are apparently security threats related to the invasion of privacy in the collection, storing, analysis, presentation of the big data framework. In the proposed system, most importantly the standardized or formal types of private information will be filtered and removed soon after data collection and before data storing, so this first- level filtering layer reduces security threats considerably.

The remaining threats can be dealt by the second security filter layer of the proposed architecture before presentation and after data analysis procedure. This makes the proposed scheme very useful and effective in a real world application.

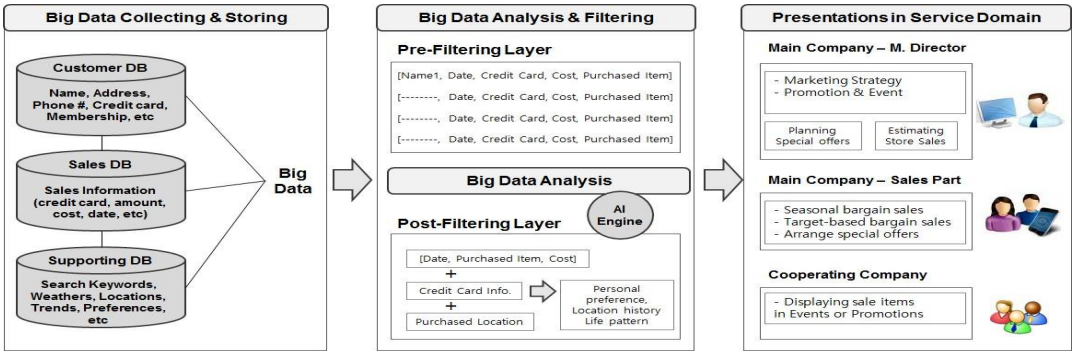

Figure 3. Application Examples: Double Privacy Layer Architecture

Privacy Enhancing with Double-Filtering

Above Figure 3 shows a sample case for the proposed scenario, which is used in the big data framework for a company marketing decision. The company has a customer database and sales database, and also collects data from the web and SNS. Those data are stored in the company big data storage. And the companies big data solution performs a big data analysis for making a reasonable inference on the customers thinking, trend, or needs.

In this company, the first privacy filter is adopted before the data analysis and the second filter is used soon after the data analysis. These two filters remove the originally protective personal information from the data sources and also mask every synthesized privacy violating stuffs from the analysis results such as personal preferences on health, disease, religion, hobby and foods, personal location history, personal life patterns, and personal behaviour patterns.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

IV. CONCLUSIONS

The proposed paper provides security architecture, the double privacy layer architecture which has two security layers easily adapted to the general framework of big data analysis, and those additional layers are used to remove or mask sensitive personal information. The Pre filtering Layer filters big data inputted by the data collector with pre-defined masks of privacy information which have been well-defined and stored as standard templates in the matching database. There could be some non standardized protective information, and also there could be some privacy information synthesized or revealed in the big data analysis procedure. For example, something like a location history of a certain person, or a residence pattern of a certain person in his/her house are commonly synthesized information in the big data analysis. The Post filtering Layer will find and remove such privacy information from them, using the artificial intelligence engine which has been trained for this goal. Consequently, in our security architecture, sensitive personal information will be removed before the storing and also before the presentation.

REFERENCES

[1] Edd Dumbill, http://strata.oreilly.com/2012/what is big data.html, "What is big data?," O'Reilly, **(2012)** November.
[2] Data, Rethinking Personal. "Strengthening Trust," World Economic Forum. **(2012)** May.
[3] Manyika, "Big data: The next frontier for innovation, competition, and productivity,"**(2011)** May.
[4] C.H. Lee, J. Hur, H.J. Oh, H.J. Kim, P.M. Ryu and H.K. Kim, "Technology Trends of Issue Detection and Predictive Analysis on Social Big Data," Electronics and Telecommunications Trends, ETRI, **(2013)**, pp. 62-71.
[5] IDG, "Worldwide Big Data Technology and Services View report", IDG Tech Report, http://www.idc.com, **(2012)**.
[6] N. MacDonald, "Information Security Is Becoming a Big Data Analytics Problem," Gartner Group,**(2012)** March.
[7] "Method for activating research and institutional assignments Big Data Services," National Information Society Agency, **(2013)**.
[8] Si-Jung Kim, "Double Secure Layers Architecture for Privacy Protection in Big Data," In proceedings of the 2014 International Conference on Platform and Service (PlatCon-14), **(2014)** January.
[9] "An Analysis of Technology Demand for Big Data based Privacy Information Protection", Industry Academic Cooperation Foundation Sungshin women's university, **(2012)** December.
[10] "Big Data Analytics Report", TDWI Research, **(2011)**.
[11] "Welcome to Apache Hadoop," http://hadoop.apache.org/.
[12] Owen, Kan Zhang, "Hadoop Security Design," Yahoo, Inc., Technical Report, **(2009**) October.
[13] Hyeong-hyo Lee, "Cloud Computing Security Trends," National IT Industry Promotion Agency, Week Technology Trends, **(2011)**. pp.12-23
[14] "The OECD Privacy Framework", OECD, **(2015).**
[15] Shoban Babu Sriramoju, "Review on Big Data and Mining Algorithm" in "International Journal of Research in Applied Science & Engineering Technology", Vol-5, Issue-XI, November 2017 [ISSN : 2321-9653].
[16] Shoban Babu Sriramoju, "OPPORTUNITIES AND SECURITY IMPLICATIONS OF BIG DATA MINING" in "International Journal of Research in Science and Engineering", http://ijrise.org/asset/archive/17Dec7.pdf, Vol-3, Issue-6, Nov-Dec 2017, 44-58 [ISSN : 2394-8299].
[17] Shoban Babu Sriramoju, "A FRAMEWORK FOR SOLVING IDENTITY DISCLOSURE PROBLEM IN COLLABORATIVE DATA PUBLISHING" in "International Journal of Research in Science and Engineering", http://ijrise.org/asset/archive/17Dec8.pdf, Vol-3, Issue-6, Nov-Dec 2017, 59-66 [ISSN : 2394-8299].
[18] *Shoban Babu Sriramoju,* "Brand Clustering Based on Social Big Data : A Case Study", International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE), *https://ijircce.com/upload/2017/october/52_Brand%20Clustering%20Based%20on%20Social%20Big%20Data%20ijircce.pdf, Volume 5 Issue 10, October 2017, 15958 – 15965 [ISSN :* **2320-9801]**
[19] *Shoban Babu Sriramoju, " Heat Diffusion Based Search for Experts on World Wide Web", International Journal of Science and Research (IJSR), https://www.ijsr.net/archive/v6i11/v6i11.php, Volume 6 Issue 11, November 2017, 632 - 635, #ijsrnet*
[20] Dr. Shoban Babu Sriramoju, Prof. Mangesh Ingle, Prof. Ashish Mahalle "Trust and Iterative Filtering Approaches for Secure Data Collection in Wireless Sensor Networks" in "International Journal of Research in Science and Engineering" Vol-3, Issue-4,July-August 2017 [ISSN : 2394-8299].
[21] Dr. Shoban Babu, Prof. Mangesh Ingle, Prof. Ashish Mahalle "HLA Based solution for Packet Loss Detection in Mobile Ad Hoc Networks" in "International Journal of Research in Science and Engineering" Vol-3, Issue-4,July-August 2017 [ISSN : 2394-8299].
[22] Shoban Babu Sriramoju. "A Framework for Keyword Based Query and Response System for Web Based Expert Search" in "International Journal of Science and Research" Index Copernicus Value(2015):78.96 [ISSN : 2319-7064].
[23] Sriramoju Ajay Babu, Dr. S. Shoban Babu. "Improving Quality of Content Based Image Retrieval with Graph Based Ranking" in "International Journal of Research and Applications" Vol-1, Issue-1,Jan-Mar 2014 [ISSN : 2349-0020].
[24] Dr. Shoban Babu Sriramoju, Ramesh Gadde. "A Ranking Model Framework for Multiple Vertical Search Domains" in "International Journal of Research and Applications" Vol-1, Issue-1,Jan-Mar 2014 [ISSN : 2349-0020].
[25] Mounika Reddy, Avula Deepak, Ekkati Kalyani Dharavath, Kranthi Gande, Shoban Sriramoju. "Risk Aware Response Answer for Mitigating Painter Routing Attacks" in "International Journal of Information Technology and Management" Vol-VI, Issue-I, Feb'2014 [ISSN : 2249-4510]
[26] Mounica Doosetty, Keerthi Kodakandla, Ashok R, Shoban Babu Sriramoju. "Extensive Secure Cloud Storage System Supporting Privacy-Preserving Public Auditing" in "International Journal of Information Technology and Management" Vol-VI, Issue-I, Feb'2012 [ISSN : 2249-4510]
[27] Shoban Babu Sriramoju. "An Application for Annotating Web Search Results" in "International Journal of Innovative Research in Computer

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

and Communication Engineering" Vol-2,Issue-3,March'2014
[ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]
[28] Shoban Babu Sriramoju. "Multi View Point Measure for Achieving Highest Intra-Cluster Similarity" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol-2,Issue-3,March'2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]
[29] Shoban Babu Sriramoju, Madan Kumar Chandran. "UP-Growth Algorithms for Knowledge Discovery from Transactional Databases" in "International Journal of Advanced Research in Computer Science and Software Engineering" Vol-4,Issue-2,February'2014 [ISSN : 2277 128X]
[30] Shoban Babu Sriramoju, Azmera Chandu Naik, N.Samba Siva Rao. "Predicting The Misusability Of Data From Malicious Insiders" in "International Journal of Computer Engineering and Applications"
Vol-V,Issue-II,Febrauary'2014 [ISSN : 2321-3469]
[31] Ajay Babu Sriramoju, Dr. S. Shoban Babu. "Analysis in Image Compression Using Bit-Plane Separation Method" in "International Journal of Information Technology and Management"
Vol-VII, Issue-X, Nov' 2014 [ISSN : 2249-4510]

[32] Shoban Babu Sriramoju. "Mining Big Sources Using Efficient Data Mining Algorithms" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol-2, Issue-1,January'2014
[ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]
[33] Mr. Ajay Babu Sriramoju, Dr. S. Shoban Babu. "Objective Quality Metric Design for Wireless Image and Video Communication" in "International Journal of Information Technology and Management"
Vol-VII, Issue-I0, Nov' 2014 [ISSN : 2249-4510]
[34] Ajay Babu Sriramoju, Dr. S. Shoban Babu. "Study of Multiplexing Space and Focal Surfaces and Automultiscopic Displays for Image Procesing" in "International Journal of Information Technology and Management"
Vol-V, Issue-I, Aug' 2013 [ISSN : 2249-4510]
[35] Dr. Shoban Babu Sriramoju. "A Review on Processing Big Data" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol-2, Issue-1, January'2014
[ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]
[36] Shoban Babu Sriramoju, Dr. Atul Kumar. "An Analysis around the study of Distributed Data Mining Method in the Grid Environment : Technique, Algorithms and Services" in "Journal of Advances in Science and Technology"
Vol-IV, Issue No-VII, November'2012 [ISSN : 2230-9659]
[37] Shoban Babu Sriramoju, Dr. Atul Kumar. "An Analysis on Effective, Precise and Privacy Preserving Data Mining Association Rules with Partitioning on Distributed Databases" in "International Journal of Information Technology and management" Vol-III, Issue-I, August'2012 [ISSN : 2249-4510]
[38] Shoban Babu Sriramoju, Dr. Atul Kumar. "A Competent Strategy Regarding Relationship of Rule Mining on Distributed Database Algorithm" in "Journal of Advances in Science and Technology"
Vol-II, Issue No-II,November'2011 [ISSN : 2230-9659]
[39] Shoban Babu Sriramoju, Dr. Atul Kumar. "Allocated Greater Order Organization of Rule Mining utilizing Information Produced Through Textual facts" in "International Journal of Information Technology and management" Vol-I, Issue-I, August'2011 [ISSN : 2249-4510]