Available online at www.worldscientificnews.com

World Scientific News

WSN 46 (2016) 65-76 EISSN 2392-2192

A survey on various applications and challenges of big data analytics and its security methods

A. Subashini[a], R, Yasotha[b]

PG Scholar, Department of CSE, Anna University Regional Campus, Coimbatore, Tamil Nadu, India

[a,b]E-mail address: subashini.4be@gmail.com , yasothar.60@gail.com

ABSTRACT

Big-data is loosely distributed processing applications that operate large amount of knowledge. The growing power of Big Data assume the significance of analyzing huge amount of data with a frequent and quick rate of growth and change in databases and data warehouse. MapReduce is most common framework for large-scale information analytics chiefly owing to its salient options like quantifiability, fault-tolerance, easy programming, and adaptability. Apache Hadoop is an emerging technology that is widely used in the data intensive applications like Big Data Analysis. This technology is currently used in the searching applications of Google, Yahoo, and Amazon. Big data applications are an incredible advantage to associations, business, organizations and numerous huge scale and little scale businesses. Cloud computing assumes an extremely essential part in protecting information, applications and the related foundation with the assistance of strategies, technologies, controls, and enormous information tools. In addition, cloud computing, big data and its applications, advantages are liable to speak to the most promising new frontiers in science. This paper presents a summary of the distinctive options that differentiate massive information from traditional datasets. Additionally, the application of massive data analytics within the E-commerce and also the varied technologies that build analytics of client information potential is mentioned. It also summarizes about the need for Big data in Health care and in Government. Issues or challenges to big data analytics are discussed in greater level of concern. Protection and Security methods such as Vormetric Encryption, Data Security Platform, Encryption and Key Management, Fine-grained Access Controls, Security Intelligence and Automation are explained which is used to exploit the challenges to Big data.

***Keywords*:** Interest Locality; Data Groupings; CloudBLAST; Cloudburst

1. INTRODUCTION

Big data belongs to datasets whose size is outside the potential of usual data software package tools to capture, store, manage, and analyze. Big data, portrayed by the uncommon volume of information, information era speed, and basic assortment of information, support for broad information examination to structure a mainly challenging task. The main goal of big data is to offer organizations some assistance with making better business choices by encouraging information researchers and clients to break down large volumes of exchange information and other data source.

a) **Volume:** several factors contribute towards increasing Volume

Volume - storing dealing information, live streaming information and information collected from sensors etc.

b) **Variety:** these days knowledge comes altogether styles of formats – from old databases, text documents ,emails, video, audio, transactions etc.,

c) **Velocity:** this suggests however how quick the info is being created and the way quick the into must be processed to fulfil the demand.

The other 2 dimensions that require contemplating with reference to massive knowledge are Variability and Complexity

d) **Variability:** in conjunction with the rate, the data flow is extremely inconsistent with periodic peaks.

e) **Quality:** Complexity of the information conjointly has to be thought of once the information is returning from multiple sources. the information should be coupled, matched, clean and reworked into needed formats before actual process.

2. CHARACTERISTICS OF HADOOP

Scalable – As per demand of further nodes those will be added while not creating any modification in knowledge formats, the way of data loading and within the method the roles or application are written.

Cost effective – Hadoop introduce the particularly parallel computing to trade goods servers. The output may be a considerable scale back in value that successively makes it inexpensive to model all the information.

Flexible – Hadoop is schema less therefore it will take any form of data like structured or not and it will accept knowledge from number of various sources. This knowledge will be joined and aggregated in random ways in which permitting elaborated analysis than any other system will provide.

Fault tolerant – If any node is lost, the system is ready to redirects work to a different node of the information and continues processing while not losing a beat.

Map reduce is standard due to its simple programming interface and extraordinary performance whereas applying an outsize range of applications. Such applications receive an outsize amount of file and additionally referred as "Big-data applications". A MapReduce job at first divides the data into individual chunks that are processed by Map jobs in parallel. The outputs of the map function sorted by the framework are then input to the reduce tasks. Generally the input and also the output of the work are each keep during a file-system. Scheduling, Monitoring and re-executing unsuccessful tasks are taken care by the framework.

3. APPLICATIONS

3. 1. TEXT MINING

This technique is extremely dependent on the use of text primarily based content from blogs and social media sites to create judgment on the connection of a problem. Text collected is filtered employing a keyword filter to retrieve relevant information. The Ecommerce seller generates list of keywords pertaining to the product being monitored. These keywords will be used to determine sentiments a couple of product.

3. 2. SENTIMENT ANALYSIS (BASED ON MACHINE LEARNING ALGORITHM)

This procedure of examination works utilizing either a machine learning calculation or artificial intelligence, to distinguish notions about a particular good on service. Basically, every word got from the huge information is broke down and labelled, after which it is referenced with a predefined word or equivalent word which deciphers whether the conclusion is certain or not. Each of these statements is then analyzed (using a predefined sentiment database) to predict the emotions of each word.

3. 3. PREDICTIVE ANALYSIS

Predictive analytics is that the use of past/historical information to predict future trends. This analysis makes use of applied mathematics models and machine learning rule to spot patterns and learn from historical information. Predictive analytics may also be outlined as a method that uses machine learning to investigate information and create predictions. This technology has been around for a moment, although the adoption has been low due to the quality and prices. Using the large information-analytical platform to investigate these information (alongside data processing and machine learning algorithm), E-commerce merchant will efficiently predict client behaviour quicker, more efficiently and at more effective price.

Predictive analytics will be utilized in E-commerce in the following ways:

- Product Recommendation
- Price Management
- Predictive Search

Typically an outsized e-commerce website offers thousands of product and services available. Navigating and finding out a product out of thousands on a website can be a significant reversal to customers. However, with the invention of recommender system, an E-Commerce application can quickly determine merchandise that closely suits the consumer's vogue. There are two main technologies that help the recommendation system for analysing the information:

- collaborative Filtering
- agglomeration rule

COLLABORATIVE FILTERING

Using a technology known as collaborative Filtering (CF), information of historical user preference is made. Once a new client access the e-commerce website, the client is matched with the information of preferences, so as to get a preference category that closely matches the customer's style. These products are then suggested to the new client.

CLUSTERING ALGORITHM

Clustering algorithm technique works by identifying groups of users that have similar preferences. These users are then clustered into one cluster and are given a completely unique symbol. New customers cluster are foreseen by calculating the common similarities of the individual members of this cluster. Therefore, a user may be a partial member of more than one cluster relying upon the load of the user's average opinion. The big data application refers to the massive scale distributed applications that typically work with large information sets. Information exploration and analysis became a tough drawback in several sectors in the span of massive information. With massive and complicated information, computation becomes troublesome to be handled by the traditional processing application that triggers the event of massive information applications. Google's Map reduce framework and apache Hadoop are the de facto software for big data applications, inside that these applications generates a massive amount of intermediate information. Producing and Bioinformatics are the two major areas of massive information applications. Big knowledge gives associate infrastructure for transparency in producing business that has the ability to unravel uncertainties like inconsistent part performance and availableness. In these large information applications, an abstract framework of prophetical manufacturing begins with information acquisition where there is an occurrence to accumulate different types of sensory information like pressure, vibration, acoustics, voltage, current, and controller information.

The mixture of sensory knowledge and historical knowledge constructs the large knowledge in producing. This generated huge knowledge from the on top of combination acts because the input into prophetical tools and preventive methods like prognostics and health management. Another vital application for Hadoop is Bioinformatics that covers future generation sequencing and different biological domains. Bioinformatics which needs an outsized scale knowledge analysis uses Hadoop. Cloud computing gets the parallel distributed computing framework along with computer clusters and net interfaces.

3. 4. DRAW

Big data parallel computing frameworks and large-scale distributed file systems are being continuously developed by academic and industrial pioneers to facilitate the high-

performance runs of data-intensive applications, such as bio-informatics, astronomy, and high-energy physics. Generally, many scientific and engineering applications have interest locality: 1) domain scientists are solely interested in a set of the whole information set, and 2) scientists are likely to access one subset more frequently than others. These correlated data have high possibility to be processed as a group by specific domain applications. Here in this algorithm, the "data grouping" is done to represent the likelihood of two or a lot of information (e.g., blocks in Hadoop) to be accessed as a bunch.

In Hadoop default information placement strategy there is no guarantee that the information accessed as a group is equally distributed. Thus, so as to process the distributed data in parallel, a MapReduce job is partitioned into several map tasks. Also, to exploit the predictability of data access patterns and improve the performance of distributed file systems, dynamic data grouping is effective. After analysis, the possibility for random data distribution is to evenly distribute the data from the same group. The observation shows that this possibility is affected by three factors: 1) the variety of duplicate for every information block in every rack (NR); 2) the utmost number of coincident map tasks on each node (NS); and 3) the information grouping access patterns. Therefore, a new Data-gRouping-AWare data placement scheme (DRAW) algorithm is developed that takes into consideration the information grouping effects to considerably improve the performance for data-intensive applications with interest neighbourhood. DRAW is designed and implemented as a Hadoop-version prototype.

DRAW that means "Data-gRouping-AWare" is the new information placement theme for information intensive applications with interest neighbourhood. Interest neighbourhood is that the method of looking for the desired specific dataset from the massive huge datasets. DRAW is the new rule developed using the Hadoop technology for the applications in bioinformatics. DRAW is intended at rack-level, to optimize the grouping distribution within a rack. During this rule, there are three parts: a knowledge access history graph (HDAG), an information grouping matrix (DGM) associated with an best data placement rule (ODPA). HDAG is employed to use the system log files learning data grouping information. The Namenode in every Hadoop cluster rack maintains system log files that records each system operation, together with the files that are accessed. DGM is employed to quantify the grouping weights among the information and generate the optimized data groupings. OPDA is employed to create the best information placement.

3. 5. CLOUDBLAST

CloudBLAST is that the mixture of MapReduce and virtualization on distributed resources for bioinformatics applications. This can be a replacement approach to parallelization, preparation and management of bioinformatics applications that integrates many rising technologies for distributed computing. This approach uses MapReduce paradigm to parallelize tools and manage their execution, machine virtualization to encapsulate their execution environments and usually used information sets into flexibly deployable virtual machines and network virtualization to attach resources behind firewalls/NATs whereas conserving the mandatory performance and the communication environment.

Cloud BLAST approach uses the subsequent techniques: machine virtualization, network virtualization and Hadoop. These three techniques will be combined to deploy necessary bioinformatics applications supported BLAST on computer clusters on distinct

body domains connected by a wide-area network (WAN). The experimental validation is completed by deploying a Xen-based virtual cluster across two sites using virtual workspaces for authenticating users and deploying VMs, and for connecting the nodes behind NATs, over a 200Mbps WAN link. in this approach, Apache Hadoop, an open supply implementation is employed for MapReduce paradigm to pose the execution of NCBI BLAST2, a sequential implementation of BLAST created publically accessible by the National Centre for Biotechnology information.

3. 6. CLOUD BURST

Cloudburst is an open offer tool accessible as a model for parallelizing algorithms with MapReduce. A huge quantity of sequence knowledge is generated using this next-generation DNA sequencing machines. This tool is also a scan mapping rule optimized for mapping next-generation sequence data to the human order and different reference genomes, to be utilized in a kind of biological analyses in conjunction with SNP discovery, genotyping and personal biological science. This tool uses MapReduce-based read mapping algorithm modelled after RMAP, however runs in parallel on multiple machines with Hadoop. An improvement of this algorithm is finished for mapping several short reads from next-generation sequencing machines to a reference order providing a user specified variety of mismatches or variations. RMAP is an algorithmic technique known as seed-and-extend to accelerate the mapping method. in this algorithmic rule, initial notice sub-strings referred to as seeds that specifically match in both the reads and also the reference sequences, and then extend the shared seeds into longer, inexact alignments employing a lot of sensitive formula that permits for mismatches or gaps. This technique additionally uses a range of strategies for locating and increasing the seeds, and has completely different options and performance. However, each of these techniques is used for execution on one computing node, and intrinsically needs an extended period of time or limits the sensitivity of the alignments they notice. Cloudburst may be a read mapping algorithm that indexes the non-overlapping k-merges within the reads as seeds whose size s = m / (k+1) is computed from the minimum length of the reads (m) and also the most variety of variations or mismatches (k). Thus, cloudburst is taken into account to be as a replacement sensitive parallel seed and extend read-mapping algorithm optimized for mapping single finish next generation sequence knowledge to reference genomes. This tool provides all alignments for every scan with up to a user-specified variety of variations as well as both mismatches and indels. The alignments is filtered to find the only best non-ambiguous alignment for every read, and output is similar to RMAPM (RMAP exploitation couple scores). As a result, cloudburst will replace RMAP during a knowledge analysis pipeline without ever-changing the results, however provides abundant larger performance by using the open-source implementation of the distributed programming framework MapReduce referred to as Hadoop

Mahout is a project developed by Apache whose aims to build a scalable machine learning libraries by virtue of Hadoop. RHIPE and Ricardo project are tools that integrate R applied mathematics tool and Hadoop to support parallel information analysis, cheetah could be a information storage tool designed on Map Reduce with virtual view on top of the star or snowflake schemas and with some improvement techniques like knowledge compression and columnar store. Osprey could be a shared-nothing information system that supports MapReduce-style fault tolerance. Osprey doesn't directly use MapReduce or GFS. However, the actual fact table in star schema is divided and replicated like GFS, and tasks are scheduled

by central runtime hardware like Map Reduce. A distinction is that osprey does not stop intermediate outputs. Instead, it uses a method known as bound de clustering that limits knowledge inaccessibility when node failures happen.

The Map reduce is employed for data intensive scientific analysis and bio science that are well studied in CloudBLAST parallelizes NCBI BLAST2 algorithmic program mistreatment Hadoop. They break their input sequences into multiple blocks with an instance of the vanilla version of NCBI BLAST2 for each block, using the Hadoop Streaming utility cloudburst may be a parallel scan-mapping tool that maps NGS read sequencing information to a reference order for genotyping in parallel.

4. NEED FOR BIG DATA ANALYTICS

4. 1. IN HEALTH CARE

To improve the standard of healthcare by considering the following:

Providing patient centric services

To provide faster relief to the patients by providing proof based medicine detecting diseases at the sooner stages supported the clinical information on the market, minimizing drug doses to avoid aspect effect and providing efficient medication supported genetic makeup. This helps in reducing admittance rates thereby reducing value for the patients.

Detecting spreading diseases earlier

Predicting the microorganism diseases earlier before spreading supported the live analysis. This will be known by analysing the social logs of the patients affected by a illness in a very specific geo-location. This helps the aid professionals to advise the victims by taking necessary preventive measures.

Monitoring the hospital's quality

Observance of whether or not the hospitals are setup consistent with the norms setup by Indian medical council. This periodical check-up helps government to take necessary measures against disqualifying hospitals.

Improving the treatment methods

Customised patient treatment monitoring the impact of medication unceasingly and supported the analysis dosages of medicines may be modified for quicker relief. Observance of patient is very important signs to provide proactive care to patients. Creating associate analysis on the information generated by the patients United Nations agency already suffered from an equivalent symptoms, helps doctor to produce effective medicines to new patients.

4. 2. IN GOVERNMENT

Big data analytics helps government in building smart cities by providing quicker and reliable services to its voters.

Addressing Basic desires quickly

These days' individuals ought to await a long time to induce EB, telephone, water, ration card and gas affiliation. These are the fundamental wants of citizen. It is the responsibility of the government to provide these services as fast as potential. Massive information analytics plays a significant role in achieving it as a result of the information are analysed on daily. those who are in need are served in real time.

Providing quality education

Education is one among the precious assets which will incline to the kids. It is the duty of government to produce quality education to youngsters. BDA provides elaborate report of kids who are within the age to be admitted to the college. This helps government to assess the academic wants for these youngsters immediately. To reduce unemployment rate: to reduce unemployment rate by predicting the work desires before primarily based the literacy rate. This will be achieved by analysis the scholars graduating every year. It allows government to arrange for special trainings so as to make young entrepreneurs

Other advantages

- To supply pension to senior citizens with none delay.
- To make sure that advantages provided by government reaches all the individuals.
- To manage traffic in peak times based on the live streaming information regarding vehicles.
- To observe the necessity for mobile ambulance facilities

5. ISSUES OR CHALLENGES TO BIG DATA ANALYTICS

Whenever new technologies evolve, they meet with new challenges in all the aspects. Once the functional challenges are in place, the next kin is the technical challenges. Big data faces many technical challenges which are on the roadway of the research.

a) **Failure handling**

Devising 100% reliable systems on the go is not an easy task. Systems can be devised in such a way that the probability of failure must fall within the permitted threshold. Fault tolerance is a technical challenge in big data. When a process started it may involve with numerous network nodes and the whole computation process becomes cumbersome. Retaining check points and fixing the threshold level for process restart in case of failure, are greater concerns.

b) **Data heterogeneity**

Big data deals with unstructured, semi-structured and structured data. Linking unstructured data with structured data, converting data from one form into another required form needs a lot of research.

c) **Data quality**

Huge amount of data pertaining to a problem is undoubtedly a big asset for both Business as well as IT leaders. For predictive analysis or for better decision making amount of

relevant data helps a lot. But the quality of such data is based on the source through which they are derived. Though big data stores large relevant data, the accuracy of data is completely dependent on the source domains. Hence, there is a question of how far the data can be trusted and it definitely requires appropriate trust agent filters.

Cloud computing comes with various security problems because it encompasses several technologies together with networks, virtualization, resource programming, transaction management, load balancing, concurrency control and memory management. Hence, security problems with these systems and technologies are applicable to cloud computing. For instance, it is very important for the network that interconnects the systems in a cloud to be secure. Additionally, resource allocation and memory management algorithms even have to be secure. The big data problems area unit most acutely felt in bound industries, such as telecoms, web promoting and advertising, retail and financial services, and bound government activities. The data explosion is going to create life troublesome in several industries, and therefore the corporations can gain significant advantage that is capable to adapt well and gain the flexibility to analyze such information explosions over those alternative corporations. Finally, data processing techniques is utilized in the malware detection in clouds. The challenges of security in cloud computing environments is categorised into network level, user authentication level, data level, and generic problems

Network level: The challenges that may be categorised below a network level influence network protocols and network security, like distributed nodes, distributed information, Internode communication.

Data level: The challenges that may be classified below data level deals with information integrity and availableness like data protection and distributed knowledge.

Generic types: The challenges that may be classified below general level are ancient security tools, and use of various technologies.

OPEN PROBLEMS

a) **Data management** in a single perfect manner is remaining an open problem for both cloud and big data.
b) **Big data** handles unstructured data also. Hence, the data structures differ from conventional SQL
c) **Databases**.

The data structures for NoSQl databases are Graph, Documents and Key value stores. Though there are technologies to manage graphics and documents, Key value store needs a lot of research.

The functionalities of handling key value stores to be improved to support on demand queries and also extended key value stores are required to manage diverse data oriented rich internet applications.

Elasticity for effective usage of unstructured as well as structured resources with consistent semantics, operating with those resources at a minimum cost and enabling autonomy in multi-tenant data systems are some of the open problems.

6. PROTECTION AND SECURITY

6. 1. Vormetric Encryption: seamlessly protects massive knowledge environments at the classification system and volume level. This Big Data analytics security resolution permits organizations to gain the advantages of the intelligence gleaned from massive knowledge analytics whereas maintaining the protection of their data –with no changes to operation of the applying or to system operation or administration.

6. 2. Data Security Platform: The Vormetric information Security Platform secures crucial information – putting the safeguards and access controls for your information along with your information. The data security platform includes sturdy cryptography, key management, fine-grained access controls and therefore the security intelligence information required to spot the most recent in advanced persistent threats (APTs) and different security attacks on your knowledge.

6. 3. Encryption and Key Management: knowledge breach mitigation and compliance regimes need cryptography to safeguard data. Vormetric provides the strong, centrally managed, encryption and key management that allows compliance and is clear to processes, applications and users.

6. 4. Fine-grained Access Controls: Vormetric provides the fine-grained, policy based mostly access controls that prohibit access to knowledge that has been encrypted permitting solely approved access to knowledge by processes and users to meet strict compliance needs. Privileged users of all types will see plaintext data on condition that specifically enabled to do therefore. System update and administrative processes still work freely – however see only encrypted knowledge, not the plaintext supply.

6. 5. Security Intelligence: Vormetric logs capture all access attempts to protected knowledge providing high price, security intelligence information which will be used with a Security Information and Event Management resolution to spot compromised accounts and malicious insiders in addition as finding access patterns by processes and users which will represent and APT attack in method.

6. 6. Automation: Use the Vormetric Toolkit to simply deploy, integrate and manage your Vormetric knowledge Security implementation with the remainder of your massive knowledge implementation.

7. CONCLUSIONS

Apache Hadoop is an open supply technology that is simply offered within the web. This technology will be employed in developing numerous applications in numerous fields. This technology is principally employed in big data analysis. However this technology is found in bioinformatics applications as it will be shown within the previous sections. The Map Reduce paradigm of Hadoop is found to be used wide for developing varied algorithms of bioinformatics. variety of those algorithms like DRAW, CloudBLAST, Cloudburst etc., are already discussed within the previous sections In future, the appliance will be developed using

the Hadoop technology with additional economical algorithmic rule in analysing the genomic datasets of Bioinformatics. Analysis of those algorithms will be done by comparison such completely different bioinformatics-Map Reduce-based algorithms. the most recent algorithmic program developed for bioinformatics applications is DRAW (Data-gRouping-Aware) with the new idea of Interest locality. In IT trade, the most needed algorithmic program is that the one that is best, less time intense and most reliable. it might be terribly difficult to prove that the performance of Data-gRouping-Aware knowledge placement algorithmic program to be additional economical comparison with the already existing algorithms like cloudBLAST, cloudburst etc.

In abundant as big data analytics hold much guarantees for providing business insights, analyzing client behaviour- it is not without its distinctive challenges. In line with analysis, the most important obstacles to big data analytics are staffing and coaching , followed by privacy constraints. Majority of customers are concerned regarding however their personal place able information is employed. Privacy knowledge believe that massive Data analytics is an infringement on privacy of our daily lives.

Recently, researchers focus on how to manage, handle and additionally process the massive quantity of data as famous a huge information deals with 3 ideas volume, variety and speed that needs a new mechanisms to manage, process and securing the massive information. As managing and process of big knowledge have several issues and needed a lot of efforts to handle these needs once deal with huge information, security is one amongst the challenges that arise once systems try to handle the idea of huge information. a lot of researches required to beat the safety of huge knowledge rather than current security algorithms and methodology.

References

[1] "Survey on Various Applications of Hadoop in Bioinformatics" Minerva Laishram, Computer Science & Engineering, Visveswaraya Technological University *International Journal of Computer Applications* 129(6), (2015).

[2] "CloudBurst: highly sensitive read mapping with MapReduce", Michael C. Schatz∗ Center for Bioinformatics and Computational Biology, University of Maryland, College Park MD 20742, USA.

[3] "DRAW: A New Data-gRouping-AWare Data Placement Scheme for Data Intensive Applications With Interest Locality" Jun Wang, Qiangju Xiao, Jiangling Yin, and Pengju Shang. Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32826 USA, *IEEE TRANSACTIONS ON MAGNETICS,* 49(6) June 2013.

[4] "Big Data Computing and Clouds: Challenges, Solutions, and Future Directions" Assuncoa, M.D. et al., 2013. arXiv, 1(1), pp. 1-39.

[5] "Big Data-State of the Art"Chalmers, S., Bothorel, C. & CLEMENTE, R., 2013. Thesis. Brest: Telecom Bretagne, Institut Mines-Telecom.

[6] "Using Big Data Analytics in Information Technology (IT) Service Delivery Internet Technologies and Applications Research", Fung, H.P, 2013. 1(1), pp. 6-10.

[7] "MapReduce: Simplied Data Processing on Large Clusters. Communications of the ACM", Dean, J. & Ghemawat, S., 51(1) (2008) 107-13.

[8] "Twister: A runtime for iterative MapReduce", J. Ekanayake et al, In Proceedings of the 19th ACM HPDC, pages 810-818, 2010.

[9] "On the energy (in) efficiency of hadoopclusters", J. Leverich et al, ACM SIGOPS Operating Systems Review, 44(1): 61-65, 2010.

[10] "Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatic applications", A. Matsunaga et al. In Fourth IEEE International Conference on eScience, pages 222-229, 2008.

[11] "Processing Theta-Joins using MapReduce", A. Okcan et al. In Proceedings of the 2011 ACM SIGMOD, 2011.

[12] A. Pavlo et al, In Proceedings of the ACM SIGMOD, pages 165-178, 2009.

[13] "A Survey of Big Data Cloud Computing Security", Elmustafa Sayed Ali Ahmed and Rashid A.Saeed, *International Journal of Computer Science and Software Engineering* 3(1) (2014).

[14] Intel IT centre, "Peer Research Big Data Analytics ", Intel's IT Manager Survey on How Organizations Are Using Big Data, AUGUST 2012.

[15] "SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING", Venkata Narasimha Inukollu, Sailaja Arsi ,and Srinivasa Rao Ravuri, *International Journal of Network Security & Its Applications* 6(3) (2014).

[16] "Security issues associated with big data in cloud computing", R. Saranya, V.P. Muthu Kumar, *International Journal of Multidisciplinary Research and Development*, (4) (2015) 580-585.

[17] "Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments." Securosis blog, version1.0 (2012).

[18] Khalid A (2010). Cloud Computing: applying issues in Small Business. International Conference on Signal Acquisition and Processing (ICSAP'10), 278-281.

[19] Kilzer, Ann, Emmett Witchel, Indrajit Roy, Vitaly Shmatikov and Srinath T.V Setty, "Airavat: Security and Privacy for MapReduce".

[20] Yanglin Ren, Monitoring patients via a secure and mobile healthcare system, IEEE Symposium on wireless communication, 2011.

[21] Dai Yuefa, Wu Bo, Gu Yaqiang, Data Security Model for Cloud Computing, International Workshop on Information Security and Application, 2009.

(Received 25 March 2016; accepted 08 April 2016)