International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 2, Issue 6 (Nov-Dec 2014), PP. 128-133

LITERATURE SURVEY ON BIG DATA AND PRESERVING PRIVACY FOR THE BIG DATA IN CLOUD

Hari Kumar.R M.E (CSE), Dr.P.UmaMaheswari Ph.d,
Department of Computer Science and Engineering,
Info Institute of Engineering,
Coimbatore,India
Harikumar990info@gmail.com, dr.umasundar@gmail.com

***Abstract*- Big data is the term that characterized by its increasing volume, velocity, variety and veracity. All these characteristics make processing on this big data a complex task. So, for processing such data we need to do it differently like map reduce framework. When an organization exchanges data for mining useful information from this big data then privacy of the data becomes an important problem. In the past, several privacy preserving algorithms have been proposed. Of all those anonymizing the data has been the most efficient one. Anonymizing the dataset can be done on several operations like generalization, suppression, anatomy, specialization, permutation and perturbation. These algorithms are all suitable for dataset that does not have the characteristics of the big data. To preserve the privacy of the large dataset an algorithm was proposed recently. It applies the top down specialization approach for anonymizing the dataset and the scalability is increasing my applying the map reduce frame work. In this paper we survey the growth of big data, characteristics, map-reduce framework and all the privacy preserving mechanisms and propose future directions of our research.**

I. Big Data and its evolution

The amount of data produced in day to day life is increasing at a very rapid rate. At the start of the computer era the size of the data was measured in KB (Kilo Byte). Later it extended to MB (Mega Byte), GB (Giga Byte), TB (Tera Byte), PB (Peta Byte), EB(Exa Byte) an now recently the digital world is dealing data of sizes in ZB(Zeta Byte). An example for the size of 1 Tera Byte is that it is equivalent to 210 single sided DVD's. 1 Zeta Byte is equivalent to 1,000,000,000 Tera Bytes. A survey says that Google processes 20 Peta Bytes of data in 2008, and Face book has 2.5 Peta Bytes of data + 15 Tera Bytes of data in April 2009[4][5].

Data can be of two types – structured and unstructured data. Structured data means those data that are displayed with titled columns and rows which can be easily ordered and processed by data mining tools. Example for such data is the text files. These structured data constitute to only about 20% of the total data in the real world. The majority of the data is unstructured. Unstructured data is the one that does not have the proper structure and cannot be identified based on the internal structure. Example for such data are audio files, video files etc. These unstructured data constitute to about 80% of the total data in the digital world. The data that is growing rapidly is also an unstructured data which is very difficult to process. All these data need to processed and mined for gathering useful information from the available data [2] [3]. So far there is no single standard definition defined for the term big data, in [3] big data is defined as it is a collection of datasets that are so large and complex that it becomes extremely difficult to process using on-hand management tools or traditional data processing applications [14]. In [18] big data is defined as large and complex datasets made up of variety of structured and unstructured data which are too big, too fast or too hard to be managed by traditional techniques.

Characteristics of Big Data

The most common characteristics of big data [2][3][18] arises from 3v's by Gartner namely

1. Volume
2. Velocity
3. Variety

In addition many papers [4][5] propose new v's other than the above 3v's by Gartner to characterize big data. They are,

4. Veracity
5. Value

Volume refers to the size of the data that is very large and in the size of tera bytes and peta bytes. It can also be said as the quantity of data. Velocity refers to the pace or the speed in which the data arrives. The time is the key factor. There has been increase in the rate of arrival of data which is mainly because of the following reasons

a. Increasing automation process
b. Increasing connectivity between the systems
c. Increase in the social interaction between the people

Variety includes any type of data - structured and unstructured data such as text, audio, sensor data, click streams, video, log files and many more. Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Value is the one that measures the usefulness of data that helps in making decisions. These are the 5 V's that are proposed in papers that deal with big data.

Evolution of Big Data

The evolution of big data is listed below as it addressed in [2],

1944	Wesleyan University published a book called "The Scholar and the Future of the Research"
1956	FICO was founded which works on the principle that data used intelligently cab be used to improve business decisions.
1958	FICO builds first credit scoring system for American Investments.
1961	Derek Price publishes "Science since Babylon"

1967	B. A. Marron and P. A. D. de Maine publish ―Automatic data compression. The paper describes―a fully automatic and rapid three-part compressor which can be used with ‗any‘body of information to greatly reduce slow external storage requirements and to increase the rate of information transmission through a computer.
1971	Arthur Miller writes in "The Assault on Privacy"[1] that ―Too many information handlers seem to measure a man by the number of bits of storage capacity his dossier will occupy.
1980	The CPG / Retail industry transitioned from bi-monthly audit data to scanner data changed the dynamics of the industry.
1981	The Hungarian Central Statistics Office starts a research project to account for the country‘s information industries, including measuring information volume inbits
1996 (October)	The world‘s leading online travel company was started
1996	Digital storage becomes more cost-effective for storing data than paper according to R.J.T. Morris and B.J. Truskowski, in ―The Evolution of Storage Systems, IBM Systems Journal, July 1, 2003.
1997 (October)	The first article in the ACM digital library to use the term ―big data was published.
1997 (October)	Michael Lesk publishes ―"How much information is there in the world?"
1998	Most visited website in the world Google was founded. It has been estimated to run more than one million servers in data centers and to process over one billion search requests and about 24 petabytes of user-generated data each day.
1998 (April)	John R. Masey, Chief Scientist at SGI, presents at a USENIX meeting a paper titled ―Big Data… and the Next Wave of Infrastress.
2001	Paper on 3D Data management by Doug laney explaining about 3 v's.
2004	Facebook is a social networking service was launched. Google published a paper on MapReduce
2005	Apache Hadoop, an open-source software framework for storage and large scale processing of data-sets on clusters of commodity hardware, was created by Doug Cutting and Mike Cafarella
2006 (July)	Twitter called as "the SMS of the Internet" an online social networking and microblogging service was launched
2007	The first generation iPhone (smart phone from Apple inc) was released
2008	Facebook reaches 100M users.
2010	Special report on Data, data everywhere by "The Econonmist", EMC buys Greenplum,IBM buys Netezza
2011	Mckinsey report on big data, oracle buys endecea,Hp buys vertica.
2012	Big Data becomes buzz word after Gartner prediction, Facebook user hits 1B
2013	Fast Data era, YouTube hits 1B users

II. CLOUD COMPUTING ISSUES

Over the past 50 to 60 years of computer era, computing technology has undergone a series of platform and environment changes. As today, there are billions of people whose internet for their day to day activities the services provided are transformed from the centralized computing to the parallel and distributed computing. Centralize computing is the term that can be termed as where all the computing resources are all centralized in one physical system. All resources (processors, memory, and storage) are fully

Shared and tightly coupled within one integrated OS. Parallel computing term is defined as all processors are either tightly coupled with centralized shared memory or loosely coupled with distributed memory. Programs running in a parallel computer are called parallel programs. The process of writing parallel programs is often referred to as parallel programming. Distributed computing is the term that can be defined as the field of computer science/engineering that studies the distributed systems. A distributed system consists of multiple autonomous computers, each having its own private memory, communicating through a computer network. Cloud computing [6][7][8][9] is defined as an Internet cloud of resources that can be either a centralized or a distributed computing system. The cloud applies parallel or distributed computing, or both. Clouds can be built with physical or virtualized resources over large data centers that are centralized or distributed. Some authors consider cloud computing to be a form of utility computing or service computing.

Figure 1 depicts the cloud landscape and major cloud players [6][7] based on the three cloud service models.
There are three cloud service models. They are

1. IaaS
2. PaaS
3. SaaS

Infrastructure as a Service (IaaS) - This model puts together infrastructures demanded by users—namely servers, storage, networks, and the data center fabric. The user can deploy and run on multiple VMs running guest OSes on specific applications. The user does not manage or control the underlying cloud infrastructure, but can specify when to request and release the needed resources.

Platform as a Service (PaaS) - This model enables the user to deploy user-built applications onto a virtualized cloud platform. PaaS includes middleware, databases, development tools, and some runtime support such as Web 2.0 and Java. The platform includes both hardware and software integrated with specific programming interfaces. The provider supplies the API and software tools (e.g., Java, Python, Web 2.0, .NET). The user is freed from managing the cloud infrastructure.

Software as a Service (SaaS) - This refers to browser-initiated application software over thousands of paid cloud customers. The SaaS model applies to business processes, industry applications, consumer relationship management (CRM), enterprise resources planning (ERP), human resources (HR), and collaborative applications. On the customer side, there is no upfront investment in servers or software licensing. On the provider side, costs are rather low, compared with conventional hosting of user applications.

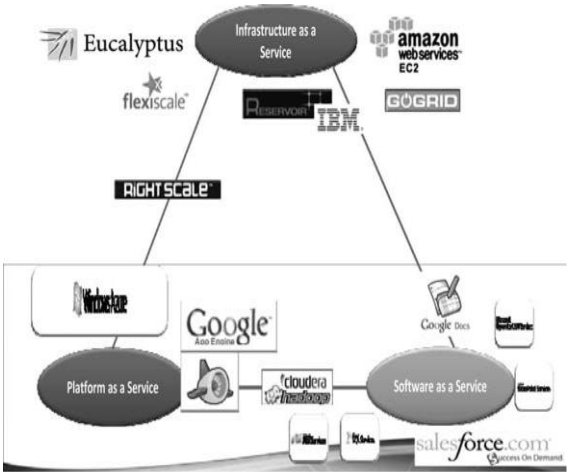

Fig 1: cloud landscape and major cloud players based on the three cloud service models.

Top 10 Obstacles for Cloud Computing

In [6], the obstacles for the cloud computing are listed. Of these obstacles, the first three affects the adoption, next five affects the growth and the last two affects the policy and business.

1. Availability/Business Continuity
2. Data Lock-In
3. Data Confidentiality and Auditability
4. Data Transfer Bottlenecks
5. Performance Unpredictability
6. Scalable Storage
7. Bugs in Large Distributed Systems
8. Scaling Quickly
9. Reputation Fate Sharing
10. Software Licensing

In – scalability and privacy [8][9] of the data are the two issues that are addressed and in our research too we are going to focus on these two issues in cloud computing.

III. PRIVACY PRESERVING TECHNIQUES

As we have said already the size and the variety of the data is growing rapidly, the tools and techniques that are used to handle such data should also be upgraded. Here we present a typical scenario of data collection and publishing in Figure 2[10]. In the data collection phase data publisher collects the data from the data owners and in the data publishing phase the data publisher publishes the data to the public or the data recipients where the data is mined for gathering useful information. Consider an example where the hospital X collects the details of all its patients who are the data owners, the hospital X acts as the data publisher. In order to get useful information from the patient details the hospital X releases the data to the data recipient. This exchange or publishing takes place in the cloud.

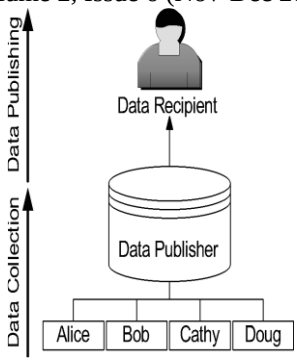

Fig 2 Data Collection and Data Publishing

There are two models in data publishers. First one is the untrusted model; the data publisher itself is not trusted and may use the sensitive information from the record owners. Several techniques using cryptography [Yang et al. 2005], anonymous communication[Chaum 1981; Jakobsson et al. 2002] and statistical methods[Warner 1965] were used to safe guard the privacy of the data owners from the untrusted data publishers. Second model is the trusted model; the data publisher is a trusted one. The record owners are willing to give their data to data publisher and the data publisher is trust worthy. This trust on the data publisher by the data owners is not visible to the data recipients. In this survey we deal with the trusted model, and we consider only the privacy issues in the data publishing phase.

In general, the data publisher has the table of the form,

A (Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes)

Here explicit identifier is the set of attributes that can uniquely identify the records and it has the explicit information about the record owners. Examples of such attributes are name and SSN (social security number). Quasi Identifier is the set of attributes that has the high probability of identifying the record owners. Examples for such attributes are Zip code, sex and date of birth. Sensitive attributes are the set of attributes that consists of sensitive person specific information such as disease, salary and disability status. Non sensitive attributes are the set of attributes that does not fall under the previous three categories and all these four attributes are disjoint in a table [19].

Anonymization [Cox 1980; Dalenius 1986] is a PPDP approach that it focuses to hide the identity and/or sensitive data of record owners from data recipients [19]. To achieve this initial way is to remove the explicit identifier from the publishing table. Even after removing the explicit identifier, there is the possibility of linking attacks. Linking attacks [10] are performed with the help of quasi identifiers to find the record owners. To prevent this linking attacks the data publisher provides the following table to the data recipient for data mining,

T (QID', Sensitive Attributes, Non-Sensitive Attributes)

QID' is an anonymous version of the original QID obtained by applying Anonymization operations to the attributes in QID in the original table A. Anonymization operations hides some detailed information so that several records become

indistinguishable with respect to QID. The Anonymization problem is to produce an anonymous T that satisfies a given privacy requirement determined by the chosen privacy model and to retain as much data utility as possible. Information metric is used to measure the utility of an anonymous table. Note that the Non-Sensitive Attributes are published if they are important to the data mining task.

Types of linking attacks

A. Record linkage

In the record linkage, some value on QID attributes identifies a small number of records in the released table which is called a group. Suppose if the victim's QID attributes matches then it comes under the group which is vulnerable to be attacked.

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

Table 1: Patient Table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

Table 2 External Table

Consider an example for the record linkage, here in Table 1 the hospital wants to release the patient table to the data recipients for the analysis purpose. Consider that the data recipient has the table 2 and knows that every person in table 1 has a record in table 2. Joining the two tables based on job, sex and age gives the record owners sensitive information. This way the privacy of the record owner is leaked. To solve this record linkage problem, Samarati and Sweeney [1998a, 1998b] proposed a technique called k-anonymity. K anonymity [12] [17] means that if one record in the published table has one QID value then at least k-1 record in the table also has the same QID value. A table that satisfies this requirement is called k-anonymous.

Job	Sex	Age	Disease
Professional	Male	[35-40]	Hepatitis
Professional	Male	[35-40]	Hepatitis
Professional	Male	[35-40]	HIV
Artist	Female	[30-35]	Flu
Artist	Female	[30-35]	HIV
Artist	Female	[30-35]	HIV
Artist	Female	[30-35]	HIV

Table 3: 3-Anonymous table on Table 1

Table 3 is obtained by anonymizing table 1 by generalization operation on QID= {Job, Sex and Age} using the taxonomy trees. This table has two indistinguishable groups that contain at least 3 records and hence it is 3-anonymous table. The two indistinguishable record groups are [professional, male, [35-40]] and [Artist, female, [30-35]]. Variants of K-anonymity are (X, Y)-anonymity and MultiRelational k-anonymity. (X, Y) –anonymity specifies that each value on X is linked to at least k-distinct values on Y. MultiRelational K-anonymity are a technique to ensure k-anonymity on multiple relational tables.

B. Attribute Linkage

In the attribute linkage, the attacker need not particularly identify the record of the victim, but can easily infer his/her sensitive values from the published data T, based on the set of sensitive values associated to the group that the victim belongs to. In case some sensitive values predominate in a group, a successful inference becomes relatively easy even if k-anonymity is satisfied. Machanavajjhala et al. [2006, 2007] proposed the diversity principle, called l-diversity, to prevent attribute linkage. The l-diversity requires every qid group to contain at least l "well-represented" sensitive values. One disadvantage of l-diversity is that it does not provide probability based risk measure.

C. Table Linkage

In the previous linkage attacks, the attacker knows that the victim's record is present in the published table. But in the table linkage attack, the attacker does not know the presence or absence of victim record. A table linkage occurs when an attacker can confidently infer the presence or the absence of the victim's record in the released table.

D. Probabilistic Attack

In this model, the attacker does not focus on record, attribute or tables for gaining the sensitive value of the victim, instead the attacker would change his/her probabilistic belief on the sensitive information after accessing the published table. Chawla et al. [2005] suggested that having access to the published anonymous data table should not enhance an attacker's power of isolating any record owner. Consequently, they proposed a privacy model to prevent (c, t)-isolation in a statistical database.

E. Anonymization Operations

Typically, the table in its original form does not satisfy the privacy requirements and so cannot be published as such. It has to be modified before being published. The modification is done by sequence of Anonymization operations to the table. Anonymization operations come in several flavors like generalization, anatomization, suppression, permutation and perturbation.

F. Generalization and Suppression

A Generalization operation is the one that replaces a particular value with the parent value in the taxonomy tree of

that attribute. The reverse operation is called specialization. In the generalized table, replacing a value with the child value of the attribute in the taxonomy tree is called specialization. A suppression operation is the one that replaces a particular value with a special value, indicating that the replaced values are not disclosed. Each generalization or suppression operation hides some details in QID. For a numerical attribute, the exact values can be replaced with the range that covers the exact values. For the categorical attribute, a specific value is replaced with the general value according to the taxonomy tree.

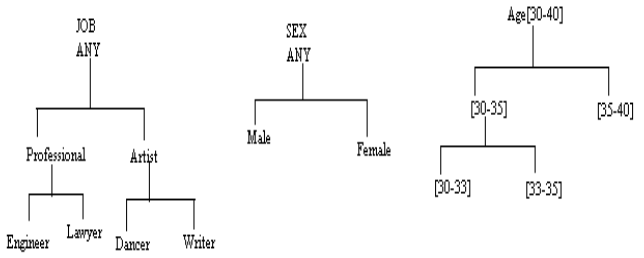

Fig 3 Taxonomy Tree for QID in Table 1

In figure 3, the node professional is more general than the engineer and lawyer nodes and is the child of Job Any node which is the parent node. In this diagram we have shown the taxonomy tree for the QID attributes in the table 1. For a categorical attribute like job, we generalize the particular value. For example, if the value of job in a record is dancer then it is generalized as artist. For a numerical attribute, we define a range of values so that the exact value comes under that range. For example if a record has the age as 36, then it can be generalized under the range [35-40]. There are several generalization schemes proposed by several authors. They are full domain generalization, sub tree generalization scheme, sibling generalization scheme, cell generalization scheme and multi dimensional generalization. There are also different suppression schemes like cell suppression. In summary, the choice of anonymization operations has an implication on the search space of anonymous tables and data distortion record suppression and value suppression.

G. Anatomization and permutation

Unlike generalization and suppression, anatomization does not alter the values in the QID attributes or sensitive attributes, but it de-associates the relationship between the two. To be precise, anatomization operation [11] releases two tables: a quasi identifier table (QIT) that contains the quasi identifier attributes and a sensitive table (ST) that contains the sensitive attributes. A group ID is the common attribute between ST and QIT to identify the records. By this way, the attribute values are not modified in the published table which is one of the advantages of this method. All records in the same group will have the same value on GroupID in both tables, and therefore are linked to the sensitive values in the group in the exact same way. For example, we take a patient table as shown below.

Age	Sex	Disease (Sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

Table 4 patient table to be published

This table has to be published to the data recipients for data analysis. Before publishing this data, the data publisher performs the anatomization approach for anonymizing the data. As a result we get two tables called quasi identifier table and sensitive table.

Age	Sex	Group Id
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

Table 5 QIT table for release

Group Id	Disease	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

Table 6 ST Table for release

H. Perturbation

The general idea of perturbation operation is to replace the original data values with some random synthetic values, so that the statistical information compute from the perturbed data does not differ much from the statistical information computed from the original data values. The perturbed data records do not correspond to real-world record owners, so the attacker cannot perform the sensitive linkages or recover sensitive information from the published data. One limitation of the perturbation approach is that the published records are "synthetic" in that they do not correspond to the real-world entities represented by the original data; therefore, individual records in the perturbed data are basically meaningless to the human recipients.

I. Information Metrics

Although privacy preservation is the primary goal of anonymization operation, it is also essential that the anonymized data remains practically useful for data analysis.

There are broad categories for measuring the usefulness of the published data. A data metric measures the quality of the anonymized table and compares it with the data quality of the original table. A search metric guides each step of an anonymization (search) algorithm to identify an anonymous table with maximum information or minimum distortion. Alternatively, information metric can be categorized by its information purposes, including general purpose, special purpose, or trade-off purpose.

In [1], privacy preserving data publishing is achieve through top own specialization approach for data anonymization. One important thing to be noted in this research is that the use of map reduce framework to successfully anonymized large datasets. As already said the size of data is growing at rapid rate and the data are shared by most users in cloud. Since the growth of the data is rapid, the existing algorithms do not anonymized the datasets efficiently so they deliberately design a group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way. The utility or the usefulness of the data is measured using the IGPL metric. In the centralized TDS approach [15] the scalability of the datasets is improved by using the TIPS data structure. This data structure stores a huge amount of Meta data to achieve high scalability. To overcome this difficulty the distributed TDS approach [16] was proposed which mainly concentrated only on successful anonymization operation and not the scalability. To overcome this difficulty in this paper they proposed a two phase top down specialization approach for data anonymization operation. This is done with the help of deliberate design of map reduce jobs to concretely accomplish the specialization operation in a scalable manner. A two phase TDS is proposed to gain high scalability via allowing specializations to be conducted on multiple data partitions in parallel.

J. Summary and future work

In this paper, we presented a survey of big data, cloud computing issues and privacy preservation. The first section describes about big data and its characteristics. It shows how the big data evolution happened. Big data characteristics show that we need different software and techniques for processing it. The second section describes about cloud computing and what are the issues that have to be addressed in the cloud computing field. The third section describes the importance of information sharing and the approach to achieve this is privacy preserving data publishing. Later the third section discusses what are the types of attacks that are possible after publishing the data. We reviewed the anonymizing operations and information metrics used to measure the usefulness of the published data. There are number of promising areas for research because privacy is a challenging area and it does not have solid definitions as it serves as a tradeoff between data utility and potential for misuse. We plan to do our research in adoption of bottom up generalization approach for data anonymization using map reduce.

REFERENCES

[1]. Xuyun Zhang, Laurence T. Yang, Chang Liu, and Jinjun Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud" ieee transactions on parallel and distributed systems, vol. 25, no. 2, february 2014

[2]. R.Devankuruchi "Analysis of Big Data Over the Years" International Journal of Scientific and Research Publications, Volume 4, Issue 1, January 2014 1 ISSN 2250-3153

[3]. Ms.Manisha saini, Ms.Pooja Taneja, Ms.Pinki Sethi "Big Data Analytics: Insights and Innovations" International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X

[4]. Steve Sonaka "Big Data and the Ag sector: More than lots of numbers" International food and agribusiness review, volume 17 issue 1, 2014

[5]. Alejandro Zarate Santovena "Big Data: Evolution, components, challenges and opportunities" pp 27-33

[6]. M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M.Zaharia, "A View of Cloud Computing," Comm. ACM, vol. 53,no. 4, pp. 50-58,2010

[7]. L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?," IEEE Trans.Parallel and Distributed Systems, vol. 23, no. 2, pp.296-303, Feb.2012.

[8]. H. Takabi, J.B.D. Joshi, and G. Ahn, "Security and Privacy Challenges in Cloud Computing Environments, IEEE Security and Privacy, vol. 8, no. 6, pp. 24-31, Nov. 2010.

[9]. D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues," Future Generation Computer Systems, vol. 28, no. 3, pp. 583-592, 2011.

[10]. B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Develop-ments," ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, 2010.

[11]. X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB'06), pp. 139-150, 2006.

[12]. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," Proc. 22nd Int'l Conf. Data Eng. (ICDE '06), 2006.

[13]. V. Borkar, M.J. Carey, and C. Li, "Inside 'Big Data Management': Ogres, Onions, or Parfaits?," Proc. 15th Int'l Conf. Extending Database Technology (EDBT '12), pp. 3-14, 2012.

[14]. J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," Comm. ACM, vol. 51, no. 1, pp. 107-113, 2008.

[15]. N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 4, Article 18, 2010.

[16]. B. Fung, K. Wang, L. Wang, and P.C.K. Hung, "Privacy-Preserving Data Publishing for Cluster Analysis," Data and Knowledge Eng., vol. 68, no. 6, pp. 552-575, 2009.

[17]. L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[18]. Katarina Grolinger, Michael Hayes, Wilson A. Higashino, Alexandra L'Heureux, David S. Allison "Challenges for MapReduce in Big Data".

[19]. Charu C. Aggarwal, Philip S. Yu "A General Survey of Privacy-Preserving Data Mining Models and Algorithms".