International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Index Copernicus Value (2015): 62.86 | Impact Factor (2015): 3.791

A Survey on Analytics for Preserving Privacy in Big Data

K. Prema

M. Tech Scholar, Sree Vidyanikethan Engineering College (Autonomous), Sree Sainath Nagar, Tirupati, India

Abstract: ***Background/Objectives: The age of big data is now coming. But the traditional data analytics may not be able to handle such large quantities of data. The question that arises now is how to develop a high performance platform to efficiently analyze big data and how to design an appropriate mining algorithm to find the useful things from big data. Findings: Limitations of big data analytics, applications of big data analytics in various fields are explained .To date; health care industry has not fully grasped the potential benefits to be gained from big data analytics. The open issues on computation, quality of end result, security, and privacy in platform/framework, data mining perspective are discussed to explain which open issues we may face. Methods: Methods for preserving privacy in big data analytics are also discussed.***

Keywords: Big data, data analytics, data mining , privacy

1. Introduction

Since past few years the huge amount of data has expanded continuously in various areas thus Big Data Analytics has become one of the most discussed topics among researchers for further research direction. The term Big Data is utilized to refer the colossal measure of datasets. Compared to the traditional datasets huge datasets includes a set of unstructured data or information which requires more significant real-time investigation.

The concept of big data allows us to know about the abstraction and in-depth understanding behind various hidden values. Now industries are more focused on big data Analysis and various government organizations are declared many plans for increasing the research towards big data analytics.

Nowadays the big data associated with various industries are growing expeditiously hence Google which is one of the leading software giants in the market processes hundreds of Petabyte amount of data, Facebook and a Chinese organization named Baidu generates 10 PB amounts of data per month respectively.

The fast development of the combination of cloud computing phase and the Internet of Things (IoT) endorses a sharp growth of data. Big data oriented cloud computing provides an ease of memory management as various applications and data resources are spread among the users of the worldwide in a distributed manner.

The model of IoT (Internet of Things) includes a huge amount of information gathering by different sensor hubs and transmission of the information over the cloud system for storing and further processing. An essential benefit of cloud computing infrastructure includes various resources such as availability, elasticity and cost reduction in the various amounts of data in a process where one has to pay before using a particular service. Cloud infrastructure gives a few advantages to the organizations where they do not have to pay the maintenance cost for the IT infrastructure. In a cloud environment, data analytics and analysis will be extremely strong as there are as there are many challenges that can be a rise.

1.1 Research Trends In Big Data

Major research trends in Big Data can be categorized as follows:

- Storage, Search and Retrieval of Big Data
- Analytics on Big Data
- Computations on Big Data

1.1.1 Storage, Search and Retrieval of Big Data

Storage is very complex and not only does it require managing capacity and finding out the best collection and retrieval methods, it also means to synchronize both the IT and the business teams and paying attention to complex security and privacy issues.

1.1.2 Analytics on Big Data

Big Data analytics comprises of tools, algorithms, and architecture that analyze and transform large and massive volumes of data. Big data analytics is a technology-enabled strategy for enabling an organization to have a competitive edge over others by analyzing market and customer trends. Analytics on real-time data, online transactional data gives deeper insights of the trends to make timely and accurate decisions.

1.1.3 Computations on Big Data

Computing is concerned with the processing, transforming, handling and storage of information. Systems such as Map Reduce, Hadoop have made writing and executing ad hoc big-data analysis and computation easy. As search engines have transformed information access, other forms of big-data computing can and will transform the activities like medical and scientific research, defense task etc.

1.2 Big Data Techniques

Big Data needs extraordinary techniques to efficiently process large volume of data within limited run times. Reasonably, Big Data techniques are driven by specified applications. For example, Wal-Mart applies machine

learning and statistical techniques to explore patterns from their large volume of transaction data. These patterns can produce higher competitiveness in pricing strategies and advertising campaigns. Taobao (A Chinese company like eBay) adopts large stream data mining techniques on users' browse data recorded on its website, and exploits a good deal of valuable information to support their decision-making.

Big Data techniques involve a number of disciplines, including statistics, data mining, machine learning, neural networks, social network analysis, signal processing, pattern recognition, optimization methods and visualization approaches. There are many specific techniques in these disciplines, and they overlap with each other hourly.

(1) Optimization Methods

Optimization methods have been applied to solve quantitative problems in a lot of fields, such as physics, biology, engineering, and economics. In several computational strategies for addressing global optimization problems are discussed, such as simulated annealing, adaptive simulated annealing, quantum annealing, as well as genetic algorithm which naturally lends itself to parallelism and therefore can be highly efficient. Stochastic optimization, including genetic programming, evolutionary programming, and particle swarm optimization are useful and specific optimization techniques inspired by the process of nature. However, they often have high complexity in memory and time consumption. Many research works have been done to scale up the large-scale optimization by cooperative co-evolutionary algorithms. Real-time optimization is also required in many Big Data application, such as WSNs and ITSs. Data reduction and parallelization are also alternative approaches in optimization problems.

(2) Statistics

Statistics is the science to collect, organizes, and interprets data. Statistical techniques are used to exploit correlation ships and causal relationships between different objectives. Numerical descriptions are also provided by statistics. However, standard statistical techniques are usually not well suited to manage Big Data, and many researchers have proposed extensions of classical techniques or completely new methods. Authors proposed efficient approximate algorithm for large-scale multivariate monotonic regression, which is an approach for estimating functions that are monotonic with respect to input variables. Another trend of data-driven statistical analysis focuses on scale and parallel implementation of statistical algorithms. A survey of parallel statistics can be found and several parallel statistics algorithm. Statistical computing and statistical learning are the two hot research sub-fields.

(3) Data mining

Data mining is a set of techniques to extract valuable information (patterns) from data, including clustering analysis, classification, regression and association rule learning. It involves the methods from machine learning and statistics. Big Data mining is more challenging compared with traditional data mining algorithms. Taking clustering as an example, a natural way of clustering Big Data is to extend existing methods (such as hierarchical clustering, K-Mean, and Fuzzy CMean) so that they can cope with the huge workloads.

Most extensions usually rely on analyzing a certain amount of samples of Big Data, and vary in how the sample-based results are used to derive a partition for the overall data. This kind of clustering algorithms include CLARA (Clustering Large Applications) algorithm, CLARANS (Clustering Large Applications based upon Randomized Search), BIRCH (Balanced Iterative Reducing using Cluster Hierarchies) algorithm, and so on.

Genetic algorithms are also applied to clustering as optimization criterion to reflect the goodness. Clustering Big Data is also developing to distributed and parallel implementation. Taking discriminant analysis as another example, researchers try to develop effective algorithm for large-scale discriminant analysis.

The emphasis is on the reduction of computational complexity. Taking bioinformatics as another example, it becomes increasingly data-driven that leads to paradigm change from traditional single-gene biology to the approaches that combine integrative database analysis and data mining. This new paradigm enables the synthesis of large-scale portraits of genome function.

(4) Machine learning

Machine learning is an important subjection of artificial intelligence which is aimed to design algorithms that allow computers to evolve behaviors based on empirical data. The most obvious characteristic of machine learning is to discovery knowledge and make intelligent decisions automatically. When Big Data is concerned, we need to scale up machine learning algorithms, both supervised learning and unsupervised learning, to cope with it. Deep machine learning has become a new research frontier in artificial intelligence.

In addition, there are several frameworks, like Map/Reduce, Dryad LINQ, and IBM parallel machine learning toolbox, which have capabilities to scale up machine learning. For example, Support Vector Machine (SVM), which is a very fundamental algorithm used in classification and regression problems, suffers from serious scalability problem in both memory use and computation time. Parallel SVM (PSVM) is introduced recently to reduce memory and time consumption.

There are many scale machine learning algorithms ,but many important specific sub-fields in large-scale machine learning, such as large-scale recommender systems, natural language processing, association rule learning, ensemble learning, still face the scalability problems. Artificial neural network (ANN) is a mature techniques and has a wide range of application coverage.

Its successful applications can be found in pattern recognition, image analysis, adaptive control, and other areas. Most of the currently employed ANNs for artificial intelligence are based on statistical estimations, classification optimization and control theory. It is generally

acknowledged, the more hidden layers and nodes in a neural network, the higher accuracy they can produce.

However, the complexity in a neural network also increases the learning time. Therefore, the learning process in a neural networks over Big Data is severely time and memory consuming. Neural processing of large-scale data sets often leads to very large networks.

Then, there are two main challenges in this situation. One is that the conventional training algorithms perform very poorly, and the other is that the training time and memory limitations are increasingly intractable. Naturally, two common approaches can be employed in this situation.

One is to reduce the data size by some sampling methods, and the structure of the neural network maybe remains the same. The other one is to scale up neural networks in parallel and distributed ways. For example, the combination of deep learning and parallel training implementation techniques provides potential ways to process Big Data.

(5) Visualization Approaches
Visualization Approaches are the techniques used to create tables, images, diagrams and other intuitive display ways to understand data. Big Data visualization is not that easy like traditional relative small data sets because of the complexity in 3Vs or 4Vs. The extension of traditional visualization approaches are already emerged but far away from enough.

When it comes to large-scale data visualization, many researchers use feature extraction and a geometric modeling to significantly reduce the data's size before the actual data rendering. For more closely and intuitively data interpretation, some researchers try to run batch-mode software rendering of the data at the highest possible resolution in a parallel way. Choosing proper data representation is also very important when we try to visualize Big Data

(6) Social Network Analysis
Social Network Analysis (SNA) which has emerged as a key technique in modern sociology, views social relationships in terms of network theory, and it consists of nodes and ties. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, history, information science, organizational studies, social psychology, development studies, and sociolinguistics and is now commonly available as a consumer tool. SNA include social system design, human behavior modeling, social network visualization], social networks evolution analysis, and graph query and mining.

Recently, online social networks and Social media analysis have become popular. One of the main obstacles regarding SNA is the vastness of Big Data. Analysis of a network consisting of millions or billions of connected objects is usually computationally costly. Two hot research frontiers, social computing and cloud computing, are in favor of SNA to some degree.

Higher level Big Data technologies include distributed file systems distributed computational systems, massively parallel-processing (MPP) systems, data mining based on grid computing, cloud-based storage and computing resources], as well as granular computing and biological computing.

These technologies will be introduced in the following sub-sections. Many researchers regard the curse of dimensionality as one aspect of Big Data problems. Indeed, Big Data should not be constricted in data volume, but all take the high-dimension characteristic of data into consideration. In fact, processing high-dimensional data is already a tough task in current scientific research.

The state-of-the-art techniques for handling high-dimensional data intuitively fall into dimension reduction. Namely, we try to map the high-dimensional data space into lower dimensional space with less loss of information as possible. There are a large number of methods to reduce dimension. Linear mapping methods, such as principal component analysis (PCA) and factor analysis are popular linear dimension reduction techniques.

Non-linear techniques include kernel PCA, manifold learning techniques such as Iso map, locally linear embedding (LLE), Hessian LLE, Laplacian eigen maps, and LTSA. Recently, a generative deep network, called auto encoder, performs very well as non-linear dimensionality reduction. Random projection in dimensionality reduction also has been well-developed.

1.3 Use of Big Data Analytics in Various Fields

1.3.1 Need for Big Data Analytics in Healthcare
To improve the quality of healthcare by considering the following:

(1)Providing patient centric services
To provide faster relief to the patients by providing evidence based medicine--detecting diseases at the earlier stages based on the clinical data available, minimizing drug doses to avoid side effect and providing efficient medicine based on genetic make ups. This helps in reducing readmission rates there by reducing cost for the patients.

(2)Detecting spreading diseases earlier
Predicting the viral diseases earlier before spreading based on the live analysis. This can be identified by analyzing the social logs of the patients suffering from a disease in a particular geo-location. This helps the healthcare professionals to advise the victims by taking necessary preventive measures.

(3)Monitoring the hospital's quality
Monitors whether the hospitals are setup according to the norms setup by Indian medical council. This periodical check-up helps government in taking necessary measures against disqualifying hospitals.

(4)Improving the treatment methods
Customized patient treatment---monitoring the effect of medication continuously and based on the analysis dosages of medications can be changed for faster relief. Monitoring patient vital signs to provide proactive care to patients.

Making an analysis on the data generated by the patients who already suffered from the same symptoms, helps doctor to provide effective medicines to new patients.

1.3.2 Need for Big Data in Government

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

(1)Addressing Basic Needs Quickly

Today people need to wait for a long time to get EB, telephone, water, ration card and gas connection. These are the basic needs of citizen. It is the responsibility of the government to provide these services as quick as possible. Big data analytics plays a major role in achieving it because the data will be analyzed on daily basis. People who are in need will be served immediately.

(2)Providing quality education

Education is one of the valuable assets that can be given to the children. It is the duty of government to provide quality education to children(9). BDA provides detailed report of children who are in the age to be admitted to the school. This helps government to assess the educational needs for these children immediately.

(3) To reduce unemployment rate

To minimize unemployment rate by predicting the job needs before based the literacy rate. This can be achieved by analysis the students graduating each year. It enables government to arrange for special trainings in order to build young entrepreneurs.

Other Benefits

- To provide pension to senior citizens without any delay.
- To ensure that benefits provided by government reaches all the people.
- To control traffic in peak times based on the live streaming data about vehicles.
- To monitor the need for mobile ambulance facilities.

1.4 Limitations

There are 5 limitations to the use of big data analytics. They are:-

1. Prioritizing correlations

Data analysts use big data to tease out correlation: when one variable is linked to another. However, not all these correlations are substantial or meaningful. More specifically, just because 2 variables are correlated or linked doesn't mean that a causative relationship exists between them (i.e., "correlation does not imply causation").

For instance, between 2000 and 2009, the number of divorces in the U.S. state of Maine and the per capita consumption of margarine both similarly decreased. However, margarine and divorce have little to do with each other. A good consultant will help you figure out which correlations mean something to your business and which correlations mean little to your business.

2. The Wrong Questions

Big data can be used to discern correlations and insights using an endless array of questions. However, it's up to the user to figure out which questions are meaningful. If you end up getting a right answer to the wrong question, you do yourself, your clients, and your business, a costly disservice.

3. Security

As with many technological endeavors, big data analytics is prone to data breach. The information that you provide a third party could get leaked to customers or competitors.

4. Transferability

Because much of the data you need analyzed lies behind a firewall or on a private cloud, it takes technical know-how to efficiently get this data to an analytics team. Furthermore, it may be difficult to consistently transfer data to specialists for repeat analysis.

5. Inconsistency in data collection

Sometimes the tools we use to gather big data sets are imprecise. For example, Google is famous for its tweaks and updates that change the search experience in countless ways; the results of a search on one day will likely be different from those on another day. If you were using Google search to generate data sets, and these data sets changed often, then the correlations you derive would change, too.

Exposing whole of the big data may yield good analytics results but at the same time can pose great security challenges. Traditional security mechanisms fail to handle big data due its large volume, variety and velocity. Among various security aspects of big data, privacy is one of the most important issues. Traditional methods like cryptography can be used but they don't prove to be efficient because of complex nature of data.

Data Anonymization or de-identification is also helpful in hiding personal information. It is the process of changing data that will be used or published in a way that prevents the identification of key information.

There are basically three data anonymization methods that are used in preserving big data privacy. They are: K-Anonymity, L-Diversity, and T-Closeness. Differential privacy is another big data privacy preservation method that is being widely used.

Data Anonymization

Data anonymization is the process of changing data that will be used or published in a way that prevents the identification of key information. It is also sometimes referred as data de-identification. In this method key pieces of confidential data are obscured in a way that maintains data privacy.

For example, Table 1 represents the data set that needs to be analyzed for obtaining income trends without disclosing individual identity.

Table 2 represents data made anonymous by removing identifier attribute Voter ID. This table may look anonymous but can be linked with external data of to re-identify individuals.

K-Anonymity

A dataset is called k-anonymized if for any tuples with given

attributes in the dataset there are at least k-1 other records that match those attributes. K-anonymity can be achieved by using suppression and generalization. In suppression, quasi identifiers are replaced or obscured by some constant values like 0,* etc. In generalization, quasi identifiers are replaced by more general values from levels up the hierarchy.

Table 3 shows 2-anonymized version of table 2 using suppression. Here, age attribute has been suppressed and k=2. K-anonymous data can still be vulnerable to attacks like unsorted matching attack, temporal attack, and complementary release attack [4]. Therefore we move towards ***l***-diversity method of data anonymization.

ℓ-Diversity

ℓ-diversity technique of data anonymization tries to bring diversity in the sensitive attribute of data. It ensures that each equivalence class of quasi identifiers has at least L different values of sensitive attribute.

In Table 1 income is a sensitive attribute. For data to be L-diverse there should be at least L different values of income associated with each equivalence class. Table 4 shows 3-diverse version of table 1 since each equivalence class has at least 3 different values for sensitive attribute income.

The problem with this method is that it depends upon the range of sensitive attribute. If we want to make data L diverse whereas sensitive attribute has less than L different values, fictitious data is to be inserted. This fictitious data will enhance the security but may result in problems during analysis. Also L-diversity method is prone to skewness and similarity attack and thus can't prevent attribute disclosure.

t – closeness

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

The main advantage of t-closeness is that it prevents attribute disclosure. Data anonymization can be applied to big data but the problem lies in the fact that as size and variety of data increases, the chances of re-identification also increase. Thus, anonymization has a limited potential in the field of big data privacy.

Differential Privacy

Differential Privacy is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protections. It aims to minimize the chances of individual identification while querying the data. The method of differential privacy is shown in fig 1

As opposed to anonymization, data is not modified in differential privacy. Users don't have direct access to the database. There is an interface that calculates the results and adds desired inaccuracies. It acts as a firewall. These inaccuracies are large enough that they protect privacy, but small enough that the answers provided to analysts and researchers are still useful.

The advantages of differential privacy over anonymization are:

- The original data set is not modified at all. There is no need for suppression or generalization.
- Distortion is added to the results by mathematical calculations based on the type of data, type of questions etc.
- The distortion is added in such a way that value hidden is useful to analysts.

2. Open Issues and Challenges

2.1 Open issues

Although the data analytics today may be inefficient for big data caused by the environment, devices, systems, and even problems that are quite different from traditional mining problems, because several characteristics of big data also exist in the traditional data analytics. Several open issues caused by the big data will be addressed as the platform/ framework and data mining perspectives in this section to explain what dilemmas we may confront because of big data. Here are some of the open issues:

2.1.1 Platform and framework perspective

(1) Input and output ratio of platform

A large number of reports and researches mentioned that we will enter the big data age in the near future. Some of them insinuated to us that these fruitful results of big data will lead us to a whole new world where "everything" is possible; therefore, the big data analytics will be an omniscient and omnipotent system. From the pragmatic perspective, the big data analytics is indeed useful and has many possibilities which can help us more accurately understand the so-called "things."

However, the situation in most studies of big data analytics is that they argued that the results of big data are valuable, but the business models of most big data analytics are not clear. The fact is that assuming we have infinite computing resources for big data analytics is a thoroughly impracticable plan, the input and output ratio (e.g., return on investment) will need to be taken into account before an organization constructs the big data analytics center.

(2) Communication between systems

Since most big data analytics systems will be designed for parallel computing, and they typically will work on other systems (e.g., cloud platform) or work with other systems (e.g., search engine or knowledge base), the communication between the big data analytics and other systems will strongly impact the performance of the whole process of KDD. The first research issue for the communication is that the communication cost will incur between systems of data analytics. How to reduce the communication cost will be the very first thing that the data scientists need to care.

Another research issue for the communication is how the big data analytics communicates with other systems. The consistency of data between different systems, modules, and operators is also an important open issue on the communication between systems. Because the communication will appear more frequently between

systems of big data analytics, how to reduce the cost of communication and how to make the communication between these systems as reliable as possible will be the two important open issues for big data analytics.

(3) Bottlenecks on data analytics system

The bottlenecks will be appeared in different places of the data analytics for big data because the environments, systems, and input data have changed which are different from the traditional data analytics. The data deluge of big data will fill up the "input" system of data analytics, and it will also increase the computation load of the data "analysis" system. This situation is just like the torrent of water (i.e., data deluge) rushed down the mountain (i.e., data analytics), how to split it and how to avoid it flowing into a narrow place (e.g., the operator is not able to handle the input data) will be the most important things to avoid the bottlenecks in data analytics system. One of the current solutions to the avoidance of bottlenecks on a data analytics system is to add more computation resources while the other is to split the analysis works to different computation nodes. A complete consideration for the whole data analytics to avoid the bottlenecks of that kind of analytics system is still needed for big data.

(4) Security issues

Since much more environment data and human behavior will be gathered to the big data analytics, how to protect them will also be an open issue because without a security way to handle the collected data, the big data analytics cannot be a reliable system. In spite of the security that we have to tighten for big data analytics before it can gather more data from everywhere, the fact is that until now, there are still not many studies focusing on the security issues of the big data analytics.

According to our observation, the security issues of big data analytics can be divided into fourfold: input, data analysis, output, and communication with other systems. For the input, it can be regarded as the data gathering which is relevant to the sensor, the handheld devices, and even the devices of internet of things. One of the important security issues on the input part of big data analytics is to make sure that the sensors will not be compromised by the attacks. For the analysis and input, it can be regarded as the security problem of such a system.

For communication with other system, the security problem is on the communications between big data analytics and other external systems. Because of these latent problems, security has become one of the open issues of big data analytics.

2.1.2 Data mining perspective

(1) Data mining algorithm for map reduce solution

As we mentioned in the previous sections, most of the traditional data mining algorithms are not designed for parallel computing; therefore, they are not particularly useful for the big data mining. Several recent studies have attempted to modify the traditional data mining algorithms to make them applicable to Hadoop-based platforms.

As long as porting the data mining algorithms to Hadoop is inevitable, making the data mining algorithms work on a map-reduce architecture is the first very thing to do to apply traditional data mining methods to big data analytics. Unfortunately, not many studies attempted to make the data mining and soft computing algorithms work on Hadoop because several different backgrounds are needed to develop and design such algorithms.

For instance, the researcher and his or her research group need to have the background in data mining and Hadoop so as to develop and design such algorithms. Another open issue is that most data mining algorithms are designed for centralized computing; that is, they can only work on all the data at the same time. Thus, how to make them work on a parallel computing system is also a difficult work.

The good news is that some studies have successfully applied the traditional data mining algorithms to the map-reduce architecture. These results imply that it is possible to do so. According to our observation, although the traditional mining or soft computing algorithms can be used to help us analyze the data in big data analytics, unfortunately, until now, not many studies are focused on it. As a consequence, it is an important open issue in big data analytics.

(2) Noise, outliers, incomplete and inconsistent data

Although big data analytics is a new age for data analysis, because several solutions adopt classical ways to analyze the data on big data analytics, the open issues of traditional data mining algorithms also exist in these new systems. The open issues of noise, outliers, incomplete, and inconsistent data in traditional data mining algorithms will also appear in big data mining algorithms.

More incomplete and inconsistent data will easily appear because the data are captured by or generated from different sensors and systems. The impact of noise, outliers, incomplete and inconsistent data will be enlarged for big data analytics. Therefore, how to mitigate the impact will be the open issues for big data analytics.

(3) Bottlenecks on data mining algorithm

Most of the data mining algorithms in big data analytics will be designed for parallel computing. However, once data mining algorithms are designed or modified for parallel computing, it is the information exchange between different data mining procedures that may incur bottlenecks. One of them is the synchronization issue because different mining procedures will finish their jobs at different times even though they use the same mining algorithm to work on the same amount of data.

Thus, some of the mining procedures will have to wait until the others finished their jobs. This situation may occur because the loading of different computer nodes may be different during the data mining process, or it may occur because the convergence speeds are different for the same data mining algorithm. The bottlenecks of data mining algorithms will become an open issue for the big data analytics which explains that we need to take into account this issue when we develop and design a new data mining algorithm for big data analytics.

(4) Privacy issues

The privacy concern typically will make most people uncomfortable, especially if systems cannot guarantee that their personal information will not be accessed by the other people and organizations. Different from the concern of the security, the privacy issue is about if it is possible for the system to restore or infer personal information from the results of big data analytics, even though the input data are anonymous. The privacy issue has become a very important issue because the data mining and other analysis technologies will be widely used in big data analytics, the private information may be exposed to the other people after the analysis process.

For example, although all the gathered data for shop behavior are anonymous (e.g., buying a pistol), because the data can be easily collected by different devices and systems (e.g., location of the shop and age of the buyer), a data mining algorithm can easily infer who bought this pistol. More precisely, the data analytics is able to reduce the scope of the database because location of the shop and age of the buyer provide the information to help the system find out possible persons. For this reason, any sensitive information needs to be carefully protected and used. The anonymous, temporary identification, and encryption are the representative technologies for privacy of data analytics, but the critical factor is how to use, what to use, and why to use the collected data on big data analytic.

2.2 Challenges

Different challenges arise in each sub-process when it comes to data-driven applications. In the following subsections, we will give a brief discussion about challenges we are facing for each sub-process.

- Data capture and storage
- Data transmission
- Data curation
- Data analysis
- Data visualization

2.3 Tables

Table 1: Base Dataset

Age	Sex	City	Income
24	M	Delhi	1,00,00
24	M	Gurgaon	18,000
24	M	Gurgaon	25,500
24	M	Delhi	12,000
26	F	Delhi	20,000
26	F	Delhi	50,000
26	M	Delhi	29,000
26	F	Delhi	48,000
32	M	Delhi	26,000
32	F	Gurgaon	45,000
32	F	Gurgaon	34,000
32	M	Delhi	34,000

Table 2: Anonymous Dataset

Age	Sex	City	Income
24	M	Delhi	1,00,000
24	M	Gurgaon	18,000
24	M	Gurgaon	25,500
24	M	Delhi	12,000
26	F	Delhi	20,000
26	F	Delhi	50,000
26	M	Delhi	29,000
26	F	Delhi	48,000
32	M	Delhi	26,000
32	F	Gurgaon	45,000
32	F	Gurgaon	34,000
32	M	Delhi	34,000

Table 3: 2-Anonymized Dataset (Using Suppression)

Age	Sex	City	Income
24	M	Delhi	1,00,000
24	M	Gurgaon	18,000
24	M	Gurgaon	25,500
24	M	Delhi	12,000
26	F	Delhi	20,000
26	F	Delhi	50,000
26	M	Delhi	29,000
26	F	Delhi	48,000
32	M	Delhi	26,000
32	F	Gurgaon	45,000
32	F	Gurgaon	34,000
32	M	Delhi	34,000

Table 4: 2-Anonymized Dataset (Using Generalization), 3-Diverse Dataset

Age	Sex	City	Income
24	Person	ncr	1,00,000
24	Person	ncr	18,000
24	Person	ncr	25,500
24	Person	ncr	12,000
26	Person	ncr	20,000
26	Person	ncr	50,000
26	Person	ncr	29,000
26	Person	ncr	48,000
32	Person	ncr	26,000
32	Person	ncr	45,000
32	Person	ncr	34,000
32	Person	ncr	34,000

2.4 Figures

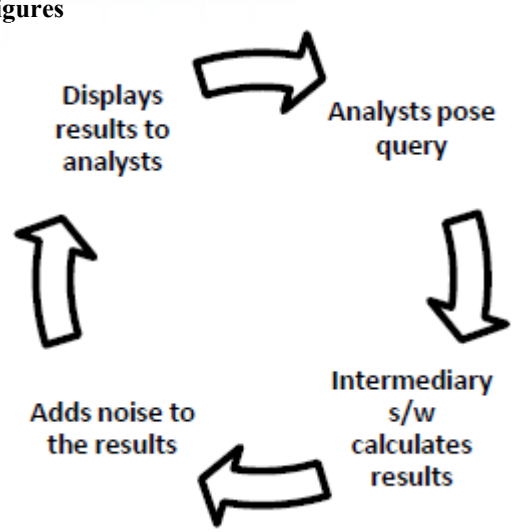

Figure 1: Differential privacy process

3. Conclusion

The open issues on computation, quality of end result, security, and privacy are then discussed to explain which open issues we may face. Last but not least we can find *solutions* to welcome the new age of big data, the possible high impact research trends are given below:

For the computation time, there is no doubt at all that parallel computing is one of the important future trends to make the data analytics work for big data, and consequently the technologies of cloud computing, Hadoop, and map-reduce will play the important roles for the big data analytics. To handle the computation resources of the cloud based platform and to finish the task of data analysis as fast as possible, the scheduling method is another future trend.

- Using efficient methods to reduce the computation time of input, comparison, sampling, and a variety of reduction methods will play an important role in big data analytics. Because these methods typically do not consider parallel computing environment, how to make them work on parallel computing environment will be a future research trend. Similar to the input, the data mining algorithms also face the same situation that we mentioned in the previous section, how to make them work on parallel computing environment will be a very important research trend because there are abundant research results on traditional data mining algorithms.
- How to model the mining problem to find *something* from big data and how to display the knowledge we got from big data analytics will also be another two vital future trends because the results of these two researches will decide if the data analytics can practically work for real world approaches, not just a theoretical stuff.
- The methods of extracting information from external and relative knowledge resources to further reinforce the big data analytics, until now, are not very popular in big data
  analytics. But combining information from different resources to add the value of output knowledge is a common solution in the area of information retrieval, such as clustering search engine or document summarization. For this reason, information fusion will also be a future trend for improving the end results of big data analytics.
- Because the metaheuristic algorithms are capable of finding an approximate solution within a reasonable time, they have been widely used in solving the data mining problem in recent years. Until now, many state-of-the-art metaheuristic algorithms still have not been applied to big data analytics. In addition, compared to some early data mining algorithms, the performance of metaheuristic is no doubt superior in terms of the computation time and the quality of end result. From these observations, the application of metaheuristic algorithms to big data analytics will also be an important research topic.
- Because social network is part of the daily life of most people and because its data is also a kind of big data, how to analyze the data of a social network has become a promising research issue. Obviously, it can be used to predict the behavior of a user. After that, we can make applicable strategies for the user. For instance, a business intelligence system can use the analysis results to encourage particular customers to buy the goods they are interested.
- The security and privacy issues that accompany the work of data analysis are intuitive research topics which contain how to safely store the data, how to make sure the data communication is protected, and how to prevent someone from finding out the information about us. Many problems of data security and privacy are essentially the same as those of the traditional data analysis even if we are entering the big data age. Thus, show to protect the data will also appear in the research of big data analytics.

References

[1] HILTON, MICHAEL. "Differential Privacy: A Historical Survey." 2002
[2] Sweeney, Latanya. "K-anonymity: A model for protecting privacy."International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002).
[3] Dwork, C.: Differential Privacy. In: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP) (2), pp. 1–12 (2006)
[4] Wong, Raymond Chi-Wing, et al. "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.
[5] Li,Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." 2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007.
[6] Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity."ACM Transactions on Knowledge Discovery from Data (TKDD) 1.1 (2007).
[7] Chiu, Chuang-Cheng, and Chieh-Yuan Tsai. "A k-anonymity clustering method for effective data privacy preservation." International Conference on Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2007.
[8] Aggarwal, Charu C., and S. Yu Philip. "A general survey of privacy-preserving data mining models and algorithms." *Privacy-preserving data mining*. Springer US, 2008.
[9] Gedik, Bugra, and Ling Liu. "Protecting location privacy with personalized k-anonymity: Architecture and algorithms." IEEE Transactions on Mobile Computing 7.1 (2008).
[10] Collins, Chris A., et al. "Early assembly of the most massive galaxies." Nature 458.7238 (2009).
[11] Zhou, Shuheng, Katrina Ligett, and Larry Wasserman. "Differential privacy with compression." 2009 IEEE International Symposium on Information Theory. IEEE, 2009.
[12] Roth, Aaron. "New algorithms for preserving differential privacy". Diss. Microsoft Research, 2010.
[13] Friedman, Arik, and Assaf Schuster. "Data mining with differential privacy." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[14] Zhou, Bin, and Jian Pei. "The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks."Knowledge and Information Systems 28.1 (2011).
[15] Doka, Katerina, Dimitrios Tsoumakos, and Nectarios Koziris. "KANIS: Preserving k-Anonymity Over Distributed Data." proceedings of the 5th International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases. 2011.
[16] Bertino, Elisa, et al. "Challenges and Opportunities with Big Data." (2011).
[17] Zhang, Ning, Ming Li, and Wenjing Lou. "Distributed data mining with differential privacy."2011 IEEE International Conference on Communications (ICC). IEEE, 2011.
[18] [19] FeiFei, Zhao, et al. "Study on Privacy Protection Algorithm Based on K-Anonymity." Physics Procedia 33 (2012).
[19] [20] Soria-Comas, Jordi, and Josep Domingo-Ferrert. "Differential privacy via t-closeness in data publishing." Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on. IEEE, 2013.
[20] Katal, Avita, Mohammad Wazid, and R. H. Goudar. "Big data: issues, challenges, tools and good practices." Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013.
[21] Singh, Neelam, Neha Garg, and Varsha Mittal. "Big Data–insights, motivation and challenges." International Journal of Scientific & Engineering Research, Volume 4, Issue 12, December-2013.
[22] [Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." Mobile Networks and Applications 19.2 (2014).
[23] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information Sciences 275 (2014).
[24] harma, Manish, et al. "Privacy Preserving Data Publishing Based on k-Anonymity by Categorization of Sensitive Values.",2014.
[25] Wang, Yue, Xintao Wu, and Donghui Hu.” Using Randomized Response for Differential Privacy Preserving Data Collection”. Technical Report, DPL-2014-003, University of Arkansas, 2014.
[26] Fouad, Mohamed R., Khaled Elbassioni, and Elisa Bertino. "A Super Modularity Based Differential Privacy Preserving Algorithm for Data Anonymization IEEE Transactions on Knowledge and Data Engineering Vol. 26." (2014).
[27] Gosain, Anjana, and Nikita Chugh. "Privacy Preservation in Big Data." *International Journal of Computer Applications* 100.17 (2014).
[28] Elabd, Emad, Hatem Abdulkader, and Ahmed Mubark. "L–Diversity-Based Semantic Anonymaztion for Data Publishing." (2015).
[29] Rahimi, Masoud, Mehdi Bateni, and Hosein Mohammadinejad. "Extended K-Anonymity Model for Privacy Preserving on Micro Data." International Journal of Computer Network and Information Security 7.12 (2015).
[30] Soria-Comas, Jordi, and Josep Domingo-Ferrer. "Big Data Privacy: Challenges to Privacy Principles and Models." (2015).
[31] Lin, Chi, et al. "Differential privacy preserving in big data analytics for connected health." Journal of medical systems 40.4 (2016)