A SURVEY ON VERIFICATION OF OUTSOURCED DATA MINING COMPUTATION

#1 G. HYMAVATHI, M.Tech Student,
#2 KARPURAPU SUDHAKARABABU, Assistant Professor,
Dept of CSE,
SRI CHUNDI RANGANAYAKULU ENGINEERING COLLEGE, GANAPAVARAM, GUNTUR,A.P, INDIA.

ABSTRACT: Cloud computing is a computing where data is processed using computing resources (e.g. Server). A Client of weak computational power outsourced their data to cloud (e.g. Server) for computation. This introduces Data mining as Service paradigm. Outsourcing data to the server facing a critical problem of verification. Whether the server returned correct and complete computation result to the client. Other issue is related to the security or privacy of outsourced data for computation. Then another issue is related to revenue generation by cloud server. Most of the previous research paper focus on different algorithms associated with frequent itemset mining and data security. Work on result integrity verification is rare as compared to the data security and efficient mining algorithms. This paper focuses on the problem of result verification of outsourced frequent itemset mining computation by analyzing numerous work but different authors in coming Section. The actual study reveals the efficiency and effectiveness of the verification approaches.

Keywords: Data mining as a service, Cloud computing, Frequent itemset mining, Result verification.

I. INTRODUCTION

Cloud computing is an Internet-based computing which provides shared computer processing resources and data to computers and other devices. On user demand access to shared computing resources (e.g., Computer Networks, Storage, Servers, Applications, and Services) which can be rapidly provisioned and released with minimal management effort. Cloud computing provides various capabilities to store and process users data in third-party data centers; Third party data centers may be located far from the user in the distance from across a city or across the world. In the simple word, cloud computing are nothing but storing and accessing the stored data and programs using the internet instead of computer's physical drive. Data mining is the process of analyzing the data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of the analytical tools for analyzing data. It permits users to analyze data from many different angles, differentiate it, and summarize the relationships identified. Technically process of data mining used for finding patterns or corelationship among dozens of fields in large relational databases. A general definition of frequent itemset mining is finding a frequent occurrence of itemset from a collection of ‘n’ no of itemset where support for that itemset is greater than or equal to the minimum support threshold. An itemset is nothing but a collection of one or more items. Item can be anything. For e.g. Milk, Butter, Eggs, Coke, etc. Cloud computing, an emerging trend of provisioning scalable computing services, provides the opportunity that data mining is offered as an outsourced service[1]. Frequent itemset mining plays important roles in data analysis from a large database and many other significant data mining task[6]. Frequent itemset mining has been proven essential in many applications such as market data analysis, networking data study, and human gene association study[1]. A frequent itemset mining algorithm runs on a set of text documents produced over a social network can display the central topic of discussion and pattern of usage of words in discussion threads and blogs[6]. But frequent itemset mining is very expensive computation task because of exponentially large data size. There for the weak computational power systems (Client) outsource their data to the computationally more powerful systems (Server). It has been shown that outsourcing data to a service provider brings several benefits to the data owner such as cost relief and a less commitment to storage and computational resources. This introduces the DataMining as Service (DMaS) paradigm[1]. Cloud computing provides a natural solution for the Data Mining as Service paradigm. In the cloud computing era, a better solution is to outsource the computations to a cloud service provider. In addition to saving on overhead and labor costs, Most of the client outsources data to improved efficiency, greater productivity and the opportunity to focus on core products and functions of the business. Outsourcing is a practice used by different companies to reduce costs by transferring portions of work to outside suppliers rather than completing it internally. Outsourcing is an effective cost saving strategy when used
properly. Although outsourcing data to the third party (server) is a viable option to the data owners (client), but client hesitate to place trust on cloud computing. Outsourcing data to the cloud raises few issues 1) First issue is about the integrity of the computed results [1]. There are many possible reasons for the cloud to cheat with computation result. The correctness integrity of mining results can be corrupted if the service provider is with random fault or not honest (e.g., lazy, malicious, etc.). Or Computation result can be damaged if the service provider is honest but makes mistakes in the mining process. 2) The second issue is related to the security or privacy of outsourced data for computation if service provider contaminates or tampers computational result [5] or server can access to valuable data of the owner and may learn or disclose sensitive information from it. 3) The third issue is related to revenue generation by cloud server [2]. A cloud would like to improve its revenue by investing fewer resources while charging for more. A cloud server may provide some fake results instead of being spending its resources in computing the correct ones. The existing research works contribute to practical data security by using different encryption scheme and mapping scheme on data content [7]. Most of the previous research paper focuses on various algorithms which are efficiently used in frequent itemset mining. Integrity verification outsourcing frequent itemset mining is rare and challenging. Out of these above mentioned issues this paper focus on the first issue i.e. result integrity verification of outsourced computation of frequent itemset mining. This article has been classified as follows Section 2 is dedicated to related work, Section 3 Related work, Section 4 Acknowledgement and Section 5 Conclude efforts of the work.

II.RELATED WORK

In [7] author focuses on the integrity and verification issue in Uncertain Frequent Itemset mining problem during the outsourcing process, i.e., how the client verifies the mining results. The author specifically explores and extends the existing work on deterministic frequent itemset out-sourcing verification to the uncertain scenario. For this purpose, the author extends the existing outsourcing frequent itemset mining work to uncertain area w.r.t. the two popular uncertain frequent itemset definition criteria and the approximate uncertain frequent itemset mining methods. Specifically, authors construct and improve the basic/enhanced verification scheme with such different uncertain frequent itemset definition respectively. To verify the outsourced computation [8] proposed different idea that client can use two or more cloud for the computation instead of using a single cloud for computation. If the client wants better assurance of the integrity of outsourced computation, then he can use more cloud for the guarantee. Minimum three clouds must be utilized for computation to get the better assurance of the integrity of computation. Client outsourced his data to the three different clouds for computation, now all cloud returns the output of computation out of that majority of the expected output to be his correct output. The disadvantage of this proposed idea is client have to pay more if the number of clouds used more. An author first introduces the rational lazy-and-partiallydishonest workers [9] in the outsourcing computation model. A new fair conditional payment scheme proposed for outsourcing computation that is only based on traditional electronic cash systems. The proposed construction uses a semi-trusted third party to get the fairness and efficiency. However, the third party is only involved in the protocol in case of any disputes. Compared with the other existing solutions the proposed solution is much more efficient. This advanced work has a disadvantage that Third Party is not fully trusted, and may collude with one party to obtain profits at the expense of the other party.

III.PROPOSED SYSTEM

Figure 1 shows the system architecture of the proposed system. In proposed system the data owner reads the dataset. The data set is used as the input for the probabilistic approach which divides the given data set in frequent and infrequent data set. The key thought of our techniques is to build an arrangement of infrequent itemsets from genuine things, and utilize these infrequent itemsets as confirmation to check the uprightness of the server's mining result. It expels genuine things from the first dataset to build manufactured proof occasional itemsets. Embed duplicates of things that exist in the dataset to develop counterfeit confirmation visit. The deterministic approach is used for the authentication check which is performed by using Merkle Hash Tree. It measure the execution of evidence development at the server side and confirmation at the customer side also, investigated different variables that effect the confirmation execution of our deterministic approach, including different mistake proportion, visit itemsets of various lengths, and diverse database sizes. Same time the data owner sends the requirement to the cloud server where the frequent item sets are computed and send to the data owner where the results are compared with the data owner’s needs. The results are verified according to the user provided budget and output is displayed to the user. The proposed work put forward an idea of verification of outsourced data mining computation of frequent itemset mining. The figure 2 shows proposed system architecture of outsourced data mining computation. The proposed system architecture describes the process of verification of outsourced data mining computation of frequent itemset. In which end user outsourced the data to the cloud for frequent
itemset mining. Then cloud returns the computation result to the end user.

Figure 1: Proposed system architecture for verification of outsourced data mining computation.

The end user generates Power set from the small set of the outsourced data. Then compute for Top N frequent and Top N infrequent itemset by considering threshold value. Calculate the support for the Top N frequent and Top N infrequent itemset. The M tree algorithm is used to form tree on the basis of the support calculated. Same support value itemsets are clustered at node end. The same tree is formed for the return result from the cloud. Then end user compares both the trees to verify the return result received from the cloud. If the both tree matched then returned mining result from the cloud is correct else return result is incorrect. At the end verification report is generated for the end user.

IV. LITERATURE SURVEY

C. S. Yeo [1] in this paper identifies various computing paradigms promising to deliver the vision of computing utilities; defines Cloud computing and provides the architecture for creating market-oriented Clouds by leveraging technologies such as VMs; provides thoughts on market-based resource management strategies that encompass both customer-driven service management and computational risk management to sustain SLA-oriented resource allocation; presents some representative Cloud platforms especially those developed in industries along with our current work towards realizing market-oriented resource allocation of Clouds by leveraging the 3rd generation Aneka enterprise Grid technology; reveals our early thoughts on interconnecting Clouds for dynamically creating an atmospheric computing environment along with pointers to future community research; and concludes with the need for convergence of competing IT paradigms for delivering our 21st century vision. W. K. Wong[2] proposed that Outsourcing association rule mining to an outside service provider brings several important benefits to the data owner. These include (i) relief from the high mining cost, (ii) minimization of demands in resources, and (iii) effective centralized mining for multiple distributed owners. On the other hand, security is an issue; the service provider should be prevented from accessing the actual data since (i) the data may be associated with private information, (ii) the frequency analysis is meant to be used solely by the owner. This paper proposes substitution cipher techniques in the encryption of transactional data for outsourcing association rule mining. After identifying the non-trivial threats to a straightforward one-to-one item mapping substitution cipher, here I have propose a more secure encryption scheme based on a one-to-n item mapping that transforms transactions non-deterministically, yet guarantees correct decryption. I have developed an effective and efficient encryption algorithm based on this method. Our algorithm performs a single pass over the database and thus is suitable for applications in which data owners send streams of transactions to the service provider. A comprehensive cryptanalysis study is carried out. The results show that our technique is highly secures with a low data transformation cost. L. Qiu[3] data mining plays an important role in decision making. Since many organizations do not possess the in-house expertise of data mining, it is beneficial to outsource data mining tasks to external service providers. However, most organizations hesitate to do so due to the concern of loss of business intelligence and customer privacy. In this paper, we present a Bloom filter based solution to enable organizations to outsource their tasks of mining association rules, at the same time, protect their business intelligence and customer privacy. Our approach can achieve high precision in data mining by trading-off the storage requirement.

C. Clifton [4] Privacy preserving data mining – getting valid data mining results without learning the underlying data values – has been receiving attention in the research community and beyond. It is unclear what privacy preserving means. This paper provides a framework and metrics for discussing the meaning of privacy preserving data mining, as a foundation for further research in this field.

IV. SYSTEM IMPLEMENTATION

The system output is mainly based on privacy preserving method. It will be evaluated using robustfrugal algorithm. The evaluation of data can be done by using calculating support value and arrange that in a descending order. Then the data have been split up into two groups to preserve the data from the third party server. Now the data have been encrypted and stored in a server side. Multiplecompanies have access the server. So, the data will be disclosed. For that purpose,
client encrypts its data and stores it in a server in some other format. Based on the mining queries server conducts mining and sends encrypted pattern to the client. Finally client decrypts the encrypted pattern and gets true support of the original transactions.

VI. OUTSOURCING DATA MINING

6.1. Privacy-preserving Mechanisms of Outsourcing

To extract useful business information from huge amounts of data with techniques such as association rule mining, the following IT expertise and facilities are essential (1) Software tools, (2) Hardware infrastructures and (3) Human resources. Not all organizations possess the above enabling resources to carry out data mining by themselves. So there is a trend of growing offshore outsourcing by shifting IT jobs to other places where everything is economical. This trend also indicates that IT outsourcing has become a market-proven cost-efficient model and is changing the landscape of the IT service industry. When an organization hands over its source database to an external party, its data privacy is under the custody of the external party which may not be fully trustworthy. A possible solution is preprocessing the source databases with some kind of masking technique before handing them over to external IT services provider, so that the source databases retain the original association information whilst the privacy contained in the data can be well preserved. This paper [10] suggests that a preferred solution should possess the following properties (1) Time efficient, (2) Storage efficient, (3) High mining precisions. And also discusses, Bloom filter-based approach has been proposed for privacy preserving whilst outsourcing computing tasks of association rule mining. Here an analysis is given of this approach based on the above properties.

6.2. (In) Security and (Im)

Practicality of Outsourcing Precise Association Rule Mining

This paper, [11] analyses both the security and costs associated with outsourcing association rule mining. This shows how to break the encoding scheme discussed in previous research [14] without using context specific information and reduce the security to a one-to-one mapping. Given that a satisfactory definition of security is lacking in the previous WCH+ algorithm, one might attempt to apply the notion of information theoretical security to encoding. It suggests, an encoding scheme is secure if the encoded database contains no information of the input database. However they will be very expensive. Here a solution property is proposed that obfuscates the frequency of the original items. On the whole this research work questions the practicality of outsourcing association rule mining in general. And also outsourcing is not efficient in many settings and still it is an open problem if there exist provably secure encoding schemes that are still practical.

6.3. Non-deterministic One-to-n Substitution Scheme

This paper [12] proposes an improvement on an existed substitution cipher encryption algorithm: non-deterministic one-to-n item mapping. The new transformation is more efficient while it is still valid and secure not be covered by one-to-one mappings. Both theoretical analysis and experiments validate this work. The non-deterministic one-to-n substitution scheme proposed previously encrypts transactional records primarily through an admissible one-to-n mapping m from the original items set I to a set B. Then an additional random subset (E) of B and “fake” items set F is produced and added in order to make sure it is secure enough not to be covered by a one-n-one mapping. In addition, some treatments need to be done on the E-selection for the purpose of promising decryption correct. An analysis is also given for the previous algorithm as well as new improved algorithms. Compared with the original algorithm, this algorithm achieved the non determinacy by selecting suitable E directly and avoided doing redundant operations that have more than 50% probability in the original one. Hence the security and efficiency are two primarily concerned aspects of the delivering process of outsourcing association rule mining.

6.4. Protecting Medical Data for Analysis

This paper [13] presents data analyses from the data protection point of view. It also proposes a solution for outsourced model-based data analyses. A formal framework for protecting the data that leave the organization’s boundary, based on relational data model abstract data type is presented. The data and the data structure are modified so that the process of data analysis can still take place and the results can be obtained, but the data content itself is hard to reveal. Once the data analyses results are returned, the inverse process discloses the meaning of the model to the data owners. The problem with medical enterprises is that employees do not posses in-house expertise for doing data analyses, but they do have domain knowledge and understand the data structures much better. Additionally, the available resources (hardware and software) may not be adequate for analyses. They have two choices: not doing data analyses at all or doing it with help from outside. The former is sometimes not an option; the latter poses a potential security threat to data. In medical environment the data security and data privacy are of a paramount importance. Misuse of a patient’s personal data can cause severe consequences. Once outside the safe-house environment of organization’s databases and data warehouses, they may be used for purposes other than specified. This paper proposes an approach where no data semantics is lost, the statistics inside the data remains intact, but the data are still protected. In our framework, we transform the (relational) database that is to be exported to the outside world. The transformations are to be performed on both data structure and data values. The approach adopts
the terminologies the abstract data type (ADT) and the relational data model (RDM) that is based on it.

VII. CONCLUSION
The presented paper is literature review on different methodology used for the verification of frequent itemset mining computation. This presents the survey on different research paper, finding advantages, disadvantages and different issues related to verification. Simple problem definition for research work has been derived from above survey. An observed conclusion is need of enriched framework for the verification. The future scope is to design and implement a framework for verification of outsourced data mining computation of frequent itemset.

REFERENCES

AUTHOR’S PROFILE:

KARPURAPU SUDHAKARABABU Received The M.Tech Degree From Jntu.K, Jawarharlal Nehru Technological University, And Kakinada In 2013. Currently He Is Working As Assistant Professor Sri Chundi Ranganayakulu Engg. College Chilakaluripet Guntur Dist. In Andhra Pradesh, India.