International Journal on Applications in Information and Communication Engineering
Volume 7 : Issue 2 : May 2021 , pp 38 –43 www.aetsjournal.com ISSN (Online) : 2394 - 6237

SHARING OF ELECTRONIC HEALTH RECORDS THAT IS BOTH RELIABLE AND PROTECTS THE PATIENT'S PRIVACY AND DOES SO USING A TWO-PHASE TOP-DOWN APPROACH WITH INCREMENTAL MAP REDUCE TECHNIQUE

Srikanth R, Arputha Ajitha Rose, Thejesh Manivel, Anishmary
Department of Computer Science and Engineering,
Rathinam Technical Campus,
Coimbatore, Tamilnadu,India

Abstract - Users of various cloud services are able to work together on the analysis of personal data by sharing the data in question. This includes a wide variety of documents, such as medical histories, financial data, and criminal histories, amongst others. In order to maintain the anonymity of the information while still complying with the particular privacy laws, the identities of the data sets have been generalized. K-anonymity is being used more and more frequently these days with the intention of keeping users' private information secure. A wide variety of cloud applications are currently taking advantage of the big data trend in order to store exceptionally massive amounts of data. The process of keeping data sets in the cloud will take a bigger amount of RAM, and users will have a difficult time storing and sharing massive datasets in the cloud. Because of the availability of this vulnerability, users are unable to collaborate on cloud-based data sets. This is a significant limitation. It has been claimed that the incremental mapreduce concept could be utilized in order to discover a solution to this problem. [Citation needed] [Citation needed] Maintaining a track of the number of repeated words that are contained in the datasets is how MapReduce is able to reduce the amount of memory that is required by doing so. It is possible to swap MongoDB datasets with one another thanks to the utilization of the incremental mapreduce technique. This is accomplished by splitting the responsibility of data collection and cutting down on the quantity of data that is replicated within the datasets. In order to simplify the process of the user sharing information with other users and to make it easier for users to do so, we are developing this feature.

Key Words –Health Records. Medical, MangoDB, Mapreduce

I. INTRODUCTION

I. The move toward cloud computing is a trend that is already causing disruptions, and it causes major difficulties to both the traditional IT sector and research groups. This is because cloud computing presents a number of unique challenges. Cloud computing is another technology that is anticipated to be a major driver of economic expansion over the course of the next few years. Cloud computing provides users with enormous capacities for data storage and processing power, and it also enables users to develop programs in a manner that is both cost-effective and does not require large upfront investments or vast physical infrastructures. Cloud computing also provides users with enormous capacities for data storage and processing power, and it provides users with both of these capabilities. Customers Who Use Cloud Computing Stand to Enjoy Both of These Advantages Customers who use cloud computing stand to gain both of these advantages. The fact that it may be difficult to retain one's privacy while working within the constraints of cloud computing is one of the most significant drawbacks connected with this kind of computing. Cloud computing works best when used in conjunction with other types of computing. There are additional concerns regarding privacy that have been around for a while, such as those around the disclosure of private medical information to a research group so that they can perform data analysis on the information. One example of an issue that has persisted for a considerable amount of time is presented here. When the information at question is of a non-interactive type, data anonymization is a common practice that is utilized for the purpose of achieving the objective of protecting the privacy of persons. The protection of individual confidentiality was the motivation behind the development of this strategy. "Data anonymization" refers to the process of concealing the identity of a person or other sensitive information from parties that are not the record's owners. This can be done both for individuals and for sensitive information in general. It is possible to proceed in this manner in order to protect the individual's right to personal privacy. It is feasible to carry out the computations that are necessary for TDS by making use of a method that is known as Two-Phase Top-Down

Specialization. This is one of the ways that this can be done. The two-phase technique is based on the two steps of parallelization, both of which can be completed by utilizing the mapreduce algorithm on the cloud. Both of these stages can be completed in the same amount of time. It does not matter in what sequence you do each of these responsibilities. The job level and the task level are the two levels, or tiers, that are available for parallelization to be implemented on. These two levels are both considered to be sub-tiers of the job level. The term "job level" refers to the simultaneous execution of a considerable number of mapreduce jobs that take advantage of all of the public clouds that are currently accessible. The phrase "job level" is used to refer to this particular concept. The ability to carry out a large number of mapper and reducer operations in parallel across a variety of data divides is what is meant when the phrase "task level" is used. The big data trend has led to a significant increase in the amount of data that has been collected, which has, in turn, led to the development of a wide variety of anonymizing algorithms that are capable of performing a number of different tasks. In other words, the big data trend has resulted in an increase in the amount of data that has been collected by a significant amount. The big data trend has directly led to a huge increase in the amount of data that has been collected, which is a direct result of the trend. This has become a challenge for the anonymization of data sets, and for the processing of massive data sets, we use Map Reduce in conjunction with cloud computing to provide high computational capabilities for application jobs that may be carried out. Anonymization of data sets has become more difficult as a result of this. In the beginning, this presented a difficulty; but, it is now something that really must be done. This has become a challenging position as it has progressed.

II. RELATED WORKS

In accordance with an examination of the data If the GUPT system is equipped with the capability, it will present the final output accuracy assurances in the form of privacy budgets. If the system does not have this capability, the assurances will not be displayed. GUPT is able to provide assurances that additional searches on the dataset can be conducted in a way that satisfies their standards for both accuracy and privacy. Because of the capabilities that they possess, this is something that they are able to accomplish at their disposal. Due to the fact that GUPT has the capability to enable the execution of extra queries, this is not only plausible but also totally doable. This is because of the fact that GUPT has the capacity to permit the execution of additional questions. GUPT is able to perform analysis on material that is kept confidential with a degree of accuracy that is considered to be reasonable. [2] Yingyi Bu Howe, John Magdalena Balazinska Michael D. Ernst et al. In this essay, we claim that software programmers should be held responsible for the increasing popularity of cloud computing because of the way in which they have contributed to the development of this technology. From our point of view, computing, storage, and networking ought to place their emphasis, not on the performance of a single node, but rather on the horizontal scalability of virtualized resources. This is because virtualized resources may be scaled up and down as needed. This is due to the fact that a single node can only carry out a limited number of activities at once. This is the case regardless of whether a cloud provider sells its services at a lower level of abstraction, such as EC2, or a higher level, such as AppEngine. For example, EC2 is a lower level. AppEngine is a higher level. For instance, EC2 is a level lower than EC3. AppEngine operates on a more advanced level. When it comes to scaling, application software must now be capable of doing so in both ways, whereas in the past, this ability was only required for scaling down. Previously, scaling down was the only direction in which this capability was required. In order for it to function properly, the software that manages the infrastructure needs to be aware of the fact that it is now operating not on bare metal but rather on virtual machines. Only then will it be able to function properly (VMs). It is recommended that hardware management systems be created on the scale of a container given that this will be the smallest size that can be purchased. This is because customers will only be able to buy items in this size when they place an order. In [3] L. Wang, J. Zhan, W. Shi, and Y. Liang, et al. The research that was presented in this article made a contribution to the creation of a dynamic service provisioning (ESP) paradigm, which has the potential to be applied to the field of cloud computing. Within the confines of the ESP model, a resource supplier has the ability to provide MTC or HTC service providers with one-of-a-kind runtime environments on demand. On the other hand, service providers have the capacity to adjust the amount of dynamic resources at their disposal. Second, we came up with an enabling system that was based on ESP, built it, and then put it into production. The name Dawning Cloud was chosen for it by us. By applying this method, it is feasible to independently manage workloads that are heterogeneous MTC and HTC. This is possible due to the method's ability to handle heterogeneous data. Thirdly, the results of our tests suggest that on a cloud platform, MTC and HTC service providers, in addition to the resource service provider, may be able to realize economies of scale when managing typical MTC and HTC workloads. This was revealed by the fact that these service providers were able to perform better than the resource service provider. Because MTC and HTC service providers might be able to benefit from economies of scale, we came to this conclusion as a result of that fact. The fact that these service providers were able to take on additional work while having fewer resources available was the determining element that led to this result. We use an analytical method to establish, in the conclusion, but this is by no means the least significant part, that Dawning Cloud is capable of reaching the feasible economies of scale on cloud platforms. This is not the least important aspect. This is true irrespective of the particular workloads that are being executed on the platform at the moment.

The ninth entry focuses on persons including W. Lou and others, in addition to other individuals including N. Cao, C. Wang, M. Li, and K. Ren. A multi-keyword ranked search strategy that also supports latent semantic search has been developed as a direct outcome of this research. An investigation will be carried out using the protocol, and it will look over all of the encrypted material that has been stored in the cloud. In order to create document indexes, we make use of vectors that are constructed from TF value elements, and we also make use of these vectors. The latent semantic analysis, commonly known as LSA, makes use of the matrix that is contained within these vectors in order to analyze the latent connections that may exist between words and passages. These vectors are stored in the same location as the matrix. By encrypting the index as well as the vector that is being searched with a secure splitting k-NN approach, we are able to attain correct ranking results. This has allowed us to achieve our goal of accurate ranking results. Not only does this ensure that the data's integrity is preserved to the fullest extent that can be reasonably achieved, but it also enables us to produce reliable ranking results. This step was taken in order to protect not only the accuracy of the information but also the confidentiality of its storage, which was the primary goal of this action. The proposed method will, in addition to returning files that are an exact match, maybe return files that contain phrases that are semantically linked to the query keyword. This is in addition to returning files that are an exact match. In addition to bringing back files that are an exact match, this will also be carried out. This action will also bring back files that are an exact match, in addition to doing what it already does.

The investigation that Ahmed E. Youssef and his co-authors carried out led to the formulation of important hypotheses and the discovery of relevant data (9). Because of the work done by GUPT, it is now possible for non-specialists to carry out data analytics while still maintaining a respect for the individuals whose data is being analyzed. The efforts put in by GUPT have made it possible for this to occur. GUPT promises that the findings will under no circumstances be disclosed to a third party, and they take this commitment very seriously. We provide a wide range of one-of-a-kind adjustments to the sample and aggregate framework that can be implemented in order to improve the usability and accuracy of analyses that are performed on privately held data. These adjustments can be implemented in order to improve the usability and accuracy of analyses that are performed on privately held data. These adjustments may be implemented in order to achieve the following goals:

2. Anonymization of data: Anonymization of data is a strategy that businesses can use to ensure the security of their data that is kept in the cloud. The National Institute of Standards and Technology is responsible for the development of this methodology (NIST). Despite the fact that it offers a better level of security, this technique enables the data to be continuously evaluated and used, despite the fact that it offers a higher level of security. The process of modifying data in such a way that it inhibits the identification of sensitive information prior to its use or distribution is known as data anonymization. This process can also be referred to as data masking. This is done with the intention of preserving the personal privacy of the people involved. The process that we refer to when we talk about the process that we refer to when we talk about data anonymization is the process that we refer to when we talk about the process of eliminating the paper or digital trail that could lead an eavesdropper to the originator of the information. Erasing the paper or digital trail that could link an eavesdropper to the person who first possessed the information is one of the steps involved in this process. It is possible that the process of anonymizing data will answer some of the problems that have been brought up in relation to the safety of data. This will make it possible to carry out cloud computing in an environment that is less prone to risk and will make it much easier to create a demilitarized zone that has an increased level of safety. Additionally, this will make it possible to create a demilitarized zone that has an increased level of safety. An electronic trail is the information that is left behind after someone sends data through a computer system. This information can be used to track down who sent the data. This information can be used to trace the data's origin and determine when it was transmitted. Those who are trained in forensics can determine who was responsible for conveying the message if tags are attached to the data. This stage is frequently achieved in scenarios involving criminal behavior; the lone exception to this rule is when companies periodically demoralize user isolation in order to gain user data. This stage is reached frequently in scenarios involving criminal action. This point is typically reached when illegal action is a factor in the situation. The conditions surrounding the conduct of a criminal offense are the ones in which this method is most frequently carried out. The disclosure of personally identifiable information such as addresses can be accomplished by corporations through the use of two different strategies: data mining and location tracking. Credit card numbers and Social Security numbers are two examples of other categories of information that can be displayed. It is possible that this is done in order to attract further advertisers, but there may also be other objectives behind it. This may be upsetting for those who place a great priority on the security of their personal information, and it bolsters the argument in favor of the utilization of processes that anonymize data.

After a file has been sent from one person to another, it is possible for the file to still contain information that can be used to identify the person who sent the file to the recipient in the first place. This information can be found in the form of metadata. After the file has been transferred, the information that was logged can be looked over in order to find the dispatcher information that was saved. After the file has been sent, you are able to proceed with this step. On the other hand, once the file has been anonymized, the data

associated with its transmission cannot be used to determine who the original sender was. At the very least, this cannot be accomplished in the manner in which the concept considers it to be achievable. The method of anonymizing data is distinct from other procedures in that, after the data has been anonymized, the field arrangement of the data is maintained, but in other processes, this aspect of the data is lost. This is a significant advantage that should not be ignored. When utilized in contexts that contain test data, the data will, as a direct outcome of the fact that they will continue to convey the idea that they are genuine, continue to give the impression that they are genuine. People who place a high value on the protection of their privacy may have reservations about the possibility that the anonymization process could be undone in certain circumstances. Because there are ways to retrieve personally identifiable information (PII) that has been removed from datasets, a significant number of the anonymization procedures that are now in use can be skipped over. PII refers to information that can be used to identify a specific individual. Information that can be used to identify a particular person is referred to as personally identifiable information, or PII. One method that may be applied to unearth this information is to perform a cross-reference search on any and all sets of papers that may still be located. This is one of the possible approaches that could be taken. The process that is being questioned here is referred to by the phrase "de-anonymization," which is the term that is used to describe the practice.

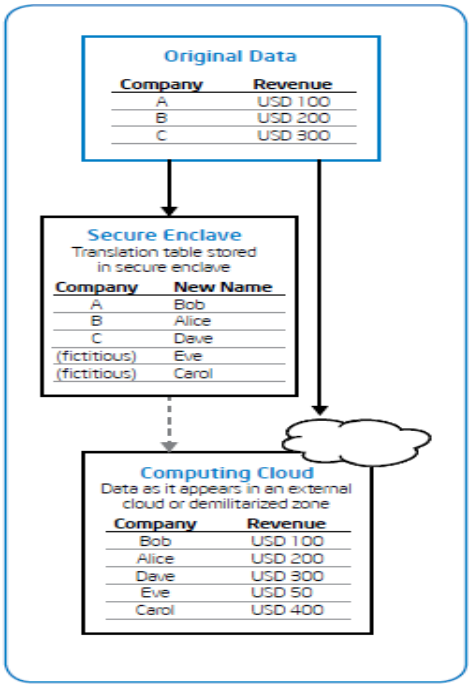


Figure 1: Data anonymization helps enable safer computing in cloud

4. K-Anonymity

The K-Anonymity model offers a method that is both formal and effective for the protection of one's anonymity. If there is an attempt made to identify the data, the goal is to make certain that each record is identical to the other K records regardless of whether or not the attempt was successful in identifying the data. One can say that a collection of data has been K-anonymized if, for each record in the collection that demonstrates a particular set of characteristics, there are at least k-1 more records that share those traits. This is the minimum number of records that must be present for the collection to be considered anonymous. The answer to this question indicates whether or not the data may be used to identify a specific person. K-Anonymity is able to carry out its functions because it assigns qualities to data characteristics and prescribes the manner in which these properties must be controlled. This allows it to execute the functions that it was designed to carry out. Using a method known as k-means clustering, k-anonymity ensures that a certain person inside a set of size k cannot be separated from a given number of other people. This is accomplished by the utilization of the word "k." It is possible to insert fake records into the data even if there are not k sequences of quasi-identifiers that are identical to one another. This is because the data may be parsed and analyzed. Having said that, it is essential to keep in mind that the effects of these fictional records will, at some point in the future, require that they be removed from existence.

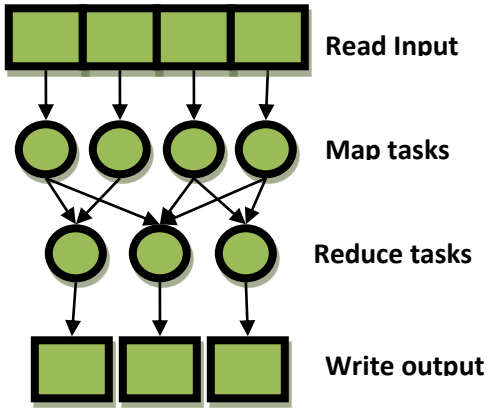


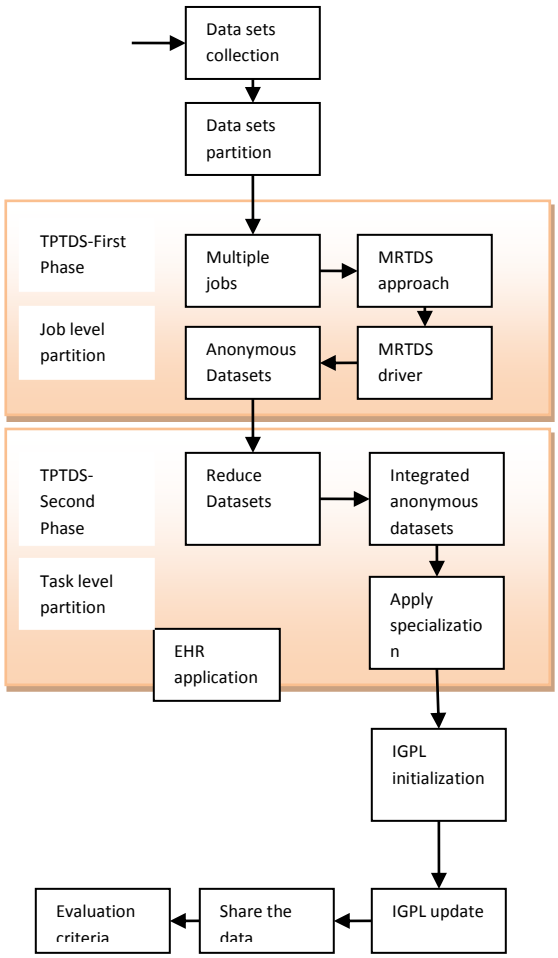
Figure 2: Incremental mapreduce diagram

3. Incremental MapReduce: Map-reduce is a data processing paradigm that condenses large volumes of data into useful collective outputs. Incremental MapReduce is an extension of map-reduce. The term "incremental map-reduce" can also refer to the form "incremental map-reduce." The mapReduce database command is designed to work exclusively with map-reduce processes and is available as an option on MongoDB.

During this particular interval of time of the map-reduce procedure, the "map" action is carried out on each document that MongoDB receives as input. This reduces the amount of work that needs to be done. The key-value pairs will be created by the map function, which is the one responsible for doing so. The reduction phase is utilized by MongoDB for keys that can be reduced to a variety of value representations. Following the completion of this phase's primary objective, the information that was acquired for it will be examined. MongoDB will wait until the operation is finished before storing the results of the operation in a collection. While the confirm function is being carried out, the output of the reduce function can, at the user's discretion, either be provided to additional concentrate or processed in accordance with the aggregate's findings. This is something that can be done even as the confirm function is being carried out. Each and every map-reduce application that is a component of MongoDB is written in JavaScript, and it is executed within the mongDB function. Input for map-reduce processing comes from the documents that belong to a certain collection. Prior to beginning the processing phase that is known as the map phase, any sorting or limiting operations that need to be performed can be carried out. mapReduce has the power to output the results of a map-reduce process either as a document or as a collection, depending on which option the user chooses. Either the input gatherings or the destination gatherings can be partitioned, depending on your preference. It is possible for map-reduce operations to feel like tedious aggregation work. This is something to keep in mind. You should keep this in mind moving forward. MongoDB provides the mapReduce command and the db.collection function in the mongo shell in order to enable the execution of map-reduce jobs. The approach for encapsulating mapReduce() calls is known as the mapReduce encapsulation pattern. In the event that perhaps the map-reduce data set is consistently growing, it is feasible that carrying out map-reduce operations on a subset of the data each time rather than doing map-reduce on the complete data set each time will yield superior results. Because incremental map-reduce can only process data that has been recently uploaded to the system, this is the predicament that has arisen.
In order to carry out incremental map-reduce, the following must be true:

+ A map-reduce task will use the current collection as its input, and the results will be exported to a separate collection. + The current collection will be utilized as input for the task. + When you have additional data to develop, run further map-reduce jobs with them. + + + The query parameter that specifies conditions that match just the recently generated documents + + + The out option that specifies the reduction action that will be used to combine the recently generated results with the output collection that was previously generated

The map-reduce approach that MongoDB makes use of has the capacity of either returning the results in a sequential form or writing them to a collection. Both of these options are available to the user. If you write the results of a map-reduce job to a collection, you will be able to run a subsequent map-reduce operation on the same input collection and then merge, replace, or reduce the results of the previous map-reduce job. This is only possible if you write the results of the map-reduce job to the collection before running the subsequent map-reduce operation. When delivering the results of a map-reduce operation in sequence, the documents that result must have a size that is either smaller than or equal to the limit for BSONDocument Size, which at the time that this article was written is 16 megabytes. If the size of the documents is greater than or equal to the limit, the results will be rejected. Users of MongoDB have the capability to carry out actions known as map-reduce when they are working with collective collections. The following is more information concerning the constraints and limitations that are imposed by the approach. In certain contexts, map-reduce algorithms come equipped with the added capacity of publishing their findings to a sharded collection.



III. CONCLUSION

However, even the most cutting-edge systems have a long way to go before they can fulfill the requirements of the vast majority of data-intensive computing applications. The Location programming model is now the focus of a substantial amount of research. A method for overcoming a number of fundamental limits that are inherent in previously created Map-Reduce systems is described here as part of this body of research with the intention of achieving the aforementioned goal. These restrictions are going to be talked about in relation to the method that has been suggested. We analyze massive volumes of data in this inquiry by utilizing the Hadoop-mongoDB approach, and we construct the mapreduce framework in a static fashion. In today's modern world, a growing number of cloud computing companies such as Cloudera are among those that offer HaaS to their customers.

5. REFERNCES:

[1]. Sreedhar, K. C., Faruk, M. N., & Venkateswarlu, B. (2017). A genetic TDS and BUG with pseudo-identifier for privacy preservation over incremental data sets. Journal of intelligent & fuzzy systems, 32(4), 2863-2873.

[2]. Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. SpringerPlus, 4(1), 1-36.

[3]. Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. Journal of Big Data, 3(1), 1-25.

[4]. Hussein, A. F., ArunKumar, N., Ramirez-Gonzalez, G., Abdulhay, E., Tavares, J. M. R., & de Albuquerque, V. H. C. (2018). A medical records managing and securing blockchain based system supported by a genetic algorithm and discrete wavelet transform. Cognitive Systems Research, 52, 1-11.

[5]. Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 571-588.

[6]. .Enaizan, O., Zaidan, A. A., Alwi, N. H., Zaidan, B. B., Alsalem, M. A., Albahri, O. S., & Albahri, A. S. (2020). Electronic medical record systems: Decision support examination framework for individual, security and privacy concerns using multi-perspective analysis. Health and Technology, 10(3), 795-822.

[7]. Kamal, S., Ripon, S. H., Dey, N., Ashour, A. S., & Santhi, V. (2016). A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset. Computer methods and programs in biomedicine, 131, 191-206.

[8]. .Zhang, X., Leckie, C., Dou, W., Chen, J., Kotagiri, R., & Salcic, Z. (2016, October). Scalable local-recoding anonymization using locality sensitive hashing for big data privacy preservation. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 1793-1802).

[9]. Hamlen, K., Kantarcioglu, M., Khan, L., & Thuraisingham, B. (2010). Security issues for cloud computing. International Journal of Information Security and Privacy (IJISP), 4(2), 36-48.

[10]. Greenhalgh, T., Potts, H. W., Wong, G., Bark, P., & Swinglehurst, D. (2009). Tensions and paradoxes in electronic patient record research: a systematic literature review using the meta-narrative method. The Milbank Quarterly, 87(4), 729-788.