**7- CONCLUSION**

CHAPTER 8- CONCLUSION

Analytics in big data is maturing and moving towards mass adoption. The emergence of analytics increases the need for innovative tools and methodologies to protect data against privacy violation. Many data anonymization methods were proposed to provide some degree of privacy protection by applying data suppression and other distortion techniques. However, currently available methods suffer from poor scalability, performance and lack of framework standardization. Current anonymization methods are unable to cope with the massive size of data processing. Some of these methods were especially proposed for MapReduce framework to operate in Big Data. However, they still operate in conventional data management approaches. Therefore, there were no remarkable gains in the performance. To address these shortcomings, this thesis proposed a framework that can operate in MapReduce environment to benefit from its advantages, as well as from those in Hadoop ecosystems. The framework provides a granular user’s access that can be tuned to different authorization levels. The proposed solution provides a fine-grained alteration based on the user’s authorization level to access MapReduce domain for analytics. By using well-developed role-based access control approaches, this framework is capable of assigning roles to users and mapping them to relevant data attributes.

The research followed a logical sequence of finding the hinders that faced the current anonymization methods of big data. The impairments were related to the concept of choosing the best Quasi-Identifier to generalize or specialize attributes. Moreover, the random split, the poor algorithms, the lack of gradual-access, and the limited number of Q-IDs caused a high degradation in anonymization performance and security. The research has moved from a fact regarding the big data equivalency. The data equivalency increases parallel with the increase number of records. This fact indicates the major difference between traditional data and big data. Generally, big data equivalency is high, specially, in Q-ID attributes that have small number of values. Having these concerns in the current anonymization methods, and knowing some equivalency facts helps us to outline the shortcomings of the current anonymization methods. Therefore, Multi-Dimensional Sensitivity-Based Anonymization methods (MDSBA) was developed to overcome the current impairments and to harness the big data anonymization operations.

MDSBA core concept was derived from probability and Q-ID aggregation. The probability concept has expedited intensive computation. The current anonymization algorithm iterates several times to determine the best generalized or specialized Q-ID. In MDSBA, the best generalized Q-ID is determined prior the anonymization operation. The lowest Q-ID probability value is intuitively generalized. Therefore, there is no need to calculate the best score for each Q-ID attribute. The anonymization process does not impose a direct iteration, instead, the grouping action is repeated several times for generalization. The process starts by grouping all Q-IDs first to filter out the fully equivalent records. The remained records are not fully equivalent. Therefore, all Q-IDs are grouped except the one with the lowest Q-ID probability. The lowest one is generalized by the interval masking or the taxonomy tree. Once again, all Q-IDs are grouped except the lowest two Q-IDs probabilities. Similarly, the lowest two Q-IDs are generalized by the interval masking or the taxonomy tree. This operation of grouping continues till the grouping aggregates only one Q-ID. The masking is finally applied to all Q-IDs except the one with the highest probability. The few left over non-equivalent groups are totally suppressed. The equivalency is measured by the ownership level *k̄* instead of k. this is important to keep the k value constant, while the value of *k̄* increases graduallyas per user’s access.

The current anonymization methods split data randomly to fit in the limited memory size. This random split reduces the gained information and increases the number of non-equivalent records. For this reason, MDSBA splits data logically as per class value. Sensitive class consists of limited number of values, and each value is aggregated in one group and processed separately. This step is initially conducted before the aggregation and anonymization start. The number of aggregated groups equals to the number of class values. This logical split is essential to avoid data overflow to the memory. The parallel distributed framework is able to manage a large data size to a certain extent, since the very large data size may increase error rates and pitfalls, which may unexpectedly terminates the process. The second core concept of MDSBA is the aggregation of Q-IDs. Every two to four Q-ID groups are horizontally aggregated and mapped to one or more of the business roles. Multi-dimensional data requires a large number of Q-IDs, which may exceed tens or even more. The current anonymization methods accept a limited number of Q-IDs, which may not exceed eight or nine. The more Q-IDs added will reduce the performance and more computation time is needed. However, the recent decade has witnessed a technology revolution in social media, which enabled adversaries to develop new attacking scenarios and techniques. This recalled a need to increase the number of Q-IDs, with keeping the processing costs low. MDBSA participated in increasing the number of Q-IDs, and keeping a low processing time. This was structured by aggregating the Q-IDs into small groups.

The thesis starts proposing MDSBA in chapter three. The chapter introduces a preliminarily definition of probabilities and aggregations of MDSBA. The research established some mathematical equations aiming to providing a gradual level of anonymization according to user’s access level. The equations find the sensitivity level for users. Thus, users with large sensitivity levels may gain less information. Moreover, the sensitivity level can be affected by the time factor. Since datasets importance degrades with the time, which causes a sensitivity level decrease as a result.

After introducing MDSBA, the research sequence experimented MDSBA algorithms in a real big data framework. The experiments examined MDSBA anonymization results, and its impact on prediction results. The classification error was used as a benchmark to determine the amount of information loss. The classification error was compared with the other anonymization methods in BUG and TDS. MDSBA failed with the small data size, and succeeded with the larger data size. BUG and TDS showed less classification error rate on anonymizing small datasets, while MDSBA showed less classification error rate on anonymizing large datasets. Measuring the classification error rate is not an accurate method of measuring the information loss, since different classification methods may output various results. Thus, another benchmark was proposed to measure the actual information loss rather than measuring the prediction level. The benchmark, known by Disruption, showed lower disruption results on large size of datasets. The disruption is even lower when data is split into smaller data files, or on using a larger value of k. Moreover, MDSBA computation cost is much lower than BUG and TDS methods in big datasets.

To accomplish one of the main objectives of this research in finding a solution to big data gradual anonymization, a complete framework is proposed. The framework is distributed between the federation service and the service provider. The framework comprises three various services: core, initializer, and anonymizer. The core service is embedded in the federation service, which converts user’s requests to assertions transmitted over SAML service. The service provider stores the requested datasets in parallel with the initializer and anonymizer services. The initializer service operates on the edge server of the service provider, which generates the anonymization script, and transfers the script and other contextual files to the MapReduce domain. The anonymizer service completes the task by conducting the anonymization process. The anonymized copy of data will be created and ready for the user’s access in a secure enclosed directory. The anonymized copy can be generated from production (live) or archived data. The archived data create a stable anonymized copy, which can be available for users as long as the obsolescence value did not expire. On the other hand, the production data expiry may last for few hours or days only.

The research investigated all possible shortcoming in MDSBA security. Two impairments are found in MDSBA structure; Obvious Guess and Across Group Unique Identifiers (AGUI). Several experiments are carried out to eliminate these two security breaches. The experiments results showed that preventing Obvious Guess is possible, by starting an early filtration step. However, AGUI cannot be totally prevented, but it can be mitigated by increasing the k value. These conclusions leaded the research to suggest a large number of k values. The chosen large k value should not affect the gained information negatively. Therefore, Chapter 6 suggested some heuristic-based approaches to find the optimal values of k. Some mathematical concepts are applied to find as large k value as possible. Cumulative frequency and linear regression are used to move upward of k value, and starting from k=20. The k value increases by 10 until the disparity between the actual value and the linear regression line is high. Moreover, the research suggested some security levels controlled by data owners. The security levels are high, medium, and low. The security parameters are controlled by the number of inhibitors and k value interval. The number of inhibitors belongs to the linear regression. Increasing the number of inhibitors will reduce the security level.

Finally, MDSBA was executed in a new parallel framework, known by Spark. It is a memory-base framework, which performs faster than MapReduce. Spark was tested and compared with MapReduce. Two similar algorithms are programmed in Pig Latin script for MapReduce and in Scala script for Spark. The results of the experiments showed a better performance in Spark processing time, when data size fits the memory size. Spark performance has degraded on increasing the data size, while keeping a fixed memory size. In contrast, MapReduce performed better on increasing the data size. MapReduce operates well with the limited resource, and unlike Spark which requires larger resources.

This thesis has answered the research questions by demonstrating that it is possible to provide the data owners with methods that enable them to control the granularity of the user’s access in big data analytics. The research has further demonstrated that it is possible to implement a framework that is able apply an access control model. The model can enforce the organizational business roles over big data. The framework is able to increase the analytics performance and reduce the information loss, which may occur as a symptom of data anonymization. MDSBA method has made some significant improvement for data obfuscation in security and performance approaches. It has proven that applying such a method is desirable in big data. However, MDSBA suffers some hiccups regarding multiple G(QID) groups, and the possibility of some records re-identification. Across Groups Unique Identifier (AGUI) may appear in some cases. Even with low chances of re-identification, AGUI remains the major hurdle in MDSBA framework. Moreover, eliminating AGUI from data records may create a hitch in MDSBA performance.

It should be noted that the experiments in this thesis were conducted in the university lab. The lab is established by virtual machines infrastructure, which was not fully qualified for MapReduce or Spark domains. Also, it contained limited resources of name nodes and workers or data nodes. Experiments were conducted on a maximum of four virtual machines with limited memory and processors. Such humble infrastructure and small scale cannot process really large datasets, say in terabytes, as a reason of resources limitation and the congested transmission medium between nodes. During the experiments, the university network was congested most of the time. The slow transmission has clearly appeared on applying Pig Latin script to read/write to/from the disks. For these reasons, gigabytes of data sets are used instead of terabytes. Moreover, many experiments were repeated several times, to avoid the arbitrary tasks failure. The failure was explained in Chapter 7.

The future of data anonymization will be directed toward finding more advanced method for real-time data. Big data access demands increase with data size increase and the pervasive of data over the cloud. This high demand imposes needs to develop better anonymization methods that are able to obfuscate data in a real-time or near real-time. This is essential to provide a fast and reliable data for various applications. MDSBA may support the real-time anonymization in the future, especially with the fast development growth of processing tool.