# CHAPTER 1 – INTRODUCTION

Big Data is a new technology trend has become a fact as a reason of the massive data growth in the recent past. Digital data can be obtained from a number of quantitative and qualitative data sources, including smart phones, radio frequency identification sensors, Internet of Things, driver trackers, smart watches, smart glasses, embeddables, video recordings, audio recordings, radar, navigation sequences, cheap storages, loud services, social websites, tablets, and others. The International Data Corporation estimates that, on a world-wide basis, the total amount of digital data created and replicated each year has grown exponentially from 1 Zettabyte (1,000,000 Petabyte) in 2010 to 15 Zettabyte in 2017 [1].

There is no rigorous definition of big data. However, the term big data refers to the massive amount of digital information [2]. Two major specifications distinguish big data from the conventional data: Online Transaction Processing scaling (OLTP), and Online Analytical Processing (OLAP). OLTP presents the storing and retrieving, while OLAP presents data analytics [3]. These two features can be gained by using a distributed environment, where many computers process the data in a parallel time [4]. Big data needs to be stored, retrieved and analyzed. Thus, data analytics is one part of big data processes. Data is beneficial when it is analyzed, so users gain more information, and are able to understand the bigger picture of the business activities. Hence, the term data analytics is involved with the big data [5].

Data analysis has a spanning multiple disciplines [6]. Analytics term is becoming an essential part of Information Technology business. Medical, financial, industrial, transportation, government intelligence, and more. Consider data analytics as a prominent tool to monetize business data. Medical organizations request medical data of patients, hospitals, tools and equipment to find the best method of improving their business, and developing medicine and tools. Commercial side is not the only part of data analytics. In 2009, the American centers for Disease Control and Prevention (CDC) has failed to track the H1N1 disease around the states in a real-time. The disease was spreading everywhere, and threatening the public health. Collected information from patients are two to three weeks lag. Google analyzed over 50 million common search terms for Americans, and accurately tracked the areas infected by the flu virus by what people searched for on the Internet [7]. Banks rely on data analytics to develop their customer relationship, mortgage management, risk assessments, and fraud inspection [8]. Data analysis manifests a new exploitation for the recent technology, which supports a real-time collaboration between customers and business. Numerous companies have established their businesses based on the collected data from the collaboration between customers and companies. Facebook, and Uber are one clear example, where customers interact recursively with the business provider. The applications provided by the smart phones induce such an interactivity [9].

Data analytics aims to provide statistical information as a whole, while protecting the privacy of the individuals in the dataset. However, privacy attacks in data analytics is a major concern, which emerged a need for protection policies and algorithms. Hence, scientists proposed several privacy models to reduce the probable attacks against data, by presenting two categories of privacy models: interactive and non-interactive categories. Interactive models tend to hide the actual data, and provide statistical results instead. Data owners provide interactive interfaces, where queries are submitted through to obtain statistical summary results. Protecting against queries is accomplished by sanitization approaches, by adding noise to the input parameters or to the output results. The perturbation is a small numerical value that can be calculated by Laplace or Gaussian equations. These privacy models are known as differential privacy models [6, 10-12]. In the non-interactive models, the data owner, publishes an anonymized copy of the collected data, termed as anonymization or de-identification. Also, data owner removes some personal identifier attributes such as names, birthdates, and social security numbers [13-15]. However, other auxiliary details cannot be removed for statistical and scientific purposes. Information such as age, gender, postcode, marital status, and education are essential information in data analytics.

Homomorphic encryption is another type of the interactive models. Its concept is similar to the differential privacy model, but Homomorphic encryption is more secure, and users cannot access encrypted data. Three types of Homographic encryption are still being developed by researchers: partially Homomorphic (PHE), somewhat Homomorphic (SWHE), and fully Homomorphic encryption (FHE) schemes. In PHE, either multiplication or addition calculation can be operated at once, but not both. SWHE can support a limited number of addition and multiplication operations. Eventually, FHE sustains both addition and multiplication, and can compute any function [16].

Interactive models are secure choice for certain tasks and firms. However, it suffers some limitations of queries. User’s queries read from a black box, so this does not provide a wide range of flexibility on working with data groups, domains, and sub-domains. On the other hand, non-interactive models provide a complete anonymized version of data, where users have the opportunity to view data and rectify the appropriate query for obtaining the statistical results.

In non-interactive models, auxiliary information may provide personal re-identification for a certain extent. These identifiers may not gain 100% of re-identification, but a risk of predicting some data remains high. For example, knowing the patient age, gender, and postcode, may lead to uniquely identifying that patient with 87% [17]. These identifiers are known as Quasi Identifiers (Q-ID). A popular anonymity model, *k-anonymity,* was formally studied by Sweeney [17]. The model suggests an anonymization for Q-ID, which tends to find a group of attributes that can identify some tuples in the database. The model hides the sensitive values by ensuring the equivalency between records with at least k times[13]. Two different techniques were developed to investigate the *K-anonymity*: top-down specialization (TDS) and bottom-up generalization (BUG). The first technique is based on walking through the taxonomy tree from the top towards the bottom, known as the Top-Down Specialization. The second technique constitutes of techniques that generalize data from the bottom of the taxonomy tree towards its top. These two technique aim to find equivalency in each data domain. Examples of BUG are proposed in *ℓ-diversity [18]*, *LKC-Model [19]*, and *(α, k)-Anonymity [20]. An e*xample of TDS is the multi-dimensional TDS (MDTDS) [21] [22].

The previously mentioned anonymization models were proposed for average size data. Big data manifests different scalable approaches, which makes anonymization imposes alternative techniques. There was a need for more relevant models in order to cope with large sizes of data. Recent proposed models, especially for big data, were proposed such as; Parallel BUG [23], Hybrid BUG and TDS [24], and Two-Phase TDS [25]. However, the recently developed anonymization techniques are quite similar to the extant mentioned models. In fact, the modifications, over the previous versions, have degraded the information usefulness.

Moreover, there is no rigorous access control framework for big data. The increased demand for big data analytics has promoted the publicity-driven business. As a result, a larger number of users from different firms are engaged to benefit from data analytics. This recalls a need for a large scale framework that is able to control user’s in a fine-grained access control. The framework should be able to manage user’s authorization, and authentication. As mentioned earlier, anonymization provides a version of anonymized data, which makes re-identification more probable. The need framework should control permit/deny privileges on the data attribute level. Also, the framework should provide gradual levels of anonymization as per user’s access privileges.

Any big data anonymization model is supposed to split the large tasks into small limited tasks. This is essential to utilize the MapReduce slave servers. Hence, each node spends less time on each process, and before transferring the rest of the job to another server. Splitting tasks method is not implemented in MDTDS. Besides, MDTDS splits the large size of data into small junks. This technique negatively affects the information gained and increased the data loss.

Big Data implies data utilization in storage and retrieval on need. Big Data analytics is where advanced analytic techniques operate on big data sets [26]. Hence, analytics is the point of interest in big data, and it may be exploited by data miners to breach privacy [27]. In the past few years several models that address the data leakage concerns have been proposed [28-30]. The data analytics process can be subjected to various attacks. Side attack is considered to be the major violation against privacy [31]. This attack is prevalent for the cases that stopping data leakage is difficult. In such cases, the sensitive information hidden from the attackers, is accessed by them indirectly. For instance, they achieve such data leakage and gain indirect access to data by comparing output results with other internal or external data that they may previously have [32].

Side attacks in data analytics can be categorized into three types: state attack, privacy attack, and timing attack. An adversary code can trigger the state attack [31]. The code changes the values of the static variables, such as the keyword. In this case, the privacy algorithms may lose the protection control. The attacker may run malicious code to transfer the other mapper’s output through the network. Another attack on privacy occurs when the adversary reads some data and compares it with their external data. It is not necessary for the adversary to access the sensitive data. They can predict it, based on the other attributes. Finally, the timing attack is mounted by planting an infinitive loop in the script or by forcing scripts to run longer than their normal cycle. The user search keyword is also prone to attacks. The search keyword can be derived by using different techniques based on the output values, the order of output keys, or the relationship between the output keys and the values [10].

A new breed Data analytics and their utilization in Big Data environments have witnessed rapid growth in the recent past. This trend imposes finding an access control model with a scale-up ability to cope with the big data growth. Many privacy models were proposed to reform data privacy concern on releasing sensitive data. The proposed security models consider a limited number of users access, Mandatory Access Control is used to provide security. Hence, Mandatory Access Control contaminates a large number of user’s access.

It is hard to precisely define big data. Sharing a common definition between academia, industry and the media is intractable. Big data is predominantly associated with two processes: data storage and data analytics [33]. However, the term big data has appeared since the world faced technical difficulties in storing, retrieving and analyzing the massive volume of data. Hardware devices of Memories and CPU technologies were overwhelmed by the numerous volume of data to the point of scalability crisis [34]. Hence, the computer’s hardware development does not cope with the rapid increase in data, which causes technical obstacles on dealing with a large data volume. There was a need to revamp the methods and tools that process big data [5].

In section four, we proposed a novel model of Multi-Dimensional Sensitivity-Based Anonymity (MDSBA). The model is proposed to resolve the above-mentioned concerns in MDTDS. The model introduces a new technique of distributing workload within the parallel distributed MapReduce paradigm. The model adopts a gradual user’s access, and it is integrated with the Role-Based Access Control model. This reduces the compression process to the minimal by anonymizing groups from the bottom to the top, and skipping the equivalent records grouping. Besides, data is not split into small chunks, instead, it distributes data to multi-nodes to reduce the data loss.

This paper is divided into five sections. Section two describes the expected anonymization models specifications in big data. Section three looks into the current anonymization models adopted in big data. Section four proposes the adoptive Multi-Dimensional Sensitivity-Based Anonymization model (MDSBA). And section five introduces the RBAC mapping in the MDSBA model.

[1] J. M. Tien, "The Sputnik of servgoods: Autonomous vehicles," *Journal of Systems Science and Systems Engineering,* pp. 1-30, 2017.

[2] J. K. Jinzhu Kong, "A Practical Approach to Improve the Data Privacy of Virtual Machines," ed, 2010, pp. 936-941.

[3] M. a. Alam, *Oracle NoSQL database : real-time big data management for the enterprise*: New York : McGraw-Hill Education, 2014, 2014.

[4] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, "Addressing big data issues in Scientific Data Infrastructure," ed, 2013, pp. 48-55.

[5] V. Mayer-Schönberger and K. Cukier, *Big data: A revolution that will transform how we live, work, and think*: Houghton Mifflin Harcourt, 2013.

[6] C. Dwork, "Differential Privacy," *Microsoft,* 2010.

[7] V. Mayer-Schönberger, *Big data a revolution that will transform how we live, work, and think*. Boston: Boston : Houghton Mifflin Harcourt, 2013.

[8] D. Bholat, "Big data and central banks," *Big Data & Society,* vol. 2, p. 2053951715579469, 2015.

[9] H. R. Varian, "Beyond big data," *Business Economics,* vol. 49, pp. 27-31, 2014.

[10] A. R. Cynthia Dwork, "The Algorithmic Foundations of Differential Privacy," *Theoretical Computer Science,* vol. 9, 2014.

[11] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, "GUPT: privacy preserving data analysis made easy," in *SIGMOD '12*, K. S. Candan, uk, Y. Chen, R. Snodgrass, L. Gravano, and A. Fuxman, Eds., ed: ACM, 2012, pp. 349-360.

[12] A. Roth and T. Roughgarden, "Interactive privacy via the median mechanism," ed, 2010, pp. 765-774.

[13] L. Sweeney, "ACHIEVING -ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 10, pp. 571-588, 2002.

[14] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets Practice on the Map," ed, 2008, pp. 277-286.

[15] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing Classification Data for Privacy Preservation," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 19, 2007.

[16] M. Ogburn, C. Turner, and P. Dahal, "Homomorphic Encryption," *Procedia Computer Science,* vol. 20, pp. 502-509, 2013.

[17] L. Sweeney, "ANONYMITY: A MODEL FOR PROTECTING PRIVACY " *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 10, pp. 557-570, 2002.

[18] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," ed, 2006, pp. 24-24.

[19] N. Mohammed, B. Fung, and M. Debbabi, "Walking in the crowd: anonymizing trajectory data for pattern analysis," ed, 2009, pp. 1441-1444.

[20] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 754-759.

[21] H. Samet, *Multidimensional and Metric Data Structure*: Morgan Kaufmann, 2006.

[22] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," ed. USA, 2005, pp. 205-216.

[23] A. Irudayasamy and L. Arockiam, "Parallel bottom-up generalization approach for data anonymization using map reduce for security of data in public cloud," *Indian Journal of Science and Technology,* vol. 8, p. 1, 2015.

[24] X. Zhang, C. Liu, C. Yang, J. Chen, S. Nepal, and W. Dou, "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud," vol. 80, ed, 2014, pp. 1008-1020.

[25] L. T. Y. Xuyun Zhang, Chang Liu, Jinjun Chen,, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud," *IEEE,* 2014.

[26] P. Russom, "Big data analytics," *TDWI Best Practices Report, Fourth Quarter,* pp. 1-35, 2011.

[27] P. Samarati, "Protecting respondents identities in microdata release," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 13, pp. 1010-1027, 2001.

[28] N. Qing Zhang, D. Koudas, D. Srivastava, and D. Ting Yu, "Aggregate Query Answering on Anonymized Tables," ed, 2007, pp. 116-125.

[29] S. Ninghui Li, S. Tiancheng Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," ed, 2007, pp. 106-115.

[30] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 1, pp. 3-es, 2007.

[31] A. A. Cardenas, P. K. Manadhata, and S. P. Rajan, "Big data analytics for security," *IEEE Security & Privacy,* pp. 74-76, 2013.

[32] R. Hariharan, C. Mahesh, P. Prasenna, and R. V. Kumar, "Enhancing privacy preservation in data mining using cluster based greedy method in hierarchical approach," *Indian Journal of Science and Technology,* vol. 9, 2016.

[33] J. S. Ward and A. Barker, "Undefined by data: a survey of big data definitions," *arXiv preprint arXiv:1309.5821,* 2013.

[34] P. Russom, "BIG DATA ANALYTICS," *TDWI RESEARCH/IBM,* vol. 4, 2011.