# 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, embeddable, 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 such as; 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 United 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. This approach is conducted 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 highly secure for certain tasks and firms. However, users may find it difficult to create relevant queries, while they read from a black box. Users are unable to access the actual data, they can only view attributes description. 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. Moreover, non-interactive models consider the background knowledge by attackers, when performing attribute linkage protection. This intuition is essential since the recent few years. Cloud services and social media play a very strong role in providing adversaries with precise background knowledge.

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 gain 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]*, and Incognito [19]*.* Example of TDS are proposed in *LKC-Model* [20], *(α, k)-Anonymity* [21], and the multi-dimensional TDS (MDTDS) [22] [23].

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. Proposed models should consider big data processing tools of parallel distributed computing, such as MapReduce. However, Recent proposed models, such as parallel BUG [24], hybrid BUG / TDS [25], and Two-Phase TDS [26], are quite similar to the extant mentioned models for average size data. In fact, the modifications, over the previous versions, have degraded the information usefulness.

Moreover, there is no rigorous access control framework for big data analytics. 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 users in a fine-grained access. The framework should be able to manage user’s authorization, and authentication. As mentioned earlier, anonymization provides a complete version of anonymized data, which makes re-identification more probable. Currently, we are unable to assign the access permission for certain attributes. The needed framework should control permit/deny privileges on the data attribute level. This permits the access of the needed data only. Also, the framework should provide gradual levels of anonymization as per user’s access privileges.

To fill the previously mentioned gaps, this research proposes a novel fine-grained access control framework. The framework follows the BUG anonymization model, with a multi-dimensional sensitivity-based anonymization (MDSBA). The framework provides a scalable anonymization approach, with a parallel distributed computation computability technique.

## Research Question

As discussed in the introduction, big data suffers from lack of a robust framework that is able to manage access control for analytics. Moving big data to the cloud network emerged multi-tenants data storage, and multi-domain of user’s access levels. Security concept has been shifted toward a larger protection scale, where unknown number of users, organizations, and applications may access big data, and from anywhere and at any time. Concurrently, current provided anonymity solutions cannot be durable for such massive growing computational costs, and security threats. For these reasons, we may seamlessly lose control over the empowerment of data analytics. Resolving these concerns may utilize a comprehensive framework that is able to control access privileges in a fine-grained paradigm, mimics the role-based access control model, and supports scalability and performance of big data.

The following question is raised and derived from the Big Data concerns:

How a framework of Access Control Model can enforce the organizational business roles over Big Data analytics, with considering scalability and performance concerns. The framework:

* 1. Should participate in resolving privacy violation of data analytics in Big Data.
  2. Should enhance the efficiency of Online Transaction Process scaling (OLTP)
  3. Can resolve the big data granular access for analytics.
  4. Can enforce the external organizational policies by delegating them access and roles permissions.
  5. Suitable for multi-tenant and multi-domain environments.
  6. Establishes a fine-grained access control, by implementing data anonymity approach.
  7. Integrates RBAC concept within the same framework.

## Thesis objective

The core objective of this thesis is to provide a fine-grained access control framework for big data anonymization. The anonymization is provided by implementing *k-anonymity* base approach, which protects the re-identification of a person on accessing data for analytics. This is essential in multi-domain environment, where users of multi-access levels need to access big data for analytics. The framework provides security by insuring an escalated level of anonymity for less privileges users/firms, and a reduced level of anonymity for elevated privileges users/firms.

The framework leverages the granular anonymity, and able to control an effective access for a certain part of data. The role-based anonymization control framework integrates organizational business roles on deciding the access permissions. Two access permission levels are provided; anonymization level and number of attributes.

## Thesis Contribution

In this thesis, a novel multi-dimensional sensitivity-based anonymization (MDSBA) framework is developed. The framework operates over the cloud network. The cloud network structure is a composition of data owner, a federation service (FS), and a service provider (SP). The framework is able to integrate the FS components and structure. On the other side of the cloud, SP accommodates multi-tenant data repository belonging to multiple owners. MDSBA enables data owner’s self-management for their own data. Moreover, organizations, who wish to participate in data analytics, are delegated to assign security privileges to their own users. Data owners protect their own authorization and authentication data on the FS side. Following this structure, the framework was divided into three main service; core, initializer, and anonymizer. The core service operates on the FS side, which stores the details of: organizations, users, big data information, business roles, organizations security levels, anonymization parameters, and Q-ID groups. The initializer is located on the SP gateway side. Data owners upload big data parameters in XML format. The initializer communicates with the core service throughout security assertion markup language (SAML) [27]. The method of the initializer is to map between FS and SP access privileges and parameters. Finally, the anonymizer operates in the MapReduce domain, where NameNodes, and DataNodes servers apply the granular anonymization by choosing one of the three masking methods: taxonomy-tree, interval, or suppression. The granular anonymization is conducted by sensitivity equation that calculate the specific amount of applied anonymization as per user’s access level.

The anonymization granularity approach relies on two main factors: dynamic value of *k* in *k-anonymity*, and dividing data vertically into small number of Q-ID attributes. MDSBA calculates the user’s appropriate *k* value based on the data owner’s relationship closeness. Business co-owners have closer relationship than business partners or even public users. The closer relationship assigns a lower *k* anonymity value, which in turn a lower anonymization level is applied. The anonymization implements masking operations, by applying data probability on taxonomy-tree, interval or suppression. Data attributes consist of different data types, and each of which is anonymized by one or more of the masking operations. Hence, data with a taxonomy tree anonymization nature can be generalized by moving from the root of the taxonomy tree toward the bottom. Numerical data anonymization can be easily masked by applying an interval. The probability values is derived from the number of times that the attribute value may possibly appear. Eventually, the value of *k* participates in computing the probability value that each masking process should apply. The gradual anonymization is controlled by the taxonomy tree generalization level, the interval range, and the number of suppressed characters. More anonymized data imposes a lower taxonomy level, and a larger interval range.

The framework adopts Hadoop ecosystems, Pig, Hive, and Spark to operate anonymization in highly scalable operations. MDSBA is constructed to fit Hadoop ecosystems operations. This is essential to avoid large size of data splitting into small blocks of data. Larger data size, as in big data, causes a computational data overflow on the temporary memory. Splitting data into small blocks of data size may degrade information gained. Thereby, an advantage of MapReduce features is gained to avoid such a data degradation. Ultimately, MapReduce adapts its own Hadoop distributed file system (HDFS). The file system is designed to read a large size of data blocks, around 128 MB. This structure revokes the need for splitting data in to small size of data blocks.

Moreover, this research contributes in the followings:

* The development of ownership levels and sensitivity levels equations. The ownership levels are affected by two factors of: ownership factor and time factor. The time factor is an optional value, which can be ignored by the data owner.
* The integration of an equation that is able to calculate the amount of anonymized data on any anonymized dataset, by using the disruption value (D). The D value calculates the masking value that was applied on each Q-ID attribute. Each anonymized data block is calculated, while the total value of anonymized blocks equals D.
* The development of a technique that is able to find the optimal solution of *k-anonymity* value. The solution provides data owners with few simple steps to calculate the most optimal *k* value.
* The comparison between MDSBA framework and other BUG and TDS models. The comparison includes performance, and level of information loss. Also, a comparison between Hadoop ecosystem tools was conducted.

## Thesis Layout

The remainder of this thesis is organised as follows:

**Chapter 2** introduces the background and challenges of this thesis. The chapter manifests the big data general definition and the difference between traditional data and big data. This shows the predominant difference is in data analytics. The chapter exhibits the big data analytics challenges, by comparing between data streaming and data batching processes. Moreover, big data analytics is prone for security threats and re-identification of a person. For this reason, the chapter explains amply several methods for preventing re-identification. All proposed security protection methods can be categorized under one of the three privacy concepts: differential privacy, Homomorphic encryption, and *k-anonymity*. Several methods were introduced as examples of differential privacy, and k-anonymity. After introducing security protection method for traditional data, similar sections are introduced for security protection methods for big data. The sections show why identical anonymization methods for traditional data are not successful in big data.

**Chapter 3** preliminarily defines the multi-dimensional sensitivity-based anonymization method (MDSBA), and then delves in the MDSBA details. The chapter starts by proposing the requirements for big data anonymization methods, including: equivalency increase, focusing on security more than information gain, anonymization algorithms should operate in a parallel environment, and gradual access. The rest of the chapter explores the MDSBA and the probability concept, which is the core part of the MDSBA method. Next sections describe data aggregation concept vertically and horizontally. Both probability and aggregation are the two main components of MDSBA structure. The last section before the summary describes the mathematical equations that are necessary to calculate the sensitivity level as per user’s access. This also includes the time factor and its impact on sensitivity value, and the masking operations in taxonomy tree and intervals. Moreover, the anonymization algorithms were described in details for the groups.

**Chapter 4** illustrates the state of the art of MDSBA framework. The first sections illustrate MapReduce and Hadoop ecosystems. The next section presents Hadoop security, and the best protection methods for Hadoop network domain. After introducing Hadoop ecosystems and security, next sections apply MDSBA algorithms by using Pig Latin scripts. The scripts execute the anonymization by applying several masking methods on taxonomy trees and discretization. The forth section compares between MDBSA and Multi-Dimensional Top-Down Specialization (MDTDS). The comparison includes some experiments that measure the prediction level through the classification error. The experiments prove that MDSBA is efficient in performance and has a very low prediction error in the large data size. The experiments were conducted on several datasets in both small and large data sizes. The last few sections focus on creating a new classification bench mark for measuring the performance of anonymization methods. The sections introduce mathematical equations by implementing disruption concept to measure how much is the anonymization impact on data.

**Chapter 5** explores the complete framework of MDSBA. The chapter starts by describing the communication method between the Federation Service (FS) and Service Provider (SP). The communication method is known by Security Assertion Markup Language (SAML). The second section proposes two types of datasets, archive and live data. Next sections present the details of access control and business roles. The sections describe the three services of MDSBA core. These services are distributed in FS and SP sides. The three services include: core, initializer, and anonymizer. The sections conclude the method of generating the Pig Latin script. The second part of the chapter discusses two problematic concerns in MDSBA; Obvious Guess and Across Group Unique Identifiers (AGUI). The chapter part provides solutions for these two concerns. The solutions for Obvious Guess is established by creating a zero filtration stage before anonymization, while resolving AGUI is established by increasing the value of *k*. The last section before the summary explains some experiments that measure the impact of the disruption values on AGUI.

**Chapter 6** discusses special topics regarding *k-anonymity* parameters. The chapter suggests a greedy-based heuristic approach toward an optimal k anonymization value. This is a guidance for data owners on assigning *k-anonymity* parameters. The suggested proposal is related to the role-based anonymization control framework. The framework provides a fine-grained access control by dividing Quasi-Identifier (Q-ID) attributes into vertical groups, with two to four attributes for each cluster. The chapter adopts two mathematical concepts to assign *k* value, which are cumulative frequency and linear regression. The linear regression requires more sophisticated calculations, so it outputs more rigorous results. The cumulative frequency is a special case of linear regression, when the high accuracy is not important. The section introduce *k* percentage parameter to manage and control the role-based anonymization control. The last sections shed light on the dynamic groups of attributes.

**Chapter 7** engages another recent parallel distributed framework, known by Spark. The chapter compares between MapReduce and Spark in processing sensitivity-based anonymization framework. The sections are divided in a sequence of Spark structure, general comparison between Spark and MapReduce, implementing MDSBA in Spark, a comparison between TDS and MDSBA, and finally a comparison between Spark in MapReduce in operating MDSBA. The section of implementing MDSBA in Spark describes the User-Defined Function (UDF) in Pig ecosystem of MapReduce, and in Spark. The section of “Implementing MDSBA in Spark” merely compares between UDF in top-down specialization algorithm, and UDF in MDSBA algorithm. The comparison shows a TDS intensive dependency on UDF through (IF statements) and iteration, while MDSBA implements UDF in much less dependency.

**Chapter 8** provides the conclusions and future work of this thesis. The chapter mainlydiscusses the way the research is developed throughout the end and highlights its contributions. It also reports on the limitations of this work. Finally, the potential future directions for this research are illustrated.

[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] K. Lefevre, D. Dewitt, and R. Ramakrishnan, "Incognito: efficient full-domain K-anonymity," ed, 2005, pp. 49-60.

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

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] K. D. L. a. E. Lewis, "Web Single Sign-On Authentication using SAML," 2009.