FAKULTAT FUR INFORMATIK

DER TECHNISCHEN UNIVERSITAT MUNCHEN

Master's Thesis in Informatics

Implementation of K-Anonymity for the Use Case of Automotive Industry in Big Data Context

Sharada Sowmya

FAKULTAT FUR INFORMATIK

DER TECHNISCHEN UNIVERSITAT MUNCHEN

Master's Thesis in Informatics

Implementation of K-Anonymity for the Use Case of Automotive Industry in Big Data Context

Umsetzung von K-Anonymität für der Automobilbranche im Kontext von Big Data

Author:	Sharada Sowmya
Supervisor:	Prof. Dr. Florian Matthes
Advisors:	Gloria Bondel and Gonzalo Munilla Garrido
Submission Date:	11.03.2021

I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 11.03.2021 Sharada Sowmya

Acknowledgments

First and foremost, I would like to thank my parents, Sasidhara Rao and Jayashree, for always being understanding and patient.

I would like to thank Prof. Dr. Florian Matthes for whom I am grateful for the opportunity provided to pursue my Master Thesis at the SEBIS chair, my supervisors Gloria Bondel and Gonzalo Munilla Garrido, who were highly supportive of me during the time of the thesis and provided a great environment to work in.

Lastly, I am grateful to all the people who provided their valuable feedback during the evaluation phase as it played a crucial part in upgrading the quality of the thesis content.

Abstract

The advancements in modern technologies have led to the creation of large amounts of data on an unprecedented level. This has resulted in challenges of different scales for most industries. The evolution of business models to make the most from such large data sets is one of them.

Automotive Industry, like most other industries, is also faced with similar challenges and opportunities. Modern automobiles consist of a network of sensors that gather a multitude of data. One of the biggest side effects of adapting business models from the perspective of the Automotive Industry is the concern regarding privacy and security. Data is continually gathered at a very granular level from users, and in many cases, the nature of the data is highly sensitive. Therefore, there is a need to ensure the privacy of users from whom the data is gathered, and data anonymization is one such way.

The primary focus of this thesis was the implementation of an approach to achieve data anonymization in the context of big data for the Automotive Industry. Prior to the implementation, an analysis of anonymization techniques such as Mondrian, Incognito, and Datafly was conducted. Discussions with research partners were held to understand the use case and requirements. The knowledge gained from the literature research combined with the list of requirements gathered from the discussions set the basis for the implementation. The initial period of the implementation phase involved the generation of multiple data sets based on the snapshot of a car data set from a European OEM (Original Equipment Manufacturer). To anonymize the large data sets in an efficient manner, data partitioning became a prerequisite that was met through the use of Apache Spark. Following this, each partitioned data set was then anonymized using ARX API. To maximize throughput, anonymization of the data chunks was executed asynchronously using Java Executor Service. The final step involved merging the anonymized data sets into a single one which could then be used for different business requirements.

Lastly, the distributed approach to data anonymization through partitioning was validated by benchmarking the prototype against data sets containing up to 15 million records. The performance of the prototype was observed to have improved through the distributed approach, wherein the results were generated in less time compared to the centralized approach.

Contents

1 Introduction

1.1 Motivation

The advancement in technology has opened a wide window of opportunities to develop unique business ideas. However, this also presents different challenges. One such challenge is the generation of a large amount of data requiring powerful analytics in order to efficiently utilize the resultant data to evolve existing business models. As Stan Lee says, "with great power, there must also come great responsibility," and along these lines, industries that own such big data have a huge responsibility to ensure the privacy and security of their consumers. To achieve this, there is an increasing demand and focus on privacy-preserving techniques to enable industries to make the most of such big data while also ensuring user anonymity.

There are various privacy-enhancing methodologies available for preserving the sensitivity of the information. Data de-identification is one such methodology involving the separation of personally identifiable data to protect the underlying entity from detection. Common strategies to achieve this range from masking and generalization of sensitive information, removal of outliers, data swapping and truncation, the addition of noise, among many others.

There are numerous theoretical solutions for applying de-identification techniques available in the form of concepts proposed in research papers [1]. In the context of software engineering, a sound, proven approach of any of the de-identification techniques which cater to the textual, numerical, and geographical form of data that could potentially work in the big data context is largely missing. This gap in the knowledge base can lead to re-identification attacks [2]. Netflix and AOL are the two examples of companies that released personal data intended to be anonymous but were prone to re-identification of individual users [3]. With the shortage of de-identification experts in the current market, the potential for a third party to re-identify individual data sources from generalized connected data indicates the need for increased privacy protection and safeguard the database containing such information. Moreover, the third parties cannot be generally assumed to be trustworthy, and the transfer of the data to the third parties poses a further risk. As a repercussion, there needs to be a substantial trade-off between individuals' data privacy and data utility [4].

From this perspective, we propose a proof-of-concept for de-identification that could potentially act as a solution in the context of big data for the automotive industry. The aim is to ensure that businesses which make use of the data generated from automobile users for their products and services are compliant with users' privacy requirements.

1.2 Problem Statement

In the current automotive sector, cars generate data about how they are used, where they are, and who is behind the wheel. With the greater proliferation of shared mobility, progress in connectivity within the vehicle itself has resulted in the amount of data from these vehicles growing exponentially [5]. According to a recent survey, these connected cars generate up to 25GB of data per hour [6]. Such collected data include driver and passenger data, driving behavior, bio-metric, health, location, and communication data. Analytical functions can be performed on such collected behavioral car data sets to provide an improved, personalized driving experience and also to achieve general product improvements as well. However, in contrast, this data can potentially reveal information about the car location or the passengers, and intern can allow remote manipulation of core car functions which can have hazardous outcomes [7].

Hence, we propose the use of k-anonymity [8] as the de-identification method for transforming the data sets with the objective of reducing the extent to which information can be associated with individual data principles. This allows data to be used without the possibility of sensitive data being identified. It also provides minimal to no information loss such that analytical functions performed on the data set return valid outcomes.

We will be exploring the existing techniques available for k-anonymity. Alongside, in a big data context, we propose to develop a distributed approach for k-anonymity implementation for the use case of the Automotive industry.

1.3 Research Questions

To achieve the objectives and address these problems, we aim to answer the research questions stated in this section. On a high level, the goal of the thesis is to implement k-anonymity in the context of big data for the use case of the Automotive industry.

RQ1: What are the properties of current k-anonymity implementations/algorithms? As part of this research question, we aim to analyze the existing implementations of k-anonymity. Alongside, we identify the pros and cons of using these implementations

and list their properties like the technique used, complexity, and the scenario for choosing any of the algorithms.

RQ2: What are the requirements for k-anonymity implementations in big data context? This research question aims to identify the additional infrastructure and strategy to be used for extending the implementation of k-anonymity chosen as part of the research question in the context of big data. Here we explore technologies (like Apache Spark) available to handle big data and explore more on techniques like data partitioning. We also conduct discussions with research partners to understand the use case and their requirements.

RQ3: How can a k-anonymity implementation in the context of big data look like? As part of this research question, we will design the system architecture for the implementation of k-anonymity for big data set the context, proceed with the implementation and test the prototype developed with the evaluation metrics designed.

1.4 Thesis Outline

This section provides an overview of the thesis structure through its chapters to provide an easier and more structured understanding for the reader.

Chapter 1: Introduction provides motivation, introduces the problem statement and the goal of the thesis. Alongside, we can define and explain the research questions, their objectives as well the outcome expected from each one of the research questions.

Chapter 2: Research approach states the approach that is used as part of this thesis. The research framework chosen as a basis explains the journey from requirement gathering to evaluation, primarily focusing on the four pillars of the thesis: literature research, design, implementation, and evaluation.

Chapter 3: Theoretical foundations introduces the basis of de-identification methods. We then further explore the concept of k-anonymity specifically. Finally, we realize the need to accommodate big data abstraction into privacy concerns of metadata.

Chapter 4: Related Work introduces related literature which focuses on similar topics and objectives aimed to be achieved as part of this thesis.

Chapter 5: Literature research explores the existing techniques of applying k-anonymity and their implementations. As a result, we aim to list down the properties of all the algorithms available for applying k-anonymity and provide a comparison. Furthermore, we explore the ARX tool and its pros and cons. Lastly, we explore the big data architecture in the current market.

Chapter 6: Use case of Automotive industry describes the general use case with respect to privacy in the Automotive industry. We identify the privacy threats. Moreover, we specifically investigate the two use cases which we are going to realize as part of the thesis. Finally, we perform a threat analysis of the use case.

Chapter 7: Designed Preparation phase inspects the attributes of data sets in the automotive industry. We also explore the concept of synthetic data sets and their properties. Finally, we enlist the requirement for the car data set. Here we define the scripts and strategy used in creating the car data set.

Chapter 8: Designed Implementation explains the system architecture designed for the use case of car tracing data. Alongside, data partitioning performed by Spark is expounded in detail. Lastly, we describe the data anonymization strategy adopted using ARX.

Chapter 9: Designed Evaluation phase captures the results of applying data anonymization techniques against various parameters and illustrates the performance analysis of the prototype.

Chapter 10: Conclusion and Future work summarizes the work done during research through answers to research questions. We also address the limitations that were faced and lastly proposed future work.

2 Research Approach

This section gives a brief overview of the research approach opted as part of the thesis to answer the proposed research questions. The design science research(DSR) approach proposed by Hevner et al. [9] and Peffer et al. [10] is used to define the research framework adopted. The DSR proposes the use of a learning model by building artifacts along the design process. The three primary cycles presented in the design research framework [9] is utilized. First, the theoretical foundation explaining the preliminary research conducted to familiarize with the key concepts and the extensive literature research sets the foundation for a research rigor. In this step, discussions with the companies and research partners are done to define the problem statement and its relevance. By combining the knowledge base thus acquired along with the list of requirements elicited, design and implementation of the prototype are done, and finally, the performance of the prototype is evaluated. Figure 2.1 provides an overview of the research approach using DSR.

2.1 Literature Research

Chapter 3, 4 and 5 explain the extensive literature review done to get an outline of the knowledge base relevant to the thesis. The below databases were used to perform the literature review:

- IEEE Xplore Digital Library.
- ACM Digital Library.
- SpringerLink.

The main topics covered as part of the literature review include setting up the theoretical foundation for anonymization, exploring the existing techniques for k-anonymity, and finally stating the related research work available. The knowledge base thus formed is utilized in the design and implementation in accordance with the DSR approach. The outcome of this part is the guideline for the usage of existing k-anonymity techniques as presented in Chapter 5 thus answering the RQ1.

2.2 Discussions with Research Partners

Chapter 6 presents the outcome of the discussions with the research partners. Meeting with two companies - a European OEM(Original Equipment Manufacturer) and European electronic manufacturing companies were held to understand the use case as described in Chapter 6. Further discussions were held to get a better overview of the target system architecture and the potential technology stack which could be used for the prototype. The resulting artifact generated is the list of requirements that were iterated through multiple meetings to perfect the relevance of the needs of the research partner and, in turn, answering RQ2.

2.3 Design and Implementation

The knowledge gained from the literature research combined with the list of requirements gathered from the discussions set the basis for the design and implementation phase as explained in Chapters 7 and 8. This part of the work is divided into the Preparation, Design, Implementation phase. In the preparation phase, the properties of the data set in the automotive industry are explored. This is followed by studying the properties of the synthetic data set and finally designing a system to create synthetic data set relevant for the use case chosen. The next phase involves designing the system architecture to perform k-anonymity to the data set created in the preparation phase. The final phase involves the implementation of the prototype in accordance with the system architecture designed. The artifacts thus generated as an outcome are the software architecture diagram and the prototype developed. Hereby, RQ3, which asks for proof-of-concept of the implementation, is answered.

2.4 Evaluation

Chapter 9 explains the experiments conducted in order to evaluate the prototype. An assessment of the prototype is done by performing experiments against varied input parameters. The data set containing up to records to 15 million records are considered for the performance analysis. The results thus gained are evaluated. Any indication of minor improvement possibility on the prototype is performed using the 'Access-Refine' approach as defined in the DSR framework [9]. Further refinements which could be performed on the prototype are also described in Chapter 10.

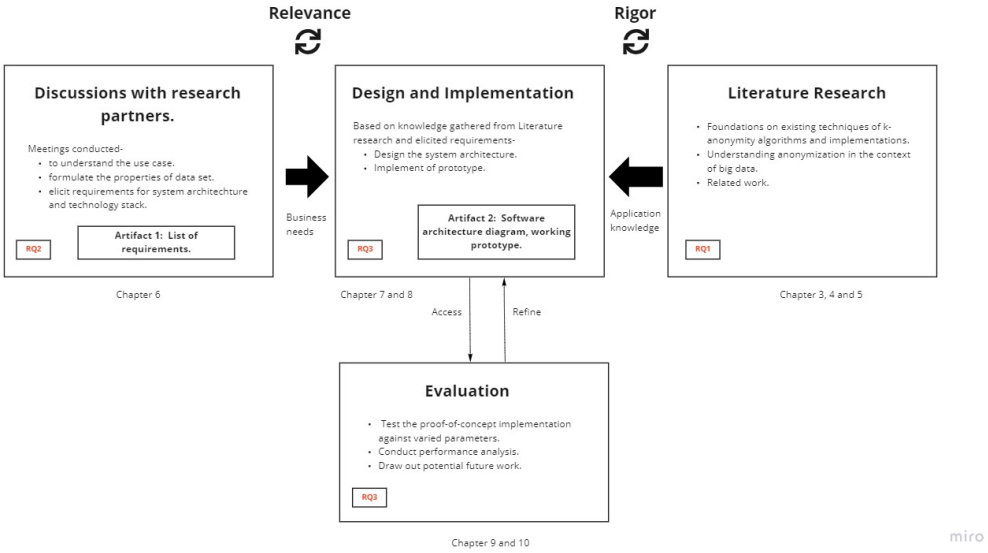

Figure 2.1: Overview of the Research Approach.

3 Theoretical Foundation

This chapter provides the theoretical foundations of the topics covered in the research and aims to give a basic understanding of specific terminologies. The most important definitions in the areas of de-identification, privacy models, and big data context that are relevant for later references are covered.

3.1 De-identification

De-identification is a strategy incorporated to prevent the personal identity of an entity from being exposed. A common example to understand this would be the use case of clinical trials wherein the data produced during the human interaction research needs to be de-identified to protect the privacy of the participating entities [11]. With digitization, there are many possible sources of data in the world. Hence, the main goal of the de-identification technique is to ultimately anonymize data such that one cannot figure out a point where in some identified data set and the de-identified data align, which would result in re-identification of the data set [7].

3.2 De-identification Laws

The most appropriate definition of any de-identification law is the scope of the information that is considered to be covered by that law. There should be a clear bifurcation between what can be considered personal and non-personal data. This clear separation is important to understand the brackets of data that can be captured without risking any entity's privacy [12]. De-identification is the process adopted to prevent an entity's identity from being connected with other related information extracted from the source data. However, there is no legitimate standard for the de-identification of data from all domains. There have been numerous instances wherein the claimed to be anonymous data was easily re-identified [12].

- New York City officials [13], accidentally revealed the whereabouts of individual taxi drivers by releasing poorly de-identified data to the public. However, only a handful of random location data points [14] could be uniquely identifiable 95% of the time. The geo-location data has always proven to be difficult to de-identify.

- Medical records have also proven to be difficult to de-identify. In 2016, the Australian government released an anonymized data set of medical billing records, including prescriptions and surgeries. Additional data sets were cross-referenced to re-identify [15] entities.

Hence, there is a need for bonafide standard de-identification laws. In the next subsection, we will have a look at the Health Insurance Portability and Accountability Act (HIPAA), as it is a very practical and often cited regulation from the US and Bill 64 unanimously approved by the Quebec National Assembly.

3.2.1 HIPAA Privacy Rule

The HIPAA Privacy Rule gives a provision for utilizing health data and providing vital statistics on the data in a responsible manner without the need of acquiring the patients' consent. This is made possible with the help of two primary HIPAA de-identification standards – Safe Harbor and the Expert Determination Method. The mechanism followed by Safe Harbor aims to remove specific patient identifiers like name and phone number. The second standard - Expert Determination Method takes advantage of the universally accepted statistical and scientific principles to make the information not identifiable individually [1].

Safe Harbor

The safe harbor [16] method uses a list approach for de-identification and has two requirements:

1. The removal or generalization of 18 elements from the data [16].

2. The external providers with whom the data is shared do not have the complete knowledge of the information individually or in combination, which could potentially map back to an individual. Safe Harbor dictates the exact rules against which de-identification can be performed. While using this approach, for example, the rule for transforming dates is that all dates need to be generalized to year. Similarly, all the zip codes need to be masked to three-digit numbers. [16]. This strategy is used irrespective of the use case.

Expert Determination

Expert Determination [17] attempts to determine the probability of an individual's identity being revealed from their personal health information catalog by conducting extensive research on current trends and best-practices of performing the de-identification

process on such data. As a result of this approach, a detailed view and practical experience in topics related to scientific and statistical paradigms are required as a pre-requisite for an individual to perform this task. The results of this process are as follows:

1. The final risk of an individual's identity being revealed from the final information alone, or when combined with other available information by the recipient of the information must be low [16].
2. A documentation of the methods and results of the analysis that justify such a determination chosen must be made available [17].

3.2.2 Bill 64

The Quebec National Assembly on June 12, 2020, approved unanimously Bill 64. Bill 64 is an Act to modernize legislative provisions such that personal information is protected under all circumstances. If enacted, this bill increases the protection of data collected by public sectors and any business in the province of Quebec [18]. The primary motivation behind Bill 64 is to introduce personal information protection law, which states that predefined and approved technologies will be allowed to transform the personal data sets with the intention of reducing the risk of any individual being identified [19]. Bill 64 introduces de-identification as a primary method to reduce the "*identifiable*" character of personal information:

"*de-identification*," which is any method that ensures that personal information "*no longer allows the person concerned to be directly identified*" [19];

3.3 Definitions related to Data

Throughout this thesis, various terms related to data are used. This subsection aims to provide appropriate definitions. The Table 3.1 representing data set example will be taken as an example to provide the definitions.

- **Data set / data table**: "[...] a collection of records, where each record is comprised of a set of attributes." [20]. The table above represents data table.
- **Attribute**: "Each column is called an attribute and denotes a field or semantic category of information that is a set of possible values [...]." [21] In the table, A, B and C are the attributes.

A	B	C
A1	B1	C1
A2	B2	C2
A3	B3	C3

Table 3.1: An example data set.

- **Attribute value**: Each cell of a table is referred to as an attribute value. In the table, A1-A3, B1-B3 and C1-C3 are attribute values.
- **Record**: "Each record is related to one data subject and is composed of a set of values[...] for each attribute [...]." [22] In the above table, A1, B1 and C1 represent a record.
- **Numerical data**: a value expressed by a number.
- **Non-numerical data**: a value expressed by characters rather than numbers.
- **Categorical data**: non-numerical data can take a limited number of values (categories).

3.4 De-identification Techniques

In this section we will have a look at the two primary identifiers and explore two commonly used strategies of de-identification.

Personal Identifiers (PID) [23] are a subset of personally identifiable information (PII) data elements, which can uniquely identify an entity and thus expose the entities identity without their knowledge or consent. [24]

`car_id`	`car_model`	mileage	`fuel_percentage`
85a0e584-12f1-47e0-8a5b-08e19c235a12	5 Series	62240	60
85a0e584-12f1-47e0-8a5b-08e19c235a12	4 Series	12020	95

Table 3.2: An example car data set with PID.

In the Table 3.2 `car_id` is Primary Identifier as the `car_id` uniquely identifies each car in the car data set table.

Quasi Identifiers (QID) [25] are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier. [26]

car_model	charging_status	mileage	fuel	timestamp	lat	long
4 Series	CHARGING_ACTIVE	6000	90.0	2018-02-14 04:29:56.009	3.2461	5.5700
5 Series	CHARGING_INACTIVE	16345	80.0	2018-02-14 06:39:56.009	3.2501	5.5743
4 Series	CHARGING_ACTIVE	26354	40.0	2018-02-14 08:49:56.009	3.2511	5.5763
5 Series	CHARGING_INACTIVE	12908	30.0	2018-02-14 09:59:56.009	3.2529	5.5772
4 Series	CHARGING_INACTIVE	7897	50.0	2018-02-14 10:09:56.009	5.2450	11.5712

Table 3.3: An example car data set with QIDs.

In the Table 3.3, `charging_status`, `fuel`, `timestamp`, `lat` and `long` are the quasi-identifiers. The combination of these attributes will reveal the identity aspect of the car. For example, from the first record in the table, we can see that a car is getting charged with a fuel percentage of 70% and the remaining charging time as 20 minutes. From this data, we can conclude that a 5 series BMW car can be found in the GPS latitude and longitude mentioned, thus revealing the identity of the car. Most customary strategies used for de-identification are concealing the personal identifiers and generalizing quasi-identifiers [27]. Pseudonymization is the commonly used technique for masking. One of the main techniques adopted for the generalization of quasi-identifiers is k-anonymity. In the following subsection, we will get a general overview of both of these techniques.

3.4.1 Pseudonymization

Pseudonymization [28] is a technique of replacing primary identifiers with pseudonyms to hide the identity of an entity. This is achieved by either deleting or masking the personal identifiers to make the entity unidentifiable. Let us take the example of a car maintenance workshop company that wants to analyze the fuel consumption of cars and find a pattern over the car types to decide which cars fuel efficiency reduces over the usage time. In the Table 3.4, the `car_id` is a personal identifier which uniquely

identifies each of the car. The analysts at the car maintenance need not know the details of this sensitive data to perform their analysis and potentially expose the provider which unnecessary risks and compliance by sharing this information.

Table 3.5 depicts an example of the same data set that has been de-identified using pseudo anonymization. The attribute values of the field `car_id` have been pseudonymized to a string of 5 masking characters so that the original values are no longer visible. This will allow the analysts to find the pattern of fuel efficiency, which is how fast the car runs out of charge as the mileage (implies more usage) increases.

Based on the scenario outlined above, we can see how personally identifiable information within the customer data set has been de-identified through a process of pseudonymization.

3.4.2 K-Anonymization

K-anonymization defines attributes that indirectly point to the individuals' identity as quasi-identifiers (QIDs) and deal with data by making at least k individuals have the same combination of QID values [29]. QI values are handled following specific standards. The next section gives a brief overview of K-anonymity and the methods used in K-anonymity.

3.5 K-Anonymity

The concept of K-anonymity as a privacy model was developed by P. Samarati in her paper [30]. For K-anonymity to be achieved in a data set, there need to be at least k-records that share the set of attributes that might reveal the identity of the entity whose properties are represented in the data set [31]. In other words, a data set is said to be k-anonymous if every combination of values of the quasi-identifiers in the data set appears at least in k different records. For example, the data set in the Table 3.6 is 2-anonymous.

In the above example the `charging_status`, `fuel`, `timestamp` and the `lat` and `long` are the quasi-identifiers as explained in the section on quasi-identifiers. Every combination of the values for these attributes appears at least 2 times in the resultant data set. Thus, the resultant data set is said to be 2-anonymous. The main intuition behind making any data set k-anonymous is that with an adequate value of k, the re-identification attack would not be possible. Even if a secondary data set is used to figure out the identity of the entity, it would be impossible to do so. Since each combination of these

car_id	car_model	charging_method	charging_status	smart_charging	mileage	fuel
85a0e584	5 Series	AC_TYPE1PLUG	CHARGING_INACTIVE	OPTIMIZED	6000	2100.0
fbf84e1d	5 Series	AC_TYPE2PLUG	CHARGING_ACTIVE	OPTIMIZED	16345	20.0
5aa3d6ee	5 Series	AC_TYPE3PLUG	CHARGING_INACTIVE	UNOPTIMIZED	26354	90.0
12920f4b	5 Series	AC_TYPE4PLUG	CHARGING_ACTIVE	UNOPTIMIZED	12908	10.0
b64f7911	5 Series	AC_TYPE1PLUG	CHARGING_INACTIVE	UNOPTIMIZED	7897	50.0
f099e7de	5 Series	AC_TYPE2PLUG	CHARGING_INACTIVE	UNOPTIMIZED	123421	40.0

Table 3.4: An example car data set to demonstrate Pseudonymization.

car_id	car_model	charging_method	charging_status	smart_charging	mileage	fuel
*****	5 Series	AC_TYPE1PLUG	CHARGING_INACTIVE	OPTIMIZED	6000	2100.0
****	5 Series	AC_TYPE2PLUG	CHARGING_ACTIVE	OPTIMIZED	16345	20.0
****	5 Series	AC_TYPE3PLUG	CHARGING_INACTIVE	UNOPTIMIZED	26354	90.0
****	5 Series	AC_TYPE4PLUG	CHARGING_ACTIVE	UNOPTIMIZED	12908	10.0
****	5 Series	AC_TYPE1PLUG	CHARGING_INACTIVE	UNOPTIMIZED	7897	50.0
****	5 Series	AC_TYPE2PLUG	CHARGING_INACTIVE	UNOPTIMIZED	123421	40.0

Table 3.5: Psuedoanonymization applied on Car data set.

car_model	charging_status	mileage	fuel	timestamp	lat	long
4 Series	CHARGING_*	6000	very full	morning	[3.245-3.250]	[5.570-5.575]
5 Series	CHARGING_*	16345	very full	morning	[3.245-3.250]	[5.570-5.575]
4 Series	CHARGING_*	26354	normal	morning	[3.251-3.256]	[5.576-5.581]
5 Series	CHARGING_*	12908	normal	morning	[3.251-3.256]	[5.576-5.581]

Table 3.6: An example 2-anonymous data set.

records is linked with at least k different identical records, the re-identification will fail. In the next section, we will have a look at how the data set is anonymized to render a 2-anonymous data set.

3.6 Methods of K-Anonymization

The two main strategies used to transform the data set to a k-anonymous data set are generalization and suppression. The next two subsections will describe these two methods in detail.

3.6.1 Generalization

Generalization [31] is the technique of reducing the precision of value of quasi-identifiers such that the records are modified to records which share same values. Consider the records in the Table 3.7

car_model	charging_method	mileage	fuel_percentage
4 Series	CHARGING_ACTIVE	6000	90.0
5 Series	CHARGING_INACTIVE	16345	80.0
4 Series	CHARGING_ACTIVE	26354	40.0
5 Series	CHARGING_INACTIVE	12908	30.0
4 Series	CHARGING_INACTIVE	7897	10.0

Table 3.7: An example car data set for methods of k-anonymization.

The attribute values of `fuel_percentage` can be mapped to numerical ranges, and each numerical range can be further transformed into a string that represents the numerical range. In the example, fuel percentage ranges are from 90.0 to 30.0. The numerical range can be [100.0 – 80.0], which can be further transformed to 'very full', [60.0 – 30.0] is another numerical range that can be further transformed to 'normal'. Generalizing a numerical value into a range is one of the most widely used techniques. Other ways include removing a value entirely, which is showcased for the attribute `charging_method`. The two values applicable for this attribute - `CHARGING_ACTIVE` and `CHARGING_INACTIVE` are transformed to `CHARGING_*` as shown the Table 3.8.

car_model	charging_method	mileage	fuel_percentage
4 Series	CHARGING_*	6000	very full
5 Series	CHARGING_*	16345	very full
4 Series	CHARGING_*	26354	normal
5 Series	CHARGING_*	12908	normal
4 Series	CHARGING_*	7897	low

Table 3.8: An example car data set post generalization.

There are two ways of applying generalization technique on a data set. Using the data set showcased in the below table as example, the two strategies – global and local generalization can be explored. Consider the data set shown in Table 3.9 on which the two generalization technique will be applied.

car_model	fuel_percentage
4 Series	90.0
5 Series	80.0
4 Series	85.0
5 Series	88.0

Table 3.9: An example car data set to demonstrate two methods of generalization.

In **global generalization**, a given value of an attribute will always be generalized the same way throughout the data set. The fuel percentage of 90.0 is will always be mapped to the numerical range [100.0 – 80.0]. Using global generalization, the data set in Table 3.9 can be transformed to the anonymized data set as shown below.

car_model	fuel_percentage
4 Series	[100.0 - 80.0]
5 Series	[100.0 - 80.0]
4 Series	[100.0 - 80.0]
5 Series	[100.0 - 80.0]

Table 3.10: An car data set post applying global generalization.

However, in **local generalization** [32], the constraint of applying the same generalization for all the values which satisfy the constraint defined is not present. This gives the liberty to choose a different generalization for each record. In the table, a value of 90.0 for fuel percentage can remain as is for one record and generalized for the other. Using local generalization, the above example table chosen can be transformed to the below-anonymized Table 3.11.

car_model	fuel_percentage
4 Series	90.0
5 Series	[100.0 - 80.0]
4 Series	90.0
5 Series	[100.0 - 80.0]

Table 3.11: An car data set post applying local generalization.

3.6.2 Supression

In the previous section, the records in the table had relatively close values. Hence achieving a 2-anonymous data set was quite easily achievable and the resultant data set is quite accurate as well. Consider the records in the Table 3.12.

In the Table 3.12, the first four records can be grouped into two pairs, each pair consisting of two records each. The fuel percentage range for this example will be [100.0 – 80.0] and [60.0 – 30.0], which take care of the first and the third record and second and fourth record, respectively. However, the last record is an out-liner. Grouping the last record with any of the other records in the data set would mean having large numerical ranges, which are fuel percentages between 60.0 to 10.0. This would be significantly reducing the utility of the resulting data set. So, a simple answer would be to simply remove them from the data. Hence, using both generalization and suppression on this example table will give the below 2-anonymous Table 3.13.

car_model	fuel_percentage
4 Series	90.0
5 Series	60.0
4 Series	85.0
5 Series	55.0
4 Series	10.0

Table 3.12: An example car data set to demonstrate suppression

car_model	fuel_percentage
4 Series	[100.0 - 80.0]
5 Series	[60.0 - 40.0]
4 Series	[100.0 - 80.0]
5 Series	[60.0 - 40.0]

Table 3.13: An car data set post applying suppression

3.7 Big Data Context

Big data began to gain popularity initially back in the 1990s, and its popularity and sheer importance have only increased exponentially over the period time. The term Big Data is used for amounts of data that cannot be processed by traditional database management systems [33]. This large amount of data can be structured or unstructured that overwhelms companies every day. The most important aspect of big data is what can be done with such large volumes of data. Analysis of such data to gain more insights and use them as a basis to make decisions and form strategies is a common idea which a company aims to do [34]. However, specific characteristics of big data can prove to be both an advantage as well as a challenge for exploring the full capability of such data. These characteristics, issues, and challenges of Big data are usually described in three Vs. which depict the dimensions and characters of Big Data that make it different from traditional data. Figure 3.1 gives an overview of the characteristics of Big Data, depicted in the 10 Vs. This approach was first proposed in the book [35]. The description provided in this book is taken as a reference to briefly explain the 10 Vs.

- **Volume** refers to the enormous amount of data sets generated by activities performed on a day-to-day basis using the components of all business ventures across the world. Storing and processing such large volumes of data is of primary importance. [36].

- **Value** refers to the practical usage of the captured data. The insights thus gained from the data play a vital role in determining the direction of any business [35].
- **Velocity** refers to the speed at which data is generated, captured, and further processed [35].
- **Veracity** refers to the reliability of the data captured and determining the accuracy of the data at the same time.
- **Viscosity** refers to the ease with which the captured data can be used for further processing. The data captured can be structured or unstructured data.
- **Variability** refers to inconsistencies mapped to the data. These outliers present in the data affect the analytics performed on the data set [37].
- **Volatility** refers to the lifespan of the captured data. This directly maps to the storage time of the data and the validity of the data over a period of time [38].
- **Viability** refers to the capability of the data set to available for usage as and when it is required.
- **Validity** refers to the captured data itself being valid post numerous manipulation done on the data.
- **Variety** refers to the ability to handle any form of data from structure, semi-structured, to unstructured data.

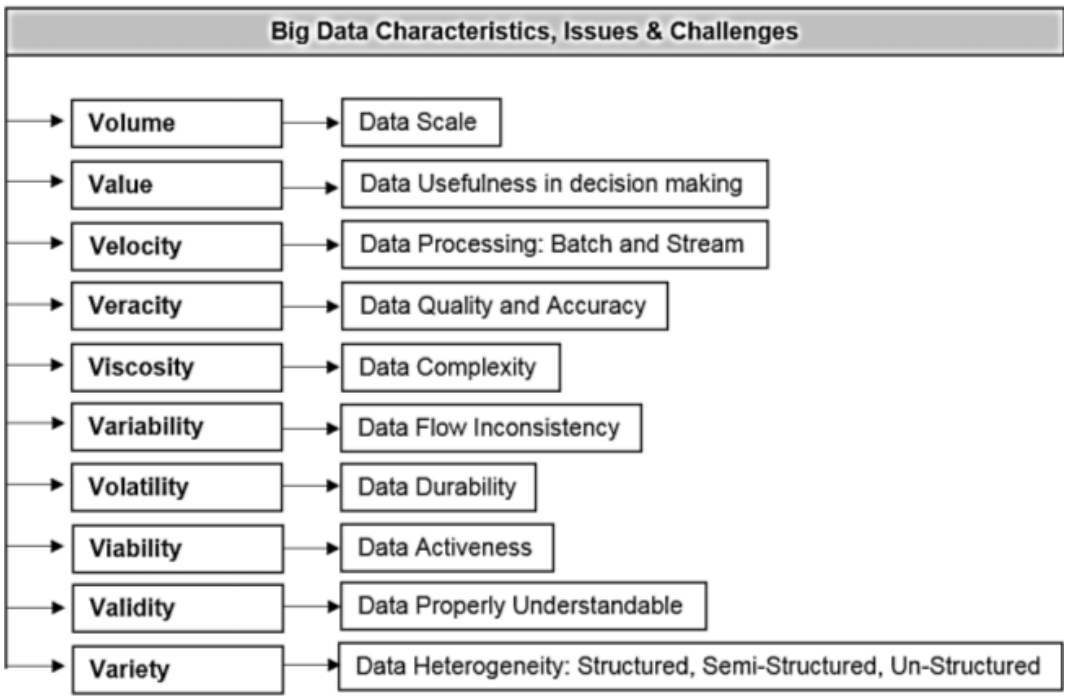

Figure 3.1: 10 Vs. of Big Data as described in the Book [35].

From the characteristics pointed out, big data today can be seen as an opportunity that can provide significant upgrades to every activity performed on a day-to-day basis, including leisure, health, environment, and so on. Hence, this data can be considered as the new fuel of the digital era or new oil of the 21st century [39]. However, irrespective of these characteristics of big data mentioned, one of the main concerns still remains to be security and privacy related to data itself [40]. Hence, it is important to take advantage of the distributed systems, and parallel processing approaches currently available to handle big data sets [40].

4 Related Work

This chapter introduces related literature that focuses on similar topics and objectives to achieve as part of this thesis.

Two-Phase Top-Down Specialization(TDS): Top-Down approach [41] for data anonymization is an iterative process that traverses through the taxonomy tree of the attributes from the topmost root node. At each iteration, it finds the most generalized values and specializes the value till the k-anonymity is violated. However, the main challenge with this approach is scalability for the use case of relatively large data sets. The research work by Xuyun Zhang et al. [42] proposes a scalable solution called the two-phase top-down specialization approach for the anonymization of large data sets using the MapReduce framework. A two-step process is adopted to achieve optimal anonymization. In the first step, the original data set is partitioned into smaller data sets. These smaller data sets are anonymized, thus producing intermediate results. In the second phase, the intermediate results are further generalized to obtain the final anonymized data set.

Parallel Bottom-up Generalization(BUG): Bottom-Up approach [43] for data anonymization is again an iterative process of generalizing the information. This involves traversing through the taxonomy tree of attributes from the bottom upwards. However, the challenge for existing anonymization techniques is to achieve privacy preservation for large-scale data sets. The research work by K.R.Pandi Lakshmi et al. [44] proposes the use of MapReduce jobs performing the data anonymization to accomplish generalization in a scalable manner. The research paper further proposes the approach - Advanced Bottom-Up Generalization where generalization is performed on different partitioned data set and thus formed intermediate anonymized result sets are merged to form the final anonymized data set.

A scalable hybrid approach of using TDS and BUG: The research work by Xuyun Zhang et al. [45] aims to combine the best features provided by TDS and BUG for efficient sub-tree anonymization for large-scale data sets. This approach tries to exploit the computation capability provided by MapReduce jobs to accomplish data anonymization of big data sets [46]. The results of this approach prove that the hybrid approach significantly improves scalability and efficiency when compared with TDS or BUG approaches taken into consideration individually.

5 Existing Implementation of K-Anonymity

As part of the literature research to figure out the existing approaches available towards k-anonymity, three primary k-anonymization algorithms were chosen to be investigated. The reason for choosing these specific algorithms is based on the following reasons: (1) these algorithms are popular in terms of the number of citations each of the algorithms received in the research papers related to k-anonymity. (2) these algorithms take into consideration the two main techniques – generalization and suppression of k-anonymity. (3) these algorithms have a valid implementation that could be used for verification. (4) the comparison of these algorithms in combination has been cited in research papers as well [47].

As a next step, each of these three algorithms can be applied in a top-down or bottom-up approach which will be explored in the second subsection. Since we are finding a solution that should work for big data context as well, the process of applying these algorithms to any data set in a centralized and distributed way is explained in detail.

5.1 Properties of K-Anonymity Algorithms

This section describes the various approaches to applying generalization and suppression techniques of k-anonymity. As an illustrative example to demonstrate the anonymization technique using Mondrian, Incognito, and Datafly anonymization, the below table of car records will be used. The attribute `car_id` is a personal identifier as it can uniquely identify each car. The combination of attributes `charging_method`, `fuel_percentage` and `zip_code` serve as the quasi-identifiers as the attributes together can potentially identify the location of the car. The attribute mileage is considered an insensitive attribute.

5.1.1 Mondrian

The paper [48] introduces the Mondrian Multidimensional k-anonymity. The basic algorithm is a two-step process. In the first step, a greedy partition algorithm is used to

Personal Identifier	Quasi Identifiers			Insensitive Attribute
car_id	charging_method	fuel	zip_code	mileage
85a0e584	CHARGING_INACTIVE	80.0	80801	60000
fbf84e1d	CHARGING_INACTIVE	60.0	80812	16345
5aa3d6ee	CHARGING_ACTIVE	30.0	80804	286354
12920f4b	CHARGING_ACTIVE	40.0	80815	12908
b64f7911	CHARGING_INACTIVE	55.0	80819	7897
f099e7de	CHARGING_ACTIVE	20.0	80802	234231

Table 5.1: An example car data set to demonstrate k-anonymization techniques.

partition the complete data set. All the attribute values which are available in the data set are partitioned into smaller subsets of the original data set. These subsets with all the attributes from the original data set are referred to as regions. The partition algorithm is recursively called till each region consisting of k records is formed. The partitioning starts with the attribute identified as a quasi-identifier in the data set chosen which has the least number of distinct values. To choose the attribute against which the partition needs to be done, Mondrian adopts the strategy of median partitioning [49]. Usually, the attribute with the largest range of values is selected. If two attributes have the same number of distinct attribute values, then the attribute which enables k-anonymity to be not violated once the partition is formed is selected. However, in the use case where the attribute has only two values (`CHARGING_ACTIVE`, `CHARGING_INACTIVE`), the partition is performed in such a way that each partition will have only one of the two values. In the second step, mapping of domain values of Quasi-identifiers to generalized or altered values is constructed using rules defined for each of the attributes in each region. The run time of the greedy partitioning algorithm is O(nlog(n)).

The result of applying the Mondrian k-anonymization technique on the car data set is displayed in Table 5.2 provided below. The input metric for the anonymization is as follows: k=2 and Quasi Identifers=`charging_method`, `fuel_percentage`. To provide a valid spatial representation of the data, the k value is chosen as 2. According to the Mondrian algorithm, the first partition starts by choosing the attribute with the least number of distinct values. In the example data set, the attribute `charging_method` has 2 distinct values, and hence the first partition is done on the attribute. Next, the attribute against which the partition needs to be done is chosen. For this, the attribute with the largest range of values is selected. In the example chosen, the attribute `fuel_percentage` is chosen, and partition is done using the median partitioning.

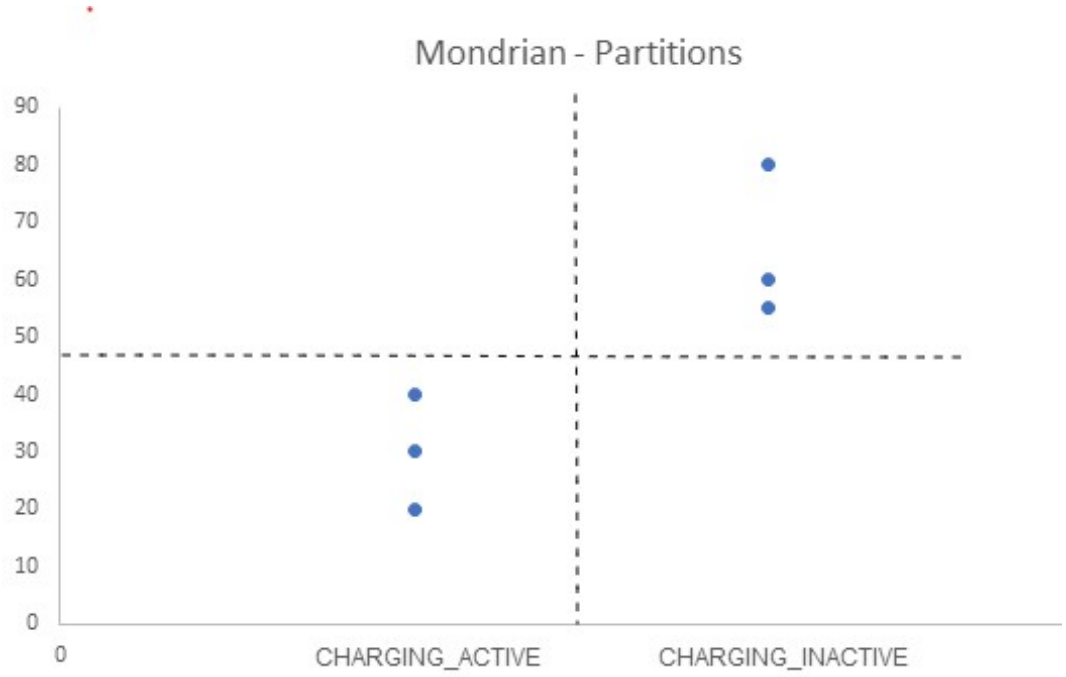

Figure 5.1: Graphical Representation of the Mondrian Partitions.

charging_method	mileage	fuel	zip_code
CHARGING_ACTIVE, CHARGING_INACTIVE	60000	(40.0-80.0]	80801
CHARGING_ACTIVE, CHARGING_INACTIVE	16345	(40.0-80.0]	80812
CHARGING_ACTIVE, CHARGING_INACTIVE	286354	[20.0-40.0]	80804
CHARGING_ACTIVE, CHARGING_INACTIVE	12908	[20.0-40.0]	80815
CHARGING_ACTIVE, CHARGING_INACTIVE	7897	(40.0-80.0]	80819
CHARGING_ACTIVE, CHARGING_INACTIVE	234231	[20.0-40.0]	80802

Table 5.2: Anonymized car data set using Mondrian.

5.1.2 Incognito

Incognito [50] is a full-domain generalization algorithm that uses the approach of dynamic programming with the help of subset property. According to the subset property, a relation T onset of attributes is said to be k-anonymous with respect to a chosen set of attributes if all the subsets of the set of attributes are k-anonymous. This is achieved in three steps. In the first step, the domain and value generalization hierarchy is defined

for all the quasi-identifiers. Figure 5.2, 5.3, 5.4 show the possible domain generalization for the attributes `charging_status`, `fuel_percentage` and `zip_code`. If each of the quasi-identifiers has distinct domains, the domain generalization hierarchy formed in step one can be combined to form a multi-attribute generalization lattice. Figure 5.5 shows the generalization lattice created for the quasi-identifiers. Each node in the lattice represents a generalization solution. In the lattice shown in Figure 5.5, the node <C0, F1, Z1> is a direct generalization of <C0, F1, Z0> and is an implied generalization of <C0, F0, Z0>. The third step is to perform the anonymization of data. Using a breadth-first search algorithm, the lattice is traversed. While traversing the lattice, each node is checked to if k-anonymity is satisfied. If a node satisfies k-anonymity then all its direct generalizations are removed as it is guaranteed that the subsets also satisfy k-anonymity. The overall complexity of the Incognito algorithm is exponential to the number of quasi-identifiers. The results of applying the Incognito k-anonymization technique on the car data set displayed in Table 5.3 are provided below. The input metric for the anonymization is as follows: k=3 and Quasi Identifiers=`charging_status`, `fuel_percentage`, `zip_code`. The result and the anonymized data set are displayed below.

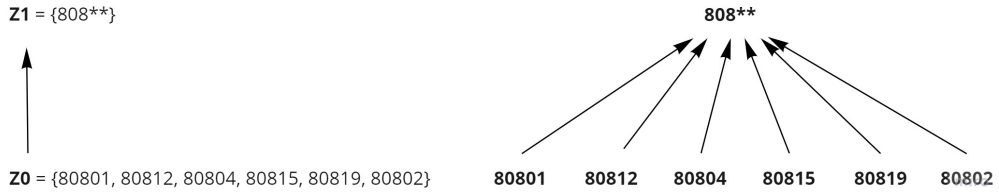

Figure 5.2: Incognito: Domain Generalization of the attribute Zip Code.

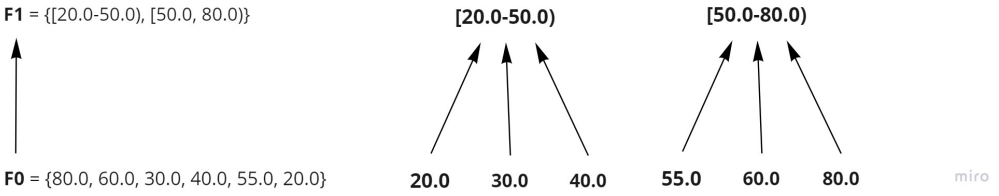

Figure 5.3: Incognito: Domain Generalization of the attribute Fuel Percentage.

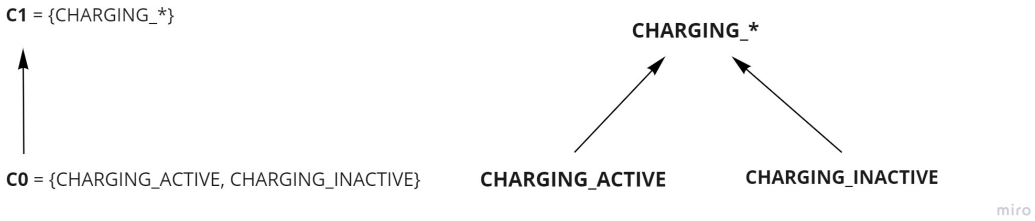

Figure 5.4: Incognito: Domain generalization of the attribute Charging Status.

charging_method	mileage	fuel_percentage	zip_code
CHARGING_*	60000	[50.0-80.0)	808**
CHARGING_*	16345	[50.0-80.0)	808**
CHARGING_*	286354	[20.0-50.0)	808**
CHARGING_*	12908	[20.0-50.0)	808**
CHARGING_*	7897	[50.0-80.0)	808**
CHARGING_*	234231	[20.0-50.0)	808**

Table 5.3: Anonymized car data set using Incognito.

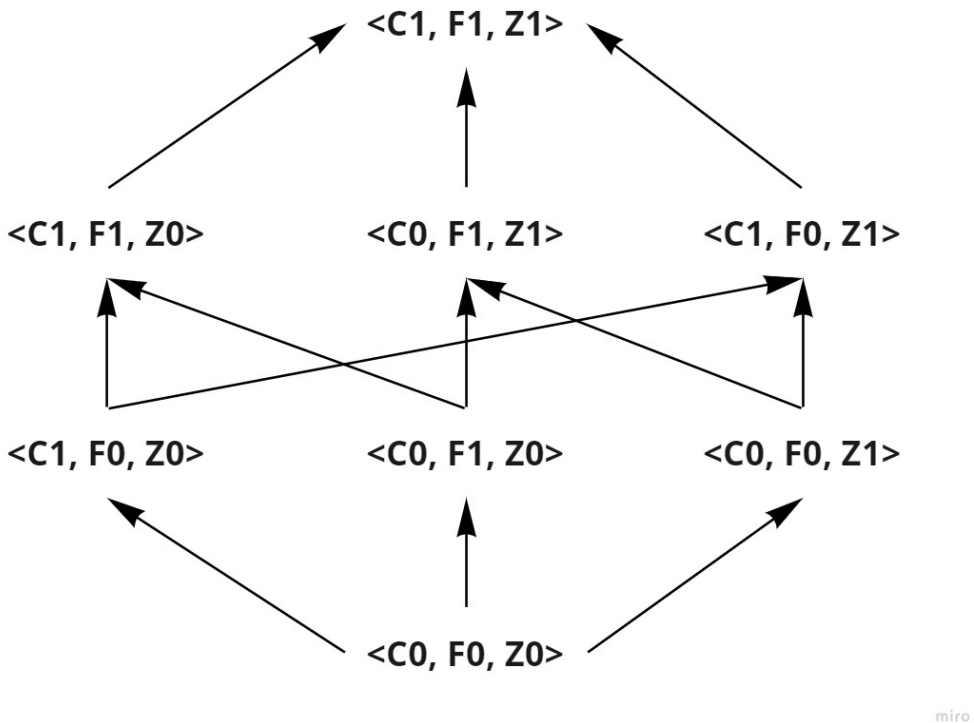

Figure 5.5: Generalization Lattice of the Domain Values of the QIDs.

5.1.3 Datafly

The main Datafly algorithm to achieve generalization and suppression of records uses a three-step process. In the first step, a frequency list is created which holds the unique combinations of the quasi identifier set created in the second step of the Datafly-prior process. Each entry in the frequency list corresponds to one or more records in the original data set. In the second step, using domain generalization defined for each of the quasi-identifiers, the generalization is done. The attribute with the most distinct values is generalized first. This step is run recursively till k or fewer records are having a unique combination of values. In the third step, all records with unique sequences which have a frequency less than k are suppressed. The complexity of the Datafly algorithm is O(N log N). For applying the Datafly algorithm on the example Table 5.1, the generalization hierarchy for the chosen quasi-identifiers are displayed in the Figure 5.6, 5.7. The input metric for the anonymization is as follows: k=3 and

Quasi Identifiers= `fuel_percentage`, `zip_code`. The result after each iteration and the anonymized data set is displayed below.

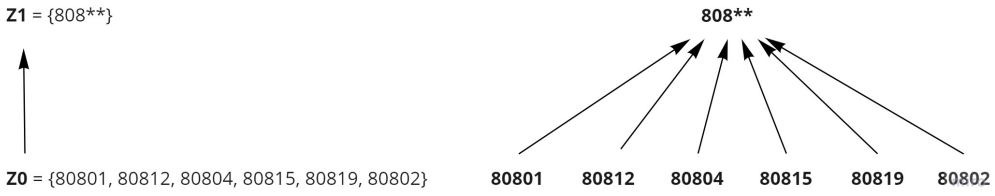

Figure 5.6: Datafly: Domain Generalization of the attribute Zip Code.

Figure 5.7: Datafly: Domain Generalization of the attribute Fuel Percentage.

charging_status	fuel_percentage	zip_code	frequency
CHARGING_INACTIVE	80.0	80801	1
CHARGING_INACTIVE	60.0	80812	1
CHARGING_ACTIVE	30.0	80804	1
CHARGING_ACTIVE	40.0	80815	1
CHARGING_INACTIVE	55.0	80819	1
CHARGING_ACTIVE	20.0	80802	1

Table 5.4: Iteration 1: Anonymization using Datafly.

charging_status	fuel_percentage	zip_code	frequency
CHARGING_INACTIVE	80.0	808**	1
CHARGING_INACTIVE	60.0	808**	1
CHARGING_ACTIVE	30.0	808**	1
CHARGING_ACTIVE	40.0	808**	1
CHARGING_INACTIVE	55.0	808**	1
CHARGING_ACTIVE	20.0	808**	1

Table 5.5: Iteration 2: Anonymization using Datafly.

charging_status	fuel_percentage	zip_code	frequency
CHARGING_INACTIVE	(40.0-80.0]	808**	3
CHARGING_ACTIVE	[20.0-40.0]	808**	3

Table 5.6: Iteration 3: Anonymization using Datafly.

charging_method	mileage	fuel_percentage	zip_code
CHARGING_INACTIVE	60000	(40.0-80.0]	808**
CHARGING_INACTIVE	16345	(40.0-80.0]	808**
CHARGING_ACTIVE	286354	[20.0-40.0]	808**
CHARGING_ACTIVE	12908	[20.0-40.0]	808**
CHARGING_INACTIVE	7897	(40.0-80.0]	808**
CHARGING_ACTIVE	234231	[20.0-40.0]	808**

Table 5.7: Anonymized car data set using Datafly.

5.2 Comparisons

5.2.1 Mondrian vs Incognito vs Datafly

The main reason for choosing to explore the three algorithms - Mondrian, Incognito and Datafly is that these three techniques were the most popular and most cited techniques in the 18 research papers studied on k-anonymity. The next main question is which of these techniques provide the best anonymization. This will mainly depend on the source data set and the system which taps into the properties of these techniques to anonymize the source data set. However, it is difficult to pin point which one

performs the best as there are scenarios where algorithms tend to perform better in some applications and not give the desired results in other scenarios [51].

Rooted from the analysis of the anonymization approach followed by Mondrian, Incognito, and Datafly algorithms, a summary of the findings can be drawn which can be used as a reference in choosing the appropriate algorithm for the requirements. In order to determine the performance of the algorithms with increase in data set size and data utility, the experiment trails done as part of the research paper [47] is referred.

- The run time complexity of Mondrian is O(nlog(n)) since it uses a greedy partition algorithm. The run time complexity of Datafly is also O(N log N). Hence, we can infer that the execution time taken for anonymization by these two algorithms is comparable.
- The run time complexity of the Incognito algorithm is exponential to the number of quasi-identifiers. Hence, the execution time for anonymization increases exponentially as the number of quasi-identifiers increases. Hence, the Incognito works better when the number of quasi-identifiers is small.
- Both Incognito and Datafly algorithm have a prior step that needs to be performed before anonymization. This prior step involves defining the domain generalization for all the quasi-identifiers. Since the generalization hierarchy is set well in advance, the anonymized data set will be compliant with the constraints defined in generalization hierarchies. This interpretation is also provided in the paper [47].
- Mondrian used the concept of median partitioning to perform generalization. Hence, the generalized values of the quasi-identifiers are generated on the go during the anonymization process. However, for categorical values, there is no constraint defined for partitions that are formed. Hence, Mondrian works better for data sets with only numerical values.
- Mondrian algorithm does not work well with quasi-identifiers which have exactly two categorical values. This results in two partitions that have only one of the given two values. Thus, if the number of records with one of the categorical values is less and further partitions cannot be formed, this will result in information loss.
- Incognito and Datafly algorithm work optimal for both uniformly and not so uniformly distributed data. However, due to the median partitioning approach of Mondrian, when the data is not uniformly distributed, optimal partitioning cannot be achieved which results in data loss and consequently, the data utility also reduces.

- The execution time for Incognito and Datafly increases as the height of the generalization hierarchy increases as this results in more search operations that need to be performed. The Samarati algorithm [52] tries to overcome this blocker by using binary search to obtain the solution in less time.
- Both Datafly and Incognito algorithm require the domain generalizations to be defined against all the quasi-identifiers. This implies that the user needs prior knowledge of the domain to formulate the generalization definitions. On contrary, the Mondrian algorithm takes advantage of the partitioning logic to perform generalization.
- Datafly and Mondrian are best-suited algorithms for data sets which large quasi-identifiers.
- Mondrian and Datafly provided promising outcomes with the increase in data size according to the experimental trials conducted in the paper [47].
- Since the Datafly algorithm guarantees to provide an anonymized data set as an outcome, the algorithm skips many nodes resulting in a data set that is more generalized than required resulting in data loss. Sometimes, suppression of all values with the same frequency can happen which reduces the data utility of the anonymized data set [53]. The Samarati algorithm [52] tries to overcome this blocker by providing the constraint of optimal generalization.
- From the experimental trials conducted in the paper [47], on comparing Datafly, Incognito and Mondrian provide better performance in terms of information loss and Mondrian provided better performance when the data was uniformly distributed.

5.2.2 Top Down Specialization vs Bottom-Up Generalization

Top-Down Specialization [41] is an iterative process that traverses through the taxonomy tree of the attributes from the topmost root node. At each iteration, it finds the most generalized values and specializes the value till the k-anonymity is violated.

The Top-Down Specialization(TDS) approach begins with removing all the non-quasi identifiers from the data set. Thus formed data set will contain only the sensitive attributes and the quasi-identifiers. The next step is to define a generalization taxonomy tree for all the quasi-identifiers. The root of this tree will have the most generalized value like ANY and as we traverse down the taxonomy tree, the attribute values get more specific. With each step down the taxonomy tree, the value of the parent node

is specialized into specific values of its child node. This process is repeated till the specialization results in the violation of the defined k-anonymity.

The Bottom-Up Generalization [43] is again an iterative process of generalizing the information. This involves traversing through the taxonomy tree of attributes from the bottom upwards.

The Bottom-Up Generalization(BUG) begins with constructing the taxonomy tree containing the domain values of the quasi-identifiers with the root having the most generalized value and the leaf nodes having the most specific values. As the tree is traversed from the leaf nodes, the k-anonymity and the data loss parameter are calculated at every iteration till the generalization results in the violation of k-anonymity. In the basic BUG approach, the taxonomy tree needs to be defined explicitly for both categorical and numerical attributes. The indexed approach of BUG as proposed by Hoang [54] performs analysis on the data set and defines the taxonomy tree for both numerical and categorical values. The figure 5.8 represents the taxonomy tree for the data set in the Table 5.1 for the quasi-identifiers `charging_method`, `fuel_percentage` and `zip_code`.

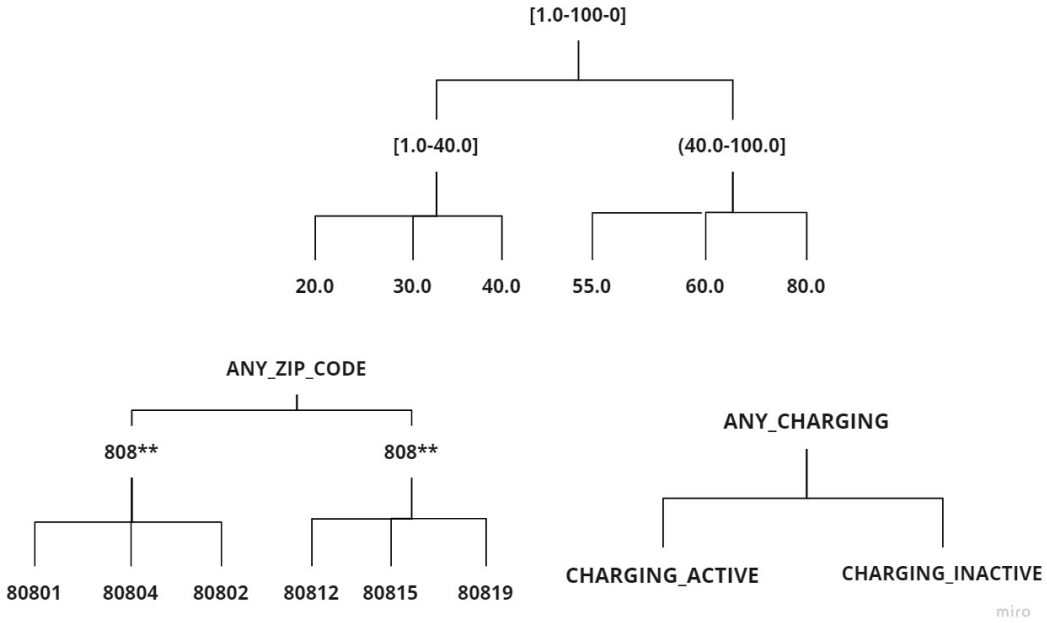

Figure 5.8: Taxonomy tree for the QIDs.

One major aspect which both TDS and BUG do not accommodate is the handling of big data. In order to apply one of these techniques to big data, the data processing needs to be done in parallel. Along with the parallel processing of big data, the technique should also cater to the needs of the constant growth of such large data. For this reason, there is an urge to explore dynamic parallelization algorithms. One such approach was proposed by Zhang [55]. This two-phase TDS approach uses the concept of MapReduce jobs made to run in parallel in each iteration. In the first step, jobs are initialized and data is iteratively specialized while calculating the information loss and anonymity parameter defined. In the second step is intermediate anonymized data sets formed are combined to form the final anonymized data set. On similar lines, Parallel BUG [56] also takes advantage of the MapReduce job to perform the job and task-level parallelization to overcome the shortcomings of basic BUG and indexed BUG approach.

Both TDS and BUG do not cater to the user-specified k-anonymity value. In fact, TDS works better for larger values of k whereas, BUG works better for smaller values of k. Consider the taxonomy tree for attribute `fuel_percentage`. Let the input k parameter be 2 and the total number of records is 6 with 2 records having fuel percentage between 1-10, 2 records with fuel percentage 10-20, and 2 records with fuel percentage 40-80. Using TDS, specialization needs to be done through the taxonomy tree to achieve a 2-anonymous data set. On the contrary, BUG will not perform any traversal as the data set is already 2-anonymous. Hence, a hybrid approach of combining TDS and BUG [57] is preferable which takes the advantages of the TDS and BUG approach.

5.2.3 Centralized vs Distributed Anonymization

A good anonymization solution is difficult to achieve for any data set as seen in Section 5.2.1. Having said that, depending on the environmental constraints one of the proposed solutions can be used according to the guidelines set in Section 5.2.1 and 5.2.2. However, from the big data context set in Section 3.7, one of the main criteria which needs to be considered is how to accommodate any of the anonymization solutions chosen in the big data context.

In recent times big data is all about capturing the information and maximizing its utilization. This can be considered as a boon as well as bane at the same time. With advancements in technology, the effective usage of big data needs to be mapped to an effective anonymization technique in order to acquire desired results. In this context, we have two anonymization approaches. One is the traditional centralized anonymization and the other one is the distributed anonymization. The next section gives a deeper overview of both these approaches and their advantages and disadvantages.

The Centralized anonymization technique can be viewed as a 'integrate then generalize' approach [58]. The main focus of this approach is the generalization of the original data set in its entirety at once. This is a two-step process. In the first step, the data generated in all the source systems are integrated into one single data set. Generalization of the complete data set is done at once in the second step. Any write or update operation on the data set triggers the anonymization step again on the whole data set. One advantage of using this approach is that during the anonymization process, the global view of the original data set is available. This is beneficial as results guarantee anonymization according to the parameters set and the more important aspect is that the data utility remains intact. However, one major downside of using this approach emerges in the use case of big data wherein the anonymization process in itself can be a heavy and expensive computational burden for a single system. The disadvantages of using this approach have led to the exploration and design of local and collaborative anonymization explained in Section 5.2.4.

In order to overcome the shortcomings of the centralized approach, a distributed approach towards generalization needs to be explored. This approach can be seen as a 'distribute, generalize and integrate' approach. This is a three-step process. In the first step, the complete original data set is horizontally partitioned into multiple smaller subsets of the original data. Each of these smaller chunks of data is fed into the nodes of the chosen distributed system. In the next step, the generalization process is instantiated at each of the nodes. The outcome of this step is the multiple anonymized chunks of data. As the last step, these chunks of data are integrated to form the final anonymized data set. One major advantage of this approach is that load on one single system is now distributed among multiple subsystems. On the contrary, an initial effort of setting up the distributed system is added. However, this approach comes with two primary challenges. The first one is that the data utility of the anonymous integrated data set is not guaranteed. The second challenge is that any data leak in the final anonymous integrated data set needs to be prevented. These two challenges can be overcome by setting up adequate sanity checks on the final data set as the last step of anonymization.

Moreover the increased computational power available today can support decentralized models through fast communication channels. Research is still needed to provide a practical implementation of such techniques using the valid use cases. There is a need for the research community to join hands with the big data analytics industry and work together towards achieving a decentralized privacy preserving analytics models. The policymakers need to encourage and promote such efforts, both at research and at practical implementation levels.

5.2.4 Local vs Collaborative Anonymization

Local anonymization tries to overcome the shortcomings of the centralized approach by performing anonymization at the source systems. When the data is generated in the source system, the source system performs basic anonymization and then hands over the intermediate data set to the central system. The central system then integrates the data set from the various source systems and performs generalizations on the integrated data set. There are two main drawbacks to this approach. The first drawback is that there is no common generalization technique that is agreed upon by all the source systems. Hence, to have consistency in the generalized data set which comes from the source systems, generalization performed in the central system will reduce the data utility drastically of the final anonymized data set. The other drawback is that anonymization done at the source system may sometimes negate the anonymization done in the central system, thus revealing the original data. The perfect example for this drawback is when randomized response [59] is used as the anonymization technique.

Collaborative anonymization aims to reduce the information loss from the previous approach by utilizing a mutual consent principle [60]. One of the source systems is chosen as the leader [58]. The job of the leader is to synchronize the anonymization process at the source systems. This reduces the possibility of inconsistent generalization of the data set and increases the data utility as well.

5.3 ARX Data Anonymization Tool

ARX [61] is a cross-platform anonymization tool for analyzing and reducing the uniqueness of records in relationally structured data sets. ARX tool has been listed as one of the standard tools to perform de-identification of data sets in the National Institute of Standards and Technology(NIST) [62]. The Windows Report [63] lists ARX as one of the best free data anonymization tools. A recent report formulated in 2020 by Tech Direct stated ARX as one of the Top 6 data anonymization tools to be used [64] based on its popularity and most wide usage. From the algorithms discussed in Section 5.2.1 and 5.2.2, ARX uses a combination of Incognito, Mondrian, and Top-Down-Specialization approach along with other optimized algorithms [65].

5.3.1 Design Specification of ARX

In the perspective of k-anonymity, ARX performs generalization and suppression by utilizing a globally optimal search algorithm for performing full domain generalization and record suppression of quasi-identifiers of the source data set. The values of the

quasi-identifiers are generalized using domain generalization hierarchies defined for each of the quasi-identifiers. As a pre-requisite step, the generalization hierarchy needs to be defined for each quasi-identifiers. For the example data set defined in the Table 5.1, generalization hierarchy trees for quasi-identifiers `charging_method` and `zip_code` is represented in Figure 5.9 and for quasi identifier `fuel_percentage` is represented in Figure 5.10.

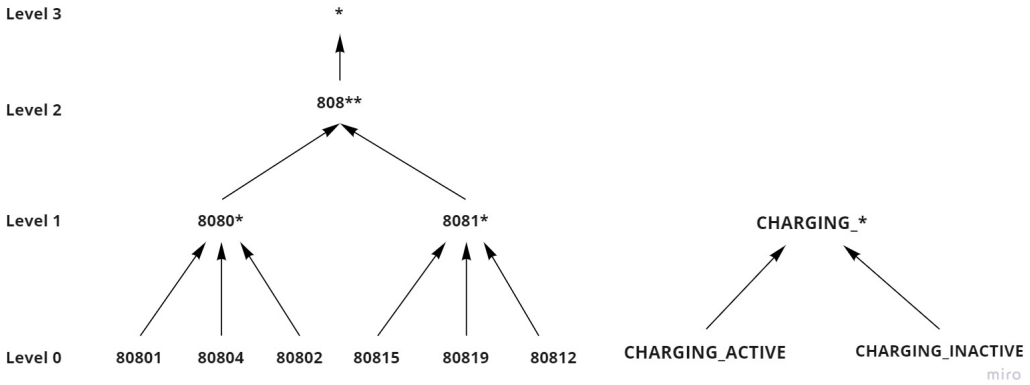

Figure 5.9: ARX: Domain Generalization of the attributes Charging Method and Zip Code.

In the ARX tool, the generalization hierarchies of the quasi-identifiers represented in the Figure 5.9 and 5.10 can be custom defined by the user for each attribute, or the hierarchy builder methods defined in the ARX library can be used to perform dynamic categorization of the domain values. On similar lines with the Incognito approach seen in the Section 5.1.2, ARX also generates a generalization lattice of all attributes depending on their hierarchy heights as represented in Figure 5.11. Two consequent nodes in the generalization lattice differ by exactly one generalization level. The bottom-most node represents the original data set and the topmost node represents the most anonymized data set. In order to convert the data set to a k-anonymous data set, ARX performs three main steps. The first step is to apply a generalization scheme at each node and check if the k value is satisfied. In step two, the outliers in the data set are suppressed. In the final step, the utility of the anonymized data set is calculated. Considering the example car data set represented in Table 5.1 as the original data set to be anonymized using ARX, the chosen quasi-identifiers are `charging_method`, `zip_code` and `fuel_percentage`. The input metric value for k is 3. The 3-anonymous car data set is represented in Table 5.8. From the generalization lattice, the anonymized data set is reached at the node <C1, F2, Z2>.

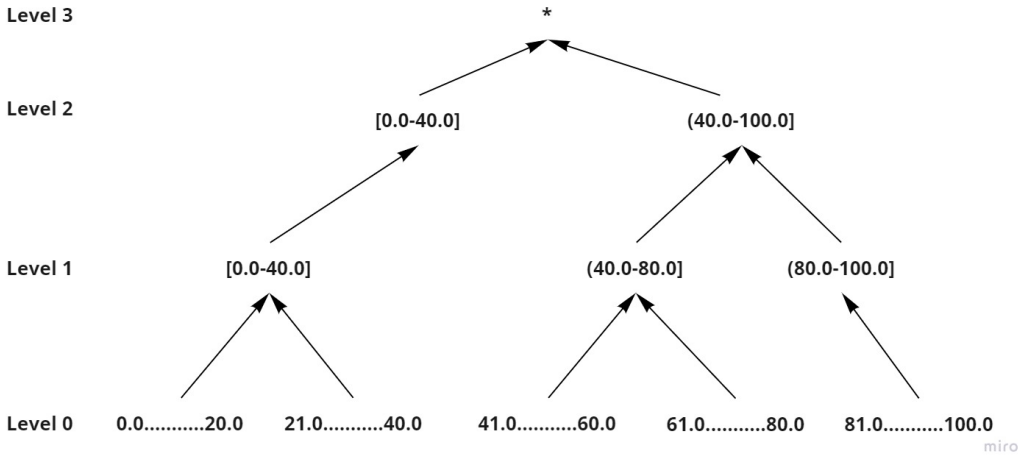

Figure 5.10: ARX: Domain Generalization of the attributes Fuel Percentage.

charging_method	mileage	fuel_percentage	zip_code
CHARGING_*	60000	(40.0-100.0]	808**
CHARGING_*	16345	(40.0-100.0]	808**
CHARGING_*	286354	[0.0-40.0]	808**
CHARGING_*	12908	[0.0-40.0]	808**
CHARGING_*	7897	(40.0-100.0]	808**
CHARGING_*	234231	[0.0-40.0]	808**

Table 5.8: Anonymized car data set using ARX tool.

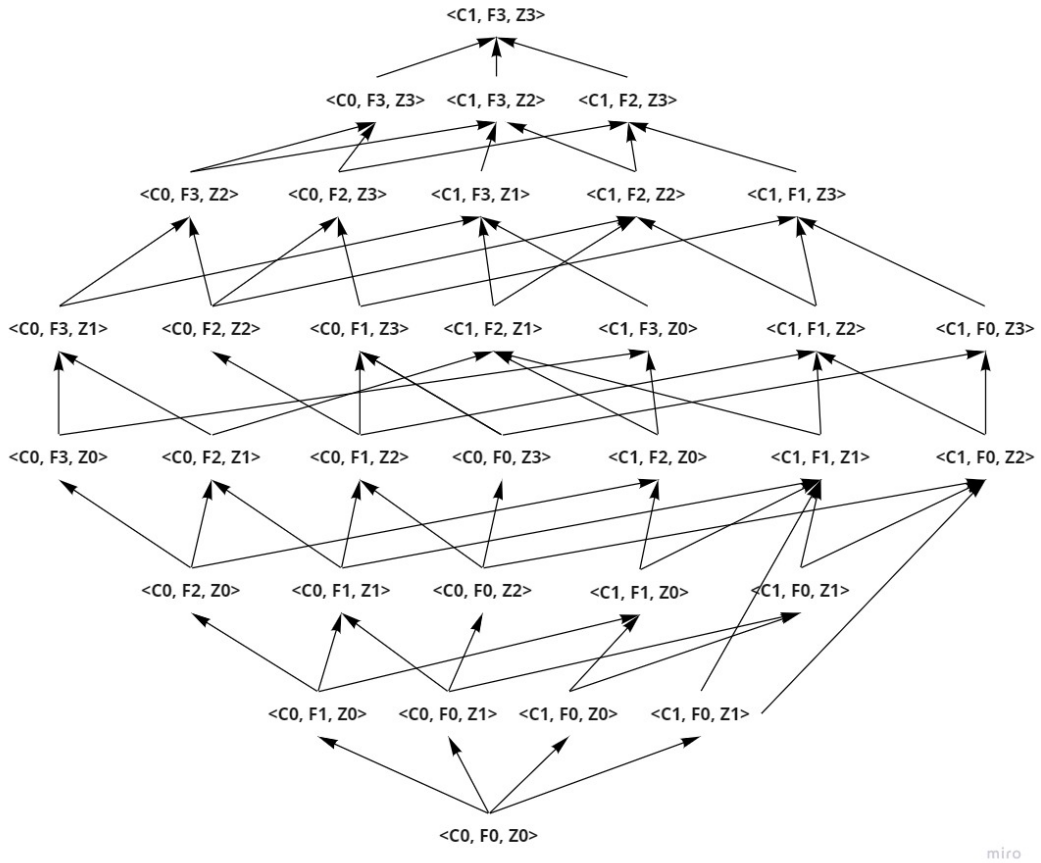

Figure 5.11: ARX: Generalization Lattice for the QIDs.

ARX uses a combination of algorithms to perform the generalization and suppression operations [65]. ARX also uses its own implementation of two algorithms namely Flash [66] and Lighting [67] which provide better performance as the number of attributes in the data set increases. The main advantage of using the ARX tool is that it a proven widely accepted solution for k-anonymization of data sets. Moreover, the also provides support to other privacy models like k-anonymity, t-closeness, l-diversity, and δ-Disclosure privacy. Two primary disadvantages of using the tool are the lack of support to all kinds of data types(for example Geo location) and since the tool uses a centralized approach of anonymization, issues mentioned in Section 5.2.3 exists.

5.3.2 Performance

Performance evaluation of the tool ARX as performed in the paper [61] is summarized in this section. For the evaluation of the performance of multi-dimensional generalization, the well-known Mondrian algorithm [48], as implemented by the open-source UTD Anonymization Toolbox version 33 [68], is used.

Datasets

Six real-world datasets were used: (1) US Census, an excerpt from the 1994 census database, (2) Competition, introduced in the KDD data mining competition in 1998, (3) Crash Statistics, NHTSA crash statistics from their Fatality Analysis Reporting System, (4) Time Use Survey, data from the American Time Use Survey, (5) Health Interviews, results from the Integrated Health Interview Series, and (6) Community Survey, responses to the American Community Survey. The sizes of the datasets on disk range between 2.52MB(US Census) and 107.56MB(Health Interviews).

Evaluation

Figure 5.12 derived from the paper [61] shows the execution times measured when performing multidimensional generalization. As described in the paper [61], ARX provided better performance with increase in the distinguishable attributes in the data set and the UTD Anonymization Toolbox provided a better performance with increase in the input privacy parameters.

In conclusion, ARX provided better performance for multidimensional generalization as the data set size increased, and UTD overall provided better data quality in the anonymized data set.

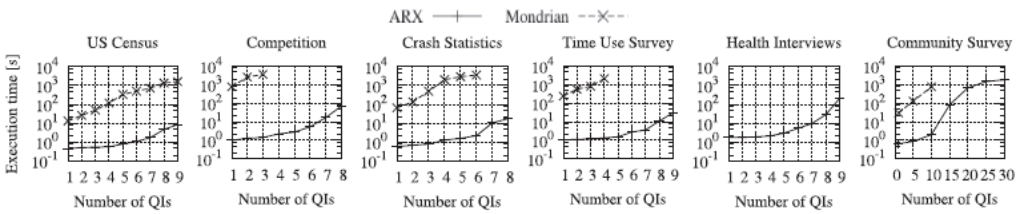

Figure 5.12: ARX: Performance Evaluation as shown in the Paper [61].

6 Use Case of Automotive Industry

This chapter aims to set the basis for the use case in the Automotive industry for connected cars. In particular, two use cases relevant for the implementation part is explored in detail. This is followed by identifying the privacy threats in the use cases and investigating which can be solved by de-identification using k-anonymity.

6.1 General Idea of Data Analytics

With the technological advancements in the current world, the amount of data being generated by any entity has increased exponentially [5]. Capturing such generated data and maximizing its utility is the direction in which all companies are targeting to reach. This abundance of information is power as well as a problem at the same time. One such problem occurs when analytics performed on the data potentially reveal sensitive information of the entity in interest, which is a huge hindrance to the privacy of the entity [3]. Hence, in this new era of analytics, it is important to map the effectiveness of any analytics performed to the secure access and processing of the information focused by the analytics.

Analytics are already happening in our day-to-day activities. In hospitals, patients' current health conditions are compared against the historical data to identify patterns that could potentially provide information on early detection and mitigation of diseases [1]. Mobile fitness apps collect data like walking and sleeping patterns along with other relevant data, which can be combined with the data captured by the hospitals to provide better robust health care programs to the user [69]. One interesting example is the use case of Smart TVs where they record choices of the user can provide suggestions accordingly [70]. Some of the social media apps as well display ads in which the user could be potentially interested in captured data of their activities on those apps [71]. From these examples, it is evident that a variety of collaborators are involved in constructing this data value chain which includes the hardware, software, operating system providers, service providers, and so on.

The data value chain of the data generated in the automotive industry will be explored in detail in the next section. The analytics performed on the data generated in the cars

will be explored, along with the roles of the stakeholders and their interactions in the data value chain.

6.2 Automotive Industry and Car data Analytics

The Automotive industry focuses on the large-scale production of automobiles and motor vehicles. The origin of this industry started with Carl Benz [72] on his invention of the first practical automobile. From then on, the automotive industry has undergone significant changes. The underlying technology has also experienced immense alternations with radical technology innovations. The current market is diversified with the increasing popularity of all kinds of vehicles, from non-renewable fuelled vehicles to fully electric vehicles along with hybrid options also available [73]. The data-gathering tools within the car can be as simple as sensors and cameras to record the behavior of the automobile system to sophisticated GPS and satellite-connected devices [74]. The main focus of these collected data is to improve the customer experience along with welfare and to increase the longevity of the automobile. The easy availability of the internet everywhere has enabled automobiles to feed critical information to various business models and has helped companies to better understand the demand-supply of products and services [75]. Using these insights, a personalized experience can be provided to the end-user. Thus, a full-fledged ecological community is built around the generation, processing, and usage of data.

Among such automobile systems, the topic of interest is the connected car. A connected car is an automobile with internet and Wi-Fi access, allowing bi-directional sharing of data with all the devices within the car as well as to the outside world. It is estimated that by 2022, around 700 million connected cars will be on the road [76]. According to a recent survey [6], these connected cars generate up to 25GB of data per hour. For entities within the car, information like the route to be taken, incoming traffic, weather predictions, and other relevant information is forecast-ed. Similarly, for the outside world, the maintenance workshop company, for example, will be provided with the status of the embodied vehicle parts which need to be serviced to determine the car service time [34]. All in all, the analytics performed on such data increases the longevity and reliability of the car as well as increases the ease of use for the user.

Car data analytics can make way to a whole new area of opportunities of service provision in varied aspects of activities taking place in the environment of the connected car. Nevertheless, while such progress is encouraged and welcomed in many cases, they often pose serious challenges to the privacy and protection of the personal data of the participating entities.

6.3 Use Case for Connected Cars

As described in Section 6.2, connected cars generate data. These data can be classified into two categories. The first category is the usage-based data like mileage. This data varies as the connected car is used over a period of time. The second category is event-based data, where the data is captured based on an action performed on the car, which triggers an event [77]. The charging status of the car is one such data where the status is captured and altered depending on the action of whether the car is being charged or not. Automobile companies capture such car data of every user of the car. Having the consent of the user, this data can be made accessible to corporate third-party service providers. Post interviews conducted at the European OEM, two primary use cases that could potentially convert this captured car data into beneficial services to the user were narrowed down. The third-party service providers who would be interested in developing these use cases into potential business models are described in the next two subsections.

6.3.1 Smart-Charging Stations

An Electric Vehicle(EV) charging station, commonly knows as an EV charging station, is an appliance that supplies the required electric energy fuel for the plugged-in electric connected cars. These charging stations can be categorized into three types [78]. The first one is the residential charging stations where the user plugs into the charging device in the vicinity of their home. No user authentication or payment is necessary in these cases. The second category is the charging stations at the parking lots. These can be free or paid services in collaboration with the owners of the parking lot. The third category involves charging stations at the rest stops, which can be used regularly by the commuters in those areas. The focus for the use case in discussion falls into the second and the third category types of charging stations. Another important parameter that needs to be considered for this use case is the charging plug types in the connected cars. In Europe, there are two types of standard plug types - type 1 and type 2 [79]. The EVs with a type 2 plug can be charged in any charging station which has a permanently connected charging cable. However, an EV with a type 1 plug needs the right adaptor for charging, and this is not commonly found in all charging stations [76].

The goal of this use case is to find out the locations where the users are most likely to run out of the charge in the EV. From the captured car data, the users most traversed locations are jotted down. From these locations, the location where most users have run out of charge in their EVs is found. Such locations will be useful for the third-party service providers looking into setting up charging stations to get an idea of the most beneficial location of placing such stations. For this purpose, the car data generated by

the car, which is captured by the European OEM, needs to be shared with the service providers. However, prior to sharing the car data with the service providers, the data needs to be anonymized, so that identity of the user or the location of the car is not revealed in the shared dataset.

6.3.2 Smart-Billboard Advertisement

Billboard advertising [80] is a strategy of using large-scale prints to advertise a product or service provided by a company. Sometimes, these are also used to spread awareness among the general crowd. These billboards are typically placed on highways, expressways, and primary avenues where one can expect more traffic such that it is accessible to a large number of pedestrians and drivers [81]. The billboards being huge and eye-catching forces people to look at the content. This results in an effective way of advertising as it reaches many people and thus tends to have the highest number of views and impressions when compared to other forms of marketing. According to a survey [82], 71% of consumers often view and process the information displayed on the billboards. Also, billboard marketing costs 80% less when compared to television marketing [83]. Based on the revenue, the top four companies which continue to invest in billboard marketing are Apple, Google, Amazon, and Netflix [84]. The location of placing these billboards is an important factor that reflects its effectiveness. Taking advantage of the environment where the billboard is placed like the primary features of the area itself like sports team and nuance has proven to increase the outcome of the advertisement strategy.

The goal of this use case is to achieve vehicle marketing using the car data collected to strategically place billboards relevant to the users. This further helps in campaign management, and relevant marketing tactics can be formulated as well [81]. From the captured car data, the users most traversed locations are noted down. Along with this, the features of the car driven by the user are also noted down. From the curated data on the location, features of the car, and the driving patterns of the drivers, insights can be extracted by the service providers. Such insights can be used to place billboards with relevant content enabling them to boost sales. This is beneficial on the users' end, as well as they will be exposed to personalized content. For this purpose, the car data generated by the car, which is captured by the European OEM, needs to be shared with the service providers. However, prior to sharing the car data with the service providers, the data needs to be anonymized, so that identity of the user or the location of the car is not revealed in the shared dataset.

6.4 Identification of Privacy Threats

This section aims to identify the privacy threats for the use case of connected cars. An analysis is provided whether the privacy threats can be mitigated by using k-anonymization techniques. In order to elicit privacy threats and model a mitigation plan for these privacy threats, the LINDDUN framework is used, which is explained in the next section.

6.4.1 LINDDUN Framework

LINDDUN [85] is a privacy threat modeling framework that provides assistance for systematic elicitation and mitigation of privacy threats in software systems. It serves as a handbook to guide users to perform the threat modeling process in an organized manner. The LINDDUN methodology has three primary steps. The first step is to model the system where analysis of the system in focus is done, and a Data Flow Diagram(DFD) is formulated. The second step is to elicit threats where the DFD components are iterated recursively to identify the privacy threats. The final step is to find solutions for the threats uncovered in the second step. The term LINDDUN is a combination of the letters where each letter specifies the privacy threats [86]. Table 6.1 displays the privacy threats along with their definitions.

System Modelling

Post re-iterating the use cases described in Section 6.2, a graphical representation of the use cases for connected cars in the form of DFD is formulated as shown in the Figure 6.1. The DFD [85] is a composition of four primary building blocks: entities (i.e. users or external service providers), processes (i.e. core functionality), data flows (i.e. representing the flow of information through the system) and data store (i.e. data storage modules). These are represented in the legend of the Figure 6.1. A brief description of the elements of the DFD represented in the Figure 6.1 is as follows:

- **User (entity):** The user describes an individual using the connected car.
- **Connected Car (entity):** The connected car is a physical device capturing driving patterns data. The car refers to the entity whose identity is aimed to be protected.
- **Platform and Services (process):** The platform and services represent the software system where the captured connected car data is aggregated and analyzed.
- **Database (data store):** The database stores all the data captured by the car in a relational or non-relational database.

Privacy threats	Definitions
Linkability	A combatant has the ability to link two entities of interest without the knowledge of the identity of the data subjects involved [87].
Identifiability	An adversary has the ability to identify an entity set of entities using an item of interest [87].
Non-Repudiation	An entity is unable to decline a claim [86]
Detectability	A combatant has the ability to differentiate if the item in focus of an entity exists or not irrespective of the fact of being able to read the contents [87].
Disclosure of Information	An adversary has the ability to read the contents of the item in focus belonging to the entity [87].
Unawareness	An entity is unaware of the capturing, processing and sharing of their personal data [86]
Non-Compliance	The process of capturing, processing and sharing of personal data does not comply with the regulations [85].

Table 6.1: LINDDUN: Privacy threats

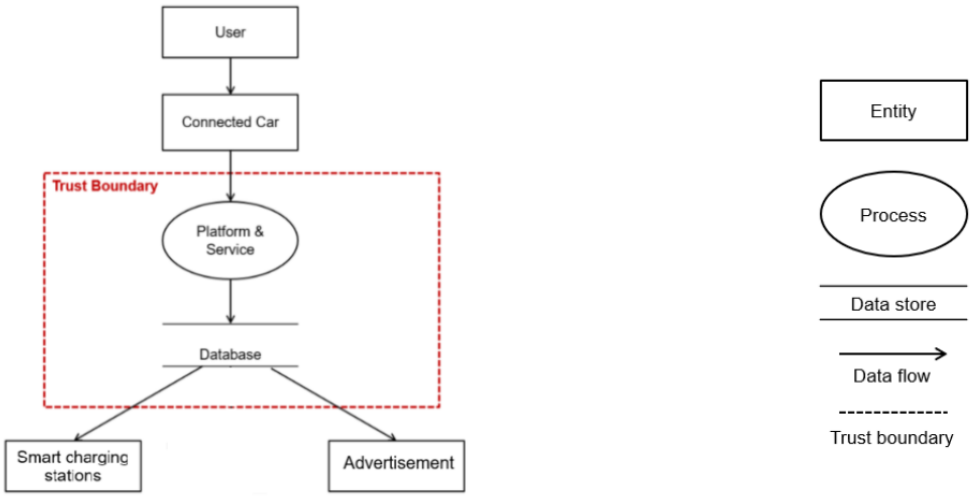

Figure 6.1: LINDDUN: DFD for the Use Case of Connected Cars.

- **User data stream (data flow):** This data flow represents the flow of data from the user to the connected car(e.g. music preference of the user captured by the audio system in the connected car).
- **Connected car data stream (data flow):** This data flow represents the flow of data from the connected car to the platform where the user can view this data in the dashboard provided to them.
- **Database stream (data flow):** This data flow represents the flow of data from the platform to a database where raw and aggregated data is stored.
- **Third party data stream (data flow):** This data flow represents the flow of data from the database to external service providers to enhance their business models.
- **Service Providers (entity):** The service providers describe the external third-party vendors interested in data generated by connected cars.

Elicitation of threats

Once the use case is modeled in the form of DFD, the next step is to map the DFD to threat categories. The Table 6.2 shows the mapping of the DFD components of the connected car use case. The 'X' in the table highlights potential threat in the system.

	Threat target	L	I	N	D	D	U	N
Data Store	Database	X	X	X	X	X		X
Data Flow	User data stream	X	X	X	X	X		X
	Connected car data stream	X	X	X	X	X		X
	Database stream	X	X	X	X	X		X
	Third party data stream	X	X	X	X	X		X
Process	Platform and Services	X	X	X	X	X		X
Entity	User	X	X				X	
	Connected Car	X	X				X	
	Service providers	X	X				X	

Table 6.2: Threat mapping for DFD elements.

Threat management

The final phase is to make a mitigation plan to tackle each of the privacy threats found. The focus for the use case is to find out which of the threats mentioned in the Section

6.4.1 can be mitigated by the usage of k-anonymity as a de-identification method. Table 6.3 shows threats which can be directly or indirectly influenced by k-anonymization. The threat targets which are marked X indicates that these threats can be mitigated through k-anonymization. The threat targets which are marked X indicates that the impact of these threats are reduced significantly with the usage of k-anonymization. 'X' represents mitigation is not possible using k-anonymization.

	Threat target	L	I	N	D	D	U	N
Data Store	Database	X	X	X	X	X		X
Data Flow	User data stream	X	X	X	X	X		X
	Connected car data stream	X	X	X	X	X		X
	Database stream	X	X	X	X	X		X
	Third party data stream	X	X	X	X	X		X
Process	Platform and Services	X	X	X	X	X		X
Entity	User	X	X				X	
	Connected Car	X	X				X	
	Service providers	X	X				X	

Table 6.3: Threat mitigation using k-anonymization.

7 Preparation Phase

This section defines the attributes of the data set generated by the connected cars at the European OEM, where interviews were held to understand the properties of the data set. This is followed by understanding the concept of synthetic data sets and their properties. Finally, the strategy adopted to formulate the car data set is defined.

7.1 Properties of Car data Set

Post interviews conducted at the European OEM, data generated by a connected car was analyzed. From the data generated, the following attributes were found to be relevant for the use case described in the Section 6. The details of these attributes are as follows:

- **Car Identifier** is the unique identifier represented as a string of numbers of letters that identifies each connected car uniquely in the car data set.
- **Car Model** is the name used by the automobile company to market similar cars. The categorization of cars based on models varies across different automobile companies.
- **Charging Method** of a connected electronic car can be two ways depending on the standard plug types - type 1 and type 2.
- **Charging Status** indicates whether the connected car is getting charging actively or not.
- **Smart Charging Status** indicates if the EV is renewable optimized to promote clean transport and low-carbon trace. This status is set by the connected car depending on the users' preference and conditions of the power system.
- **Fuel Percentage** shows the fuel percentage of the connected car.
- **Mileage** displays the distance traveled by the connected car.
- **Time Stamp** is the digital record of the time of occurrence when the rest of the attributes in the data set was recorded.

- **Latitude** notes down the latitude captured by the automotive navigation system of the connected car using Global Positioning System(GPS) sensors.
- **Longitude** notes down the longitude captured by the automotive navigation system of the connected car using GPS sensors.
- **External Temperature** indicates the temperature of the environment in which the connected car is present.

7.2 Mapping Car Data Set with Use Cases

The two primary use cases: Smart charging stations and Smart advertisement billboard placement focused as part of the thesis, are already described in the Section 6.3.1 and Section 6.3.2. This section focuses on the mapping of the defined use cases with the properties from the car data set. The subset of the properties of the car data set will provide insights to the external service providers to understand the demand-supply of products and services. Using these insights, a personalized experience can be provided to the end-user. As described in Section 6.4, the privacy threats from sharing this car data set to external entities need to be mitigated by applying the de-identification technique: k-anonymization to preserve the identity of the user as well the connected car. Table 7.2 describes the use case, the corresponding focus and threat related to the use case, and lastly, the properties from the car data set relevant for the use case.

	Smart Charging stations
Focus	Smart placement of charging stations based on users driving patterns like the most frequent places where the car is likely to run out of charge.
Threat	The optimal location of the smart charging station is derived without revealing the location or identity of the car or the user.
Attributes	Car Identifier, Car model, Charging Method, Smart Charging Status, Fuel Percentage, Time Stamp, Latitude and Longitude.

Table 7.1: Use case of smart placement of charging stations.

	Smart billboard advertisement
Focus	Smart placement of billboards based on users driving patterns like frequent traversed location and more stoppage time.
Threat	The most traversed path is extracted from dataset without revealing location or identity of car or the user.
Attributes	Car Identifier, Car model, Fuel Percentage, Time Stamp, Latitude and Longitude.

Table 7.2: Use case of smart billboard advertisement.

7.3 Formulation of Car Data Set

The synthetic data set is data that is usually artificially created rather than being generated by actual events. There are two primary reasons for generating a synthetic data set. The first reason is when the privacy requirements limit the data availability. The second reason is when the data needed for testing a functionality either does not exist or available for usage. As part of the thesis, synthetic car data set is generated using the properties of the data set observed at the European OEM. The created data set similar to the car data generated at the European OEM in the following lines listed below.

- Size of the data set varies from 100,000 to 15 million records.
- Journeys of up to 10 to 30 cars are synthesized.
- Car identifiers(`car_id`) is generated using custom Universally Unique Identifier (UUID) generator.
- Car Model(`car_model`) consists of all electric cars with model series from the production year 2019 to present.
- Charging Method(`charging_method`) was chosen to be `AC_TYPE1PLUG` or `AC_TYPE2PLUG` depending on the standard plug type used by the car model chosen.
- Charging Status(`charging_status`) can be `CHARGING_ACTIVE` or `CHARGING_INACTIVE`. Once the fuel percentage of the car reaches 0, charging status is changed from `CHARGING_INACTIVE` to `CHARGING_ACTIVE`.
- Smart Charging Status(`smart_charging_status`) can be `RENEWABLE_OPTIMIZED` or `RENEWABLE_UNOPTIMIZED`. One of these values is assigned to the record based on the car model selected.

- Fuel Percentage(`fuel_percentage`) of every car starts with 100% when the car journey is instantiated. This is logically reduced as the car journey progresses.
- Once the fuel percentage of the car reaches 0%, the car is assumed to be charged completely till fuel percentage is 100% and remaining charging time is 0 minutes.
- Initial mileage(`mileage`) of the car is randomly generated and incremented periodically.
- Initial timestamp(`isc_timestamp`) is the randomly generated and incremented periodically.
- GPS Latitude(`gps_lat`) and GPS Longitude(`gps_long`) is captured every 5 miles.
- The temperature(`temperature_external`) ranges from -10 to 40 degree Celsius.

The data set structure for applying k-anonymity on the generated car data set is represented in Table 7.3. This structure shows the name of the attribute and the unit in which the attribute is measured. Along with this, attributes are classified as Personal Identifiers, Quasi Identifiers, and Insensitive attributes. Finally, an example attribute value is laid out in the table, which gives an idea of the value generated by the generated data set.
The Java-based spring boot application is implemented for the car data set generation. This car generator is exposed as a POST service that can be used for data set generation. The code corresponding to the data set generator is documented in the GitHub repository [88].

Attribute Name	Attribute value	Attribute unit	Attribute type
car_id	B1	NaN	Personal Identifier
car_model	5 Series	NaN	Insensitive Attribute
charging_method	AC_TYPE1PLUG	NaN	Insensitive Attribute
charging_status	CHARGING_ACTIVE	NaN	Quasi Identifier
smart_charging_status	RENEWABLE_OPTIMIZED	NaN	Insensitive Attribute
fuel_percentage	72.5	NaN	Quasi Identifier
mileage	22400	miles	Insensitive Attribute
isc_timestamp	2018-02-16 03:49:56	NaN	Quasi Identifier
gps_lat	3.2485078566258347	WGS84	Quasi Identifier
gps_long	5.583555183545527	WGS84	Quasi Identifier
temperature_external	16	degreescelsius	Insensitive Attribute

Table 7.3: Data set Structure(World Geodetic System 1984-WGS84).

8 Implementation Phase

This chapter gives a detailed explanation of the system architecture chosen for the anonymization of the connected car data set in the context of the big data set. Alongside, the data processing approach for partitioning the data sets in Apache Spark is described in detail. Lastly, the applying k-anonymity techniques to the connected car data sets using ARX API is demonstrated.

8.1 System Architecture Overview

The connected car data set generated in the Preparation phase described in Section 7 needs to be anonymized before sharing it with the external service providers. For applying the techniques of k-anonymity to the data set, the ARX tool is chosen from the analysis done as part of the literature research conducted, which is described in detail in Section 5. The ARX tool is not only a proven solution for anonymizing data sets but also is listed as one of the top anonymization tools [64]. The ARX tool is exposed as a Java library that can be plugged in as an external dependency to any Java-based project, and the methods from the library can be used for performing required anonymization.

Since the solution should cater to the needs of big data sets as well, Apache Spark was chosen for data processing. This framework performs the task of processing large data sets and also can distribute the data processing across multiple core engines. Apache Spark uses a master-slave architecture consisting of two primary daemons. The Master daemon is the Driver program, and the Slave daemon is the worker node processes. Resilient Distributed Data sets (RDD) which hold the collection of data items in Spark, are used extensively to split the car data set into partitions and stored in-memory on the worker nodes available in the Spark Cluster.

Spark Driver is the first point of entry in Apache Spark. This consists of the Driver program, which runs the main() function of the application, which in turn instantiates the Spark Context. The Executors or the also known as the worker nodes, are responsible for the execution of the tasks. These executors perform data processing. The cluster manager is responsible for allocating the available resources to spark job defined. The Executor Service is used to perform anonymization of the partitioned data set

asynchronously. The Rest APIs to generate data set and make the data set k-anonymous are documented in the Github Repository [89]. Figure 8.1 represents the software architecture diagram of the application.

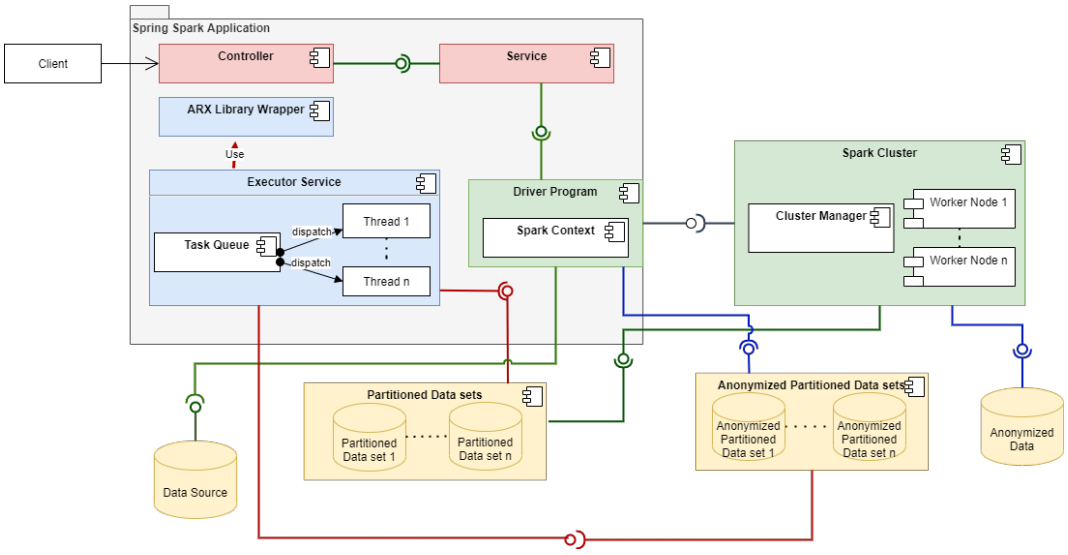

Figure 8.1: Software Architecture Diagram for the Use Case of Connected Cars.

The detailed step by step process overview of the connected car data set anonymization is as follows:

1. Data Generator Rest API is called to start the synthetic data set generation process.
2. Car data set this created data is stored in the Data Source.
3. Client makes a request to start the anonymization process.
4. The Driver program will implicitly convert the code containing the read functionality of the data set from the source database, the partition logic of the data set into a logical directed acyclic graph(DAG).
5. Default optimizations are performed by the driver program, and then logical DAG is converted to a physical execution plan with a set of stages.
6. From the execution plan, tasks are created. Data partition task is sent to the Spark cluster.

7. Driver program communicated to the cluster manager about the resources needed for the tasks to be performed.
8. Cluster manager launches executors on the worker nodes.
9. Depending on the instructions in the task for data partition cluster manager reads data from the source database and distributes the data in a specified number of worker nodes.
10. Data set is partitioned into smaller chunks of data by each of the worker nodes.
11. This partitioned data set is stored in an intermediate partitioned data source for anonymization.
12. The Executor service construct taps into the partitioned data set and maps each partition to thread. In case the number of partitions is more than the number of threads, a blocking queue is maintained by the Executor service in a FIFO(First In First Out) manner.
13. The anonymization of the partitioned data set on each thread is performed asynchronously in parallel using the functions provided by the ARX library.
14. Once the anonymization is complete on all threads; the executor service is shutdown.
15. The Spark context is instantiated again to merge the anonymized partitioned data set to a final anonymized data set.
16. Anonymized dataset is stored into the anonymized data source.

8.2 K-Anonymity Implementation Overview

As explained in Section 5.3, one of the pre-requisite steps before applying the k-anonymization function of the data set using ARX is to set up the configurations required to perform the functionality itself. The configuration definition is a three-step process.

1. The first step is to define the types of attributes. The attributes in the data set need to be mapped to one of the following attribute types defined by the ARX library:
    - `IDENTIFYING_ATTRIBUTE` which indicates that the attribute is a Personal Identifier(refer Section 3.4).

- QUASI_IDENTIFYING_ATTRIBUTE which indicates that the attribute is a Quasi Identifier.
- INSENSITIVE_ATTRIBUTE indicates that the attribute is not prone to any re-identification techniques.

2. The second step is to define the generalization hierarchies to each of the attributes marked as personal as well as quasi-identifiers. Once the generalization hierarchy is defined, the definition needs to be mapped to the respective attributes.
3. The last step is to define the privacy model, which is k-anonymity, and set the suppression limit. The suppression limit is defined in percentages of the complete data set.

For the use case of connected cars, the Table 8.1 shows the attribute type definition performed during implementation.

Attribute	Attribute Type
car_id	IDENTIFYING_ATTRIBUTE
car_model	INSENSITIVE_ATTRIBUTE
charging_method	INSENSITIVE_ATTRIBUTE
charging_status	QUASI_IDENTIFYING_ATTRIBUTE
smart_charging_status	INSENSITIVE_ATTRIBUTE
remaining_charging_time	QUASI_IDENTIFYING_ATTRIBUTE
fuel_percentage	QUASI_IDENTIFYING_ATTRIBUTE
mileage	INSENSITIVE_ATTRIBUTE
isc_timestamp	QUASI_IDENTIFYING_ATTRIBUTE
gps_lat	QUASI_IDENTIFYING_ATTRIBUTE
gps_long	QUASI_IDENTIFYING_ATTRIBUTE
temperature_external	INSENSITIVE_ATTRIBUTE

Table 8.1: ARX: Attribute Type Definition.

The generalization hierarchy for each of the quasi identifiers are defined as follows:

- charging_status attribute can have two possible values: CHARGING_ACTIVE or CHARGING_INACTIVE which will be transformed to CHARGING_*. The Figure 8.2 displays the hierarchy defined for the attribute charging_status.
- fuel_percentage attribute will be transformed to ranged based values like very low, low, normal, full, very full. The Figure 8.3 displays the hierarchy defined for the attribute fuel_percentage.

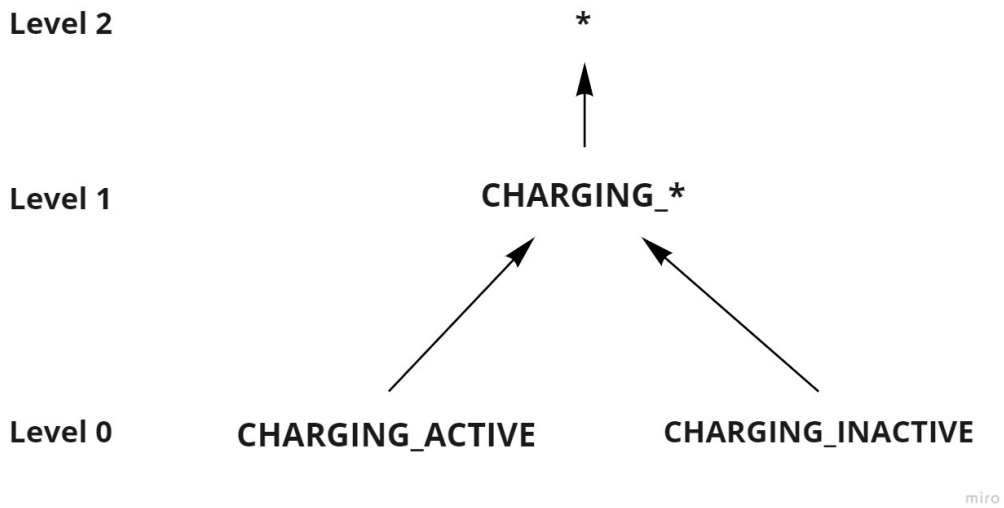

Figure 8.2: ARX: Generalization Hierarchy of the attribute Charging Status.

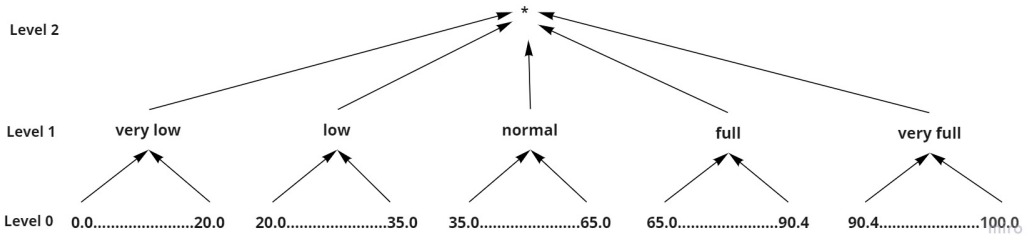

Figure 8.3: ARX: Generalization Hierarchy of the attribute Fuel Percentage.

- `isc_timestamp` attribute will be transformed to the following granular values: hours-day-month-year, day-month-year and year of the captured timestamp. The Figure 8.4 displays the hierarchy defined for the attribute `isc_timestamp`.

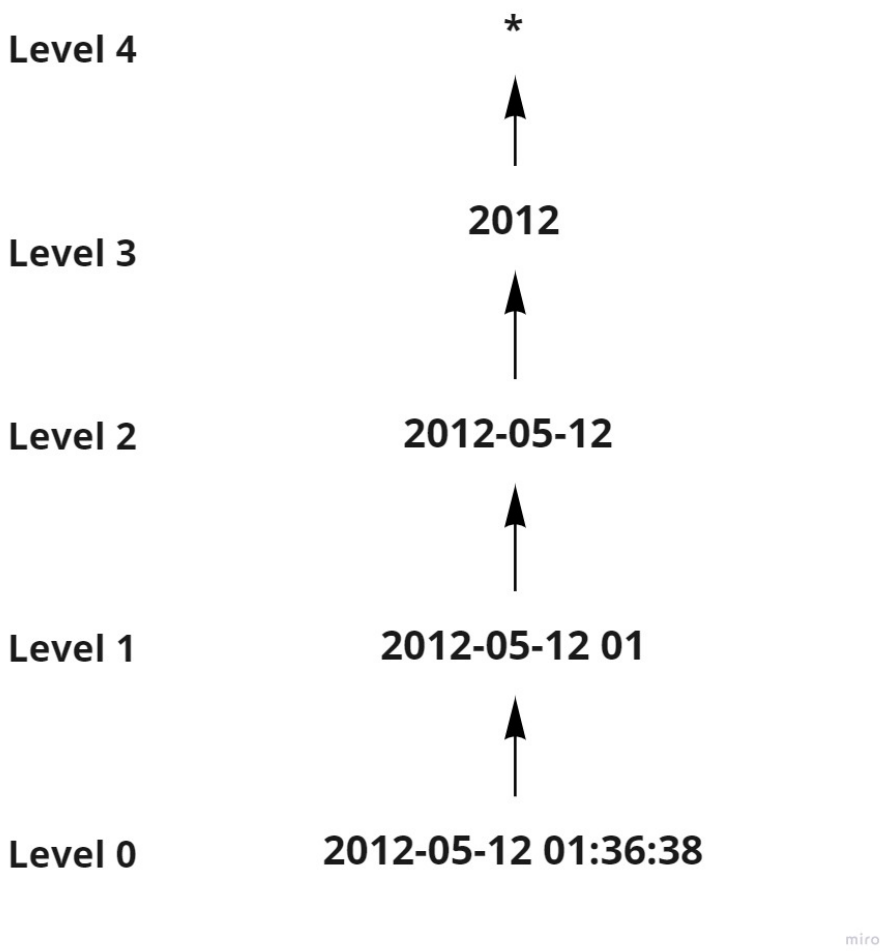

Figure 8.4: ARX: Generalization Hierarchy of the attribute Timestamp.

- `gps_lat` and `gps_long` will be transformed to range of GPS latitude and longitude values. The range of GPS latitude and longitude consists of all points which fall within the 2.5 miles circumference from the captured point. The Figure 8.5 displays the hierarchy defined for the attribute `gps_lat` and `gps_long`.

The last step is to provide a k-value to generate k-anonymous data set and suppression

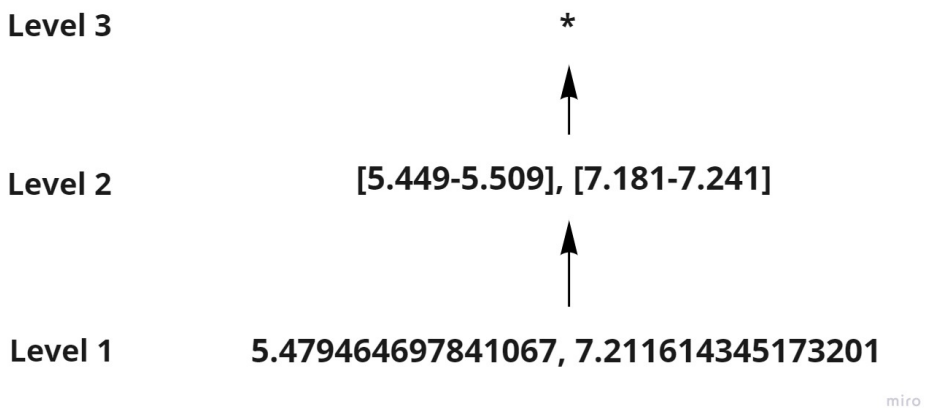

Figure 8.5: ARX: Generalization hierarchy of the attributes GPS Latitude and GPS Longitude.

limit of the records, which could be suppressed as part of the anonymization algorithm in percentage.

9 Evaluation Phase

This chapter captures the results of applying k-anonymization techniques against various parameters and illustrates the performance analysis of the prototype.

9.1 Experimental Setup

A number of tests are performed by varying the k-values against the same data set and also data sets with a varying number of records. The evaluation data set consists of records up to 15 million records. All the experiments were performed using Java 11.0 with ARX library - libarx v1 on Windows 10 Education edition with Intel Core i7-8550U processor with 16GB of installed RAM.

9.2 Evaluation of K-Anonymity

In order to evaluate the performance of the prototype, six different variations in the setup were made. This is explained in detail in the upcoming sections. The five primary parameters which varied across the experiment variations are as follows:

1. **Quasi Identifiers:** K-anonymity is applied to data sets consisting of 1 to 5 quasi-identifiers. The experiment is varied to check the performance based on types of quasi-identifiers.
2. **Number of Records:** K-anonymity is applied to data sets consisting of records up to 15 million records.
3. **Number Partitions:** K-anonymity is applied to data set with 1 million records, and the number of partitions is varying from 1 to 50.
4. **Generalization Height:** K-anonymity is applied to the data set with the generalization height of quasi-identifiers varying from 1 to 6.

9.2.1 Evaluation based on varied Types of Quasi-Identifiers

The different types of quasi identifiers chosen for this evaluation are: `charging_status`, `fuel_percentage`, `isc_timestamp`, `gps_lat`and `gps_long`. These are represented as

1, 2, 3, 4 and 5 respectively in the Figure 9.1. The execution time for the QID `charging_status` and `fuel_percentage` are very close to each other as they share the same generalization hierarchy height(1). The execution time of QID `isc_timestamp` increases exponentially as the generalization hierarchy height is 4. The execution time for QID `gps_long` and `gps_lat` is more as more computation needs to be performed to generalize this values.

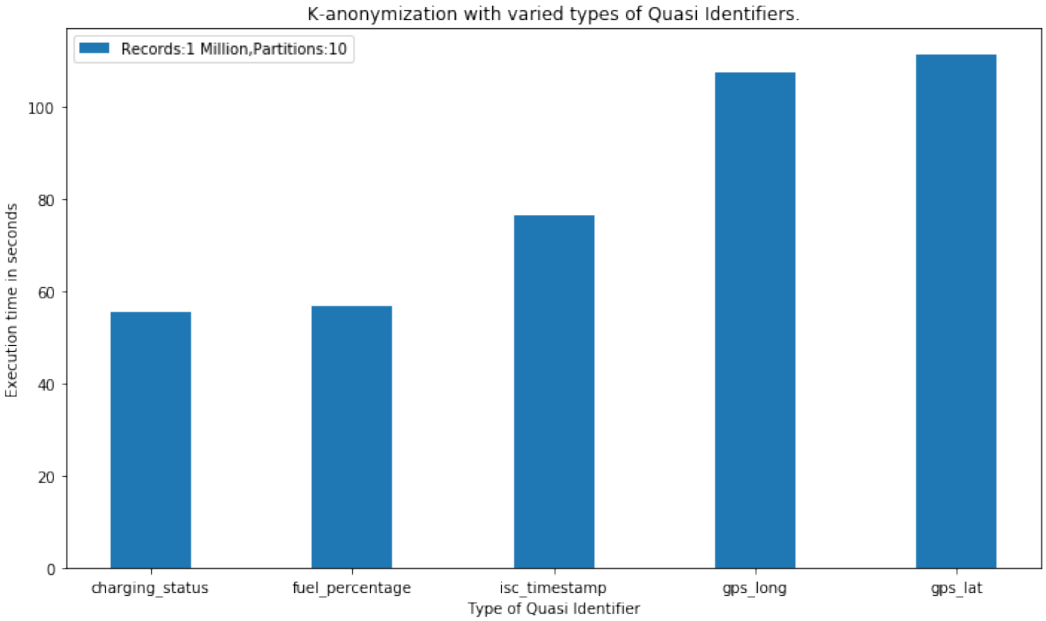

Figure 9.1: Evaluation based on varied Types of Quasi-Identifiers.

9.2.2 Evaluation based on varied Number of Quasi-Identifiers

The quasi-identifiers chosen for this evaluation are: `charging_status`, `fuel_percentage`, `isc_timestamp`, `gps_latand` `gps_long`. The experiment is conducted by varying the number of quasi-identifiers given as input parameters to the anonymization process. The execution time is noted for the number of quasi-identifiers ranging from 1 to 5, which is displayed in Figure 9.2. The execution time is substantially less when the number of quasi-identifiers in the data set is 1 and 2 as the transformation which the quasi-identifiers - `charging_status`, `fuel_percentage` can undergo is less. However, there is an increase in the execution time when the number of quasi-identifiers is increased to 3. The reason for this behavior is that the quasi identifier(`isc_timestamp`) added has a larger generalization height, as a result of which more computation time is

required. Finally, with the increase in the number of quasi-identifiers, a linear increase in the execution time is observed.

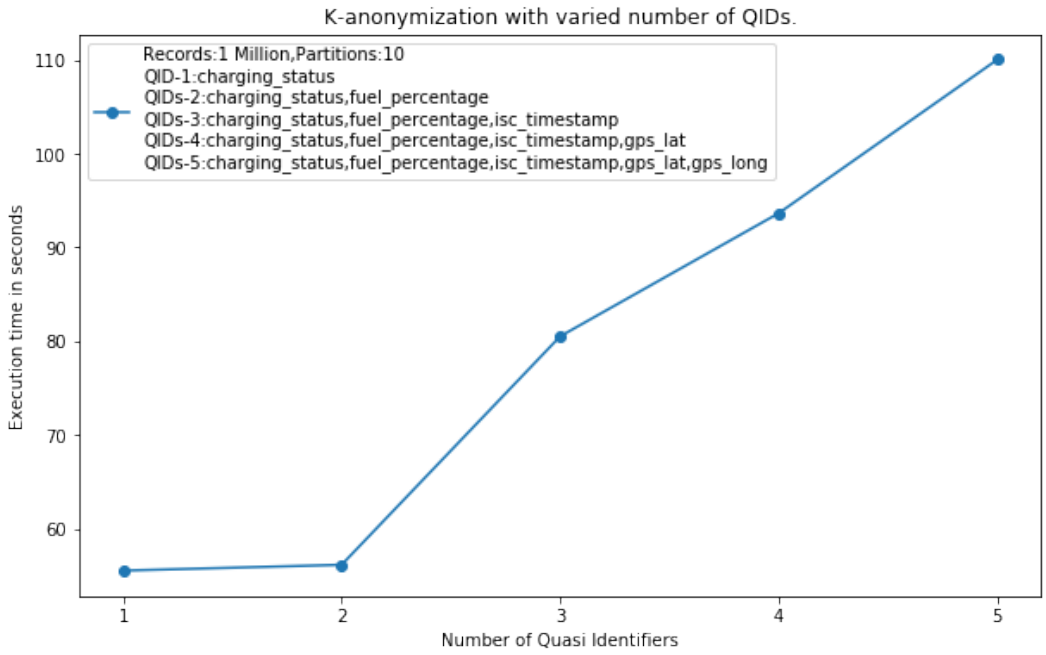

Figure 9.2: Evaluation based on varied Number of Quasi-Identifiers.

9.2.3 Evaluation based on varied Generalization Height

The quasi-identifiers chosen for this evaluation are: `charging_status`, `fuel_percentage`, `isc_timestamp`, `gps_lat`and `gps_long`. The number of quasi-identifiers is 5. The data set chosen 1 million records. The number of partitions performed is 10. The quasi identifier `isc_timestamp` is chosen as the attribute whose generalization hierarchy height is changed. The experiment is conducted with an initial generalization height of 1; then, the height is increased up to 6. From the Figure 9.3 it can be observed that the execution time consistently increases with the increase in the generalization height of the QID.

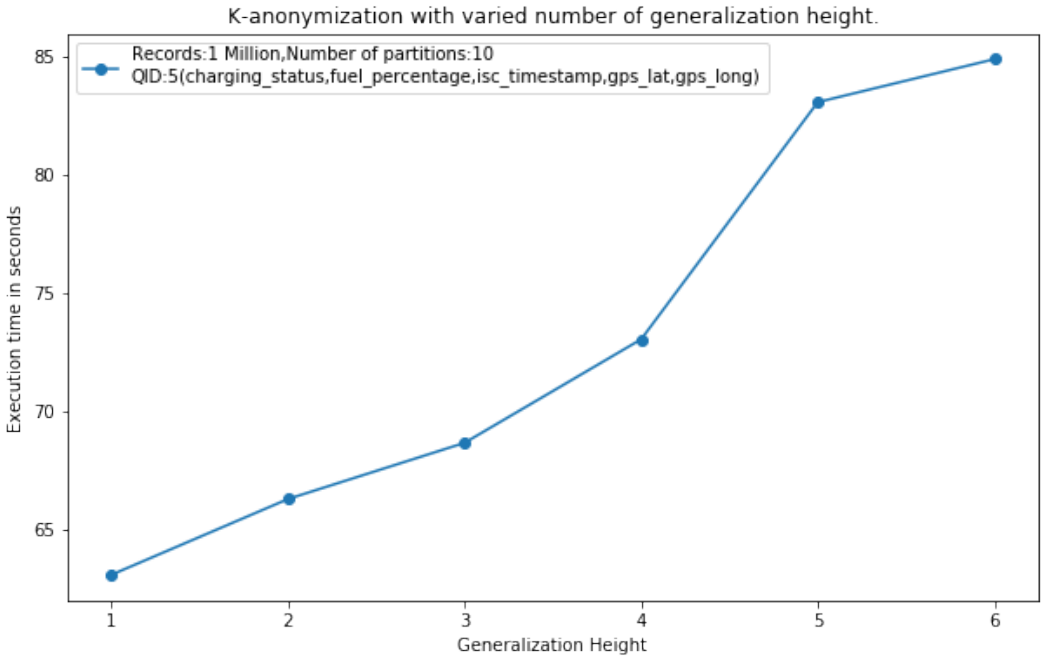

Figure 9.3: Evaluation based on varied Generalization Height.

9.2.4 Evaluation based on varied Number of Partitions.

The quasi-identifiers chosen for this evaluation are: `charging_status`, `fuel_percentage`, `isc_timestamp`, `gps_lat`and `gps_long`. The number of quasi-identifiers ranges from 3 to 5. The data set chosen 1 million records. The experiment is conducted with no partitions, and then the number of partitions is increased up to 50. Each experiment is re-run by varying the QIDs from 3 to 5. The execution time is when no partitioning is done, i.e., when the data set as a whole is anonymized as shown in the Figure9.4 is really high. Once the partitioning is done, the execution is less than half of the time taken when the partitioning was not performed. However, the execution time remains almost constant with an increase in the number of partitions due to three primary reasons: (1) the bottleneck created by the IO operations performed by the Executor service used to run the anonymization in parallel and apply ARX functions to blobs of data in each partition. (2) the amount of time taken for partitioning and merging the data set increases with the number of partitions. (3) if the number of partitions is more than the threads allocated in Executor service, then the tasks of anonymization get to queue in the Executor Service and are dispatched in a First-In-First-Out manner. Figure 9.4 shows the execution time mapped against the number of partitions when the

number of records is 1 million. Figure 9.5 shows the execution time against the number of partitions when the number of records is 5 million. For this use case, anonymization could be performed by the proposed approach. When the number of partition was 25, an exponential increase in the execution time was observed. Further partitioning of the data set where the number of partitions was less than 25 could not be performed due to technical limitations.

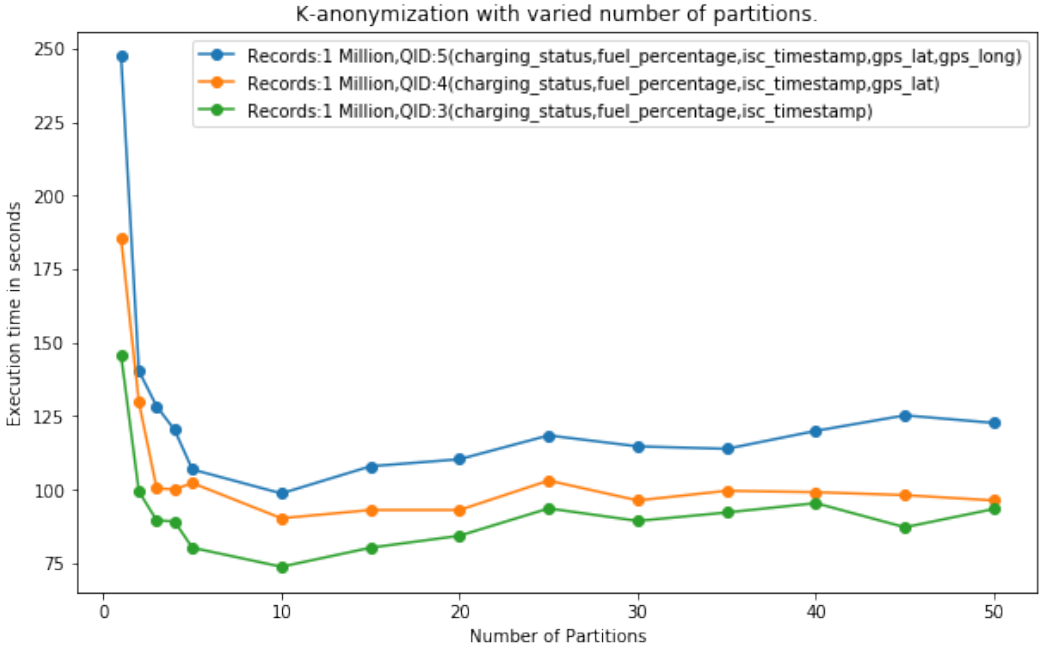

Figure 9.4: Evaluation based on varied Number of Partitions-1 million records.

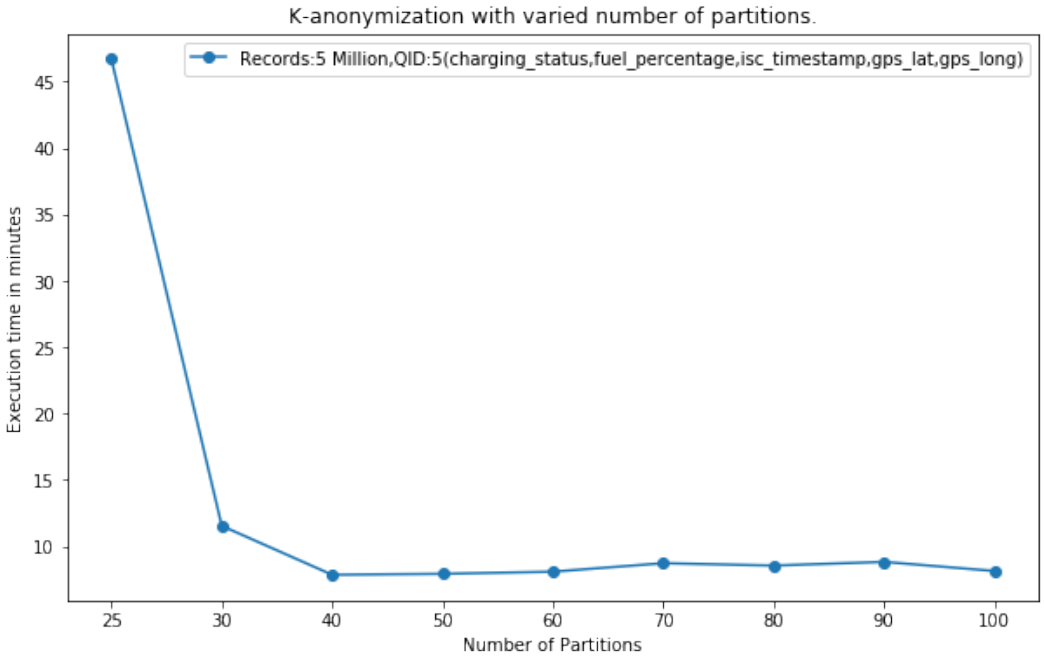

Figure 9.5: Evaluation based on varied Number of Partitions-5 million records.

9.2.5 Evaluation based on varied Number of Records.

The quasi-identifiers chosen for this evaluation are: `charging_status`, `fuel_percentage`, `isc_timestamp`, `gps_lat`and `gps_long`. The number of quasi-identifiers is 5. The number of partitions performed is 10. The size of the data set is varied from 1 million records to 15 million records, and the execution time is noted against the number of records. The execution time is linear throughout, as shown in Figure 9.6. Further anonymization for larger data sets could not be performed due to technical limitations.

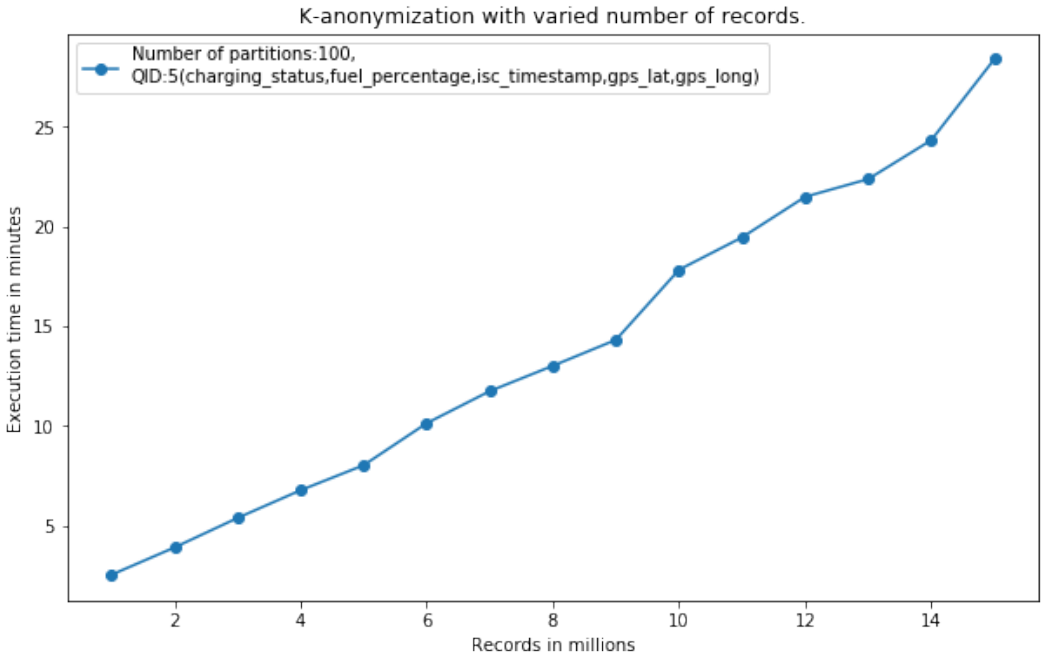

Figure 9.6: Evaluation based on varied Number of Partitions against Resultant K-Value.

9.2.6 Evaluation of Resultant K-Value for varied Number of Partitions

The quasi-identifiers chosen for this evaluation are: `charging_status`, `fuel_percentage`, `isc_timestamp`, `gps_lat` and `gps_long`. The number of quasi-identifiers is 5. The number of partitions performed vary from 5 to 50. The size of the data set is 1 million. Figure 9.7 represents the k value variation with a varied number of partitions. For experimentation purposes, two versions of the data set were considered. In the first version, the data set is used as-is and given as input for anonymization. In the second version, the data set is sorted according to `fuel_percentage` since it is a numerical value, and sorting operation on this attribute makes sense. The k value from each partition is always equal to or more than 10(which is the k-value provided as input). However, this value only increases when the partitions are merged. This is acceptable with an increase in k-value; the privacy level increases as well. When the ordered data set is anonymized, the partitions get anonymized with the resultant K-value equal or more than 10, similar to the previous case. However, sorting fuel percentage increases the possibility of similar values available in the same partition. Hence, when merged, the resultant k-value is even though more, it is less compared to the K-value of merged anonymized data set from the first case.

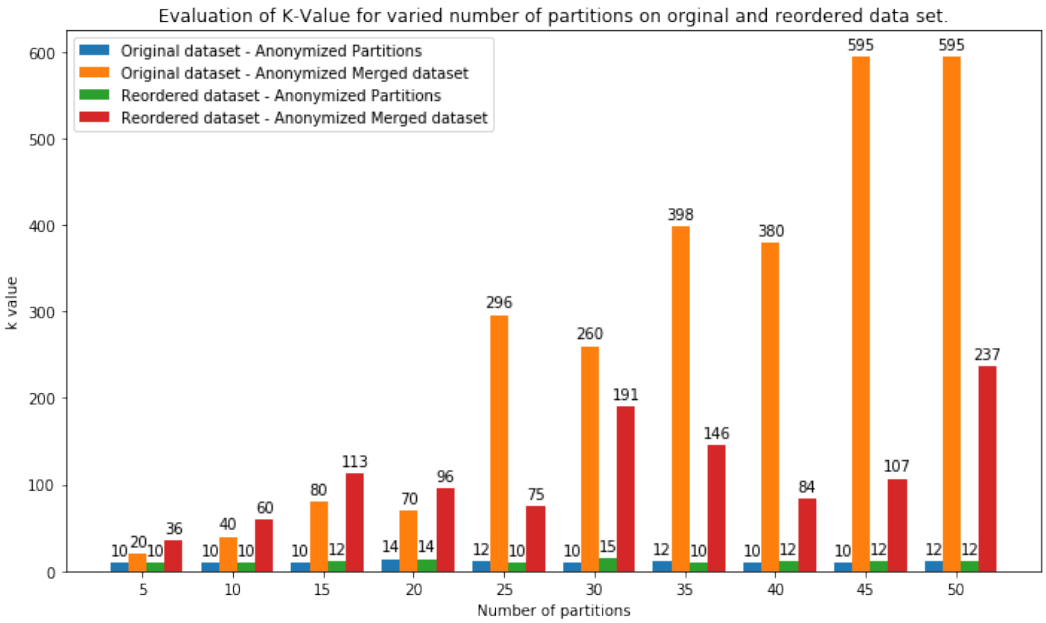

Figure 9.7: Evaluation of Resultant K-value for Varied Number of Partitions.

9.2.7 Insights from the Anonymized Data set

In order to map the evaluation to the use cases, the anonymized data set was analyzed for further interpretation. The use case of placement of smart charging stations 6.3.1 was chosen. In order to find the location where the user is most likely to run out of charge in the vehicle, queries on the anonymized data set where `fuel_percentage` was '*very low*' was chosen. The most common latitude and longitude range from the result were chosen, and the location was mapped. The Figure shows the location where the user is most likely to run out of charge, and the charging station could be potentially placed in the location.

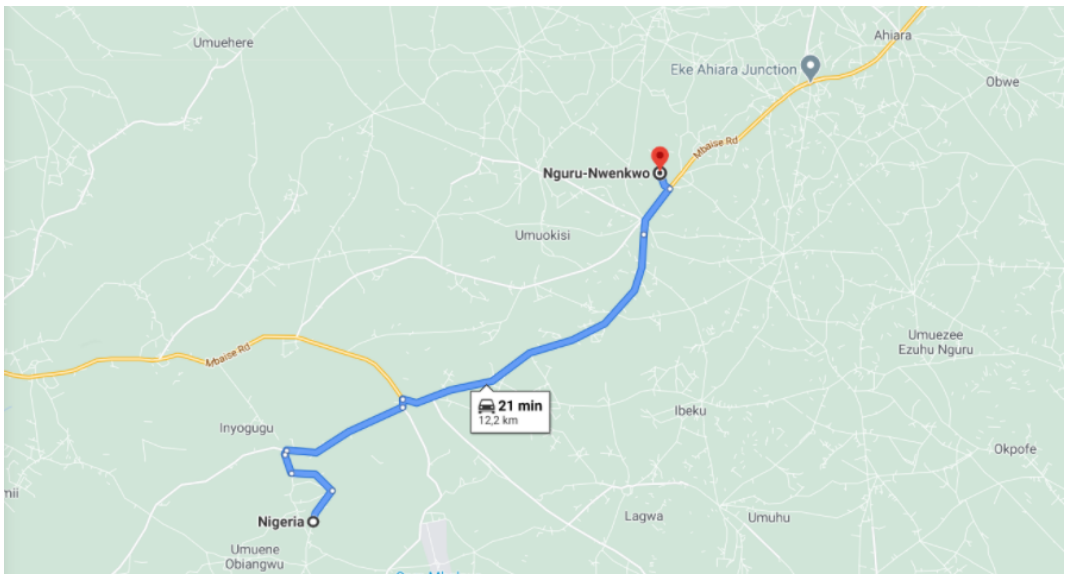

Figure 9.8: Insights from the Anonymized Data set.

10 Conclusion and Future Work

This final chapter summarizes the research outcomes of the thesis. Additionally, limitations of the research as well as the potential future research work related to the thesis are presented.

10.1 Summary

In this section, the research questions formulated as part of the thesis is presented along with the answers and contribution this work provides.

RQ1: What are the properties of current k-anonymity implementations/algorithms? From the extensive literature research conducted, three primary k-anonymization techniques: Mondrian, Incognito, Datafly, were chosen post reviewing 18 research papers as these three techniques were most popular as well as most cited. Each of these algorithms was further analyzed systematically by reviewing the behavior individually and applying the logic behind the algorithm to the example car data set chosen. From these experiments, a better understanding of the pros and cons of using these algorithms was noted down. Furthermore, a guideline was formulated which could be used as a reference in choosing the appropriate algorithm according to the requirements. Alongside, approaches: Top-Down Specialization and Bottom-Up Generalization for applying generalization to any data set were explored. A recommendation on the usage of these techniques based on scenarios was mapped out. Lastly, a new approach of applying anonymization: Local and Collaborative anonymization technique was investigated as well. ARX tool, which is listed as one of the standard tools to perform de-identification of data sets, was investigated thoroughly, and the decision was made to use the functions exposed by the tool in the implementation phase based on the initial experiments conducted to understand the feasibility of the tool.

RQ2: What are the requirements for k-anonymity implementations in big data context? From the detailed literature review conducted, the context of what big data could mean for the anonymization of the data set was studied in detail. Furthermore, the advantages and disadvantages of using the centralized and distributed approach for the anonymization process were extensively researched. Post literature review, a

decision of applying the k-anonymization technique in a distributed manner such that the prototype would work in a big data context was made. On those lines, technologies like Apache Spark and the data partitioning approach were explored by performing experiments using those approaches, which proved to be beneficial in formulating the system architecture. Interviews were conducted in parallel at European OEM and European electronics manufacturer to formulate the use case and well as getting an understanding of their requirement on k-anonymity implementation in the big data context. As an outcome, the list of requirements in the form of use cases and properties of connected car data set were drawn along with the use of Apache Spark to apply k-anonymity using the ARX tool in the big data context.

RQ3: How can a k-anonymity implementation in the context of big data look like? The system architecture was designed to take advantage of the ARX library as well as provide data anonymization in a distributed manner with the help of Apache Spark. The execution phase of the thesis was divided into three phases: Design, Implementation, and Evaluation. In the design phase, using the properties of the connected car data set studied at the European OEM, a data generator application was developed, which provided the source data set similar to the one at the European OEM. In the Implementation phase, the ARX library was integrated into the base application and pre-requisites required to use the ARX library like data processing to match the data types used by the source data set, and the steps to use the ARX library itself like hierarchy definition and other steps as described in the Section 8 were performed. The prototype was then developed, which takes the source data set as input, applies data partitioning using Apache Spark, and the anonymization of the partitioned data set was performed asynchronously using Executor Service and the functions exposed by ARX. Lastly, the partitioned anonymized data set generated is merged using Apache Spark to form the final anonymized data set. Lastly, as part of the evaluation phase, the prototype was tested against various parameters as explained in Section 9. From the results, it is proven that the prototype works well for larger data sets. Another important finding was that the proposed distributed k-anonymization approach works way better than the centralized approach as the execution time is almost halved by performing anonymization for medium to large data sets, and for even larger data sets, the anonymization could only be performed with the proposed solution.

10.2 Limitations

From the results derived from the work, the two limitations found will be explained in this section.

Firstly, the data set used for testing the prototype was generated synthetically using the properties noted down in the interviews conducted at the European OEM. Though the data set generator provided with data set that resembles the car data set at the European OEM, few of the properties could not be replicated due to confidential reasons. Moreover, the geographical validity of the GPS latitude and longitude with respect to valid locations on a map is not taken into consideration. Also, the synthetic data set itself contains limited outliers. Secondly, the experiments run as part of the evaluation phase were run on the local machine, which reduces the computational capacity provided to run the prototype.

10.3 Future Work

This work provides some topics which could be taken into consideration for further investigations.

The current prototype performs k-anonymity for large data sets containing up to 15 million records. This solution can be further extended by deploying the application in a high computing cloud environment, which would provide high computational power to process even bigger data sets. The solution can also be extended to other privacy models like l-diversity, t-closeness, and differential privacy as ARX supports these privacy models as well. From the inputs provided by Rohde and Schwarz, the data generator part of the prototype can be extended as follows: A provision could be added in the data generator, where it runs through the source data set, prepares statistical models for each attribute in the data set and does data set generated from the formed statistical models. This will provide a synthetic data set that is more in sync with the source data.

List of Figures

List of Tables

Bibliography

[1] L. Arbuckle. *De-identification 201*. Tech. rep. Privacy Analytics, 2020.

[2] O. Tomashchuk, D. Van Landuyt, D. Pletea, K. Wuyts, and W. Joosen. "A Data Utility-Driven Benchmark for De-identification Methods." In: *Trust, Privacy and Security in Digital Business*. Ed. by S. Gritzalis, E. R. Weippl, S. K. Katsikas, G. Anderst-Kotsis, A. M. Tjoa, and I. Khalil. Cham: Springer International Publishing, 2019, pp. 63–77. ISBN: 978-3-030-27813-7.

[3] B. Schneier. *Why 'Anonymous' Data Sometimes Isn't*. 2007. URL: https://www.wired.com/2007/12/why-anonymous-data-sometimes-isnt/ (visited on 12/12/2007).

[4] B. Z. H. Zhao, M. A. Kaafar, and N. Kourtellis. "Not One but Many Tradeoffs: Privacy Vs. Utility in Differentially Private Machine Learning." In: CCSW'20. Virtual Event, USA: Association for Computing Machinery, 2020, pp. 15–26. ISBN: 9781450380843. DOI: 10.1145/3411495.3421352.

[5] S. DMITRIEV. *Autonomous cars will generate more than 300 TB of data per year*. 2017. URL: https://www.tuxera.com/blog/autonomous-cars-300-tb-of-data-per-year/ (visited on 11/28/2017).

[6] kdespagniqz. *Connected cars will send 25 gigabytes of data to the cloud every hour*. 2015. URL: https://qz.com/344466/connected-cars-will-send-25-gigabytes-of-data-to-the-cloud-every-hour/ (visited on 02/13/2015).

[7] B. Horne. *Protecting Connected Cars from Cyberattack*. 2020. URL: https://securityboulevard.com/2020/10/protecting-connected-cars-from-cyberattack/ (visited on 10/07/2020).

[8] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang. "An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing." In: KDD '06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, pp. 754–759. ISBN: 1595933395. DOI: 10.1145/1150402.1150499.

[9] A. R. Hevner, S. T. March, J. Park, and S. Ram. "Design Science in Information Systems Research." In: *MIS Q.* 28.1 (Mar. 2004), pp. 75–105. ISSN: 0276-7783.

[10] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee. "A Design Science Research Methodology for Information Systems Research." In: *J. Manage. Inf. Syst.* 24.3 (Dec. 2007), pp. 45–77. ISSN: 0742-1222.

[11] I. of Medicine.Washington (DC). *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. 2015. URL: https://www.ncbi.nlm.nih.gov/books/NBK285994/.

[12] J. Jerome. *De-Identification Should Be Relevant to a Privacy Law, But Not an Automatic Get-Out-of-Jail-Free Card*. 2019. URL: https://cdt.org/insights/de-identification-should-be-relevant-to-a-privacy-law-but-not-an-automatic-get-out-of-jail-free-card/ (visited on 04/01/2020).

[13] D. GOODIN. *Poorly anonymized logs reveal NYC cab drivers' detailed whereabouts*. 2014. URL: https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts (visited on 06/23/2014).

[14] de Montjoye, Y.-A. Hidalgo, C. A.Verleysen, M. Blondel, and V. D. "Unique in the Crowd: The privacy bounds of human mobility." In: *Scientific Reports* 3.1 (Mar. 2013), pp. 2045–2322.

[15] H. Bennett. *Research reveals de-identified patient data can be re-identified*. 2017. URL: https://about.unimelb.edu.au/newsroom/news/2017/december/research-reveals-de-identified-patient-data-can-be-re-identified (visited on 12/18/2017).

[16] B. Malin. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. 2010. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

[17] J. D. Cook. *HIPAA De-identification Expert*. URL: https://www.johndcook.com/blog/expert-hipaa-deidentification/.

[18] E. Schachter. *Bill 64: Quebec Seeks to Dramatically Reform the Province's Privacy Policy*. 2020. URL: https://www.groupetcj.ca/en/news/646-bill-64-quebec-seeks-to-dramatically-reform-the-provinces-privacy-policy.html (visited on 11/25/2020).

[19] W. Deneault-Rouillard. *Bill 64 and Act to modernize legislative provisions as regards the protection of personal information*. 2020. URL: https://www.fasken.com/en/knowledge/projet-de-loi-64/2020/09/21-apercu-techno-juridique-renseignements-depersonnalises-anonymises (visited on 09/21/2020).

[20] ISO2018. *Privacy enhancing data de-identification terminology and classification of techniques*. Standard. International Organization for Standardization, Nov. 2018.

[21] L. SWEENEY. "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY." In: *World Scientific Publishing Co.* 10 (Oct. 2002), pp. 2045–2322.

[22] J. C. Julien Debussche. *Big Data and Issues and Opportunities: Anonymisation and Pseudonymisation*. Jan. 2019. URL: https://www.twobirds.com/en/news/articles/2019/global/big-data-and-issues-and-opportunities-anonymisation-pseudonymisation.

[23] J. H. University. *PROTECTING IDENTIFIERS IN HUMAN*. 2021. URL: https://guides.library.jhu.edu/protecting_identifiers/definitions (visited on 01/19/2020).

[24] U. of Pittsburgh. *Guide to Identifying Personally Identifiable Information (PII)*. URL: https://www.technology.pitt.edu/help-desk/how-to-documents/guide-identifying-personally-identifiable-information-pii.

[25] S. d. C. d. Vimercati and S. Foresti. "Quasi-Identifier." In: *Encyclopedia of Cryptography and Security*. Ed. by H. C. A. van Tilborg and S. Jajodia. Boston, MA: Springer US, 2011, pp. 1010–1011. ISBN: 978-1-4419-5906-5. DOI: 10.1007/978-1-4419-5906-5_763.

[26] OECD. *Glossary of statistical terms*. 2005. URL: https://stats.oecd.org/glossary/detail.asp?ID=6961 (visited on 11/10/2005).

[27] C. Kushida, D. Nichols, R. Jadrnicek, R. Miller, J. Walsh, and K. Griffin. "Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies." In: *Medical care* 50 Suppl (July 2012), S82–101. DOI: 10.1097/MLR.0b013e3182585355.

[28] S. TRANQUILLINI. *How to implement pseudonymization: your key questions answered*. 2019. URL: https://www.chino.io/blog/how-to-implement-pseudonymization-key-questions/ (visited on 04/26/2019).

[29] K. Ito, J. Kogure, T. Shimoyama, and H. Tsuda. "De-identification and Encryption Technologies to Protect Personal Information." In: *Fujitsu Scientific and Technical Journal* 3.52 (July 2016), pp. 28–36.

[30] P. Samarati. "Protecting Respondents' Identities in Microdata Release." In: *IEEE Trans. on Knowl. and Data Eng.* 13.6 (Nov. 2001), pp. 1010–1027. ISSN: 1041-4347. DOI: 10.1109/69.971193.

[31] P. Samarati and L. Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. 2003. URL: https://dataprivacylab.org/dataprivacy/projects/kanonymity/index3.html.

[32] H. G. Vediramana Krishnan, Y. Chen, S. Shoham, and A. Gurfinkel. "Global Guidance for Local Generalization in Model Checking." In: *Computer Aided Verification*. Ed. by S. K. Lahiri and C. Wang. Cham: Springer International Publishing, 2020, pp. 101–125. ISBN: 978-3-030-53291-8.

[33] N. Chakraborty and G. K. Patra. "Functional encryption for secured big data analytics." In: *International Journal of Computer Applications* 16.107 (2014), pp. 19–22.

[34] L. Kaith. *Connected Cars: How Is Data Strategy Important for the Growth of Your Automotive Business?* 2019. URL: https://customerthink.com/connected-cars-how-is-data-strategy-important-for-the-growth-of-your-automotive-business/ (visited on 07/05/2019).

[35] N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. A. Abbasi, and S. Salehian. *The 10 Vs, Issues and Challenges of Big Data*. Vol. 4. ICBDE '18. New York, NY, USA: Association for Computing Machinery, 2018. ISBN: 9781450363587. DOI: 10.1145/3206157.3206166.

[36] *Big data: Issues, challenges, tools and Good practices*. 2013, pp. 404–409. DOI: 10.1109/IC3.2013.6612229.

[37] A. Katal, M. Wazid, and R. Goudar. "Big data: issues, challenges, tools and good practices." In: An optional note. 2013 Sixth International Conference. IEEE, 2013.

[38] N. Khan, I. Yaqoob, I. A. T. Hashem, and Z. e. a. Inayat. "Big Data: Survey, Technologies, Opportunities, and Challenges." In: *The Scientific World Journal* 2014 (2014).

[39] Y. JORIS TOONDERS. *Data Is the New Oil of the Digital Economy*. 2014. URL: https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/.

[40] M. Alloghani, M. M. Alani, D. Al-Jumeily, T. Baker, J. Mustafina, A. Hussain, and A. J. Aljaaf. "A systematic review on the status and progress of homomorphic encryption technologies." In: *Journal of Information Security and Applications* 48 (2019), p. 102362. ISSN: 2214-2126.

[41] B. C. M. Fung, K. Wang, and P. S. Yu. "Top-Down Specialization for Information and Privacy Preservation." In: *Proceedings of the 21st International Conference on Data Engineering*. ICDE '05. USA: IEEE Computer Society, 2005, pp. 205–216. ISBN: 0769522858. DOI: 10.1109/ICDE.2005.143.

[42] X. Zhang, L. T. Yang, C. Liu, and J. Chen. "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud." In: *IEEE Transactions on Parallel and Distributed Systems* 25.2 (2014), pp. 363–373. DOI: 10.1109/TPDS.2013.48.

[43] Ke Wang, P. S. Yu, and S. Chakraborty. "Bottom-up generalization: a data mining solution to privacy protection." In: *Fourth IEEE International Conference on Data Mining (ICDM'04)*. 2004, pp. 249–256. DOI: 10.1109/ICDM.2004.10110.

[44] K. Pandilakshmi and G. Banu. "An Advanced Bottom up Generalization Approach for Big Data on Cloud." In: *International Journal of Communication and Networking System* 003 (June 2014), pp. 12–15. DOI: 10.20894/IJCNES.103.003.001.003.

[45] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen. "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud." In: *Journal of Computer and System Sciences* 80 (Aug. 2014). DOI: 10.1016/j.jcss.2014.02.007.

[46] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen. "Combining Top-Down and Bottom-Up: Scalable Sub-tree Anonymization over Big Data Using MapReduce on Cloud." In: *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. 2013, pp. 501–508. DOI: 10.1109/TrustCom.2013.235.

[47] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy. "A Systematic Comparison and Evaluation of K-Anonymization Algorithms for Practitioners." In: *Trans. Data Privacy* 7.3 (Dec. 2014), pp. 337–370. ISSN: 1888-5063.

[48] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. "Mondrian Multidimensional K-Anonymity." In: *Proceedings of the 22nd International Conference on Data Engineering*. ICDE '06. USA: IEEE Computer Society, 2006, p. 25. ISBN: 0769525709. DOI: 10.1109/ICDE.2006.101.

[49] J. H. Friedman, J. L. Bentley, and R. A. Finkel. "An Algorithm for Finding Best Matches in Logarithmic Expected Time." In: *ACM Trans. Math. Softw.* 3.3 (Sept. 1977), pp. 209–226. ISSN: 0098-3500. DOI: 10.1145/355744.355745.

[50] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. "Incognito: Efficient Full-Domain K-Anonymity." In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. SIGMOD '05. Baltimore, Maryland: Association for Computing Machinery, 2005, pp. 49–60. ISBN: 1595930604. DOI: 10.1145/1066157.1066164.

[51] M. E. Nergiz and C. Clifton. "Thoughts on k-anonymization." In: *Data and Knowledge Engineering* 63.3 (2007). 25th International Conference on Conceptual Modeling (ER 2006), pp. 622–645. ISSN: 0169-023X. DOI: https://doi.org/10.1016/j.datak.2007.03.009.

[52] P. Samarati. "Protecting respondents identities in microdata release." In: *IEEE Transactions on Knowledge and Data Engineering* 13.6 (2001), pp. 1010–1027. DOI: 10.1109/69.971193.

[53] M. S. Simi, K. S. Nayaki, and M. S. Elayidom. "An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity." In: *IOP Conference Series: Materials Science and Engineering* 225 (July 2017), p. 012279. DOI: 10.1088/1757-899x/225/1/012279.

[54] A. Hoang, M. Tran, A. Duong, and I. Echizen. "An Indexed Bottom-up Approach for Publishing Anonymized Data." In: *2012 Eighth International Conference on Computational Intelligence and Security*. 2012, pp. 641–645. DOI: 10.1109/CIS.2012.148.

[55] X. Zhang, L. T. Yang, C. Liu, and J. Chen. "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud." In: *IEEE Transactions on Parallel and Distributed Systems* 25.2 (2014), pp. 363–373. DOI: 10.1109/TPDS.2013.48.

[56] A. Irudayasamy and L. Arockiam. "Parallel Bottom-up Generalization Approach for Data Anonymization using Map Reduce for Security of Data in Public Cloud." In: *Indian journal of science and technology* 8 (2015), pp. 1–9.

[57] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen. "Combining Top-Down and Bottom-Up: Scalable Sub-tree Anonymization over Big Data Using MapReduce on Cloud." In: *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. 2013, pp. 501–508. DOI: 10.1109/TrustCom.2013.235.

[58] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C.-K. Lee. "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data." In: *ACM Trans. Knowl. Discov. Data* 4.4 (Oct. 2010). ISSN: 1556-4681. DOI: 10.1145/1857947.1857950.

[59] S. L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." In: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69. ISSN: 01621459.

[60] J. Domingo-Ferrer, J. Soria-Comas, and O. Ciobotaru. "Co-Utility: Self-Enforcing Protocols without Coordination Mechanisms." In: *CoRR* (2015). arXiv: 1503.02563.

[61] F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn. "Flexible data anonymization using ARX—Current status and challenges ahead." In: *Software: Practice and Experience* 50 (2020), pp. 1277–1304.

[62] N. I. of Standards and Technology. *De-identification Tools*. Tech. rep. Washington, D.C.: U.S. Department of Commerce, 2001. DOI: 10.6028/nist.fips.140-2.

[63] M. Dinita. *Best free data anonymization software to use in 2020*. 2019. URL: https://windowsreport.com/data-anonymization-software/ (visited on 11/21/2019).

[64] B. Curtis. *6 BEST DATA ANONYMIZATION TOOLS*. 2020. URL: https://www.yourtechdiet.com/blogs/6-best-data-anonymization-tools/.

[65] arx-deidentifier. *An overview of methods for data anonymization*. 2015. URL: https://www.slideshare.net/arx-deidentifier/prasser-methods?ref=https://arx.deidentifier.org/ (visited on 03/27/2015).

[66] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn. "Flash: Efficient, Stable and Optimal K-Anonymity." In: *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT)*. Los Alamitos, CA, USA: IEEE Computer Society, 2012, pp. 708–717. DOI: 10.1109/SocialCom-PASSAT.2012.52.

[67] F. Prasser, R. Bild, J. Eicher, H. Spengler, F. Kohlmayer, and K. A. Kuhn. "Lightning: Utility-Driven Anonymization of High-Dimensional Data." In: *Trans. Data Privacy* 9.2 (Aug. 2016), pp. 161–185. ISSN: 1888-5063.

[68] D. M. Kantarcioglu, A. Inan, and M. Kuzu. *UTD Anonymization ToolBox*. URL: http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php.

[69] R. Koch. *Fitness apps are good for your health, but often bad for your privacy*. 2019. URL: https://protonvpn.com/blog/fitness-apps-are-good-for-your-health-but-often-bad-for-your-privacy/ (visited on 10/31/2019).

[70] Z. Schiffer. *Smart TVs are data-collecting machines, new study shows*. 2019. URL: https://www.theverge.com/2019/10/11/20908128/smart-tv-surveillence-data-collection-home-roku-amazon-fire-princeton-study (visited on 10/11/2019).

[71] C. Boyd. *How social media platforms mine personal data for profit*. 2020. URL: https://blog.malwarebytes.com/privacy-2/2020/04/how-social-media-mine-data-sell-personal-information-for-profit/.

[72] Wikipedia. *Carl Benz*. URL: https://de.wikipedia.org/wiki/Carl_Benz.

[73] C. Woodford. *History of cars*. 2021. URL: https://www.explainthatstuff.com/historyofcars.html (visited on 01/31/2021).

[74] G. Meiring and H. Myburgh. "A Review of Intelligent Driving Style Analysis Systems and Related Artificial Intelligence Algorithms." In: *Sensors* 15 (Dec. 2015), pp. 30653–30682. DOI: 10.3390/s151229822.

[75] J. Woodruff. *Effects of Technology on Supply and Demand Curves*. 2019. URL: https://smallbusiness.chron.com/effects-technology-supply-demand-curves-30626.html (visited on 03/20/2019).

[76] B. Manning. *INTERNET CONNECTION BY 2020*. 2015. URL: https://centricdigital.com/blog/internet-of-things/connected-cars/ (visited on 04/22/2015).

[77] A. Kara. *Connected Cars: How IoT, Streaming Data and Real-Time Analytics are Disrupting the Rental Car Industry*. 2019. URL: https://blogs.informatica.com/2019/08/21/connected-cars-how-iot-streaming-data-and-real-time-analytics-are-disrupting-the-rental-car-industry-part-1/ (visited on 08/21/2019).

[78] S. BLANCO and B. NICHOLS. *EV Charging Stations: Where to Find Them, What Type You Need, How to Pay*. 2019. URL: https://www.caranddriver.com/news/a30031153/ev-charging-guide/ (visited on 12/03/2019).

[79] C. Lilly. *EV connector types*. 2020. URL: https://www.zap-map.com/charge-points/connectors-speeds/ (visited on 04/03/2020).

[80] K. Siddiqui, S. Sher, A. Tarani, S. Fatani, A. Raza, R. Butt, N. Azeema, M. Jinnah, and U. Karachi. "EFFECT OF SIZE, LOCATION AND CONTENT OF BILLBOARDS ON BRAND AWARENESS." In: *Journal of Business Studies Quarterly* 8 (Dec. 2016), pp. 40–57.

[81] M. Wroblewski. *The Advantages and Disadvantages of Billboards As an Advertisement Tool*. 2018. URL: https://smallbusiness.chron.com/advantages-disadvantages-billboards-advertisement-tool-16143.html (visited on 10/19/2018).

[82] V. Outdoor. *The facts about billboard advertising*. URL: http://www.visionsoutdoor.net/billboard-marketing-statistics/.

[83] P. INMAN. *Billboard Advertising costs 2021*. 2021. URL: https://75media.co.uk/blog/billboard-costs/ (visited on 01/27/2021).

[84] statista. *Largest outdoor advertising companies worldwide in 2019, by revenue.* 2019. URL: https://www.statista.com/statistics/323692/revenue-outdoor-advertising-companies/.

[85] K. Wuyts, D. V. Landuyt, L. Sion, and W. Joosen. *LINDDUN framework*. URL: https://www.linddun.org/linddun.

[86] A. Robles-González, J. Parra-Arnau, and J. Forné. "A LINDDUN-Based framework for privacy threat analysis on identification and authentication processes." In: *Computers and Security* 94 (2020), p. 101755. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2020.101755.

[87] L. Sion, K. Wuyts, K. Yskout, D. Van Landuyt, and W. Joosen. "Interaction-Based Privacy Threat Elicitation." In: *2018 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*. 2018, pp. 79–86. DOI: 10.1109/EuroSPW.2018.00017.

[88] S. Sowmya. *Spark K Anonymity*. URL: https://github.com/SharadaSowmya14/spark-k-anonymity/.

[89] S. Sowmya. *Spring Spark K Anonymity*. URL: https://github.com/SharadaSowmya14/spring-spark-kanonymity.