NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY
[IJIERT] ISSN: 2394-3696 Website: ijiert.org
VOLUME 8, ISSUE 2, Feb.-2021

TECHNOLOGY ASSESSMENT OF DUMONIA

Lee Chung
Cheju Tourism College – Bukjeju County, Jeju
South Korea

Lei Bom
Kyoto Sangyo University – Japan

EXECUTIVE SUMMARY

The following report aims to highlight how Dumnonia company will implement the k-anonymity model for maintaining and preserving the privacy of big data. This company is serving citizens in New Zealand and Australia with different types of insurance services such as vehicle insurance, death insurance, and travel insurance. As the organization is dealing with the medical data of citizens so it is important to implement effective security controls for maintaining data confidentiality and privacy. In order to deal and manage with massive amount of medical and insurance data, the business executives of company have decided to opt big data. The traditional data management models are unable to process a huge volume of big data because this data has been gathered from unstructured and unstructured data sources. Hence, the organization needs to implement effective security management tools that would promote preservation of data confidentiality and privacy along with high system performance.

The CSO, CEO, and CIO of the company are showing concerns about the privacy preservation of the medical and insurance data of end-users. In this report, the organization has been recommended to opt k-anonymity model. The first section of this report is providing information about various organizational drivers which should be considered when implementing the k-anonymity model. In this report, data models, privacy models and quality metrics are discussed in detailed manner. Two different implementations of k-anonymity techniques have been compared on the basis of different parameters. The report provides a detailed implementation guide that will allow the organization to implement k-anonymity technique to preserve the confidentiality of its insurance data.

1 ORGANIZATIONAL DRIVERS FOR DUMNONIA

The latest advancements in business analytics and data mining techniques help enterprises to effective use huge and massive data sets. The main purpose of this report is to analyze how the k-anonymity model can be implemented by the Dumnonia corporation for protecting the privacy of its sensitive business data. Dumnonia as a famous insurance player that serves its clients in New Zealand and Australia. For managing the huge volume of consumer data, the organization has adopted a big data technology (Ganz, 2015). It is difficult for the senior management of the organization to manage the confidentiality and privacy of private information of its customers. Currently, the senior management has been decided to implement k-anonymity model for preserving the privacy of sensitive part of customer's data. In the following section, 3 drivers are presented that will allow the organization to solve existing problems with the adoption of the anonymization technique:

- ✓ **Data privacy and anonymity** of customers' private information and medical data is the major organizational driver (Wang et al., 2017).
- ✓ **Existing security challenges** (like a corrupted file system, untrusted data generation, etc.) can be considered as organizational drivers that might be encountered by the organization (Ganz, 2015).
- ✓ **Encryption** should not be used by the organization to overcome existing privacy and security challenges because it slows down the speed of big data.

2 TECHNOLOGY SOLUTION ASSESSMENT

To protect the identity of customers is essential for Dumnonia company as it is dealing with sensitive data including customer information and medical data. Otherwise; the corporate image could be affected adversely (Ganz, 2015). The traditional data management models are unable to process a huge volume of big data because this data has been gathered from unstructured and unstructured data sources. Data anonymities model,

the k-anonymity model, notice and consent and differentia privacy are main models that can be used to preserve the confidentiality and privacy of sensitive information in the field of big data (Zhang et al., 2012). K-anonymity model has been recommended to Dumnonia company to adopt due to its cost-effectiveness and simplicity. This model supports clustering of information on the basis of publicly known management each data set. In addition to this, the sensitive and critical part of the data set can be kept a secret with the use of this model. In order to implement K-Anonymity, the CIO of the company discussed 3 models named as quality metrics, data model and privacy model (Wang et al., 2017)

2.1 PRIVACY MODEL

This privacy model can be utilized for protecting the privacy of information. These models are based on the following discussed principles:

1. **Suppression** - In this method, each element or attribute in the data set has been replaced by suitable symbols and signs for masking the attribute's identity. This mechanism is widely utilized for the implementation of k-anonymity technique (Wang et al., 2017)
2. **Generalization** – In this method, the data set and data attribute are modified by using other attributes that do not reveal actual data but give the same meaning as the actual data (Ganz, 2015).

Zip Code	Age	Disease
110**	6*	Malaria
110**	6*	Malaria
110**	6*	Malaria
111**	7*	Heart Disease
111**	7*	Heart Disease
111**	5*	Cancer

Figure-1 Generalization of patient's information (Ganz, 2015)

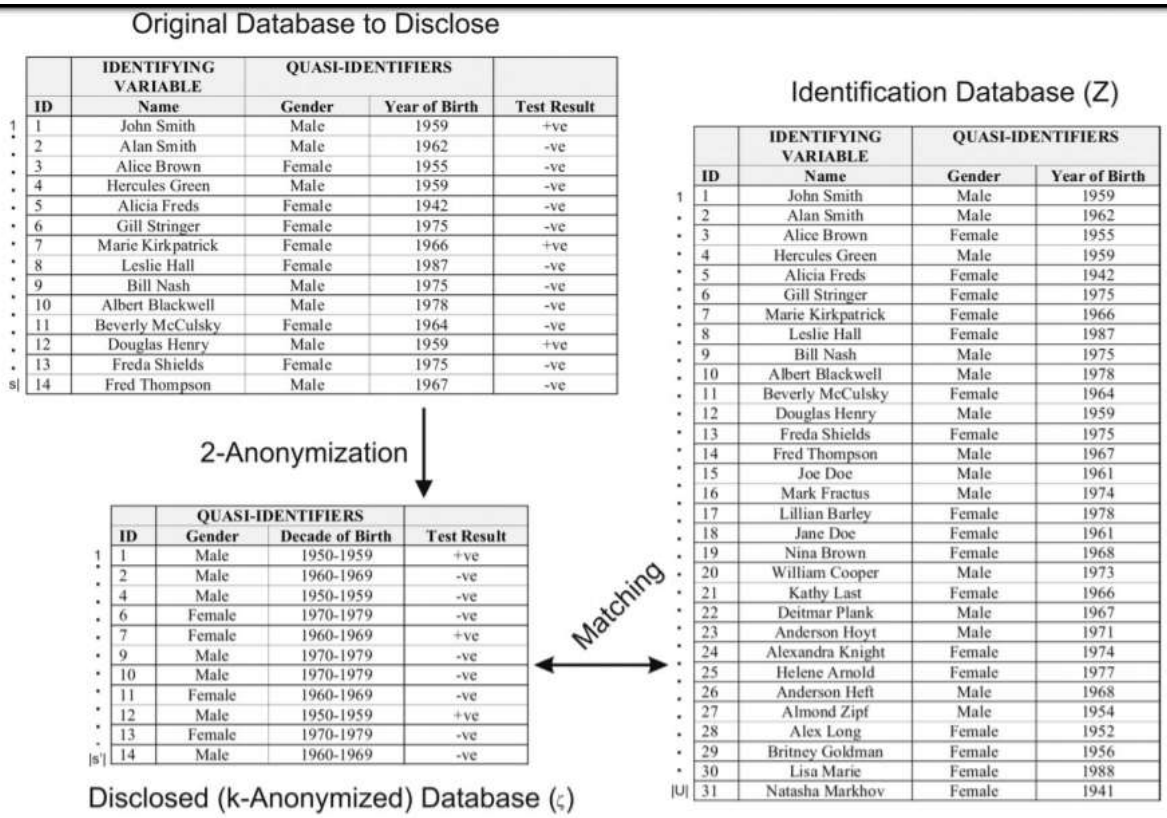
In the following section, various privacy models are discussed in a detailed manner:

K-Anonymity

For data anonymity, k-anonymity is one of the easiest and the simplest model. As a quasi-identifier # or * symbols have been used in this model for hiding the actual identity of a specific attribute or data set. In simple words, the suppression principle and quasi identifier are main principles of privacy models (Zhang et al., 2012).

	Suppression			
Generalization	*Tuple*	*Attribute*	*Cell*	*None*
Attribute	AG_TS	AG_AS ≡ AG_	AG_CS	AG_ ≡ AG_AS
Cell	CG_TS not applicable	CG_AS not applicable	CG_CS ≡ CG_	CG_ ≡ CG_CS
None	_TS	_AS	_CS	_ not interesting

Figure -2 Classification of k-anonymity technique (Devi, 2011)

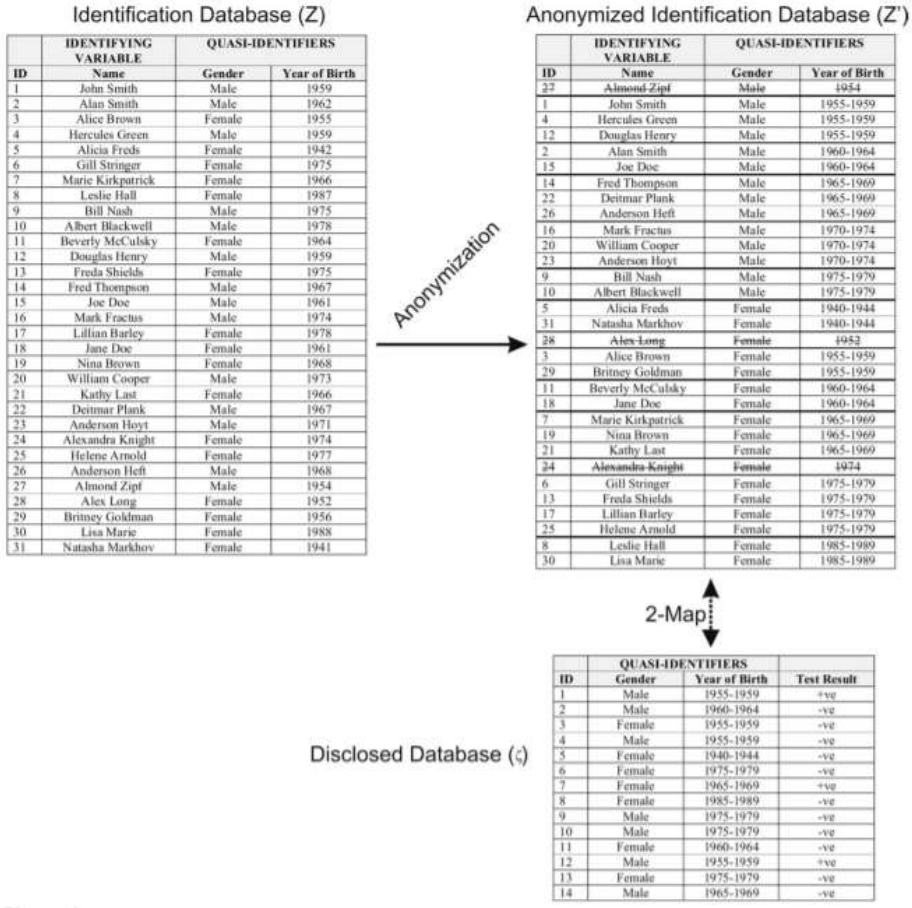


	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
ID	Name	Gender	Year of Birth	Test Result
1	John Smith	Male	1959	+ve
2	Alan Smith	Male	1962	-ve
3	Alice Brown	Female	1955	-ve
4	Hercules Green	Male	1959	-ve
5	Alicia Freds	Female	1942	-ve
6	Gill Stringer	Female	1975	-ve
7	Marie Kirkpatrick	Female	1966	+ve
8	Leslie Hall	Female	1987	-ve
9	Bill Nash	Male	1975	-ve
10	Albert Blackwell	Male	1978	-ve
11	Beverly McCulsky	Female	1964	-ve
12	Douglas Henry	Male	1959	+ve
13	Freda Shields	Female	1975	-ve
14	Fred Thompson	Male	1967	-ve

	QUASI-IDENTIFIERS		
ID	Gender	Decade of Birth	Test Result
1	Male	1950-1959	+ve
2	Male	1960-1969	-ve
4	Male	1950-1959	-ve
6	Female	1970-1979	-ve
7	Female	1960-1969	+ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve
11	Female	1960-1969	-ve
12	Male	1950-1959	+ve
13	Female	1970-1979	-ve
14	Male	1960-1969	-ve

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

Figure 3: A systematic process to create anonymity by using K-anonymity (El Emam and Dankar, 2008)



	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
~~27~~	~~Almond Zipf~~	~~Male~~	~~1954~~
1	John Smith	Male	1955-1959
4	Hercules Green	Male	1955-1959
12	Douglas Henry	Male	1955-1959
2	Alan Smith	Male	1960-1964
15	Joe Doe	Male	1960-1964
14	Fred Thompson	Male	1965-1969
22	Deitmar Plank	Male	1965-1969
26	Anderson Heft	Male	1965-1969
16	Mark Fractus	Male	1970-1974
20	William Cooper	Male	1970-1974
23	Anderson Hoyt	Male	1970-1974
9	Bill Nash	Male	1975-1979
10	Albert Blackwell	Male	1975-1979
5	Alicia Freds	Female	1940-1944
31	Natasha Markhov	Female	1940-1944
~~28~~	~~Alex Long~~	~~Female~~	~~1952~~
3	Alice Brown	Female	1955-1959
29	Britney Goldman	Female	1955-1959
11	Beverly McCulsky	Female	1960-1964
18	Jane Doe	Female	1960-1964
7	Marie Kirkpatrick	Female	1965-1969
19	Nina Brown	Female	1965-1969
21	Kathy Last	Female	1965-1969
~~24~~	~~Alexandra Knight~~	~~Female~~	~~1974~~
6	Gill Stringer	Female	1975-1979
13	Freda Shields	Female	1975-1979
17	Lillian Barley	Female	1975-1979
25	Helene Arnold	Female	1975-1979
8	Leslie Hall	Female	1985-1989
30	Lisa Marie	Female	1985-1989

	QUASI-IDENTIFIERS		
ID	Gender	Year of Birth	Test Result
1	Male	1955-1959	+ve
2	Male	1960-1964	-ve
3	Female	1955-1959	-ve
4	Male	1955-1959	-ve
5	Female	1940-1944	-ve
6	Female	1975-1979	-ve
7	Female	1965-1969	+ve
8	Female	1985-1989	-ve
9	Male	1975-1979	-ve
10	Male	1975-1979	-ve
11	Female	1960-1964	-ve
12	Male	1955-1959	+ve
13	Female	1975-1979	-ve
14	Male	1965-1969	-ve

Figure-4 Disclosure of original data (El Emam and Dankar, 2008)

The privacy of medical records and private information of patients will be protected effectively with the use of the k-anonymity model. In addition to it, the overall implementation and installation cost of K-anonymity is less than other privacy models. However, this model is suitable for Dumnonia company (FeiFei et al., 2012). In this model, there is very little differentiation between quasi-identifiers which make this model highly vulnerable to some security attacks like homogeneity attack. For example, all entries of diabetes patients are masked with '#' symbol, if unauthorized entity gets the idea that '#' is used for masking sensitive information of diabetic patients then he could disclose sensitive information and compromise the privacy of all patients (Zhang et al., 2012).

I-diversity

The challenges associated with K-anonymity model can be eradicated successfully by using I-diversity value (Wang et al., 2017). Due to the homogenous nature of the model, the values are being changed frequently. Moreover, the changed values are suppressed to maintain the confidentiality and privacy of sensitive information. Hence, I-diversity helps in preserving data confidentiality by making it secret. It avoids the presence of homogenous attacks by replacing a particular data set with suitable quasi-identifier (FeiFei et al., 2012). The model's overall performance and security level are dependent on more than one models that support anonymity. I-diversity process is very time-consuming which can be considered as its biggest limitation.

T-closeness

As compared to i-diversity and k-anonymity model, T-closeness is more secure in which all data entries are arranged in a tabular format in a systematic and logical way so that it would be difficult for the end-user to interpret the meaningful and useful data from the table (Mehmood et al., 2016). In simple language, the hackers could not understand the association between the original data and quasi-identifiers which lead to high data security. This process is difficult to execute due to its high complexity.

3 DATA MODELS

For successful adoption and deployment of the k-anonymity model, there is a need to thoroughly understand and identify the data models and data sets (Wang et al., 2017). It could not be avoided because datasets including private information of patients are different from each other due to dissimilarity in data attributes. The data model consists of 3 kinds of data as discussed in the following section:

1. **Numeric**

   This type of data model consists of a number such as a date, age, postcode, etc. In the following snapshot, the numerical data is postcode and age which can be made anonymous without compromising the original data sets and its attributes. It has been found that numerical data is easier to present and understand as compared to other forms of data.

	Quasi identifier			Other attributes	Sensitive attributes
	Gender	Age	Postcode		Diagnosis
1	male	25	4350	...	depression
2	male	27	4351	...	depression
3	male	22	4352	...	flu
4	male	28	4353	...	flu
5	female	34	4352	...	depression
6	female	31	4352	...	flue
7	female	38	4350	...	cancer
8	female	35	4350	...	cancer
9	male	42	4351	...	cancer
10	male	42	4350	...	cancer
11	male	45	4351	...	cancer
12	male	45	4350	...	cancer

Figure-5 Numeric data (Li, Jin, and Yong, 2006)

2. Categorical data

In K-anonymity model, the categorical data type is also be used in which data/information can be partitioned into different segments (Zheng et al., 2018). This type of data type does not include any kind of numeric value. For example, In the following represented snapshot, the categorical data model used for categorizing gender based on male and female (sub-categories). It is revealed that numbers are not utilized in the categorical data model. In addition to this, categorized data models could be divided into different sub-values named as a single value and merged value.

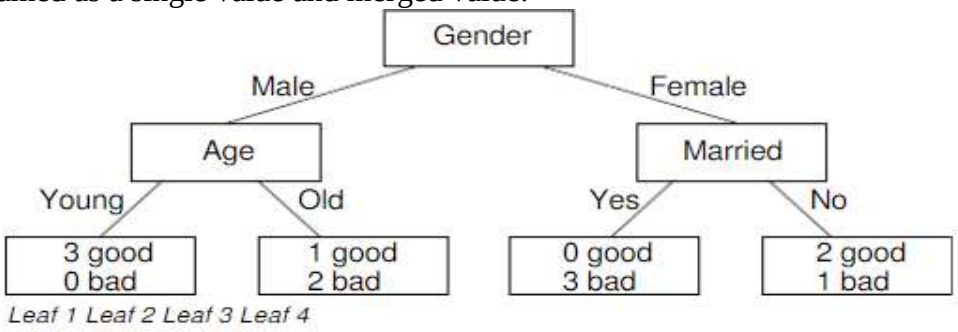


Figure-6 Categorial data for gender (Friedman, Wolff and Schuster, 2007)

Generalization hierarchies are widely used in the k-anonymity model for presenting the data sets in a more effective and precise manner. However, the concept of generalization and hierarchical structure has been followed to show categorical data set.

3.1 QUALITY METRIC

On data models and privacy models, the quality of big datasets could be ensured with the use of quality metric (Patel and Barot, 2018). This metric helps organizations to identify critical challenges related to data security, data quality, and data model efficiency. In the k-anonymity model, several types of quality metrics are utilized. The following section is going to present some quality metrics that can be used to determine the efficiency, quality and security level of the data models:

1. **Efficiency –** In recent times, efficiency has become an important quality metric that has been used for measuring and evaluating the overall quality of data model from the buyer's and end user's perspective (Mehmood et al., 2016).
2. **Data utility –** With the use of this metric, the data loss can be examined in the context of the k-anonymity privacy protection model. In addition to this, the risk of information leakage and associated leakage can also be examined effectively with the help of data utility metric (Sadhwani and Silakari, 2017).
3. **Privacy metrics –** The overall level of data confidentiality and privacy can be measured effectively with the use of privacy metrics in the context of the k-anonymity approach.

In the end, it could be summarized that anonymity model's quality can be assessed effectively based on the above-discussed metrics (Wang et al., 2017).

4 GUIDE FOR ANONYMITY IMPLEMENTATION

2 different implementations of k-anonymity

In the following section, the implementation of incognito and data fly has been compared on the basis of various parameters:

TABLE 11: ORIGINAL DATASET FOR DATAFLY ALGORITHM

BIRTH DATE	SEX	ZIP CODE	NO. OF OCCURS	TUPLE NO.
12/01/1984	M	4601	1	T1
04/04/1988	F	4888	1	T2
19/09/1989	F	4601	1	T3
27/02/1990	M	4700	1	T4
13/03/1984	M	4601	1	T5
17/05/1990	M	4700	1	T6
6	2	3		

Figure-7 Original Datafly data set (Sadhwani and Chourasia, 2017)

BIRTH DATE	SEX	ZIP CODE
1984	M	4601
1984	M	4601
1990	M	4700
1990	M	4700

Figure-8 Generalized Datafly dataset
(Sadhwani and Chourasia, 2017)

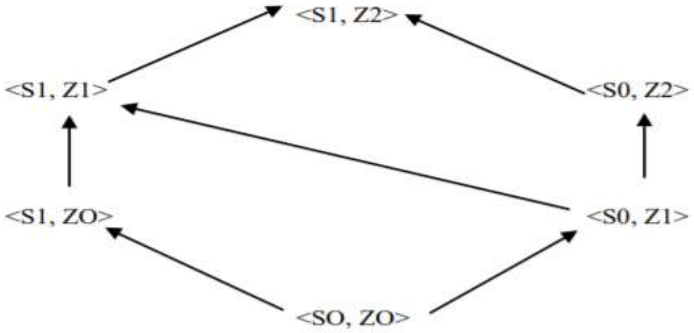


Figure-9 Output of Incognito algorithm's 2nd step
(Sadhwani and Chourasia, 2017)

Parameters	Incognito	DataFly
Dimensions	This algorithm is following only one (single) dimension. However, it is very easy to use and simple.	This model has the ability to resist and successfully deal with various types of attacks such as a homogenous attack. Hence, it is highly effective (Sadhwani and Chourasia, 2017)
Weakness	It is challenging to achieve the aims of this model.	Information loss is the major weaknesses of this model.
Property utilization	This model is abiding the concept of monotonicity that is based on two principles named as subset and generalization. Subset principle stated that if k-anonymity has not been held by specific data, it means k-anonymity will not hold by its super-set. Generalization showed that if specific dataset holds anonymity then it means its ancestor will also hold it (Zheng et al., 2018).	Generalization principles have been used by the Datafly in which specific attribute or data entry is holding anonymous data (Sadhwani and Chourasia, 2017)
Steps	It consists of only two steps. Linear fashion has been followed in the 1st step in the context of one dimension (Sadhwani and Chourasia, 2017)	This algorithm mainly consists of 6 steps as discussed below: 1. Fixation of data frequency 2. For data sets, frequencies should be stored. 3. Apply the principle of generalization 4. Continue 3rd step until distinction level has been achieved. 5. Apply suppression principle (Sadhwani and Chourasia, 2017)

5 GUIDE FOR THE IMPLEMENTATION OF K-ANONYMITY IN DUMNONIA

The critical concerns discussed by CSO, CIO, and CIO of the Dumnonia company should be considered prior to the implementation of the model. In the following section, a basic technique (algorithm) has been discussed that would be used by the company for the k-anonymity implementation:

Privacy & security

For Dumnonia company, location privacy can be considered as the biggest challenge as it performs its business operations all across New Zealand and Australia (FeiFei et al., 2012). Location privacy could affect

information sharing by the organization among different business stakeholders (Kuang et al., 2018). The organization should use k-anonymity model as it has potential to eradicate the problem of location privacy by allowing users to use same account credentials for accessing the data from anywhere. In addition to this, the extension of this model will be used by the organizations if they require high range between end-users. For example, Dumnonia organization will use p-sensitive (the extension of the k-anonymity model) to allow the users in New Zealand to access the data in a secure way (FeiFei et al., 2012).

Dumnonia company may encounter several types of attacks such as damage to corporate image, generation of fake data, etc. in the field of big data. Along with this, the overall speed of the model will be decreased if encryption technique is used by the organization to secure the data in terms of ciphertext or cryptographic codes (FeiFei et al., 2012). There is no doubt that encryption is a valuable and effective security measure but degradation of data processing speed is the major concern that could not be ignored by the organization. Without compromising the speed and efficiency of big data model, the data confidentiality and privacy could be preserved with the use of suppression principle in k-anonymity model (Patel and Barot, 2018).

Knowledge leakage

The chances of knowledge leakage are high as the organization is sharing information while operating all across New Zealand and Australia. With the adoption of the big data model, the organization will be provided with an online platform that will support smooth and instant sharing of data all across Australia but it might raise security complications. By implementing the k-anonymity model, these security and data leakage related challenges will be eradicated completely (Mehmood et al., 2016).

Cost-effectiveness

As per the case, the business executives of Dumnonia company want to implement a tool that should be affordable (Wang et al., 2017). From the performed market research, it has been found that k-anonymity is the most affordable and simplest model that does not need heavy implementation cost. However, business executives can adopt this tool due to its reasonable price and simplicity (Sadhwani and Silakari, 2017).

Extensions

To deal with the k-anonymity model's weaknesses, the extension of this model has been used. The sensitive data of customers will be protected with the utilization of the model's extension. The extension is highly customizable, simple and easy to use (FeiFei et al., 2012). The major limitation of the model extension is that it could not preserve the confidentiality of the micro datasets (Zheng et al., 2018).

REFERENCES

1) Devi, S. (2011). K-ANONYMITY: The History of an IDEA. International Journal of Soft Computing and Engineering (IJSCE), [online] 2(1). Available at: http://www.ijsce.org/wp-content/uploads/papers/v2i1/A044902211%E2%80%8B2.pdf [Accessed 20 Sep. 2019].
2) El Emam, K. and Dankar, F. (2008). Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association, 15(5), pp.627-637.
3) FeiFei, Z., LiFeng, D., Kun, W. and Yang, L. (2012). Study on Privacy Protection Algorithm Based on K-Anonymity. Physics Procedia, 33, pp.483-490.
4) Friedman, A., Wolff, R. and Schuster, A. (2007). Providing k-anonymity in data mining. The VLDB Journal, 17(4), pp.789-804.
5) Ganz, N. (2015). Data Anonymization and its Effect on Personal Privacy. [online] Available at: https://www.albany.edu/faculty/hong/pub/ganz.pdf [Accessed 20 Sep. 2019].
6) Kuang, L., Zhu, Y., Li, S., Yan, X., Yan, H. and Deng, S. (2018). A Privacy Protection Model of Data Publication Based on Game Theory. Security and Communication Networks, 2018, pp.1-13.
7) Li, J., Jin, H. and Yong, J. (2006). Current Developments of k-Anonymous Data Releasing. Proceedings of the National e-Health Privacy and Security Symposium. [online] Available at: https://eprints.usq.edu.au/1307/1/13.pdf [Accessed 20 Sep. 2019].
8) Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G. and Guo, S. (2016). Protection of Big Data Privacy. IEEE Access, 4, pp.1821-1834.

9) Patel, N. and Barot, M. (2018). Observations on Anonymization Based Privacy Preserving Data Publishing. International Journal of Recent Technology and Engineering, [online] 7(4). Available at: https://www.ijrte.org/wp-content/uploads/papers/v7i4/E1838017519.pdf [Accessed 20 Sep. 2019].
10) R Alugubelli, "DATA MINING AND ANALYTICS FRAMEWORK FOR HEALTHCARE", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.6, Issue 1, pp.534-546, February 2018, Available at :http://www.ijcrt.org/papers/IJCRT1134096.pdf
11) Sadhwani, D. and Chourasia, M. (2017). Preserving Privacy during Big Data Publishing using K-Anonymity Model – A Survey. International Journal of Advanced Research in Computer Science, [online] 8(5). Available at: https://www.ijarcs.info/index.php/Ijarcs/article/viewFile/3426/3427 [Accessed 20 Sep. 2019].
12) Sadhwani, D. and Silakari, D. (2017). Preserving Privacy during Big Data Publishing using K-Anonymity Model – A Survey. International Journal of Advanced Research in Computer Science, [online] 8(5). Available at: https://www.ijarcs.info/index.php/Ijarcs/article/viewFile/3426/3427 [Accessed 20 Sep. 2019].
13) Wang, H., Huang, H., Qin, Y., Wang, Y. and Wu, M. (2017). Efficient Location Privacy-Preserving k-Anonymity Method Based on the Credible Chain. ISPRS International Journal of Geo-Information, 6(6), p.163.
14) Zhang, J., Gong, X., Han, Z. and Feng, S. (2012). An Improved Algorithm for K-anonymity. Communications in Computer and Information Science, pp.352-360.
15) Zheng, L., Yue, H., Li, Z., Pan, X., Wu, M. and Yang, F. (2018). k-Anonymity Location Privacy Algorithm Based on Clustering. IEEE Access, 6, pp.28328-28338.