Data Anonymization using Big Data Frameworks

Mounica.B[1], Butti Pavithra[2], Tejawini.V[3], Kokila.N[4]
[1]Asst. Professor, Department of Information Science Engineering, New Horizon College of Engineering
[2,3,4]Student, Dept. Of ISE, New Horizon College of Engineering, Bangalore, India

Abstract-Privacy is the main issue that is faced in Big data especially which includes the sensitive data. When it comes for sharing or publishing of these data then the privacy breach is occurred to overcome this issue we are focusing on the anonymization of the data. There are two generally-embraced ways to overcome privacy breach they are Data Encryption and Anonymization. Our goal is to achieve zero tolerance security on a significant amount of data. This paper mainly speaks about the data anonymization using the bigdata frameworks and Cassandra database.

Keywords-Anonymization, Bigdata, Privacy, Cassandra, Spark, Kafkha.

1.Introduction

Big data has brought revolution in the world of data analytics. Data which was discarded few years ago is now considered to be a powerful asset[1].This poses a very serious security concern. As big data contains individual specific information privacy is a major security concern. In the recent years because of increase in ability to store personal information regarding the user the problem of privacy-preservation has tend to become more important. There are number of anonymization techniques to perform privacy-preservation in data mining.

According to recent Forrester study, 80 percent of breaches in data security who put information at risk involve employees or those with internal access to an organization. The biggest challenge today for companies is to preserve data form such people. For example, database users are assigned a Database Administrator (DBA) role and are granted with different system privileges. To ensure the privacy and integrity of corporate information today many companies are using a much comprehensive security approach. A database having a huge dataset with high dimension data should be secure to maintain personal data private from the world. The most important concern today is to protect sensitive information from getting disclosed or misinterpreted [5]. There are two types of

disclosures- Identity Disclosure and Attribute Disclosure. Attribute Disclosure and Identity Disclosure. Attribute disclosure is the one in which crucial information is inferred regarding a person from published data. When an individual entity can be distinguished from the published data then it is Identity disclosure [2]. Data anonymization techniques enables publication of detailed information. In this study we focus on Rjindeal Algorithm as it is one of the best algorithm that can be used for the encryption of the data, here we are Encrypting the sensitive data that has been sent from one end to the other, so that the sensitive information won't be available and it can't be theft as well**.**

The Rijndael algorithm is a new generation symmetric method of encrypting text in which a cryptographic key and algorithm are applied to a block of data at once as a group rather than to one bit at a time.

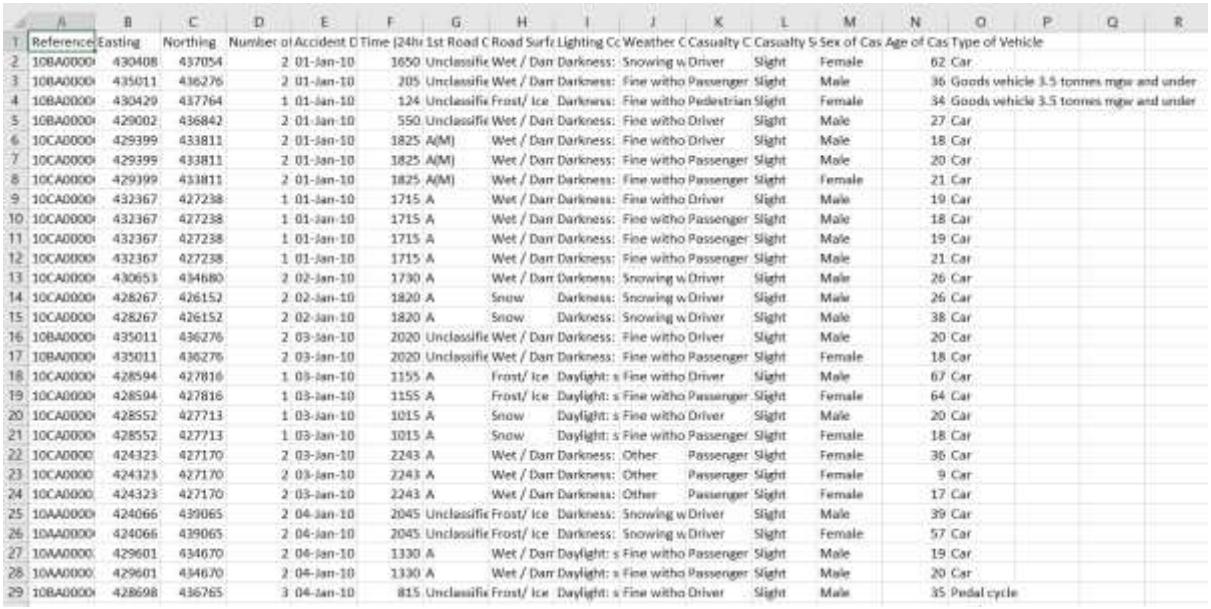
Rijndael uses a variable number of rounds, depending on key/block sizes, as follows:

9 rounds if the key/block size is 128 bits
11 rounds if the key/block size is 192 bits
13 rounds if the key/block size is 256 bits

Rijndael is a substitution linear transformation code, not requiring a Feistel network. It use three dstrategict invertible uniform transformations. these three tranforms are: Linear Mix Transform; Non-linear Transform and Key Addition Transform.

2. Data Analysis:

The dataset that we have used in this is an accident traffic data set. As the name of the dataset describes the type of the data set it is. As it is a very Huge Data Set, which will have a collection of the list of the accident that have been occurred. It is a structured dataset, in this dataset there are many columnar, where it described the details of the accidents that have been occurred, it also gives the description of the accident occurred. There is few sensitive information in this dataset like the driver license that must not be available for everyone, even the Driver Name, Casualty Name, Casualty License Number the car details of the both, here these Data Columns should be anonymized, where we are Encrypting the sensitive dataset using the appropriate algorithms.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Reference	Easting	Northing	Number of	Accident D	Time (24hr	1st Road C	Road Surfa	Lighting Cc	Weather C	Casualty C	Casualty S	Sex of Cas	Age of Cas	Type of Vehicle			
2	10BA0000	430408	437054	2	01-Jan-10	1650	Unclassifie	Wet / Dan	Darkness:	Snowing w	Driver	Slight	Female	62	Car			
3	10BA0000	435011	436276	2	01-Jan-10	205	Unclassifie	Wet / Dan	Darkness:	Fine witho	Passenger	Slight	Male	36	Goods vehicle 3.5 tonnes mgw and under			
4	10BA0000	430429	437764	1	01-Jan-10	124	Unclassifie	Frost/ Ice	Darkness:	Fine witho	Pedestrian	Slight	Female	34	Goods vehicle 3.5 tonnes mgw and under			
5	10BA0000	429002	436842	2	01-Jan-10	550	Unclassifie	Wet / Dan	Darkness:	Fine witho	Driver	Slight	Male	27	Car			
6	10CA0000	429399	433811	2	01-Jan-10	1825	A(M)	Wet / Dan	Darkness:	Fine witho	Driver	Slight	Male	18	Car			
7	10CA0000	429399	433811	2	01-Jan-10	1825	A(M)	Wet / Dan	Darkness:	Fine witho	Passenger	Slight	Male	20	Car			
8	10CA0000	429399	433811	2	01-Jan-10	1825	A(M)	Wet / Dan	Darkness:	Fine witho	Passenger	Slight	Female	21	Car			
9	10CA0000	432367	427238	1	01-Jan-10	1715	A	Wet / Dan	Darkness:	Fine witho	Driver	Slight	Male	19	Car			
10	10CA0000	432367	427238	1	01-Jan-10	1715	A	Wet / Dan	Darkness:	Fine witho	Passenger	Slight	Male	18	Car			
11	10CA0000	432367	427238	1	01-Jan-10	1715	A	Wet / Dan	Darkness:	Fine witho	Passenger	Slight	Male	19	Car			
12	10CA0000	432367	427238	1	01-Jan-10	1715	A	Wet / Dan	Darkness:	Fine witho	Passenger	Slight	Male	21	Car			
13	10CA0000	430653	434680	2	02-Jan-10	1730	A	Wet / Dan	Darkness:	Snowing w	Driver	Slight	Male	26	Car			
14	10CA0000	428267	426152	2	02-Jan-10	1820	A	Snow	Darkness:	Snowing w	Driver	Slight	Male	26	Car			
15	10CA0000	428267	426152	2	02-Jan-10	1820	A	Snow	Darkness:	Snowing w	Driver	Slight	Male	38	Car			
16	10BA0000	435011	436276	2	03-Jan-10	2020	Unclassifie	Wet / Dan	Darkness:	Fine witho	Driver	Slight	Male	20	Car			
17	10BA0000	435011	436276	2	03-Jan-10	2020	Unclassifie	Wet / Dan	Darkness:	Fine witho	Passenger	Slight	Female	18	Car			
18	10CA0000	428594	427816	1	03-Jan-10	1155	A	Frost/ Ice	Daylight: s	Fine witho	Driver	Slight	Male	67	Car			
19	10CA0000	428594	427816	1	03-Jan-10	1155	A	Frost/ Ice	Daylight: s	Fine witho	Passenger	Slight	Female	64	Car			
20	10CA0000	428552	427713	1	03-Jan-10	1015	A	Snow	Daylight: s	Fine witho	Driver	Slight	Male	20	Car			
21	10CA0000	428552	427713	1	03-Jan-10	1015	A	Snow	Daylight: s	Fine witho	Passenger	Slight	Female	18	Car			
22	10CA0000	424323	427170	2	03-Jan-10	2243	A	Wet / Dan	Darkness:	Other	Passenger	Slight	Female	36	Car			
23	10CA0000	424323	427170	2	03-Jan-10	2243	A	Wet / Dan	Darkness:	Other	Passenger	Slight	Female	9	Car			
24	10CA0000	424323	427170	2	03-Jan-10	2243	A	Wet / Dan	Darkness:	Other	Passenger	Slight	Female	17	Car			
25	10AA0000	424066	439065	2	04-Jan-10	2045	Unclassifie	Frost/ Ice	Darkness:	Snowing w	Driver	Slight	Male	39	Car			
26	10AA0000	424066	439065	2	04-Jan-10	2045	Unclassifie	Frost/ Ice	Darkness:	Snowing w	Driver	Slight	Female	57	Car			
27	10AA0000	429601	434670	2	04-Jan-10	1330	A	Wet / Dan	Daylight: s	Fine witho	Passenger	Slight	Male	19	Car			
28	10AA0000	429601	434670	2	04-Jan-10	1330	A	Wet / Dan	Daylight: s	Fine witho	Passenger	Slight	Male	20	Car			
29	10BA0000	428698	436765	3	04-Jan-10	815	Unclassifie	Frost/ Ice	Daylight: s	Fine witho	Driver	Slight	Male	35	Pedal cycle			

Figure1: The Dataset Diagram

3.Result and Analysis:

As privacy is one of the breach that we having in the big data. Thus, we have come up with data anonymization technic, where we are using the latest technologies so that the process can be easy, fast, reliable. Here we are using the Rjindael algorithms to anonymize the data

There are two ends named Producer and the Consumer, where the data from the Producer is transferred to the Consumer in the Apache Kafka.

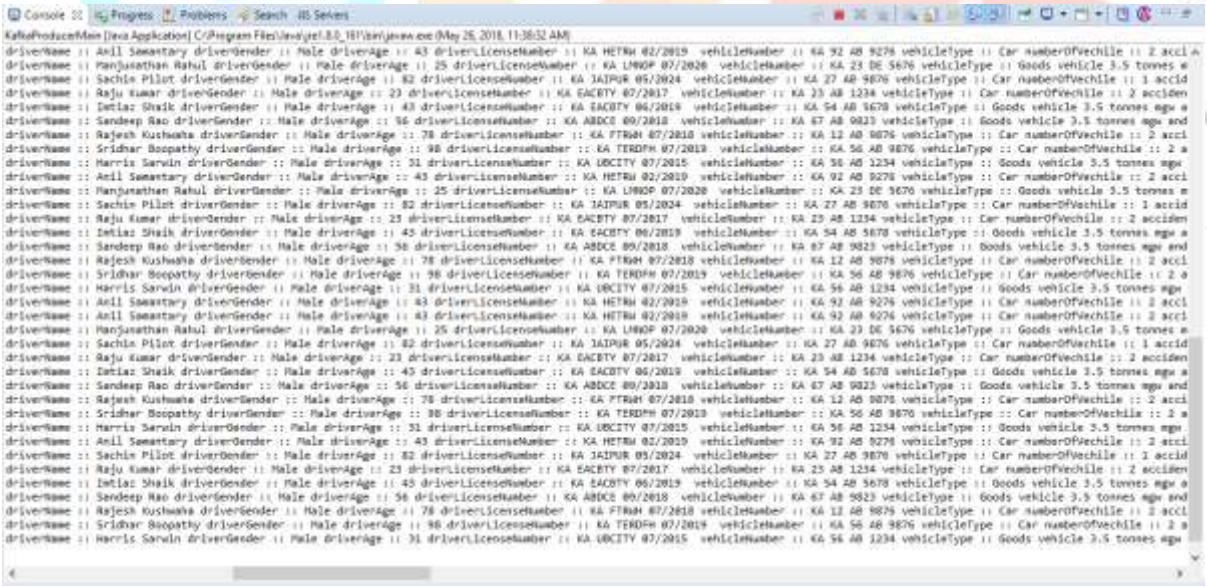


Figure2: the data sent from the Producer

The data sent from the producer is been sent as the way it is been stored in the dataset

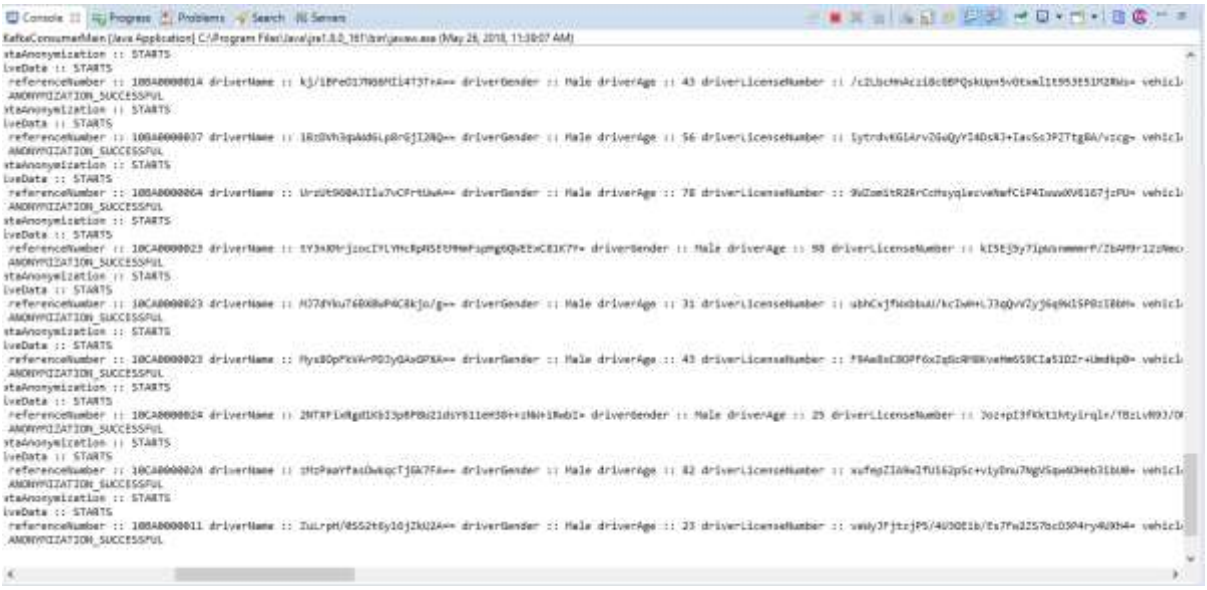


Figure3: the data received at the consumer end

The data received at the Consumer, has the sensitive data which have been Encrypted.The Encrypted data are being stored in the same format in the database at Cassandra.

4.Conclusion:

From the present study we can conclude that, the privacy breach can be minimized with the help of the encryption techniques which we have used. With the help of this the concern of each client on his data, sensitive data's can be safe from the data theft or misuse of the data can also be reduced. As well these changes can also be implemented on the real time object. Where the latest technologies can be used to make it faster and more flexible to use.

References:

1. Latanya Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557570, 2002
2. Turban and J.E. Aronaon. Decision support Systems and Intelligent Systems, Prentice-Hall, New Jersey, USA, 2001
3. A Study on the Impact of Data Anonymization on Anti-discrimination ->S. Hajian Dept. of Comput. Eng. & Math., Univ. Rovira i Virgili, Tarragona, Spain J. Domingo-Ferrer Dept. of Comput. Eng. & Math., Univ. Rovira i Virgili, Tarragona, Spain 2012 IEEE 12th International Conference on Data Mining Workshops
4. PRIVACY preservation in big data using anonymization techniques- Tanashri Karle Vidyalankar Institute of Technology, Mumbai, India Deepali Vora Vidyalankar Institute of Technology, Mumbai, India 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)
5. Sensitivity-Based Anonymization of Big Data -Mohammed Al-Zobbi Sch. of Comput., Eng. & Math., Western Sydney Univ., Sydney, NSW, Australia Seyed Shahrestani Sch. of Comput., Eng. & Math., Western Sydney Univ., Sydney, NSW, Australia Chun Ruan
6. Privacy preserving data publishing and data anonymization approaches: A review-Puneet Goswami SRM University, Delhi-NCR Campus, Sonepat, India Suman Madan JIMS, Sec-5, Rohini, Delhi India 2017 International Conference on Computing, Communication and Automation (ICCCA)
7. BigCrypt for big data encryption-Abdullah Al Mamun Computer Science and Engineering Department, Qatar University, Qatar Khaled Salah Electrical and Computer Engineering Department, Khalifa University of Science, Technology & Research, UAE Somaya Al-maadeed Computer Science and Engineering Department, Qatar University, Qatar 2017 Fourth International Conference on Software Defined Systems (SDS)
8. Hadoop eco system for big data security and privacy-Pradeep Adluru Computer Science Department, College of Staten Island, CUNY, 2800 Victory Blvd, Staten Island, NY 10314
9. Srikari Sindhoori Datla Computer Science Department, College of Staten Island, CUNY, 2800 Victory Blvd, Staten Island, NY 10314 Xiaowen Zhang Computer Science Department, College of Staten Island, CUNY, 2800 Victory Blvd, Staten Island, NY 10314 2015 Long Island Systems, Applications and Technology
10. Large-Scale Encryption in the Hadoop Environment: Challenges and Solutions- Debnath Bhattacharyya Department of Computer Science and Engineering, Vignan's Institute of Information Technology, Visakhapatnam, India
11. Sudipta Roy Department of Computer Science and Engineering, U.V. Patel College of Engineering, Ganpat University, Mehsana, India
12. Raj R. Parmar Department of Computer Science and Engineering, U.V. Patel College of Engineering, Ganpat University, Mehsana, India
13. Study on encryption methods to secure the privacy of the data and computation on encrypted data present at cloud-H. R. Nagesh Department of Computer Science & Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, India

14. L Thejaswini Department of Computer Science & Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, India 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)
15. Comparative study of encryption algorithm over big data in cloud systems-M. Padmavathamma S.V.U College of Commerce Management &Computer Sciences K. Sekar
16. S. V. Engineering College for Women, Tirupat 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)
17. A Secure Cloud Computing Based Framework for Big Data Information Management of Smart Grid- Joonsang Baek Electrical and Computer Engineering Department, Khalifa University of Science, Technology and Research, Abu Dhabi, UAE
18. Quang Hieu Vu Etisalat BT Innovation Centre (EBTIC), Khalifa University of Science, Technology and Research, Abu Dhabi, UAE
19. Joseph K. Liu Department of Info comm Security (ICS), Institute for Info comm Research, Singapore IEEE Transactions on Cloud Computing (Volume: 3, Issue: 2, April-June 1 2015)
20. An Identity-Based Security Scheme for a Big Data Driven Cloud Computing Framework in Smart Grid-Feng Ye Dept. of Electr. & Comput. Eng., Univ. of Nebraska-Lincoln, Lincoln, NE, USA Yi Qian Dept. of Electr. & Comput. Eng., Univ. of Nebraska-Lincoln, Lincoln, NE, USA Rose Qingyang Hu Dept. of Electr. & Comput. Eng., Utah State Univ., Logan, UT, USA 2015 IEEE Global Communications Conference (GLOBECOM)