sciendo

VOLUME 17, N° 1 2023
Journal of Automation, Mobile Robotics and Intelligent Systems

Automated Anonymization of Sensitive Data on Production Unit

Submitted: 11th January 2022; accepted: 7th September 2022

Marcin Kujawa, Robert Piotrowski

DOI: 10.14313/JAMRIS/1-2023/5

Abstract:
The article presents an approach to data anonymization with the use of generally available tools. The focus is put on the practical aspects of using open-source tools in conjunction with programming libraries provided by suppliers of industrial control systems. This universal approach shows the possibilities of using various operating systems as a platform for process data anonymization. An additional advantage of the described approach is the ease of integration with various types of advanced data analysis tools based both on the out-of-the-box approach (e.g., business intelligence tools) as well as customized solutions. The discussed case describes the anonymization of data for the needs of sensitive analysis by a wider group of recipients during the construction of a predictive model used to support decisions.

Keywords: *Data anonymization, Sensitive data, open-source tools, Industrial data processing, Historian data anonymization, Honeywell DCS, IT/OT integration, Operational technology*

1. Introduction

In recent years, cybersecurity has been gaining increasing importance both in the everyday life of an average Facebook user as well as, if not mainly, in the industrial world. Companies often standing at the forefront of the industrial revolution 4.0 and handling large volumes of the data must ensure that data is adequately protected against unauthorized access or analysis.

Research activities related to data anonymization nowadays have big impact on 4.0 industry. Previous studies reported various structures and technologies of data anonymization [1, 3], such as encryption [2], continued data delivery [4], and adaptive predictive models [5]. Various techniques of data protection, such as disturbance, anonymization, and cryptography, are described in the literature [6]; however, these present a different approach from the one presented in this paper. Samad et al. [7] also describes the use of anonymized data and their use in the context of artificial intelligence and machine learning methods.

This article presents an approach to data anonymization with the use of publicly available tools and physical equipment, using the example of process data from a petrochemical plant's production installation. In addition to legal safeguards related to data protection, it is worth spending some time to maintain a competitive advantage. In this case, the focus is on cryptographic techniques that hinder the processing of unauthorized data. When working on a project with sensitive data, it is often necessary to transfer data to external entities, which results in a specific approach to data. The aim of the discussed case is to make the data available to a research university by a national-scale critical production facility.

The analysis of various types of solutions was determined by reservations as to the completeness of the security of data collected and secured with the use of accessible services on the market. When you use cloud services, for example, one of your platform's managed disks handles encryption and decryption in a fully transparent manner, using envelope encryption. This encrypts data using a data encryption key, which is in turn protected with keys. However, when an attempt is made to verify by an unauthorized user, no access to data is permitted; similarly, full access to data by authorized users is permitted. It is not possible to verify whether the access to data is limited, and the data is stored in open form, whether or not the data is encrypted.

The remainder of this paper is organized as follows. The method of downloading raw data is described in Section 2. The process of data anonymization is presented in Section 3. The result of the analysis is illustrated in Section 4. Concluding remarks are listed in the last section.

2. Technical Aspects

Data are collected from an industrial facility using the available information technology (IT) systems included in the IT / operational technology (OT) layered model shown in Figure 1. It is a general structural outline for the Honeywell Distributed Control System (DCS) system on which the analyses were performed.

The value measured by the analog/digital (A/D) converter (see Level 0) is processed by the DCS controllers (see Level 1). Depending on the system configuration, it can be stored by the system server for the purposes of visualization and ongoing data analysis at operator stations (see Level 2). If long-term data archiving is necessary from the company's point of view, it is necessary to implement the so-called

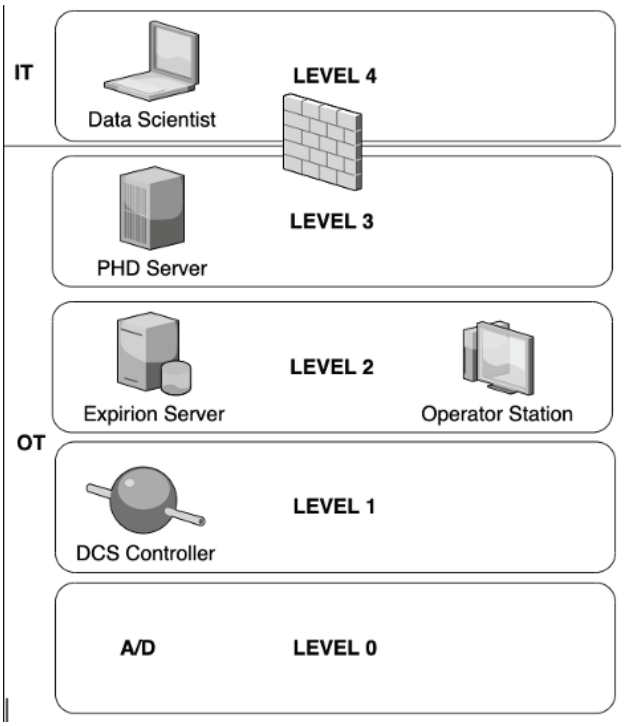

Figure 1. Industrial network layers

"historian" in Level 3. For the purposes of artificial intelligence (AI) and machine learning (ML), significant amounts of data should be processed. The exact amount depends on the methods used. The data scientist can operate on data at level 4 in the IT environment.

3. Data Anonymization

The article presents one of the possible approaches to data anonymization. In the following sections, the tools necessary to download and process data, as well as process automation, will be discussed.

3.1. Environment

To properly approach the topic of collecting, anonymizing, and processing data for the purposes of building models of the analyzed objects, many tools need to be applied. The main ones are:

- Visual Studio 2019: C# development environment,
- PyCharm: Python development environment, to generate the Python code from which the publicly available libraries for data encryption come from, which is described in more detail in Section 3.4,
- application integration library for data historization from the historian application supplier,
- Anaconda: an environment for running Python applications on the Windows operating system.

Linux or Mac-based programs can be used without installing any additional software if the company policy allows it.

3.2. Data pre-processing

Regardless of the vendor and architecture of the system from which we obtain the data, special attention should be paid to the appropriate preparation of data in several aspects:

- Data ownership,
- Data security,
- Data processing.

To obtain data in a proper way, it is necessary to regulate the approach to data ownership. In large companies, often there are internal procedures regulating legal aspects depending on the type of data, for example, sensitive data, or personal data. Moreover, to make sure that the approach to sharing data is correct, it is worth consulting the SOC Department, if the company has one. The Security Operational Center (SOC) opinion is also important when it comes to data classification and possible restrictions related to General Data Protection Regulation (GDPR).

The production plant from which data have been collected is based mainly on Honeywell production solutions, therefore process data have been collected via a historian (Honeywell Uniformance Process History Database, a non-sql database). To download data a dedicated library has been used which enables the use of the C# language to create an application that aggregates data in an appropriate manner. In this way, the obtained data in an open form are transferred for encryption. It was decided not to use the available programming tools in the .Net environment for data anonymization due to the gradual expansion of the system towards a decision support system and direct integration with DCS. To maintain consistency in the target solution, data anonymization was performed using Python.

3.3. Data Encryption Algorithm and Use of Open-Source Libraries

In the discussed case, publicly available Werkzeug [8, 9] libraries were applied as starting elements. These libraries are primarily used for user authentication in web applications created in Python.

The unconventional use of libraries allows one to obtain code ensuring data anonymity by using cryptographic techniques in a quick and transparent manner.

The diagram of data handling is presented in Figure 2. The applied algorithm guarantees checking the system status, which includes:

- correct communication between the application server and the historian server;
- reliability of the data determined by the historian.

In the first step, starting from the block described as "start," we download data from the historian (PHD – Process History Database). In the next step, the system status is checked from the side of the correctness of system connections, and when it is correct, we can proceed to the next steps leading to saving anonymized data based on data download. The data confidence level should then be checked as an internal quality parameter for the PHD system. If the collected data has sufficiently high confidence, we can proceed to the steps characteristic for proper data anonymization. In the next steps, we collect the hardware key used in the code, the process point name

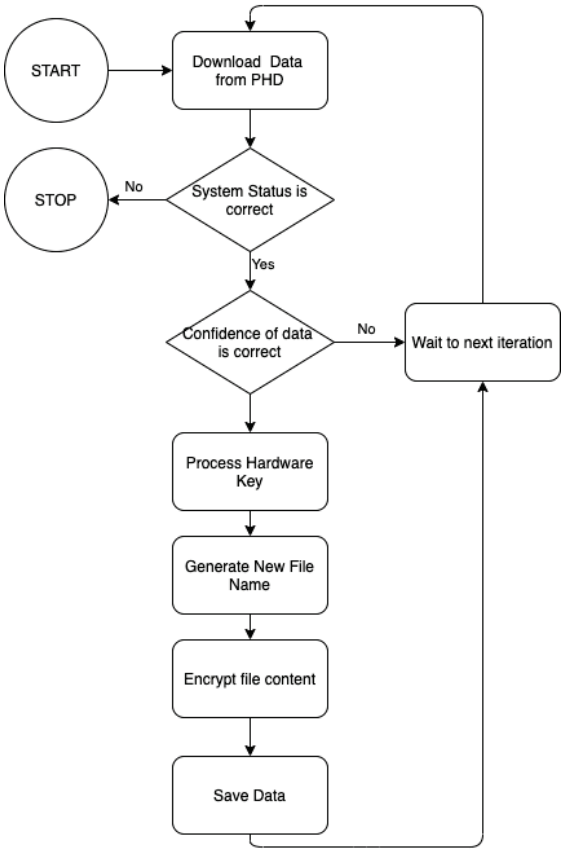

Figure 2. Overall data processing schema

is generated (in the first run, in the next steps). The last step is to encrypt the file content and save it in an anonymous form. Then we wait for the next iteration and start the process anew. Such anonymization of data ensures two-dimensional verification as part of reading anonymous data. It is necessary to know the hardware key for the processed data and to have a list of points because the generated variable name cannot be processed backwards to the input string. It is only possible to determine whether the anonymous variable string comes from a given input string.

3.4. Automatization of Data Anonymization

The main advantage of automating data encryption is reducing the amount of data processed at a time, which allows using anonymization on an ongoing basis. The use of the presented algorithm allows for download of current data and updating the output files, despite the variable file name at each application run. Changing the file name with each access, consisting in adding successive lines containing encrypted measurement data, additionally increases data security.

The described operations take place on the application server presented in Figure 3.

In the OT section, you can see three elements: the operator station, DCS server, and PHD server. Data is downloaded via the PHD server, that is, the server responsible for long-term archiving of process data. This approach guarantees the additional separation of the external system from the critical systems such

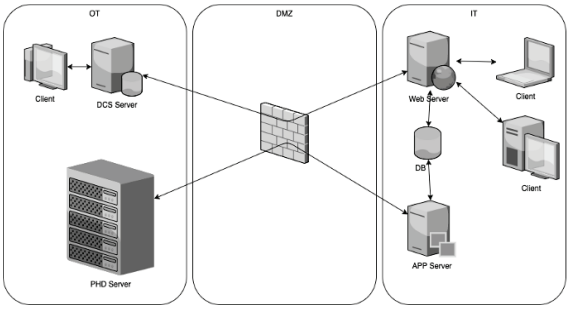

Figure 3. Diagram of the data anonymization and decision support system

as distributed control systems. An additional isolating element is a centrally located firewall in the DMZ zone separating IT from OT. Data separated in this way is downloaded by the application server (APP Server), whose purpose is to continuously download the current data and subject it to anonymization and saving to flat files to make it available to external entities. Another function of the application server is the use of data to supply predictive models where the prediction result is saved in a database (DB). To build a universal visualization layer, a visualization of data previously read from the database was implemented on the prepared server hosting the web applications. An additional advantage is the possibility of using the same server to service the operator station, that is, the client on the OT side.

3.5. Data Provided for Experiment

In production installations, each of the measurements is defined in an unambiguous way that allows its identification on the scale of the entire production plant and the type of measured physical quantity. During the experiment, the focus was on representative temperature measurements. These measurements have been renamed for the article. In the naming convention used, the first three digits symbolize the number of the production installation, the letter in the fourth position symbolizes a measured physical quantity (T – temperature, P – pressure, F – flow), the letter in the fifth position indicates the DCS point type (I – Indication, C – controller), the character after underscore is a unique point number within the production facility. Table 1 presents the values of data increase concerning the original files for representative measurements. All discussed measurements saved in flat files represent the same time range (they have the same number of lines). The initial volume of files may be different due to different measuring points and different scale of measured values. Based on the presented Table 1, a non-linear data increase rate can be noticed.

3.6. Changing the Length of Encryption Key

Changing the file content encryption type may affect the smooth operation of the system and its resource capacity. Figure 4 shows the time needed to encrypt the content of representative files in relation to various encryption algorithms (Secure Hash

Table 1. File attributes

Name before anonymization	Name after anonymization	Number of lines / File Size Before Anonymization [KB] / File Size After Anonymization [KB] / Size increase [%]
001TI_001A	qel6sW7q$027183e8f1289e8d31bd53262ddc4be8418a013517 d83507c09832b8e10c35e5	8126/299/ 992/331,78
001TI_001B	HbWQjw9W$134b5e2b3417d529bf40c392722c08c7e970b784fb 64ad4380ef5479fc1e0334	8126/283/ 992/350,53
001TI_002	YqWNmnUI$4d76684336497c7eaffaa70dcdecb548d4e69274a14 93010cf0d4e2ef056c889	8126/275/ 992/360,73
001TI_003	E50OuzmP$e55ad3917b6586527c2e143e67fcc183e39c9057023 d0384b65ce5d7550a8f5b	8126/275/ 992/360,73
001TI_004	SjYI5v72$b7864a9b8d42f9365ba99a057e6bb14da13e25735de dcebe29792051c3c0a31b	8126/275/ 992/360,73
001TI_005	6tm5q8Wo$de72d08788095917f16f95cac3e6010512c16929935 78db9dde91e1ff98a5bf1	8126/275/ 992/360,73
001TI_006	G7mTL5ZW$7fe898c919e78e47e195ab9f42d27c2bdddb0ab1b94 1f0df968ce2437261b6b7	8126/275/ 992/360,73

Table 2. SHA algorithms

Variant	Algorithm	Output Size in bits/ Internal State Size in bits/ Block size in bits
SHA 256	SHA 2	256/256/512
SHA 1	SHA 1	160/160/512
SHA 224	SHA 2	224/256/512
SHA 384	SHA 2	384/512/1024
SHA 512	SHA 2	512/512/1024
SHA 3-224	SHA 3	224/512/1152
SHA 3-256	SHA 3	256/512/1088
SHA 3-385	SHA 3	385/1024/832
SHA 3-512	SHA 3	512/1024/576
SHA 512-224	SHA 2	224/512/1024
SHA 512-256	SHA 2	256/512/1024

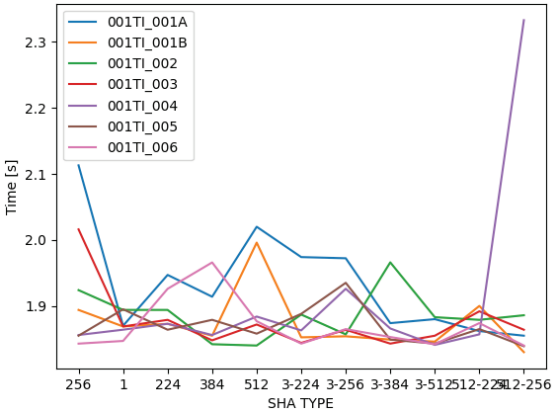

Figure 4. Change the type of data encryption inside a batch file

Algorithm – SHA). The influence of the encryption algorithm on the time of anonymization operations was presented in Table 2.

By introducing different encryption methods into the algorithm, we obtain different algorithm circulation times, which has a direct impact on the file processing time. An interesting phenomenon is the different effectiveness of encryption algorithms depending on the input data, which can be seen in Figure 4.

4. Protection of Access

The data encryption itself can be treated as one of the layers of data access security. However, in many cases it is worth using many-factor authentication based on software and hardware security. For this reason, data is stored on an encrypted drive, in accordance with corporate policy. Moreover, possible access to a physically encrypted data storage device is granted by RestApi using authentication methods based on a registered user using a variable token with a defined validity for double authentication.

5. Conclusion

When anonymizing data, attention should be paid to the selection of appropriate tools for the current needs and the data being processed. Care should be taken that the methodologies used are not redundant in relation to the risks they are intended to counteract. This approach will allow you to choose the optimal solution in terms of financial as well as acceptable from the side of security control.

When selecting the appropriate encryption algorithm, pay attention to the following factors:

- Regardless of the scientific aspect, the guidelines of the internal security services play a decisive role in choosing the method of securing sensitive data in enterprises of critical infrastructure. Therefore, the development team and the security team should work closely together.
- Costs related to the consumption of computing resources – Despite the relatively efficient work of the algorithm with data incrementally (continuously),

when we process large volumes of data, initial costs may accumulate.

- Costs related to the storage and transfer of data should be obtained, regardless of whether the model is cloud-based or based on local infrastructure.

Current IT solutions allow the use of scalability of systems and applications to optimize costs. Moreover, thanks to the possibility of a wide use of open-source tools and applications, we can use resources more effectively thanks to the experience of the community. In this case, the data encryption itself should be seen only as one of the security layers, because the IT security department will have the decisive opinion, and the discussed algorithms can only be a proposal to supplement one of the data security layers.

AUTHORS
Marcin Kujawa – Gdańsk University of Technology, Faculty of Electrical and Control Engineering, Poland, e-mail: marcin.kujawa2@pg.edu.pl.
Robert Piotrowski* – Gdańsk University of Technology, Faculty of Electrical and Control Engineering, Poland, e-mail: robert.piotrowski@pg.edu.pl.
*Corresponding author

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science and Higher Education (now: Ministry of Education and Science) under the "Implementing Doctorate" programe, No. DWD/3/41/2019. The authors wish to express their thanks for the support.

References

[1] Sánchez, D., J. Soria-Comas, and J. Soria-Comas. *Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings*. New Zealand, 2019.

[2] Nabywaniec, D. *Anonymisation and masking of sensitive data in companies*. ISBN: 978-83-283-5681-8 (in Polish), 2019.

[3] Devaux, E. *How "anonymous" is anonymized data?* https://medium.com/statice/how-anonymous-is-anonymous-c92ad265a3e3 (2021-11-20).

[4] Dholakia, J. *Building Python apis with flask, flask-restplus and swagger ui*. https://medium.com/analytics-vidhya/swagger-ui-dashboard-with-flask-restplus-api-7461b3a9a2c8 (2021-11-20).

[5] Grus J. *Data Science from Scratch: First Principles with Python*, ISBN: 978-83-283-4603-1, 2018.

[6] Murthy, S., A. Abu Bakar, F. Abdul Rahim, and R. Ramli. *A Comparative Study of Data Anonymization Techniques*. 2019 IEEE 5th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), 2019, pp. 306–309, doi: 10.1109/BigDataSecurity-HPSC-IDS.2019.00063.

[7] Samad, A.A.; Arshad, M.M.; Siraj, M.M. *Towards Enhancement of Privacy-Preserving Data Mining Model for Predicting Students' Learning Outcomes Performance*. 2021 IEEE International Conference on Computing (ICOCO), 2021, pp. 13–18, doi: 10.1109/ICOCO53166.2021.9673544.

[8] Kenedy, P. *What is werkzeug?*. https://testdriven.io/blog/what-is-werkzeug/ (2021-11-20).

[9] Grinberg, M. *Flask Web Development: Developing Web Applications with Python, 2nd edition*, ISBN: 978-83-283-6384-7, 2020.