A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-81

HS.Register – An Audit-Trail Tool to Respond to the General Data Protection Regulation (GDPR)

Duarte GONÇALVES-FERREIRA [a1], Mariana LEITE [a,b] and
Cátia SANTOS-PEREIRA [b] Manuel E. CORREIA[b,c] Luis ANTUNES[b,c]
Ricardo CRUZ-CORREIA[a,b]
[a]*CINTESIS, Porto, Portugal*
[b]*HLTSYS - HealthySystems LDA, Portugal*
[c]*Faculty of Sciences University Porto, Portugal*

Abstract. Introduction The new General Data Protection Regulation (GDPR) compels health care institutions and their software providers to properly document all personal data processing and provide clear evidence that their systems are inline with the GDPR. All applications involved in personal data processing should therefore produce meaningful event logs that can later be used for the effective auditing of complex processes. Aim This paper aims to describe and evaluate HS.Register, a system created to collect and securely manage at scale audit logs and data produced by a large number of systems. Methods HS.Register creates a single audit log by collecting and aggregating all kinds of meaningful event logs and data (e.g. ActiveDirectory, syslog, log4j, web server logs, REST, SOAP and HL7 messages). It also includes specially built dashboards for easy auditing and monitoring of complex processes, crossing different systems in an integrated way, as well as providing tools for helping on the auditing and on the diagnostics of difficult problems, using a simple web application. HS.Register is currently installed at five large Portuguese Hospitals and is composed of the following open-source components: HAproxy, RabbitMQ, Elasticsearch, Logstash and Kibana. Results HS.Register currently collects and analyses an average of 93 million events per week and it is being used to document and audit HL7 communications. Discussion Auditing tools like HS.Register are likely to become mandatory in the near future to allow for traceability and detailed auditing for GDPR compliance.

Keywords. GDPR, Audit log, HL7, ATNA

1. Introduction

The General Data Protection Regulation (GDPR) is a set of regulations for strengthening data protection laws in Europe. It becomes effective on May of 2018 and applies to organizations processing personal data in the EU, with a special mention to data concerning health. Under the GDPR, institutions have the obligation of demonstrating accountability for the fulfilment of the regulation requirements, which relies on their ability to demonstrate that appropriate procedural an security measures are being applied and, most importantly, that they are compliant with GDPR. This creates great pressure

[1] Corresponding Author, Duarte Gonçalves-Ferreira, E-mail: dferreira@fe.up.pt

on health care institutions, namely hospitals, and software producers to provide auditable traceability mechanisms for their current and legacy systems [1].

Traceability in Portugal is particularly difficult as there are on average more than 21 software applications per hospital, leading to great heterogeneity and need for interoperability [2]. Also, at least until 2013, the existing logs had severe quality issues that made it difficult to guarantee traceability. Existing audit trail standards (e.g. ASTM:E2147, ISO/TS 18308:2004, ISO/IEC 27001:2006) were still not broadly used back then [3] and, to the best of our knowledge, this has not yet changed. The same scenario is observable in inter-institution information flows[4].

The exchange of patient data among healthcare institutions is also very relevant in the context of the GDPR. In 2016, Pinto et al. built a collective vision of existing institutions at the Portuguese National Health Service and their Information Systems interactions. This study allowed the identification of about 50 recurrent interaction processes, which were classified into four different varieties, in accordance with the nature of their information flow: administrative, clinical, identity and statistical. The authors considered an effort should be made to provide the various institutions with guidelines/interfaces regarding communication and prompt such institutions to elaborate upon these [4].

Moreover the stability and availability of systems and applications are also an important issue. IT staff needs to have access to immediate secure and reliable sources of information regarding how systems are working. Specially information about problems with the core systems that support the services inside the hospital and that can directly affect the treatment of a patient, including security issues like data breaches or unauthorized access to data. Hospitals also need to have a better understanding of who, when, why, how and what data was accessed both by humans and other systems.

HS.Register is a software that was designed to cope with these problems inside hospitals, namely to tackle the lack of knowledge about what goes on and who does what with the systems and legacy applications that the hospitals use. The main objective was to centralize at scale various data sources so that Information Technologies (IT) teams could readily analyse in detail their systems and applications, by tracing how data is accessed and shared and being able to audit the systems that support the Hospitals. This paper aims to describe and evaluate HS.Register, a system developed to cope with GDPR auditing requirements in highly heterogeneous environment like the health care sector.

2. Methods

HS.Register was initially designed as a solution to store HL7 messages, logs and systems events inside an hospital infrastructure. To accomplish this objective: data should not be updatable and be analysable; it should be highly scalable and performant, other systems should not be affected; registered events should be non-refutable and non-removable; and data should be auditable and traceable.

With this requirements in mind we designed a system based on Elasticsearch [5] and other open-source components. Its high-level architecture can be seen on Figure 1. Data is gathered by dedicated agents installed throughout the hospital infrastructure and then securely sent to HS.Register using the TLS protocol in an IHE-ATNA [6] compliant way.

Data stored on the Elasticsearch repository can be searched and accessed using web applications at the dashboard node.

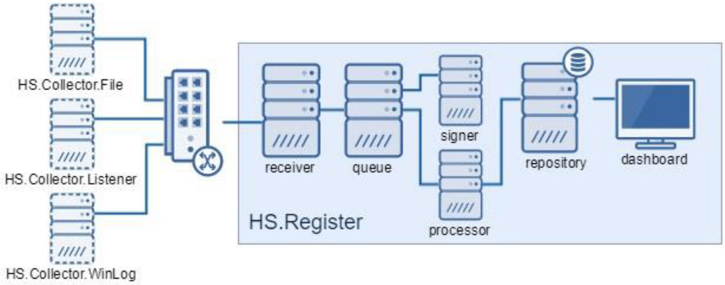

Figure 1 High-level architecture design

HS.Register is divided into two components groups, the repository, where data is processed and stored and the data collectors which are the agents that gather data from systems or directly from the network into the HS.Register repository, which is itself composed of 6 separate components:

1. Receiver for data gathering;
2. Queue to control back pressure and temporary storage;
3. Processing for data processing;
4. Signer to cryptographically sign and strongly bind events in temporal order.
5. Repository for long term storage;
6. Dashboards for reporting and visualization.

Currently there are 3 types of collectors based on the beats library, developed by Elastic: a network sniffer; a file reader; and a Windows EventLog viewer. With these agents we can gather at scale data from network communications, files and Windows eventlogs registries.

HS.Register is currently being used in 5 hospitals from Portugal hereby named as H.A, H.B, H.C, H.D and H.E for privacy purposes. Three are large general hospitals, one is a small province hospital and the other is a medium sized oncology hospital.

3. Results

The main results are presented in table 1. It shows the total number of events collected per week at the 5 hospitals. The total number of events that were gathered per week where around 93.543.000, and ranged from application logs and HL7 communication between various systems to external accesses. There was a total of 68.850.000 log events from applications, 21.672.000 from external accesses and 3.019.000 from HL7 communications. H.A and H.C have events from all types in the system, H.B does not have data regarding the application logs and H.D and H.E only have data about HL7 communications available.

Table 1. Total number of events (in thousands) per week grouped by type of system audited and hospital

	H.A	H.B	H.C	H.D	H.E	Total
Application Logs	68 826	*	24	*	*	68 850

Emergency Department	57 131					57 131
RIS	11 695		24			11 719
Communication using HL7	1 581	327	961	106	44	3 019
ACK	453	161	475	47	21	1 157
OML	344	32	202		11	589
SIU	239	11	72			322
ORU	116	100	53	28		297
ADT	50	11	67		1	129
ORM	78	11				89
Other	301	1	92	31	11	436
External Accesses	8 800	9 383	3 489	*	*	21 672
LDAP / AD	8 800	9 383	3 487			21 670
Remote Desktop			2			2
Total	79 208	9 710	4 475	106	44	93 543 000

* : Events not collected in this hospital

Most of the data (n=68.850.000) are from application logs. These events are very heterogeneous and range from logins and connections to external services and databases, to error stack traces. The emergency application logs from H.A are not yet being filtered as the hospital is still selecting which events are relevant in the context of the GDPR. H.B, H.D and H.E are in the process of gathering information with the vendors for how to implement and deploy agents to gather more data. H.C does not have an emergency department, but the events from the RIS application are already being stored.

The events from the LDAP group includes logins (successful and failed attempts) on the various systems and workstations inside the hospitals. This includes accesses from automatic monitoring tools and other systems. H.B is also storing logout events and H.A and H.C are in the process of analysing the inclusion of these events because many of the staff members may leave sessions logged-in when they leave the workstation and they do not think the data is accurate enough to be o use. H.D and H.E are not storing LDAP event at this point. H.C has a centralized point for remote access for maintenance by vendors and is also storing login and logout events.

HL7 messages are being stored at all hospitals. This information allows hospitals to analyse information flows. HS.Register stores each message received by each system where a HS.Collector agent is installed. Most of these messages are directly gathered by a network sniffer agent or directly sent to HS.Register by the hospital central HL7 BUS.

4. Discussion

For the GDPR, more importantly than the information itself, it is the knowledge of who, when and with what purpose the information is accessed and where it is stored and used. HL7 messages alone can give some insight, but due to some applications not using the HL7 standards it can be hard to gather a complete traceable data exchange process. However, if we add information about which users are using what workstation, what application is accessing a specific database or even if a vendor is doing some maintenance operation, crossing these events with the HL7 events, we can often identify critical policy violations like: account sharing by staff, unauthorized access by vendors, access to data by unauthorized systems or applications, illegal deletions or undesirable tampering of data. With the data collected by the HS.Register it is possible to comply with two of the GDPR requirements: traceability and auditability. On top of this, since all the events are timestamped and cryptographically signed and bonded together when the HS.Register receives them, the chain of events cannot be easily tampered with. This makes it for example very difficult for the receiving and sending systems to refute the existence of HL7 messages, adding a strong non-refutable characteristic to the system.

With the correct level of integration, the HS.Register could help comply with GDPR requirements, put the IT team back in control, identify problems sooner, identify the source of the problem and improve the overall quality of every Hospital Information System (HIS) and ultimately improve patient care.

Acknowledgement

The authors would like to acknowledge the project NanoSTIMA(NORTE-01-0145-FEDER-000016), which is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[1] E Parliament. Regulation (eu) 2016/679 of the european parliament and of the coucil of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. and repealing directive 95/46/ec (general data protection regulation), 2016.

[2] Lucas Ribeiro, João Cunha, and Ricardo João Cruz-correia. Information systems heterogeneity and interoperability inside hospitals - A Survey. In HealthInf 2010, pages 337–343, Valencia, Spain, 2010.

[3] Ricardo Cruz-Correia, Isabel Boldt, Luís Lapão, Cátia Pereira, Pedro Rodrigues, Ana Margarida Ferreira, and Alberto Freitas. Analysis of the quality of hospital information systems audit trails. BMC medical informatics and decision making, 13(1):84, 2013.

[4] Eduardo Pinto, António Carvalho Brito, and Ricardo João Cruz-Correia. Identification and characterization of inter-organizational information flows in the portuguese national health service. Applied Clinical Informatics, 7(4):1202, 2016.

[5] Clinton Gormley and Zachary Tong. Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. " O'Reilly Media, Inc.", 2015.

[6] IHE IT Infrastructure Technical Framework Supplement 2004-2005 Audit Trail and Node Authentication Profile. Technical report, ACC/HIMSS/RSNA, 2005.