



isec

Engenharia

MESTRADO EM INFORMÁTICA E
SISTEMAS

Resilience to Cyber-attacks in Critical Infrastructures of Portugal

DEFINITIVO

Autor

Any Keila Fortes Pereira

Orientadores

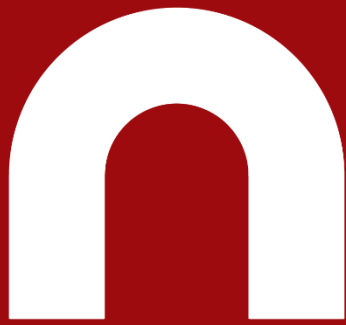
Cristina Margarida Chuva Costa

Afonso Araújo Neto

INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA

Coimbra, Abril 2021



isec

Engenharia

DEPARTAMENTO DE INFORMÁTICA E SISTEMAS

Resilience to Cyber-attacks in Critical Infrastructures of Portugal

Relatório de Estágio de Natureza Profissional para a obtenção do grau de Mestre em Informática e Sistemas

Especialização em Tecnologias da Informação e do Conhecimento.

Autor

Any Keila Fortes Pereira

Orientadores

Cristina Margarida Chuva Costa

Afonso Araújo Neto

Supervisor na empresa Dognædis

Afonso Araújo Neto

INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, Abril 2021

This work is dedicated to my parents, Salvador Pereira and Madalena Fortes for always being with me and for supporting me at every moment of my life.

ACKNOWLEDGEMENTS

The accomplishment and conclusion of this work represent the achievement of a huge personal goal, which was a difficult and challenging journey of my life. I would not have been able to successfully complete it without the aid and support of the following individuals/entities.

I would like to express my very great appreciation to ISEC for providing me with the necessary conditions, and to all my teachers for all the knowledge transmitted along my academic journey, without which this work would not be possible.

To all the Dognædis team, I am particularly grateful for the opportunity they gave me to embrace such a challenging project, for the incredible reception, help and the working conditions provided.

I would like to thank Shodan and Censys Support Team to believe in this work and for providing access to their Academic Account for collection of my data.

To my advisor, Cristina Chuva, I would like to express my very great appreciation for her advice, the useful critiques and all the valuable suggestions given, in order to carry out this work successfully. To my supervisor, Afonso Neto, I express my deep gratitude for his patient guidance, enthusiastic encouragement and his valuable technical support. I appreciate his willingness to give his time so generously so that this work could be done in the best possible way.

To my friends and colleagues, I would like to express my appreciation for their friendship, the exchange of knowledge and teamwork, especially my friends António Oliveira, Paula Denise, and my boyfriend Evandro Pereira for the encouragement and motivation in the less easy moments.

To my parents, Salvador Pereira and Madalena Fortes, I have no words to appreciate all the effort they made to make sure that my only concern was to study. I appreciate from the bottom of my heart the constant encouragement, love and affection, and to all my brothers for their unconditional support.

This written acknowledgement forces me to a formal dryness that leaves almost everything unsaid.

To all, my eternal gratitude.

ABSTRACT

Critical infrastructures are always a potential target for cyberattacks, as they control and allow access to the major infrastructure services of a nation. The repercussions of a successful attack on these systems can be catastrophic. One of the critical infrastructures of a country is the Industrial Control Systems (ICSs), used to automate and control vital functions of the various industrial infrastructures.

ICSs used to be operated in isolated environments. However, over time, to meet the demands of the modern market they began to be connected with the outside world. This has brought many benefits, but also increased their level of exposure and vulnerability. Although these systems are vital for the proper functioning of a country, there is no public work that evaluates their state visibility and vulnerability of ICSs in Portugal.

This work aims to identify the ICSs exposed to the Internet in Portugal and investigate their level of risk in terms of security. To achieve that objective, a methodology to identify the exposed ICSs on the Internet in Portugal and their level of risk was developed. The proposed methodology implied the identification of the ICSs, the calculation of their risk level according to their characteristics, and the development of a data warehouse to combine and organise the data into one comprehensive database, as well as, to enable the easy analysis of it.

Within the analysis of the data, we reached the following findings. There are many ICSs exposed and easily found on the public Internet in Portugal. The majority of them are located in Lisboa and have at least one feature that presents a High Risk to the security of the system. Most of them do not have an algorithm of encryption available, to secure the connection. From those that have, a huge percentage are using algorithms not considered secure. The majority of the identified systems have a port running the HTTP protocol, which is a concern because HTTP connections are not considered secure. From the systems running ports with High Risk associated, most of them are running the protocol FTP, a protocol not built to be secure. Most of the organisations do not have their own infrastructures to maintain the network routing policies of their systems. In this situation, it is not possible to identify the organisations because they are hidden behind the ISPs. This can be advantageous as they can be immediately identifiable by attackers, however, the organisation becomes dependent on the ISP, in such a way that if the ISP suffers an attack and goes offline, all the organisations behind it would be severely affected.

The results of this work enable Dognædis to have a knowledge base of the exposed ICSs in Portugal and their associated level of risk, making it possible to suggest improvements in their security. It also allows the industry and all organisations that have ICSs to be aware of how exposed and vulnerable their systems are, enabling them to devote more attention to the systems that may be at risk of a cyberattack and mitigate their vulnerabilities.

Keywords: ICS, SCADA, DCS, Critical Infrastructures, Security, Data warehouse

RESUMO

As infraestruturas críticas são sempre um potencial alvo para ciberataques, uma vez que a repercussão de um ataque bem-sucedido pode ser catastrófica, visto que esses sistemas controlam e permitem o acesso aos principais serviços do país. Um dos sistemas que fazem parte deste grupo de infraestruturas críticas de um país são os Sistemas de Controlo Industrial (ICSs), utilizados para automatizar e controlar os processos das várias infraestruturas industriais.

No passado, os ICSs eram utilizados em ambiente isolado, no entanto, com o passar do tempo e para satisfazer as exigências do mercado moderno, começaram a estar ligados com o ambiente externo. Isto trouxe muitos benefícios, mas também aumentou o nível de exposição e vulnerabilidade dos mesmos. Embora estes sistemas sejam vitais para o bom funcionamento de um país, não há nenhum trabalho público que avalie o estado de segurança destes sistemas em Portugal.

Este trabalho teve como maior objetivo, identificar os ICSs expostos na Internet em Portugal e investigar o nível de risco dos mesmos em termos de segurança. Com base nisso, foi desenvolvido uma metodologia que implicou a identificação dos ICSs, o cálculo do risco dos mesmos de acordo com as características que apresentam, e o desenvolvimento de uma data warehouse para juntar e organizar os dados, e permitir uma análise de forma fácil.

Ao analisar os resultados verificamos que existem muitos ICSs expostos e facilmente encontrados na Internet em Portugal. A maioria deles estão localizados em Lisboa e têm pelo menos uma característica que apresenta um risco elevado à segurança do sistema. A maioria dos sistemas não têm disponível um algoritmo de encriptação para assegurar a segurança da ligação. Dos que têm, uma enorme percentagem utiliza algoritmos que não são considerados seguros. A maioria dos sistemas identificados têm pelo menos uma porta a correr o protocolo HTTP, uma ligação que há muito tempo já não é considerada segura. Dos sistemas que estão a correr portas com risco elevado, a maioria está a correr o protocolo FTP, um protocolo não construído para ser seguro. Muitas das organizações não possuem infraestruturas próprias para gerir as políticas de rede dos seus sistemas. Nesta situação, não é possível identificar as organizações porque escondem atrás dos ISPs. Isto pode ser vantajoso porque as organizações não são facilmente identificadas pelos hackers, no entanto, ficam dependentes dos ISPs, no sentido de que, se este sofrer um ataque, todas as organizações ligadas a ela podem ser severamente afetadas.

Os resultados encontrados neste trabalho permitem à Dognædis ter uma base de conhecimento sobre o estado dos ICSs expostos na Internet em Portugal, tornando possível sugerir melhorias de segurança. Também permite que a indústria e todas as organizações que têm ICSs estejam conscientes de quão expostos e vulneráveis estão os seus sistemas, de forma a dedicarem mais atenção aos sistemas que possam estar em risco de um ataque cibernético.

Palavras-Chave: ICS, SCADA, DCS, Infraestruturas Críticas, Segurança, Data Warehouse

TABLE OF CONTENT

Acknowledgements	iii
Abstract.....	v
Resumo	vii
Table of Content	ix
List of Figures.....	xiii
List of Tables	xvii
List of Equations.....	xix
Acronyms and Definitions.....	xxi
1. Introduction	1
1.1 Instituto Superior de Engenharia de Coimbra.....	1
1.2 Host Company - Dognædis	2
1.3 Scope.....	3
1.4 Motivation.....	3
1.5 Objectives	4
1.6 Outline	4
2 State of the Art.....	7
2.1 Industrial Control Systems.....	7
2.1.1 Types of Industrial Control Systems	8
2.1.2 Components of an ICS Environment.....	10
2.1.3 Industrial Control Systems Protocols	14
2.2 Basic concepts and terminologies related to the identification and analysis of systems connected to the Internet.....	15
2.2.1 Cipher	15
2.2.2 SSL/TLS	15
2.2.3 Cipher Suite	16
2.2.4 Port	17
2.2.5 CVE	18
2.2.6 Autonomous System.....	20
2.2.7 Internet Service Provider	21
2.3 Port Scan	22
2.3.1 Port scanning tools	22

2.3.2	Port Scanning Projects.....	25
2.4	Data Warehouse	33
2.4.1	Characteristics of a data warehouse	33
2.4.2	Architectures of a data warehouse.....	33
2.4.3	ETL (Extract, Transform and Load).....	36
2.4.4	Open-source Data warehousing Tools.....	36
2.5	Security Principles	43
2.5.1	Key principles of security.....	43
2.6	Related Work	45
	Conclusions	46
3	Proposed Methodology.....	49
3.1	Phases of the methodology	50
3.1.1	Phase 1: Find reliable and updated databases with information about ICSs	50
3.1.2	Phase 2: Identify the Industrial Control Systems	51
3.1.3	Phase 3: Identify the features that are or may represent vulnerabilities to the ICSs	51
3.1.4	Phase 4: Define the calculation of the level of risk of the systems	53
3.1.5	Phase 5: Build a Data Warehouse to join and organise all the information	63
3.1.6	Phase 6: Analyse the data, draw possible conclusions and discuss the results	73
	Conclusions	74
4	Analysis of the data and discussion of the results	75
Q1.	Number of network addresses running High/Medium/Low Risk ports.....	75
Q2.	Number of network addresses using High/Medium/Low Risk SL/TLS versions	76
Q3.	Number of network addresses using High/Medium/Low Risk cipher suites	76
Q4.	Number of network addresses with High/Medium/Low Risk CVEs.....	77
Q5.	Level of risk per network address/most vulnerable network addresses.....	78
Q6.	Number of exposed ports per network address.....	79
Q7.	Number of exposed ports running standard and non-standard ICS	82
Q8.	Most common protocols exposed	82
Q9.	Number of CVEs per network address	83
Q10.	Most common CVEs identified on the systems	84
	Further analysis	85
	Conclusions	91

5	Conclusion and Future Work.....	93
	References	96
	Appendix	105
	Appendix A - Industrial Control Systems Standard Protocols and their port numbers.....	107
	Appendix B - Assignment of the classes to the exposed ports.....	109
	Appendix C - Assignment of the classes to the cipher suites.....	111
	Appendix D - Internship Proposal	115

LIST OF FIGURES

Figure 1 - Logotype of ISEC	2
Figure 2 - Logotype of Dognædis.....	2
Figure 3 - Evolution of the Dognædis brand	3
Figure 4 - A basic architecture of an ICS (Trend Micro, 2019)	8
Figure 5 - General layout of a SCADA system (Chokalingam, 2019).....	9
Figure 6 - Traditional DCS (Inst Tools, 2019)	9
Figure 7 - Example of a PLC system (Unitronics, 2019)	11
Figure 8 - Typical RTU hardware configuration (Belmonte, Boulanger, Schön, & Berkani, 2006).....	12
Figure 9 - An HMI interacting with a SCADA Server (Trend Micro, 2019).....	12
Figure 10 - Examples of MTU in a SCADA system architecture (Jiang, 2013).....	13
Figure 11 - An IED connecting with a sensor, actuator, and a transmission system (Morris & Pavurapu, 2010).....	13
Figure 12 - Example of a handshake between a server and a client using cipher suites (Outspoken Media Inc., 2020).....	16
Figure 13 - Example of a typical cipher suite.....	17
Figure 14 - Example of a cipher suite with omitted protocol.....	17
Figure 15 – Example of device information retrieved from Shodan (Ceron et al., 2019).....	20
Figure 16 - Example of a network with several Autonomous Systems (Cloudflare, 2020).....	20
Figure 17 - Example of connection between Tier 1, Tier 2, and Tier 3 providers (Datapath.io, 2016).....	21
Figure 18 - A simple Nmap on Metasploit 2 scan using the command-line	24
Figure 19 - A simple Nmap scan on Metasploit 2 using the interface GUI (Zenmap)	24
Figure 20 - Main page of the Censys website	26
Figure 21 - IPv4 result page	27
Figure 22 - Query result when searching for the tag “scada”.....	28
Figure 23 - Query result when searching for country code “PT” (Portugal).....	28
Figure 24 - Query result when searching for autonomous system code 2860.....	28

Figure 25 - Query to search for hosts tagged as “camera” and located in Portugal29

Figure 26 - Query to search for OpenSSH servers running on Debian across Europe.....29

Figure 27 - Censys Interface of Websites search option29

Figure 28 - Censys interface of Certificates search option.....29

Figure 29 – Access to Censys data through Google BigQuery 30

Figure 30 - Main page of Shodan website 31

Figure 31 - How Shodan works (Bodenheim, Butts, Dunlap, & Mullins, 2014) 31

Figure 32 – Shodan’s results when searching for protocol Modbus 32

Figure 33 - Example of a star schema (Guru99, 2020a)..... 34

Figure 34 - Example of a snowflake schema (Guru99, 2020a) 35

Figure 35 - Example of a Constellation schema (Guru99, 2020a) 35

Figure 36 - Schema of an ETL process (Folly, 2018) 36

Figure 37 - Balance between the basic principles of security 44

Figure 38 - Overall structure of the proposed methodology 49

Figure 39 - Risk matrix classification based on Impact × Likelihood (M.Talabis & Martin, 2013)
..... 55

Figure 40 - Aggregate risk scores of applications used in a health system (M.Talabis & Martin,
2013)..... 56

Figure 41 - Diagram of the data warehouse 66

Figure 42 - Transformation of the field Organisation for the dimension Dim_NetworkAddress
..... 69

Figure 43 - ETL process of the dimension Dim_NetworkAddress 71

Figure 44 - Process to perform the periodic ETL of the dimension Dim_NetworkAddress 73

Figure 45 - Number of exposed network addresses per risk level of ports 75

Figure 46 - Number of exposed network addresses per SSL/TLS version risk classification . 76

Figure 47 - Number of network addresses per risk classification of the used cipher suite 77

Figure 48 - Number of network addresses per qualitative severity of the CVEs identified 78

Figure 49 - Systems with the highest level of risk per month 78

Figure 50 - Top 10 network addresses with the highest level of risk 79

Figure 51 - Top 10 of the most exposed systems 80

Figure 52 - Evolution of the number of exposed ports of the most exposed network address	80
Figure 53 - Number of ports running ICS standard and non-standard protocols	82
Figure 54 - Most Exposed protocols.....	83
Figure 55 - Exposed database protocols	83
Figure 56 - Number of distinct CVEs per network address	84
Figure 57 - Top 10 of the most identified CVEs	84
Figure 58 - Top 10 organisations with the oldest CVEs.....	85
Figure 59 - Sum of the level of risk of some critical organisations.....	86
Figure 60 - Exposed network addresses of Credibom.....	87
Figure 61 - Exposed network addresses of Aeroportos de Portugal.....	87
Figure 62 - Exposed network addresses of TAP	88
Figure 63 - Exposed network addresses of INE	88
Figure 64 - Exposed network addresses of REN.....	89
Figure 65 - Exposed network addresses of Administração Central de Sistema de Saude, I.P.	89
Figure 66 - Exposed network address of Secretaria Geral do Ministério da Administração Interna.....	89
Figure 67 - Exposed network addresses of CMVM	90
Figure 68 - Most exposed educational institutions.....	90
Figure 69 - Level of risk of the educational institutions.....	91

LIST OF TABLES

Table 1 - List of standard protocols used in ICSs (Ceron et al., 2019)	14
Table 2 - Commands to install Nmap on the most popular Linux distributions and Mac OS platforms (Lyon, 2008).....	23
Table 3 - Commands to install Zmap on the most popular Linux distributions and Mac OS platforms.....	25
Table 4 - Prices and the features that each plan gives access (Shodan, 2019)	32
Table 5 - Features supported by Pentaho Business Analytics and Metabase for free (Hitachi Vantara LLC., 2020a; Metabase, 2020)	42
Table 6 - Classes of the SSL/TLS versions	58
Table 7 - Risk matrix classification for the feature Port.....	60
Table 8 - Risk matrix classification for the feature SSL version.....	61
Table 9 - Risk matrix classification for the feature Cipher	62
Table 10 - Risk matrix classification for the feature CVE	63

LIST OF EQUATIONS

Equation 1 - Formula to calculate the Risk score54

ACRONYMS AND DEFINITIONS

ACID - acronym for the set of database transaction properties (Atomicity, Consistency, Isolation, Durability) designed to ensure the accuracy, integrity, and completeness of data (IAN, 2016; Tutorialspoint, 2020).

API - acronym for Application Programming Interface. It is a set of definitions and protocols for building and integrating applications (MuleSoft LLC, 2020; Red Hat, 2020a).

AS - acronym for Autonomous System. It is a group of one or more IP prefixes (lists of IP addresses accessible on a network) controlled by a network operator that maintains a unified and well-defined routing policy (Cloudflare, 2020).

CMVM - acronym for Comissão do Mercado de Valores Mobiliários, the organisation in Portugal responsible for supervising and regulating the operation of markets in financial instruments and the activity of all agents acting in them.

CVE - acronym for Common Vulnerabilities and Exposures. It is a publicly well-known and documented information-security vulnerability or exposure (MITRE Corporation, 2020b).

CVSS - acronym for Common Vulnerability Scoring System owned and managed by FIRST.Org, Inc. It is a free and open industry standard that assesses the severity of the vulnerability on the systems (FIRST, 2020).

DW - acronym for Data Warehouse. It is an information system technology that aggregates structured data from single or multiple sources, to conduct data analyses that help in the decision-making processes (Tutorialspoint, 2019).

ICS - acronym for Industrial Control System. It stands for the control components used in the industry to automate industrial processes and improve efficiency. It includes devices, systems, and networks used for controlling, monitoring, and managing vital functions of the various industrial infrastructures (Mirian et al., 2016; Serbanescu, Obermeier, & Yu, 2015; Trend Micro, 2019).

INE - acronym for Instituto Nacional de Estatística. It is the official organisation in Portugal responsible for creating and publishing high-quality official statistical information.

IP - acronym for Internet Protocol. It is a numerical label used to uniquely identify systems connected to a computer network.

IPN - acronym for Instituto Pedro Nunes. It is a private non-profit organisation that promotes innovation and the transfer of technology, by establishing the connection between the scientific and the industry environment (Instituto Pedro Nunes, 2020).

ISEC - acronym for Instituto Superior de Engenharia de Coimbra - Coimbra Institute of Engineering. It is one of the teaching units of the Instituto Politécnico de Coimbra, responsible for providing higher education for the exercise of professional activities in the field of Engineering (ISEC, 2018).

ISP - acronym for Internet Service Provider. It stands for the organisation that provides Internet services. It can be commercial, community-owned, non-profit, or privately owned.

NA - acronym for Not Available. The term is utilized in situations where the information is not available or it is not provided.

NIST - acronym for National Institute of Standards and Technology. It is a non-regulatory governmental agency of the United State Department of Commerce, that promotes innovation and industrial competitiveness by advancing measurement science, standards, and technology (NIST, 2020a).

REN - acronym for Redes Energéticas Nacionais. It is a Portuguese company of electricity and natural gas transportation, responsible for the management of the national electric and natural gas system in Portugal.

SSL - acronym for Secure Sockets Layer. It is a standard security technology developed to secure and safeguard the data transferred over the Internet connection, ensuring that any data transferred remains unreadable for unauthorized users (Digicert, 2020b).

TAP - acronym for Transportes Aéreos Portugueses, the oficial portuguese airlines.

TCP - acronym for Transmission Control Protocol. It is a standard that defines how to establish and maintain a network connection between the devices before transmitting data (Doyle, 2018).

TLS - acronym for Transport Layer Security. It is a more secure and updated version of SSL.

UDP - acronym for User Datagram Protocol. It is a connectionless communication protocol that does not guarantee that the data sent will arrive intact and in the correct order (Doyle, 2018).

1. INTRODUCTION

This report intends to document the curricular internship of the Master in Informatics and Systems, resulting from the collaboration between the Coimbra Institute of Engineering (Instituto Superior de Engenharia de Coimbra - ISEC) and Dognædis. This work aims to develop an investigation on the current state of the resilience of the critical infrastructure in Portugal, more specifically Industrial Control Systems (ICSs), whose exploitation by malicious users would cause damage to the country. The result of this work will allow the organisations to have a measure of how exposed their systems, allowing them to know how resilient they are in terms of cyber security, and devote more attention to the most exposed and vulnerable systems.

This chapter presents an overview of the organisations involved in this work, ISEC and Dognædis. It also presents the scope, motivations, and the main objectives of this work. Finally, it presents the overall structure of this report.

1.1 Instituto Superior de Engenharia de Coimbra

ISEC (Instituto Superior de Engenharia de Coimbra) - Coimbra Institute of Engineering (ISEC, 2018), is one of the teaching units of the Instituto Politécnico de Coimbra, responsible for providing higher education for the exercise of professional activities within the field of Engineering.

The vision of ISEC is to be a reference for excellence in education, recognized nationally and internationally for its high-quality education with flexible, creative, and innovative practices, based on rigorous theoretical knowledge. It also intends to have social relevance, being a privileged partner for business organisations and families in its region.

The organisation aims to create, transmit and diffuse culture, science, and technology, and promote the development of the region, guided by the fundamental values of citizenship, quality, the continuous search for the valorisation, motivation, and updating of its pedagogical, scientific and technological resources. It also values good relationships with the students and partner organisations.

The aim of the Master in Informatics and Systems at ISEC is to qualify Masters in Informatics and Systems capable of carrying out their professional activity with a high level of technical and scientific competence. It is organised into 2 specializations: Applications Development and Information and Systems. These specializations are profession-oriented and aim to train specialists with skills in very concrete fields of application. Figure 1 presents the logotype of ISEC.



Figure 1 - Logotype of ISEC

1.2 Host Company - Dognædis

This internship was developed at Dognædis (commercially referred as “Cipher, a Prosegur company”), a company focused on information security. Dognædis was founded in 2010 by a group of researchers from the University of Coimbra and CERT-IPN, hosted at the IPN (Instituto Pedro Nunes), which after five years of activity became a private entity.

Since 2018, the company presents itself commercially as Cipher, a Prosegur company, just like all the cybersecurity companies acquired by the security giant Prosegur. However, the company still has its own personality and still uses the name Dognædis for non-commercial purposes and still maintains its website (dognaedis.com).

Dognædis is focused on offering security services, which includes audits, software assurance, network, and design management. It also develops products internally. Figure 2 presents the logotype of Dognædis.



Figure 2 - Logotype of Dognædis

The term Dognædis arises from the joining of the Latin words: Dognitas (Quality) + Aedis (Temple). Thus, the company presents itself as “an efficiency-oriented service provider and innovative cybersecurity technology vendor” and aims to be “at the forefront of security technologies and devoted to bringing information security to organisations and individuals, helping to make the world a safer place”.

The company focuses on excellence and innovative solutions. In 2011 one of their core products, CodeV (Dognaedis, 2015), intelligent software to detect software security flaws, was awarded by BES Innovation 2011 in the area of Information Technologies and Services.

Since 2016, Dognædis is part of the Prosegur group, the worldwide leader in security. In 2018 Prosegur also acquired Cipher, a global cybersecurity company. In 2019 all companies within

the Prosegur cybersecurity division, including Dognædis, suffered a rebranding, becoming: “Cipher, a Prosegur company”. Despite the rebranding, the company still has its own personality and still uses the name Dognædis internally and for non-commercial purposes. Figure 3 illustrates the evolution of the Dognædis brand.

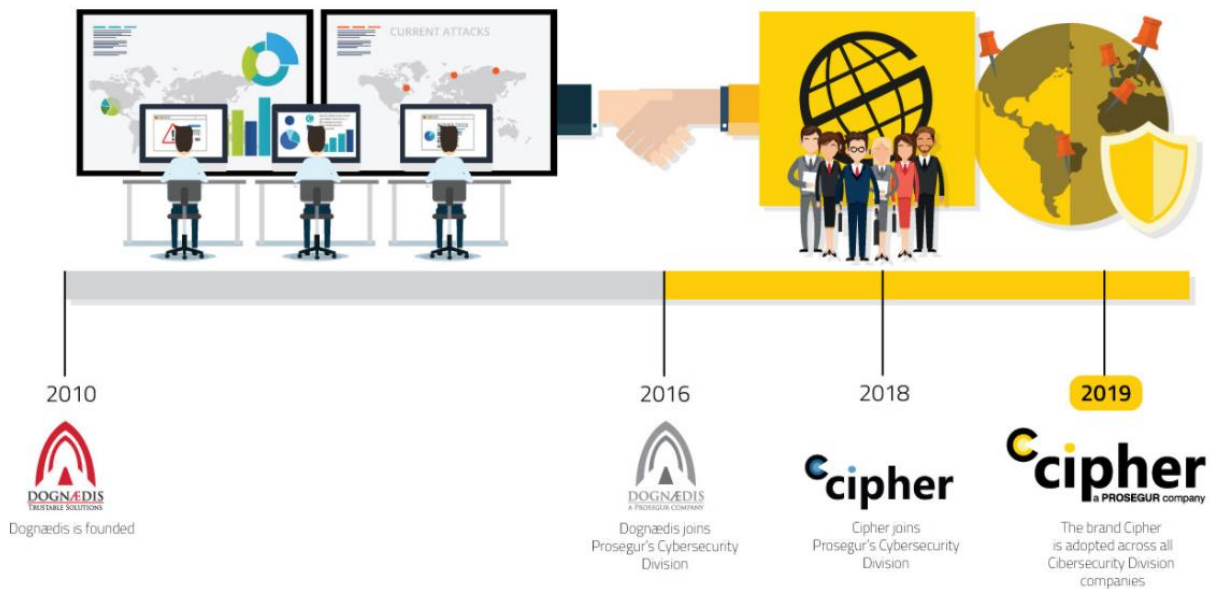


Figure 3 - Evolution of the Dognædis brand

1.3 Scope

Industrial Control Systems (ICSs) are used to automate and control vital functions of the various industrial infrastructures, ranging from small building fire control systems to large citywide power plant monitoring systems or, air traffic control systems. Such systems represent critical infrastructures of a country and are always a potential target for cyberattacks.

The progressive informatisation and, consequently, the interconnectivity provided by the Internet can put those infrastructures under the direct visibility of various malicious agents. The level of vulnerability and easy access to the critical infrastructures can be the difference between the normal life of a country and a national disaster. Currently, the state of visibility and vulnerability of the critical infrastructures in Portugal is not public knowledge.

Within the critical infrastructures, this study will focus on the identification of ICSs exposed on the Internet in Portugal and the investigation of their vulnerabilities. This will enable organisations to mitigate the vulnerabilities they present and thus reduce the risk of an attack.

1.4 Motivation

In 2005, it was estimated that more than 3 million of ICSs were in operation around the world - an increase of almost 9% per year (Wang et al., 2019). These systems used to operate in isolated environment. However, with the evolution of technologies, the necessity to reduce operational costs and to increase efficiency contributed to the connectivity between ICSs and

the outside world. This has brought many benefits, but also increased their level of exposure and vulnerability.

Although ICSs are vital for the proper functioning of a country, there is no public work, in Portugal, that allows the identification of those systems and the estimation of how vulnerable they are.

This work will enable Dognædis to have a knowledge base of the exposed ICSs in Portugal and their associated level of risk, making it possible to suggest improvements in their security and alert those organisations, which appear to be the most vulnerable. It will also allow industry and all organisations that have ICSs to be aware of how exposed and vulnerable their systems are, allowing them to know how resilient they are in terms of cyber security, and devote more attention to the most exposed and vulnerable systems.

1.5 Objectives

One of the main results of this work is the investigation and proposal of techniques that allow the identification of ICSs exposed on the Internet in Portugal and their associated vulnerabilities. It aims to achieve the following specific objectives:

- Research and proposal of public tools and databases to identify ICSs connected to the Internet;
- The development of a methodology to identify devices that are part of a network of critical infrastructure services and to estimate their associated level of risk;
- The application of the proposed methodology in practice, through a complete investigation or proof of concept in the scope of Industrial Control Systems in Portugal;
- The analysis and evaluation of the results.

No matter how much similar work has been done in the confidential field, the objective is to determine the information available to anyone with access to public information on the Internet and therefore to all malicious agents that exist.

1.6 Outline

The remainder of this document is organised as follows:

- The second chapter, State of the Art, presents the entire study of the literature performed to support the work. This includes the topics related to ICSs, Port Scan, Data Warehouse, the basic principles of information security, and other basic concepts and terminologies used throughout this work. Finally, it presents an overview on some of the most relevant studies related to the identification of the ICSs and the analysis of their vulnerabilities.
- Chapter 3 presents the methodology proposed and performed to achieve the objectives of this work. It describes and explains the main decisions taken, in each of the 6 phases of the proposed methodology.

- Chapter 4 presents the analysis and discussion of the results, corresponding to phase 6 of the proposed methodology, using reports generated in Metabase.
- Chapter 5 presents the problem statement, the overview of the work developed to achieve the objectives, and the main results and achieved conclusions. Finally, it presents the main limitations of this work and direction for future work.

2 STATE OF THE ART

This chapter presents the entire study of the literature performed to support the work. It introduces the concept of ICSs, Port Scan, Data Warehouse, and the key principles of information security. It also provides a brief explanation of some of the main concepts and terminologies, used throughout this work, in order to allow users to better understand it. Finally, it presents the existing and most relevant public studies related to the identification and analysis of ICSs.

2.1 Industrial Control Systems

Industrial Control Systems (ICSs) stands for the control components used in the industry to automate industrial processes, in order to improve their efficiency. These components include devices, systems, and networks used for controlling, monitoring, and managing vital functions of the various industrial infrastructures. Such infrastructures can be critical, ranging from small building fire control systems to large citywide power plant monitoring systems, whose failure could potentially endanger human lives. Depending on the industry, each ICS operates differently to manage different tasks (Mirian et al., 2016; Serbanescu et al., 2015; Trend Micro, 2019).

In the past, the ICSs usually operated in isolated environments. However, in order to decrease the costs and improve the performance of the daily activities of the industry, many computing, and telecommunication capabilities have been integrated with ICSs, replacing and/or complementing the existing physical control mechanisms. This change gave rise to completely or partially automated processes (Stouffer et al., 2015).

The mentioned evolution promoted the connection of several ICSs to the outside world. Consequently, their control networks have become integrated with the corporate networks, allowing the management of industrial processes in real-time. However, while this increased the connectivity of these systems, it also increased their criticality, opening up discussions about their adaptability, resilience, and security (Stouffer et al., 2015).

Figure 4 shows the basic architecture of an ICSs environment. There is a flow of information between the ICSs in the industrial environment (on the right side of the image) and between them and the corporate environment, through the integration of control networks and corporate networks.

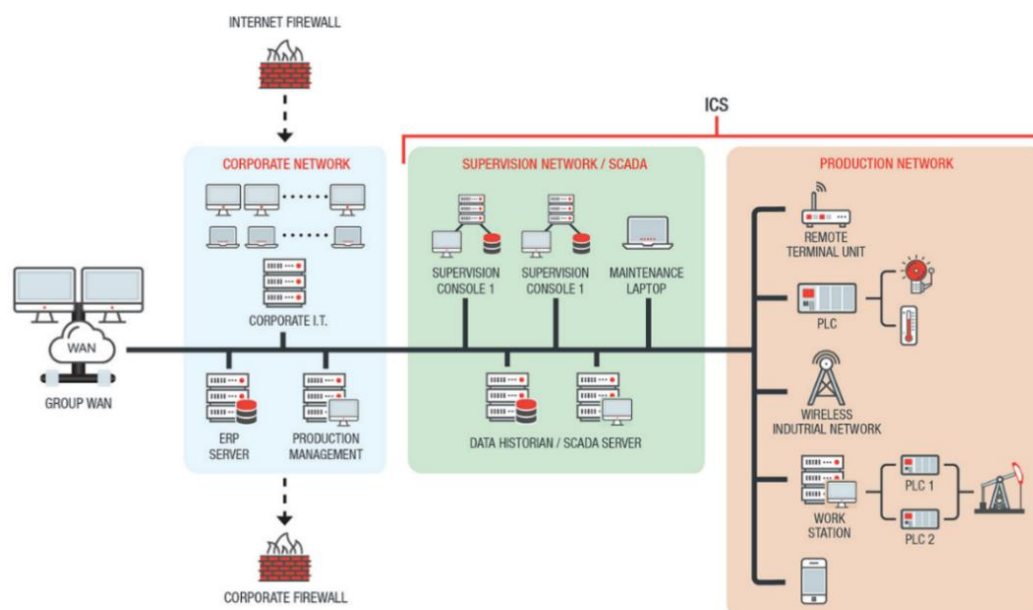


Figure 4 - A basic architecture of an ICS (Trend Micro, 2019)

2.1.1 Types of Industrial Control Systems

The most common types of ICSs are Supervisory Control and Data Acquisition, also known as SCADA systems, and Distributed Control Systems (DCS) (Trend Micro, 2019). Below, it is presented the explanation of each one.

2.1.1.1 Supervisory Control and Data Acquisition

Supervisory Control and Data Acquisition (SCADA) are computer-based control systems used to monitor and control the flow of data among the industrial physical processes. They are usually a set of interconnected devices, such as controllers, sensors, actuators, and communication devices. These systems are responsible for collecting data on the functioning of the daily activities of the industry, in the field, transfer it to a central computer, and display it to an operator. This allows the monitoring and control of the industrial processes from a central location almost in real-time (Stouffer et al., 2015; Trend Micro, 2019; Tsang, 2009).

SCADA systems provide control at the supervisory level. It allows operators to supervise, maintain, control, and collect data from industrial infrastructures through a centralized control system. A key advantage of these systems is the ability to perform supervisory operations over a variety of devices (Samtani et al., 2018; Trend Micro, 2019; Tsang, 2009; Vavra & Hromada, 2016).

Figure 5 shows a general layout of a SCADA system, where the master server communicates with an Intelligent Endpoint Device (IED), a Programmable Logic Controller (PLC), and a Remote Terminal Unit (RTU) through a WAN connection. It collects data or carries a function (e.g., close a valve) on the field devices.

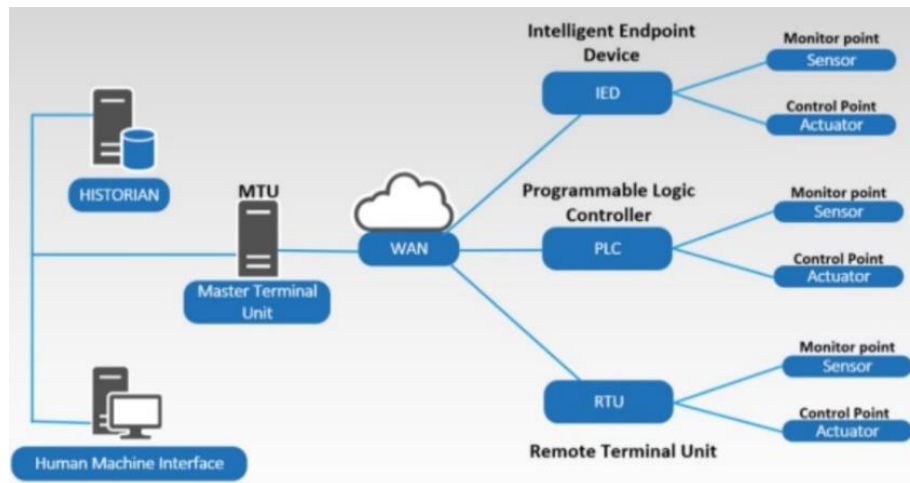


Figure 5 - General layout of a SCADA system (Chokalingam, 2019)

2.1.1.2 Distributed Control Systems

Distributed Control Systems (DCSs) control equipment in the same geographic location. They are responsible for controlling the operational details of the industrial components. DCSs usually control processes among the field devices or are small parts of a control system (Stouffer et al., 2015).

Each DCS uses a centralized supervisory control system to manage the various local controllers that are part of the overall system. This gives the industry the ability to quickly access operating data, thus allowing rapid problem resolution in the event of failures (Trend Micro, 2019).

Figure 6 shows a traditional DCS, where we can see 2 controllers managing the processes among the field devices. They are connected to the corporate network, allowing the managers and other operators to have a view of how the functioning of the field processes function in the industrial environment.

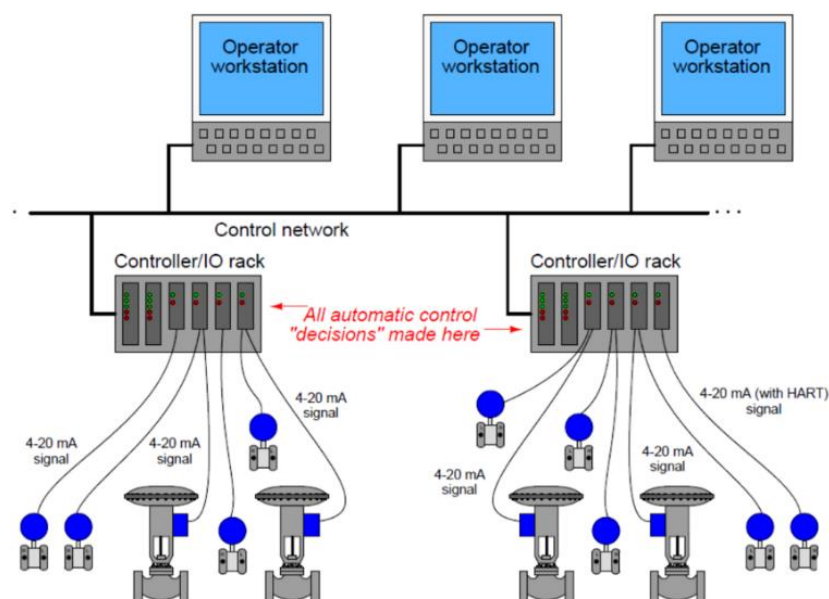


Figure 6 - Traditional DCS (Inst Tools, 2019)

SCADA systems are ideal to monitor and control the flow of data among the industrial processes, while DCSs are ideal for controlling the operational details of the industrial components. Nowadays, an ICS environment includes attributes from both, DCSs and SCADA systems (Trend Micro, 2019).

2.1.2 Components of an ICS Environment

As it is possible to observe in Figure 5 and Figure 6, there are several components that constitute a SCADA system and a DCS, and many of these components are common to both systems. Below, we present the most common components of an ICS environment.

2.1.2.1 IT and OT

IT (Information Technology) refers to all the computing and telecommunication technology responsible for managing the flow of digital information between the devices. It deals with information, focusing on storage, recovery, transmission, manipulation, and protection of data (Trend Micro, 2019). On the other hand, all the hardware and software responsible for managing the operation of physical processes refers to OT (Operational Technology). The latter deals with machines, focusing on monitoring and controlling the physical devices, processes, and/or events and their tasks in the field (Coolfire, 2019; Webranded, 2019).

The concepts of IT and OT address different perspectives. However, lately, the two technologies tend to converge and are increasingly interconnected with each other and with the Internet (Trend Micro, 2019). The mentioned convergence provides industries greater integration and visibility of their processes, but on the other hand, it allows easy access to their components, making them targets of cybercriminals (Webranded, 2019).

2.1.2.2 Programmable Logic Controller

Programmable Logic Controller (PLC) is an industrial computer used in both DCSs and SCADA systems, that allow the controlling and management of the processes running on control devices (e.g., sensors and actuators) (Stouffer et al., 2015; Trend Micro, 2019). In a SCADA system a PLC has the same functionality as an RTU (Remote Terminal Unit), while in a DCS, it works as a local controller within a scheme of supervisory control.

A PLC works as follows: it receives and processes the command from input devices, then it activates the output based on the pre-programmed settings. The memory system of a PLC stores the received instructions to implement. The instructions include monitoring and recording the execution time, start and stop processes automatically, and generating alarms in case of machine failure (Stouffer et al., 2015; Unitronics, 2019).

Figure 7 shows an example of a PLC system, connected to a module, where it receives the input and a module where it sends the output. It contains an ethernet input, a Central Processing Point that processes the instructions received, and a memory used to store the received instructions.

In addition, an application uses the PLC to program output behaviour according to the input received. A power supply provides the necessary energy for its operation. For example, the PLC receives an input from a sensor on the temperature of an industrial component, if the component exceeds a certain temperature it generates an alarm as output.

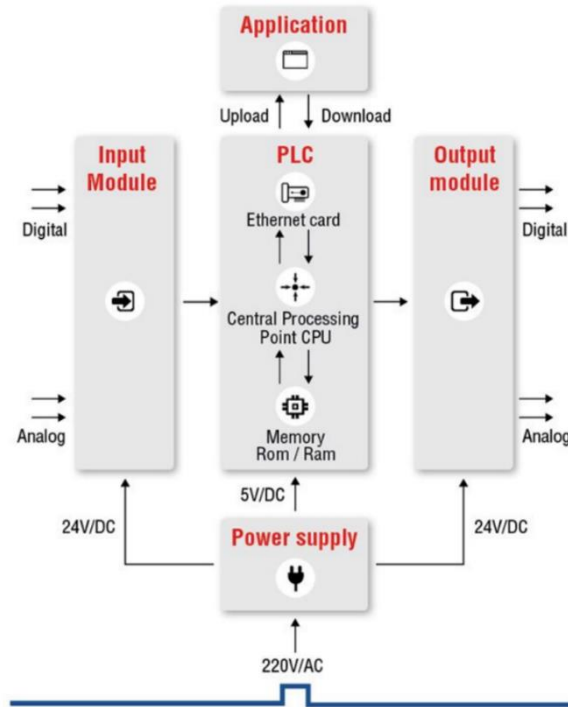


Figure 7 - Example of a PLC system (Unitronics, 2019)

2.1.2.3 Remote Terminal Unit

A Remote Terminal Unit (RTU), also known as Remote Telemetry Unit, is a data acquisition and control unit used to monitor and control equipment in remote locations. It links industrial equipment to the DCSs or SCADA systems by transferring the data of the industrial environment processes to the Master Terminal Unit (MTU) and vice-versa (Clarke et al., 2004).

The RTU traditionally communicates back to a central station, but they can also communicate within each other and sometimes act as a relay station to the RTUs, which may be inaccessible from the central station (Clarke et al., 2004). It is programmable and it comes with setup software that helps to configure the communications and inputs into outputs streams.

One of the biggest advantages of the RTUs is environmental tolerance, as it withstands extreme temperatures and can be used in remote locations. Thus, some of them have backup batteries to allow continual operation in case of a power failure (RealPars, 2019). In Figure 8, we can see several RTUs transferring the data of the field devices processes to MTU through a communication network. The MTU treats this received data and displays it to a human operator through Human Machine Interfaces (HMI).

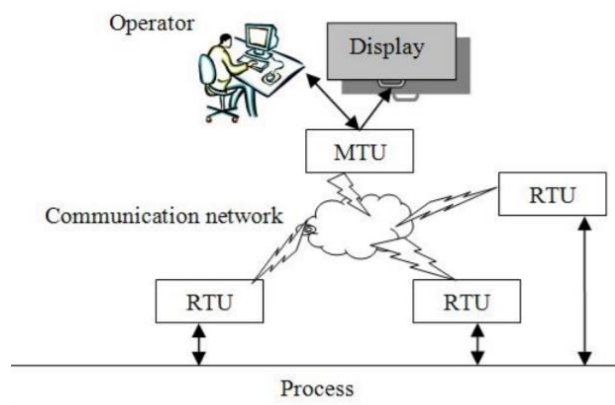


Figure 8 - Typical RTU hardware configuration (Belmonte, Boulanger, Schön, & Berkani, 2006)

2.1.2.4 Human Machine Interface

Human Machine Interface (HMI) is a graphical user interface or dashboard that connects and allows the human operator to interact with the controller devices (e.g., PLC/RTU). It is used to visualise the data gathered by the devices allowing the operator to monitor, control, configure, and adjust parameters in the industrial environment (Trend Micro, 2019).

An HMI can vary from a simple control panel to a computer with dedicated HMI software and colour graphics display (Stouffer et al., 2015). Figure 9 shows an HMI interacting with a SCADA Server. For example, the SCADA Server receives the alarm from the PLC due to the increase of temperature in some industry component and it displays it to the operator through the HMI. Therefore, the human operator can send a command to stop the equipment from functioning through the HMI.

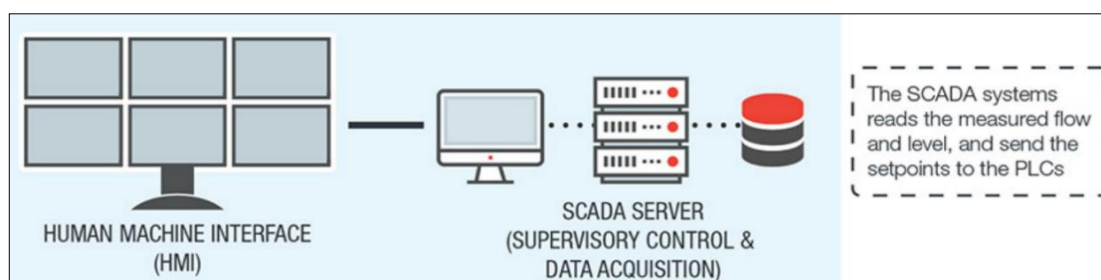


Figure 9 - An HMI interacting with a SCADA Server (Trend Micro, 2019)

2.1.2.5 Master Terminal Unit

Master Terminal Unit (MTU) also referred to as SCADA Server, Control Server or Supervisory Control, hosts the supervisory control software that communicates with the control devices over the ICS network (Stouffer et al., 2015; Trend Micro, 2019). In a SCADA system, the MTU sends the instructions to the RTUs located at remote places, collects, stores, and processes the required information, and then displays it to the HMI to assist human operators in decision making and control. In Figure 10 we can see an example of this communication.

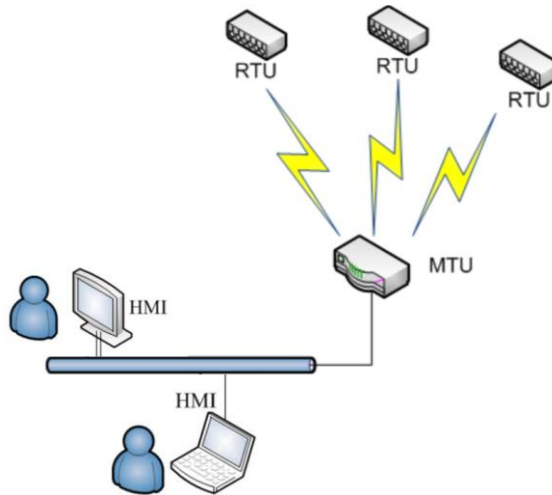


Figure 10 - Examples of MTU in a SCADA system architecture (Jiang, 2013)

Even though the fact that the communication between the MTUs and RTUs is bidirectional, only the MTUs can start the conversation. The RTUs are limited to collecting and storing the data from the devices in the industrial environment (Teja, 2011).

2.1.2.6 Intelligent Electronic Device

An Intelligent Electronic Device (IED) is a smart device able to communicate with other devices and perform at a local level processing and control activities. In the industrial environment, it receives data from sensors and power equipment and automatically executes control commands (Trend Micro, 2019).

The use of IEDs has increased a lot in the past years, and the replacement of RTUs to IEDs is more and more evident since the latter executes the same functions as the RTUs, and it presents integration and interoperability features. It also allows data access at many levels, self-monitoring of external circuit and events in real-time. It also allows to perform PLC functionalities and it has a range of features for commissioning, testing, reporting, and analysing faults (Csanyi, 2019).

Figure 11 shows an IED connected to a sensor of temperature, an actuator (valve), and a transmission system. The IED communicates with the sensor and the valve, allowing the valve to operate according to the temperature registered in the sensor, at the same time that it communicates with a transmission system.

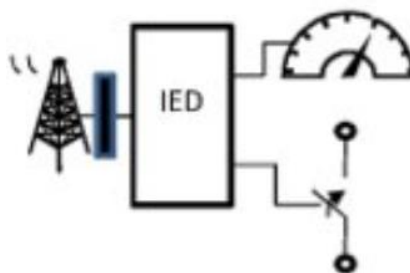


Figure 11 - An IED connecting with a sensor, actuator, and a transmission system (Morris & Pavurapu, 2010)

2.1.2.7 Data Historian

A data historian is a centralized database that gathers information from all the industrial processes and then exports it to the corporate systems. It used to be fed with data primarily by the process control system, however, it is very efficient when used in every department of the industry, centralizing all the data. A data historian records data from one or more locations and one of its strengths is the ability to correlate data over time. Data historians help in process analysis, statistical process control, and enterprise-level planning (Automated Results, 2019; Trend Micro, 2019). Nowadays, with the evolution of technologies, many external systems can send and access information from the data historian (W. Matthews, 2017).

2.1.3 Industrial Control Systems Protocols

Communication protocols are standards that define the semantics, syntax, and rules regarding the transmission of data between communication systems (Collantes & Padilla, 2015). The protocols specify how the source sends data and the receptor receives it, as well as the methods of recovery in case of failures.

In the past, ICS manufacturers usually adopted proprietary protocols in a line of production to allow communication among systems. However, this required the use of additional connectors for those systems to communicate with equipment of other production lines, or specific drivers to enable communication with systems from other manufactures. To solve these problems, manufacturers have started to use industry-accepted standards protocols.

There are numerous communication protocols. Nevertheless, there are some considered to be standard used in ICSs. Table 1 presents, in alphabetical order, a list of the most common standard protocols used by the manufacturer of ICSs (Ceron et al., 2019).

The list of the standard ICSs protocols and their default port number can be consulted on [Appendix A](#).

Table 1 - List of standard protocols used in ICSs (Ceron et al., 2019)

Standards Protocols used in ICS		
ANSI C12.22	ICCP	PC Worx
BACnet	IEC 60870-5-104	PowerLink Ethernet
Beckhoff-ADS	IEC 61850 / MMS	ProConOS
CANopen	Keyence KV-5000	Profibus
CIP	Koyo Ethernet	Profinet
CodeSys	LS Fenet	Quick Panel GE
Crimson 3	Melsec Q	Saia S-Bus
Danfoss ECL apex	Modbus	Schleicher XCX 300
DNP3	Moxa	Siemens S7
EtherCAT	Niagara Tridium Fox	Simatic
EtherNet/IP	Omron fins	Unitronics Socket1
ISA95	OPC	Yaskawa MP series ethernet
GE-SRTP	Panasonic FP	Yaskawa MP2300Siec
HART-IP	Panasonic FP2	Yokogawa FA-M3
Hitachi EHV Series		

Unfortunately, many of the protocols presented in Table 1 were not designed with security built-in. As explained above, ICSs were typically used in isolated environments and many of them have not been adapted to face the security problems that have arrived with the connection between the ICSs and the outside world.

2.2 Basic concepts and terminologies related to the identification and analysis of systems connected to the Internet

Below, we present a brief introduction and explanation of some of the main concepts and terminologies used throughout this work, in order to allow users to better understand it. It includes the concepts of cipher, SSL/TLS, cipher suite, ports, CVEs, Autonomous Systems (AS) and Internet Service Providers (ISPs).

2.2.1 Cipher

Ciphers (Rouse, 2007) are algorithms used to perform cryptographic functions (encryption, decryption, hashing, or digital signatures). They are a set of well-defined steps that use fixed rules and mathematical functions to transform legible messages (plaintext), into ciphertext (random sequence of characters), in order to make it unintelligible to anyone except the reliable receptor.

Nowadays, in addition to the cipher algorithm, modern cipher implementations depend on a piece of auxiliary information, a secret key, used by the algorithm to encrypt the data (Rouse, 2007). The key changes the detailed operation of the cipher algorithm. Thus, the encrypting procedure operates differently depending on the used key.

The key is an essential part of a strong cipher algorithm, so it is often confidential. Even if the users know the algorithm, it is almost impossible to decipher an encrypted message without the appropriate key (Rouse, 2007). Therefore, before using a cipher, both the sender and receiver must have the key or set of keys to encrypt/decrypt the message.

Ciphers using longer keys can be safer since, although the strength of the cipher does not depend entirely on the length of the key, longer ciphers are more likely to be secure against brute force attacks, as more attempts are needed to decode the message. Thus, experts recommend, depending on the algorithm and the case of use, that modern ciphers use keys with at least 128 bits (Nohe, 2019).

2.2.2 SSL/TLS

SSL (Secure Sockets Layer) is a standard security technology developed by Netscape to securely send information over the Internet (Digicert, 2020b). Its aims to ensure that any data transferred remains unreadable for not-addressed users. It uses encryption algorithms to encrypt data in transit, preventing criminals from accessing transferred information, including sensitive or personal information, such as names and addresses, social security numbers, login

credentials, and credit card numbers. To establish this secure connection, the browser and server require an SSL certificate (Digicert, 2020b).

TLS (Transport Layer Security) is an updated and more secure version of SSL. Its first version (TLS version 1.0) was initially developed as SSL version 3.1, but the protocol name was modified before publication to indicate that it is no longer associated with Netscape. The Internet Engineering Task Force (IETF) adopted the protocol and renamed it Transport Layer Security. Despite the name change, the term SSL is still commonly used to refer to TLS (Digicert, 2020a; IETF, 2020; Mocan, 2018).

SSL is used to secure HTTP connections. However, this layer is also able to protect other Internet protocols, such as SMTP, for sending e-mails, and NNTP, for newsgroups. When you access an URL starting with "HTTPS", the "S" indicates that the website is secure. These websites use SSL certificates to verify their authenticity (TechTerms, 2020).

2.2.3 Cipher Suite

A cipher suite is a set of methods/algorithms needed to secure a network connection that uses SSL/TLS. It combines a list of cryptographic algorithms and functions used in the process of secure data through the traffic (Nohe, 2019).

The name of each set is representative of the specific algorithms that comprise it and that work together to perform the handshake (the process where the server and client agree on a mutually supported cipher suite). During the handshake, the client and server switch a prioritized list of supported cipher suites to choose the one that is best supported by both. This is an important ceremony of the “handshake”, as the security of any SSL/TLS connection is highly dependent on the chosen cipher suite (Outspoken Media Inc., 2020). In Figure 12, we can see an example of the handshake between a web server and a browser, where they are comparing their list of prioritized cipher suites and then they chose the best supported by both.

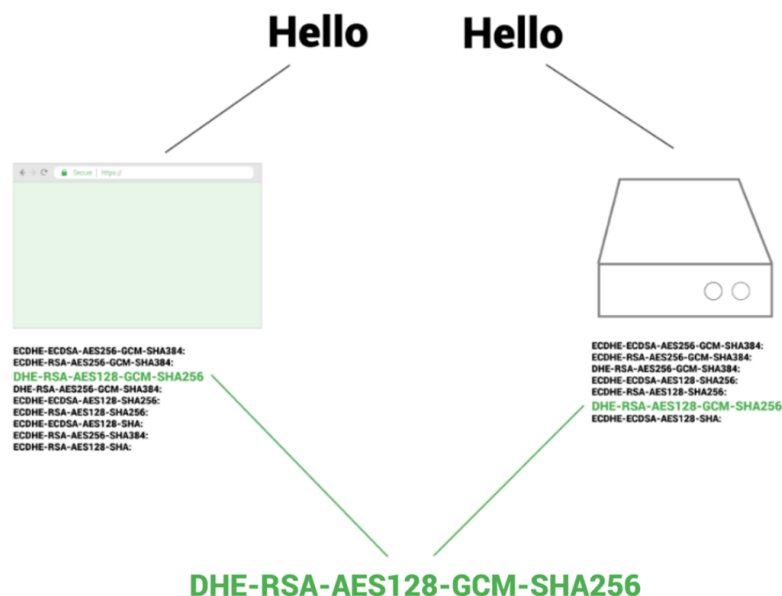


Figure 12 - Example of a handshake between a server and a client using cipher suites (Outspoken Media Inc., 2020)

The following algorithms compose a typical cipher suite (Microsoft, 2018):

- Key exchange algorithm - determines how the symmetrical keys (the type of encryption that uses a single key for both encryption and decryption process) will be exchanged;
- Authentication algorithm - determines how the server and client (if necessary) authentication will be performed;
- Bulk encryption algorithm - specifies the symmetric key algorithm used to encrypt the data;
- Message Authentication Code (MAC) algorithm - determines the method that the connection will use to perform the data integrity verifications.

A typical cipher suite is composed of 1 key exchange, 1 authentication, 1 bulk encryption, and 1 MAC algorithm, as we can see in Figure 13, presented below:

TLS_ECDHE_ECDSA_WITH_AES_256_CBC_SHA384

Figure 13 - Example of a typical cipher suite

In the example presented above TLS indicates the protocol, ECDHE specifies the key exchange algorithm, ECDSA indicates the authentication algorithm, AES_256_CBC designates the bulk encryption algorithm, and SHA384 indicates the MAC algorithm. In some cases, the protocol can be omitted, as we can see in the Figure 14, presented below:

DHE_RSA_AES256_SHA

Figure 14 - Example of a cipher suite with omitted protocol

In the presented example, DHE indicates the key exchange algorithm, RSA specifies the authentication algorithm, AES256 shows the bulk encryption algorithm, and SHA indicates the MAC algorithm.

2.2.4 Port

A port is a logical access that enables the communication between devices. The information in the port allows the system to identify the process to which the received traffic should be directed. A computer can, at the same time, receive data for different processes through multiple ports (Speedguide, 2020).

Every port contains a number that specifies the process running on the host, as well as the destination IP address. The port number is also included in the header of each packet when the communication is established. Thus, when the packet is received, the system knows which specific process it will be sent to. For example, when the system receives a packet with port number 80 in the header, the packet will be directed to the HTTP process, when with port number 25 it will be redirected to the mail process, and so on.

There are port numbers that are normally reserved for specific services to facilitate the forwarding of data to the processes. For instance, port numbers between 0 and 1023 are known as well-known port numbers, these ports identify the most historically used services (e.g., Port

80 - HTTP, Port 443 - HTTPS, Port 21 - FTP, Port 22 - SSH). Port numbers between 1024 and 49151 are known as Registered ports. These ports can be registered by companies to use on specific protocols (e.g., Port 1029 - Microsoft DCOM services and Port 1194 - OpenVPN). Port numbers between 49152 and 65536 are known as Dynamic or Private ports. They are available for use by any application to communicate with any other application (e.g., ports between 60000 and 61000 normally allocated by Mosh - a remote terminal application similar to SSH - for open sessions between Mosh servers and Mosh clients) (Speedguide, 2020; WhatIs MyIPAddress, 2018).

Still, within the communication between the devices, it is important to know that the data on the Internet is organised in standard TCP (Transmission Control Protocol) and/or UDP (User Datagram Protocol) packets. UDP does not guarantee that the data sent will arrive intact and in the correct order. It allows very fast communication, so it is ideal for applications where speed is essential and minimal data loss is not a disadvantage. On the other hand, TCP establishes the connection between the devices before transmitting data. This guarantees the integrity and delivery of the data in the order they were sent. TCP is ideal when data integrity is essential (Doyle, 2018; WhatIs MyIPAddress, 2018).

2.2.5 CVE

CVE stands for Common Vulnerabilities and Exposures. It is a public list of well-known information-security vulnerabilities and exposures. Each entry of the list contains an identification number, a description, and at least one public reference of a publicly known vulnerability (Red Hat, 2020b). The CVE is managed by MITRE Corporation (MITRE Corporation, 2020a) and funded by the Cyber Security and Infrastructure Agency, a division of the United States Department of Homeland Security (Department of Homeland Security, 2020).

When a CVE is mentioned, it typically refers to the identification number (ID) assigned to a security flaw. The CVE ID has this format: CVE-YYYY-NNNNN (e.g., CVE-2017-18911). The YYYY does not necessarily indicate the date when the vulnerability was discovered, it refers to the year when the CVE ID was assigned or the year when the CVE was published. Some examples to express this are:

- If a vulnerability is discovered in 2018, and its CVE ID is requested in 2018. The CVE ID will have the following format: "CVE-2018-NNNN";
- If a vulnerability is discovered in 2017 and published in 2018. If the CVE ID is requested in 2018, the CVE ID will have the following format: "CVE-2018-NNNNN";
- If a vulnerability is discovered in 2018, a request for its CVE ID is published in 2018, but the CVE was not made public. The vulnerability will be assigned with the format: "CVE-2018-NNNN" although it will appear in the CVE List as "Reserved";
- If a vulnerability is discovered and made public in 2018 without having a CVE ID assigned to it. If its CVE ID is requested in 2019, the vulnerability is given "CVE-2018-XXXX" since it had been first made public in 2018.

Not all vulnerabilities receive a CVE ID. To assign a CVE ID the failures must meet criteria, such as (Red Hat, 2020b):

- Independent correctness. The flaw must be able to be fixed independently from other flaws;
- Recognized or documented by the affected vendor. The vendor of the system must recognize the vulnerability and its negative impact on the security of the affected system;
- Affects only one codebase. Vulnerabilities that affect more than one solution receive separate CVEs. Each affected codebase or solution receives a unique CVE.

Each CVE has a score named Common Vulnerability Scoring System (CVSS) owned and managed by FIRST.Org, Inc. (FIRST, 2020). This score assesses the severity of the vulnerability on the systems. It ranges from 0 to 10, where 0 means that the exploitation of the CVE does not present any severity and 10 means that the exploitation of the CVE presents high severity for the system. Based on the score of each CVE, NIST also provides a qualitative severity ranking of "Low" (score from 0.0 to 3,9), "Medium" (score from 4.0 to 6,9), and "High" (score from 6,9 to 10.0). The metrics that influence the calculation of this score are (MITRE Corporation, 2020c; NIST, 2020b):

- Access vector - describes the level of access that is necessary to have to exploit the vulnerability on the system;
- Attack complexity - describes the ease of exploitation of the vulnerability;
- Authentication - describes the number of times that is necessary to authenticate to exploit the vulnerability;
- Confidentiality - describes the impact of the vulnerability exploitation on the confidentiality of data processed by the target system;
- Integrity - describes the impact of the vulnerability exploitation on the integrity of the target system;
- Availability - describes the impact of the vulnerability exploitation on the availability of the exploited system.

To help in better understanding some of the concepts and terminologies used in this study, in Figure 15 we present an example of system information retrieved from Shodan. The example shows a system, identified by IP address (130.89.14.205), which is managed by the UTWENTE Autonomous System, with AS number 1133. The system is running the protocols SSH, HTTP, and HTTPS on ports 22, 80, and 443, respectively. It was identified nine known vulnerabilities in the port running the protocol HTTP. The vulnerabilities are identified by their CVE numbers.

IP Address	ASN	AS name	Device	Ports	Protocols	Vulnerabilities
130.89.14.205	AS1133	UTWENTE	<not available>	22	SSH	-
				80	HTTP	CVE-2018-1302 CVE-2017-15710 CVE-2018-1301 CVE-2018-1283 CVE-2018-1303 CVE-2017-15715 CVE-2018-1333 CVE-2018-11763 CVE-2018-1312
				443	HTTPS	

Figure 15 – Example of device information retrieved from Shodan (Ceron et al., 2019)

2.2.6 Autonomous System

Autonomous System also referred with the acronym AS is a group of one or more IP prefixes (lists of IP addresses accessible on a network) owned by a network operator that maintains a unified and well-defined routing policy (Cloudflare, 2020).

Every AS controls a specific set of IP addresses and every computer or device that accesses the Internet is connected to an AS. In general, an AS is operated by a large organisation, such as an Internet Service Provider (ISP), a big technology company, a university, or a governmental agency (Thousandeyes, 2020).

Every Autonomous System has an identification number known as Autonomous System Number (ASN), assigned by the Internet Assigned Numbers Authority (IANA). This number enables us to manage and control the routing within their networks and to exchange routing policies with other ISPs. Figure 16 presents an example of a network with several AS represented by their ASNs.



Figure 16 - Example of a network with several Autonomous Systems (Cloudflare, 2020)

2.2.7 Internet Service Provider

Internet Service Provider (ISP) is the term used for the organisation that provides Internet services, enabling users to access, use, and participate in the Internet (Besen & Israel, 2013; Chapin & Owens, 2005). It can be commercial, non-profit, community-owned, or privately owned.

There are 3 levels of ISPs that depend on the type of Internet services provided: Tier 1, Tier 2, and Tier 3 providers (GeeksforGeeks, 2020b).

- A Tier 1 ISP is an Internet provider that only exchanges Internet traffic, on a non-commercial basis, with another Tier 1 provider. It is also known as Back Bone Service Provider, providing traffic to all other Internet providers, not end-users. They are responsible to build infrastructure such as the Atlantic Internet sea cables, allowing the exchange of Internet traffic between continents and countries. These ISPs exchange traffic strictly through peering agreements (agreement between Internet networks on a connection to exchange traffic).
- A Tier 2 ISP is an Internet provider that engages in the practice of peering with other networks. It utilizes a combination of paid transit and peering to reach some destination(s) within an Internet Region. Tier 2 ISPs are typically regional or national providers and are connected to Tier 1 and Tier 3 ISPs.
- A Tier 3 ISP is an Internet provider that strictly purchases Internet transit. It is engaged in delivering Internet connections to end customers such as residential homes and businesses.

Figure 17 presents an example of a connection between Tier 1, Tier 2, and Tier 3 Internet providers.

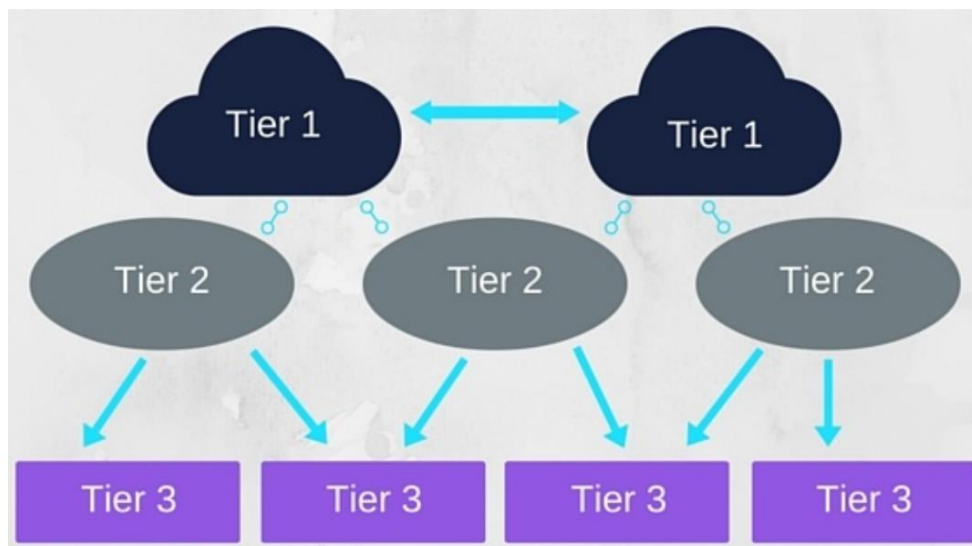


Figure 17 - Example of connection between Tier 1, Tier 2, and Tier 3 providers (Datapath.io, 2016)

Throughout this section were presented the main concepts and terminologies used throughout this report. This will allow the reader to better understand this report.

2.3 Port Scan

A port scan is a technique used in the context of cybersecurity to obtain information from systems (Ceron et al., 2019). When it is performed in a system, it sends back metadata, also known as banners, which contains information that describes the characteristics of the system. This data can be costumed messages configured to cause misinformation or sensitive information concerning the system, which can potentially compromise its security (Shodan, 2019). While security professionals use this technique to identify open ports and other vulnerabilities, malicious users can use it to obtain information that may allow them to improperly access and interact with the systems.

When someone scans a system, a request connection is sent for each port. If the connection is successful, the port is determined as open. Nevertheless, there are port scanning tools that provide more granular information. Scanned ports can be classified as (Nmap, 2019a; WhatIsMyIPAddress, 2019):

- **Open** - The port is actively accepting connections and packets. Usually, administrators try to close or protect them with firewalls, while attackers want to exploit them;
- **Closed** - The port does not accept connections and ignores all packets directed at it;
- **Filtered** - The port scanning tool cannot determine if the port is open or not. When it is possible to determinates that it is open, limited information is provided;
- **Unfiltered** - The port is accessible. However, the port scanning tool cannot determine if the port is open or closed;
- **Open | filtered** - The port scanning tool cannot determine whether the port is open or filtered. This happens when open ports do not respond;
- **Closed | filtered** - The port scanning tool is unable to determine whether the port is closed or filtered.

It is necessary to keep in mind that all the information retrieved when performing port scanning technique is based on what the scanned system sends back. The system can have firewalls, which prevents the retrieve of information from the systems or, as said before, the information can be customized messages configured to cause misinformation.

2.3.1 Port scanning tools

Port scanning tools or port scanners are applications used to perform a port scan technique, to identify open ports and the services running on them (SecurityTrails, 2018). There are many port scanning tools emerging, but the most popular is Nmap (SecurityTrails, 2018), having already appeared in several movies, for example, *The Matrix Reloaded*, *The Bourne Ultimatum*, and *Die Hard 4*. Its combination of versatility and usability make it the standard port scanning tool and the most ever used. Its use in the context of cybersecurity is very strong and popular (Lyon, 2008; Margea & Margea, 2011).

Although Nmap is the most used tool to scan ports, scanning the entire public address space with it requires weeks of time or many machines. Taking into account this limitation, a team at

the University of Michigan developed Zmap as an alternative with higher performance. The aim of this technology was not to compete or replace Nmap but to perform comprehensive Internet-wide research scans more rapidly.

The following subsections, describe Nmap and Zmap.

2.3.1.1 Nmap

Nmap, short for Network Mapper, is a free and open-source application launched in 1997, by Gordon Lyon, to perform port scanning techniques. His goal was to have a tool that would allow the efficient implementation of the various port scanning techniques that were previously offered separately by different port scanners into a single tool (Lyon, 2008).

Nmap has grown to become the most popular port scanner of the world, thanks to the power of open-source development, which has led to the addition of several features over the years by millions of users from around the world (Lyon, 2008; SecurityTrails, 2018). This promoted the progressive reduction of bugs and the introduction of features, following the growing demands of the cybersecurity world.

Nmap discovers devices on a network and it determines the states of its ports, the services running on them, the operating system used, and many other information related to the system. It supports all the major Unix, Windows, and Mac OS platforms, with both console and graphical versions. Table 2 shows how to install Nmap on the most popular Linux distributions, as well as Mac OS platforms.

Table 2 - Commands to install Nmap on the most popular Linux distributions and Mac OS platforms (Lyon, 2008)

Operation system	Command to install Nmap
CentOS/RHEL based distributions	yum install nmap
Ubuntu/Debian	apt-get install nmap
MacOS	fink install nmap sudo port install nmap

To install Nmap on Windows, users have two options: use the windows self-installer or Command-line Zip Binaries (Nmap, 2019b). Most Windows users choose the first option since it is easier and it provides the option to install Zenmap (the graphical version of Nmap). In the Command-line Zip Binaries installation option, there is no graphical interface included, so it is necessary to run nmap.exe from a DOS/command window or download and install a superior command shell. More information and instructions for installing and executing Nmap can be consulted on the Nmap Official Guide.

The Figures below show examples of a simple Nmap scan to a specific target (Metasploit 2 - a test environment with many vulnerabilities). Both Figures show the results for the same port scan: Figure 18 shows the results of the scan using the command-line and Figure 19 uses the interface GUI (Zenmap).

The above Figures show a short example of how Nmap can be used to discover network services and scan remote ports.

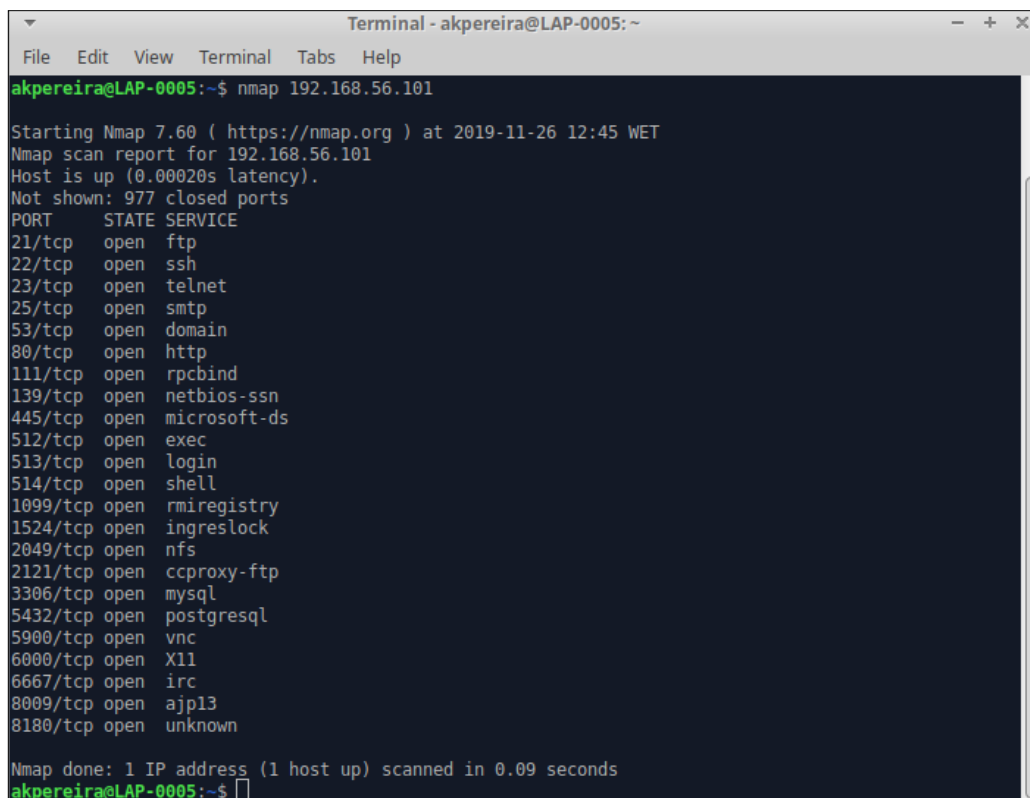


Figure 18 - A simple Nmap on Metasploit 2 scan using the command-line

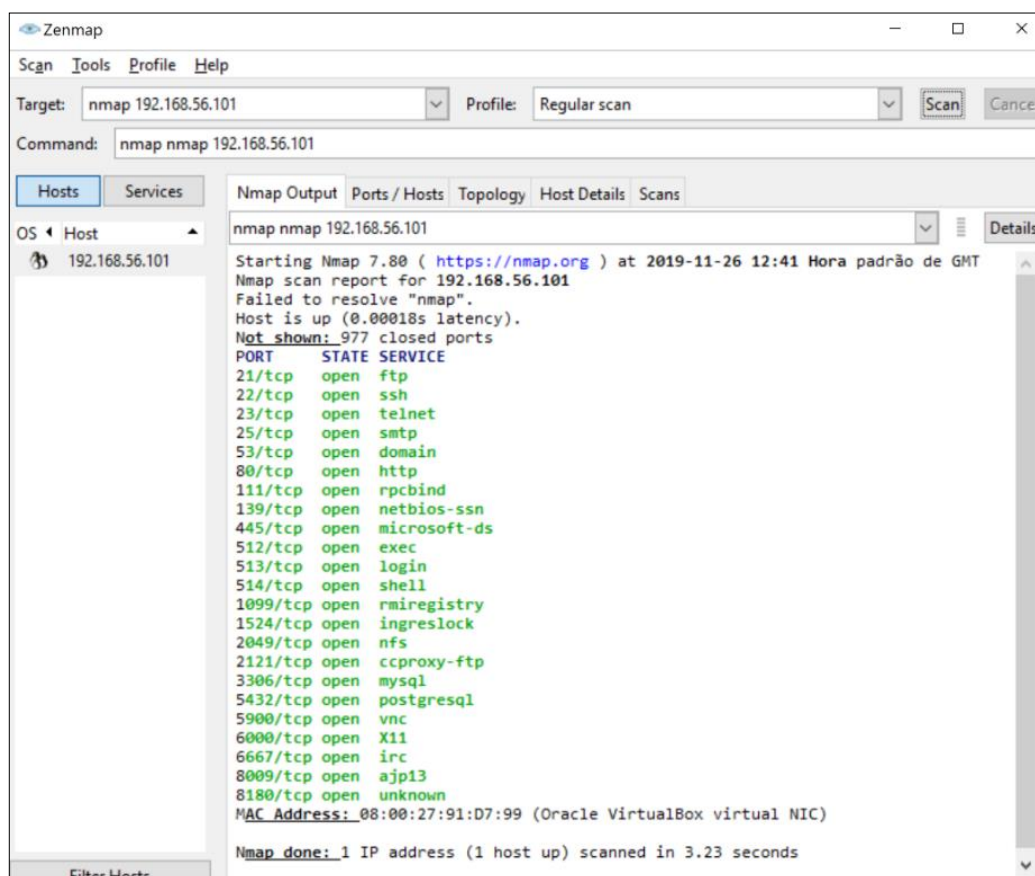


Figure 19 - A simple Nmap scan on Metasploit 2 using the interface GUI (Zenmap)

There are many techniques for doing different types of scans on Nmap. More Nmap techniques can be read on the official documents on the website (<https://nmap.org/docs.html>) or on the Official Nmap Project Guide (Lyon, 2008).

2.3.1.2 Zmap

ZMap is a free and open-source port scanner tool released in 2013, designed by researchers at the University of Michigan, for performing comprehensive Internet-wide research scans, as a faster alternative to Nmap (Durumeric et al., 2013).

It was built specifically to scan a single port or service in the entire Internet. Zmap can scan a single port, on the entire IPv4 address space, in less than one hour. This is 1300 times faster than the Nmap. Previous Internet scanning tools have needed weeks or even months to scan the entire IPv4 address space (Ceron et al., 2019).

Zmap also has an open-source project that provides open-source tools to help researchers to study, on large-scale, the public Internet (Zmap, 2019). It supports all the major Linux distributions and Mac OS platforms. Table 3 shows how to install Zmap on the most popular operation system.

Table 3 - Commands to install Zmap on the most popular Linux distributions and Mac OS platforms

Operation system	Command to install Zmap
Fedora/EPEL	<code>sudo yum install zmap</code>
Ubuntu/Debian	<code>sudo apt install zmap</code>
Gentoo	<code>sudo emerge zmap</code>
macOS (using Homebrew)	<code>brew install zmap</code>
Arch Linux	<code>sudo pacman -S zmap</code>

There are already several projects that perform port scan techniques to the entire IPv4 address space daily, and store the information gathered. These projects can be used as alternatives to obtain information about systems connected to the Internet, avoiding performing port scan techniques again on a system whose metadata information has already been stored and thus, avoiding generating unnecessary network traffic against it.

In addition, performing port scanning techniques implies an ethical discussion, since there are reports of systems that crashed when port scanned. For example, when a vulnerability on heart bleed was discovered in 2014, several systems crashed because many researchers started to perform port scan techniques, intending to find vulnerable systems (Ceron et al., 2019; Partridge & Allman, 2016).

2.3.2 Port Scanning Projects

Port Scanning Projects are search engines that perform port scan techniques, using port scan tools, against systems directly connected to the Internet and store the metadata that they send back (Ceron et al., 2019; Shodan, 2019). The two most well-known port scanning projects are Censys and Shodan.

2.3.2.1 Censys

Censys is a search engine created in 2015 at the University of Michigan in 2015, by the same researchers that developed Zmap (Ceron et al., 2019). It continuously scans every host and digital certificates on the Internet. Censys has helped in the discovery and analysis of many significant Internet vulnerabilities. It aims to help organisations become more secure against attacks and improve their systems (Censys, 2019b).

Censys regularly scans every public Internet Protocol (IP) address and make the information about them intelligible through the search engine website. It also provides access to all of their data through an API. The data provided by the API is the same as those accessed through the web interface.

Censys claims that its mission is to make security driven by data, giving cybersecurity professionals information to discover, monitor, and analyse devices connected to the Internet.

The endpoints of the API are hosted at <https://censys.io/api/v1/>. To have access to them, the user needs to authenticate and provide the API ID and Key presented in his user account.

Censys has 3 datasets that store information from daily ZMap scans of the Internet, available field-based filters. Figure 20 shows the layout of the main page of Censys website. The central part of the webpage has a search box where the user can search for: IPv4 Hosts, Websites, and Certificates. To access the information related to the IPv4 address space, it is necessary to choose the IPv4 Hosts option.

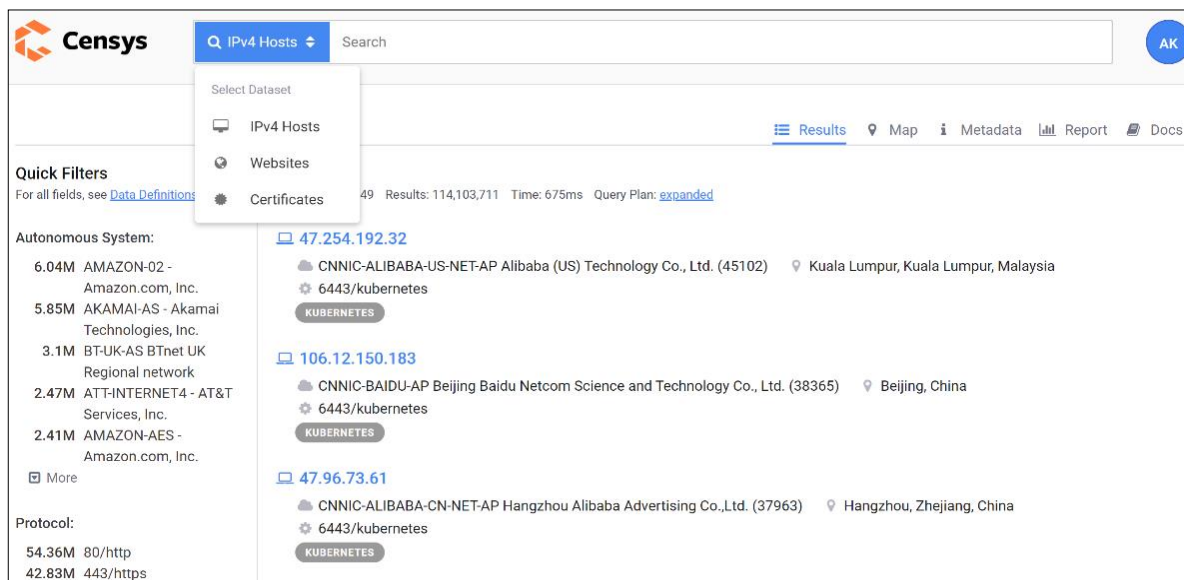


Figure 20 - Main page of the Censys website

By choosing one of the displayed IP addresses when we select the IPv4 Hosts in the search box, we can see all the information about the network address. Figure 21 illustrates the information gathered when we select the IP address (123.7.172.6). As we can see, the host has three services open: 80/HTTP, 23/Telnet, and 22/SSH. Censys also provide information such as the AS, the organisation, the location, and other information of the network address retrieved on the metadata.

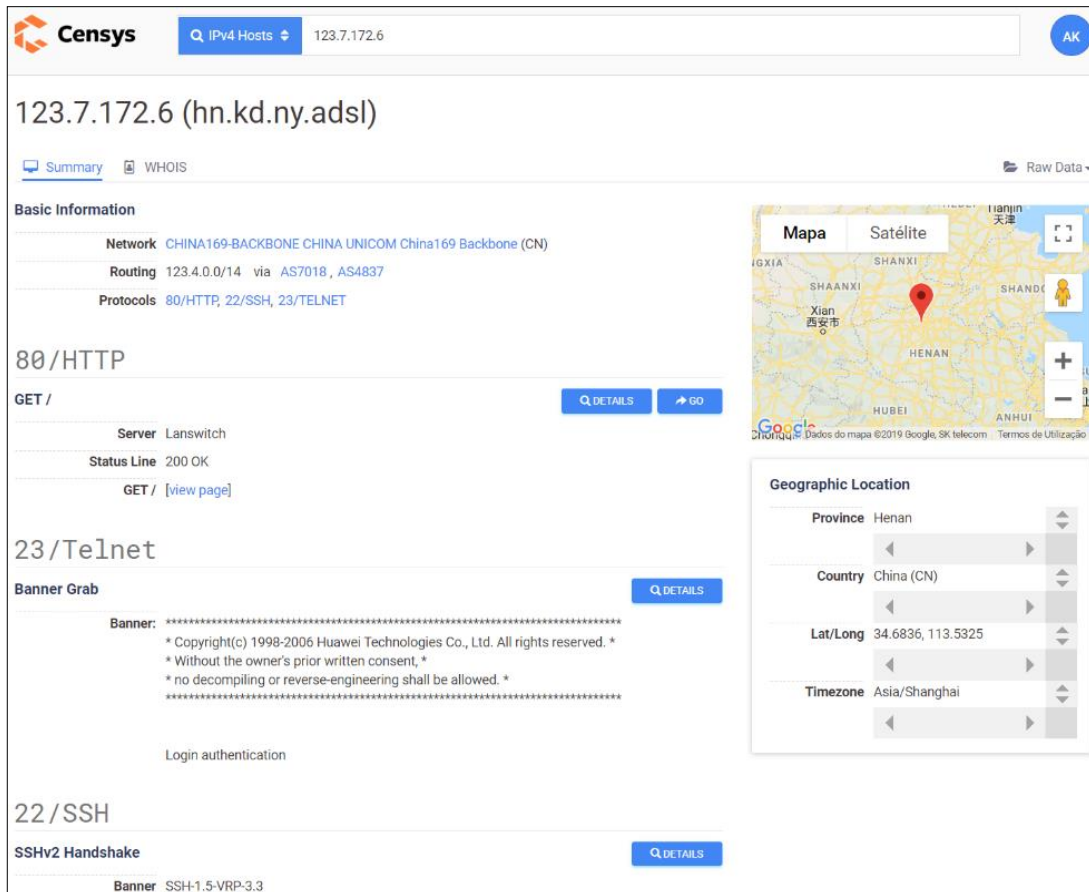


Figure 21 - IPv4 result page

It is also possible to search for specific results by using specific restrictions on the queries. Scan by metadata includes searches by Tags (specific values attached to some hosts), namely: Location, Manufactures, Autonomous System, and many other characteristics based on the banner information (Hudak, 2018). In (Censys, 2019a) it is possible to see the list of all tags that we can search for. Below, we present some examples of this type of research.

In Figure 22 we show results of the search for the tag “scada”. In Figure 23, we present the output of the search for country code “PT” (Portugal). In Figure 24, we show the results of the search for autonomous system code 2860 (NOS Comunicações).

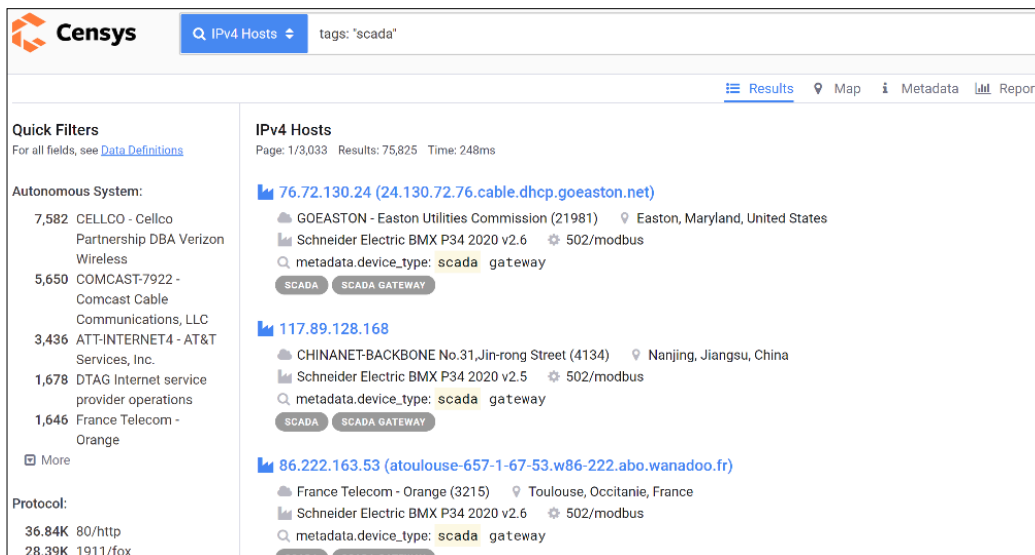


Figure 22 - Query result when searching for the tag “scada”

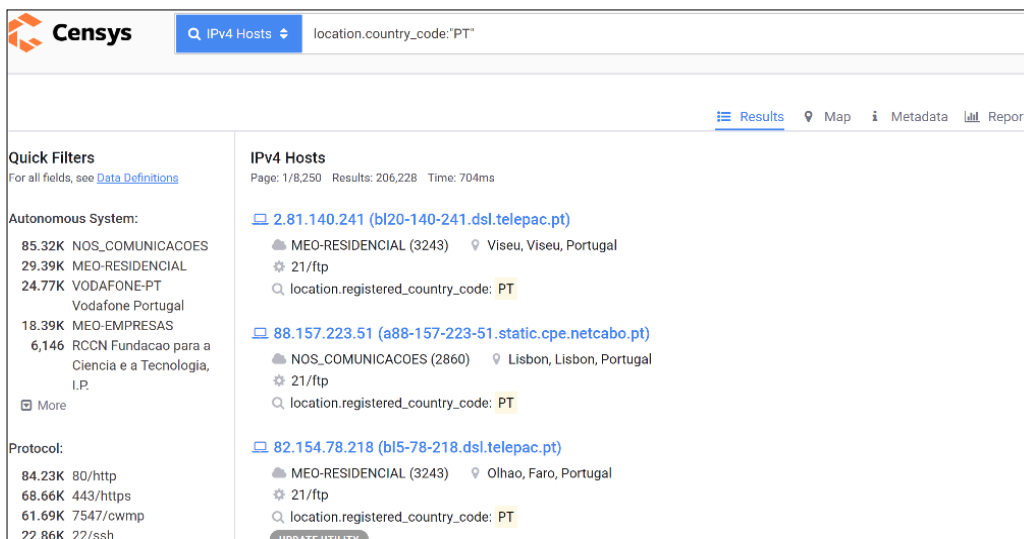


Figure 23 - Query result when searching for country code “PT” (Portugal)

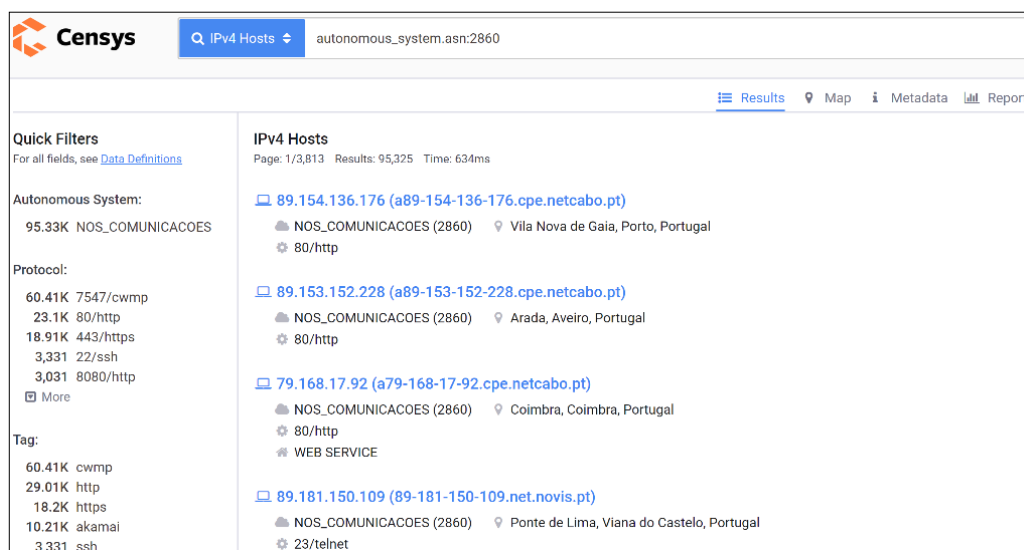


Figure 24 - Query result when searching for autonomous system code 2860

It is also possible to perform more complex searches introducing "AND" or "OR" in the query. For instance, Figure 25 shows a search, for hosts located in Portugal and tagged as camera, and Figure 26 shows a query to search OpenSSH servers running on Debian across Europe.

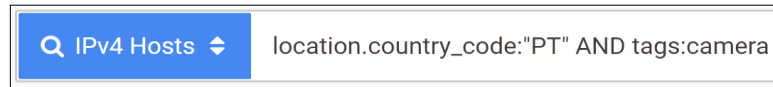


Figure 25 - Query to search for hosts tagged as “camera” and located in Portugal



Figure 26 - Query to search for OpenSSH servers running on Debian across Europe

As mentioned before, in addition to the search for "IPv4 Hosts”, Censys also makes it possible to search for websites (this option is visible in Figure 20), which provides almost the same view as IPv4 hosts. In Figure 27, we can see the interface of the Website search option.

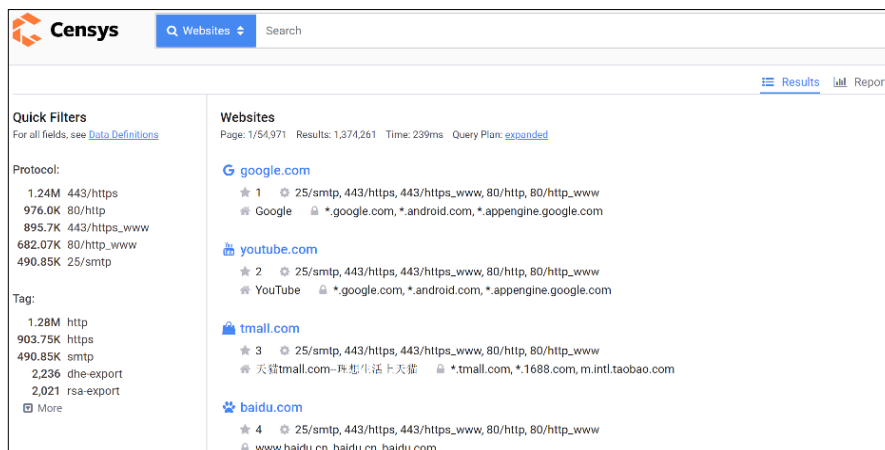


Figure 27 - Censys Interface of Websites search option

Selecting the option “Certificates” on the search box, it is possible to obtain information about the existing digital certificates on the scanned hosts, nearly in real-time. In Figure 28, we can see the interface of the Certificate search option. It shows a list of several digital certificates found by Censys. By selecting one of them it is possible to see more detailed information.

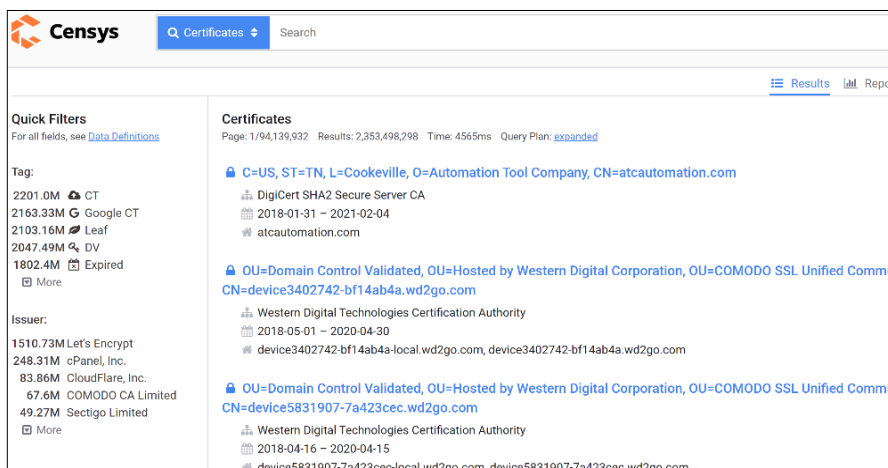


Figure 28 - Censys interface of Certificates search option

In 2018, Censys became a private initiative. Nowadays, clients must pay for full access to the data according to two plans: Censys Pro and Censys Enterprise. The former addresses companies that use Censys data internally, but for commercial use. It provides threat hunters, penetration testers, and security analysts to discover and understand the threat landscape of the company or their client's digital infrastructures. It allows performing up to 25000 queries per month, which may return up to 25000 results each. Users can access data through the Censys website or via the API (Censys, 2019b). The Enterprise plan covers the needs of organisations that intend to access all Censys data or need a certain type of information that can only be achieved with personalized queries. These accounts have access to Censys datasets, for offline access, through a Google BigQuery account and to the Censys API (Censys, 2019b).

Censys remains committed to providing free data to the research community. Thus, it provides researchers access to their data through a Google BigQuery account. This allows researchers to run SQL commands and access all the data provided in the Enterprise plan (Censys, 2019a).

In Figure 29 we can see the layout of the Google BigQuery to access Censys data, where we can see the total number of IPv4 network addresses scanned in the latest Censys scan.

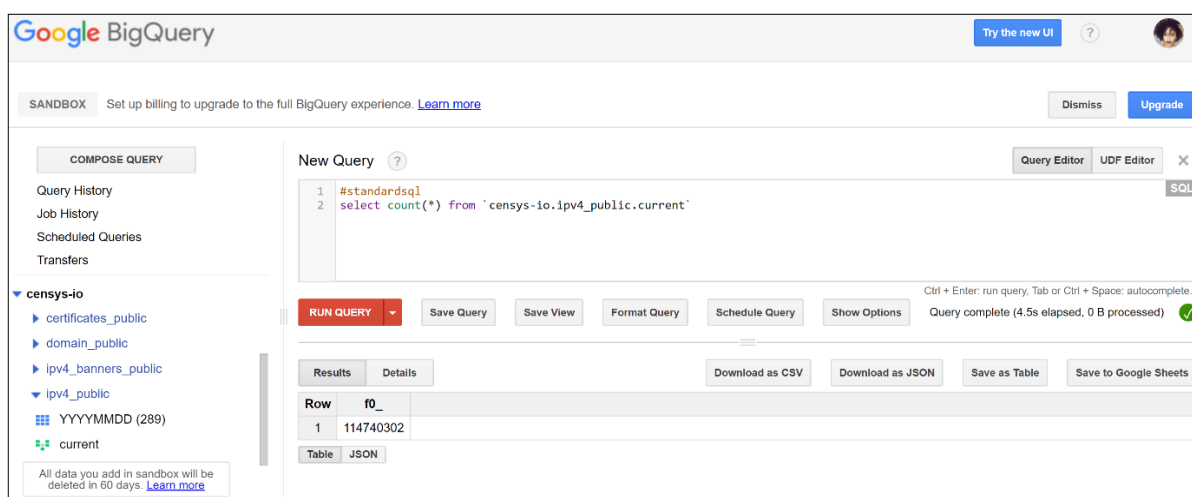


Figure 29 – Access to Censys data through Google BigQuery

2.3.2.2 Shodan

Shodan is a search engine that allows users to obtain information about devices directly connected to the Internet. It was created in 2009 by John Matherly. Shodan website advertises the tool as “the world’s first search engine for Internet-connected devices” (Shodan, 2019).

In 2013 Shodan was able to scan the entire IPv4 address space, and since then, it updates its database every day, in real-time. Shodan does not reveal the used port scanning tool (Ceron et al., 2019). This search engine also provides a public API that allows users to access and interact with their data. Figure 30 shows the main page of the Shodan website.



Figure 30 - Main page of Shodan website

Although Shodan’s greatest value is to help defenders to find vulnerable devices on their networks, it is not responsible for the use given to the information obtained on it. The following is written on their website page: “We provide the platform that ensures accurate, consistent and up-to-date information on Internet-facing devices - it's up to you to decide what type of information you're most interested in” (Shodan, 2019).

Shodan works as follows: it starts by generating a random IP address and randomly selects a single service port to scan. If it succeeds, then it stores the IP address and the banner received from the host. If the scan is unsuccessful, then it generates a new random IP address and service port. Figure 31 illustrates this cycle.

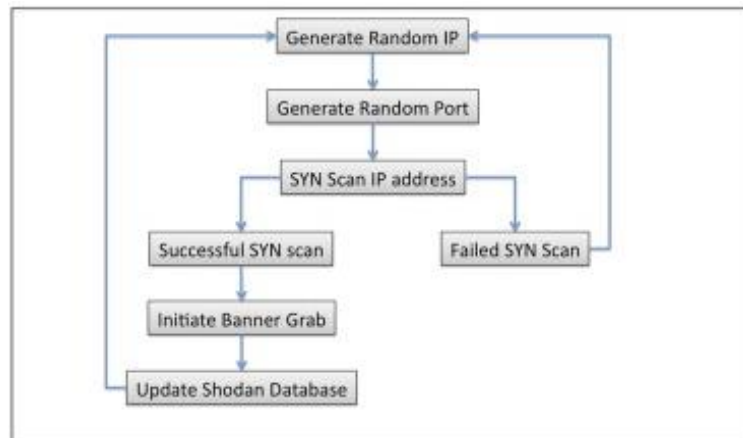


Figure 31 - How Shodan works (Bodenheim, Butts, Dunlap, & Mullins, 2014)

In Figure 32, we can see the results obtained by Shodan when we search for the protocol Modbus (one of the ICS standard protocols). It found 352 results for this query, which is subdivided into 10 pages of results. For each of the results, it shows the information gathered from the metadata.

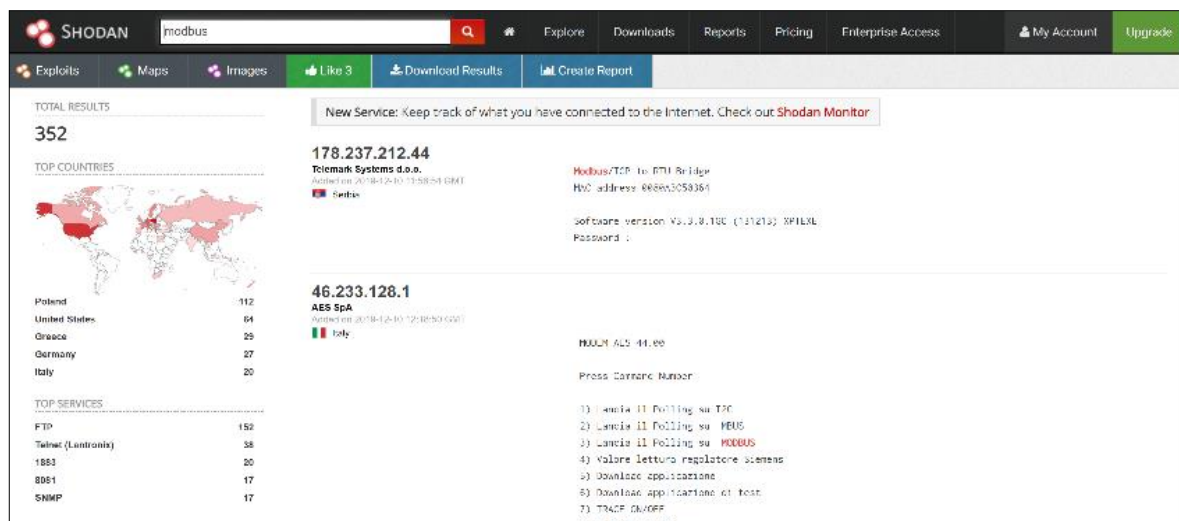


Figure 32 – Shodan’s results when searching for protocol Modbus

Although Shodan provides 10 pages of results for the search present in Figure 32, we only have access to 2 pages using the free version. Shodan is a private initiative, and although it allows access to some basic data, for additional data, it is necessary to pay. It has 3 payment plans: Freelancer, Small Business, and Corporate, all of them with no contracts, no setup fees, and users can cancel the plans at any time. Table 4 presents the prices and the features that each plan gives access to. The information about the level of access was retrieved from Shodan official website, and the prices are in USA dollars (Shodan, 2019).

Table 4 - Prices and the features that each plan gives access (Shodan, 2019)

Price and Features	Freelancer	Small Business	Corporate
Cost per month	\$59	\$299	\$899
Number of results per month	Up to 1 million results	Up to 20 million results	Unlimited results
Number Scan per month	Up to 5,120 IPs	Up to 65,536 IPs	Up to 300,000 IPs
Number of IPs for Network monitoring	5,120 IPs	65,536 IPs	300,000 IPs
Access to the filters	Most of the filters	Most of the filters	All filters
Paging through results	yes	yes	Yes
Access to the API	yes	yes	Yes
Commercial use	yes	yes	Yes
Support	E-Mail support	E-Mail support	Premium support; Complimentary membership upgrades
Vulnerability search filter	no	yes	Yes

Instead of buying the plans presented above, Shodan also has the option to purchase credits for specific use. The company offers three credit options: Export credits, Query credits, and Scan credits.

- Export credits allow users to export data from Shodan. They are single-use, and each one allows to download up to 10000 results;
- Query credits allow users to perform queries through the API. One query credit allows to download 100 results;

- Scan credits allow users to request network scans via API. One scan credit allows to scan one IP address. Both query credits and scan credits renew every month.

Shodan also offers a free plan to academic users. The academic account permits: download up to 10000 results per month; scan up to 100 IPs per month, and network monitoring for 16 IPs. To have access to the academic account, users should create a normal account in Shodan using their university email and request for the upgrade to the academic account, by sending an email to the Shodan support service with their username and email associated with the Shodan account.

2.4 Data Warehouse

A data warehouse (DW) is an information system that aggregates structured data from single or multiple sources, to conduct data analyses that help in the decision-making processes (Tutorialspoint, 2019). It operates as a central repository of information, to support analytical reporting, structured and unstructured queries, and decision making (Almeida, 2017).

2.4.1 Characteristics of a data warehouse

A data warehouse has 4 main characteristics: subject-oriented, integrated, time-variant, and non-volatile. Below we explain each of the characteristics (Malinowski & Zimányi, 2008):

- It is subject-oriented as it offers information regarding a specific and defined theme. It provides a simple and concise view, by excluding data that is not necessary for decision making;
- It is integrated as similar data from different sources are converted to a common unit of measure. For example, naming conventions, format, and coding must be consistent;
- It is time-variant as the data in a data warehouse provides information from a historical point of view. Thus, it is identified, implicitly, or explicitly, with a particular time. An additional aspect of the temporal variation is that, once the data is inserted, it can no longer be updated, changed, or deleted;
- It is non-volatile because previous data in the data warehouse is permanent. This means that it is not deleted when new data is added. Therefore, data is read-only and periodically refreshed. The frequent changes in operational databases are not reflected in the data warehouse.

2.4.2 Architectures of a data warehouse

Before understanding how each of the architectures works, it is necessary to understand the concepts of fact and dimension tables (Malinowski & Zimányi, 2008). The former stores data about events associated with a certain business process, including measurements, metrics, and facts. Each row represents a unique event and contains the measurement data associated with

it. Fact tables use two types of data: foreign keys to dimension tables and numerical facts (measures).

A dimension table provides structured information that describes the events recorded in the fact table. The dimension tables store information about the events on the fact table, allowing users to build queries more easily and intuitively to answer business questions. While fact tables contain information about events, such as login or sales, dimension tables contain references information about the events, such as date, users, customers, etc.

The most well-known data architectures used in the context of data warehouses are (Guru99, 2020a):

- Star schema;
- Snowflake schema;
- Fact Constellation schema.

A star scheme is composed of one, or multiple fact tables, referencing a set of dimension tables. Its diagram resembles a star with a fact table on the centre, with points radiating from the centre to the associated dimension tables. It is the simplest schema to build data warehouses. Figure 33 presents an example of a star schema, where we can see the fact table Revenue referencing to the dimension tables: Dealer, Branch Dim, Date Dim, and Product.

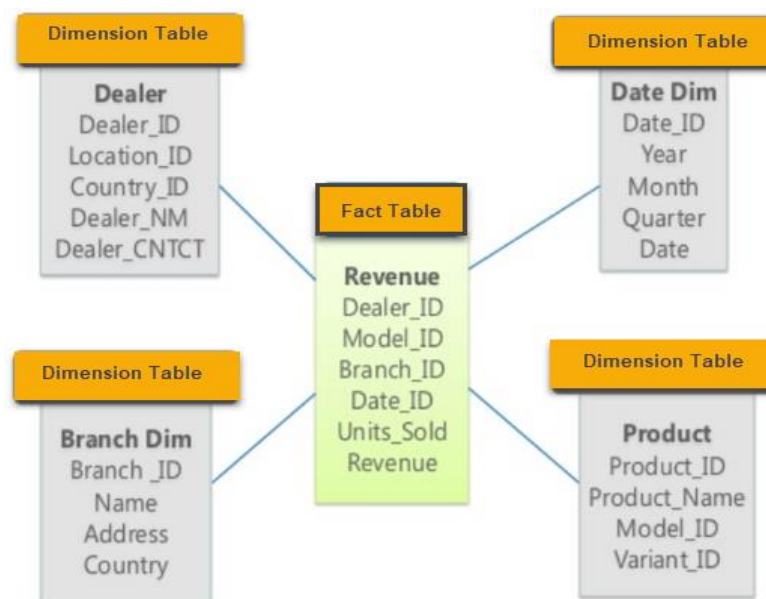


Figure 33 - Example of a star schema (Guru99, 2020a)

The snowflake schema is a variant of the star schema. One of the main differences between them is the normalization of the dimension tables in the snowflake schema, which splits data into additional tables. The process is called snowflaking. The snowflake effect only the dimension tables and does not have any impact on the fact tables (Guru99, 2020a). In Figure 34, we can see an example of a snowflake schema, where the dimension Dealer is normalized, point to the other two dimensions table: Location and Country.

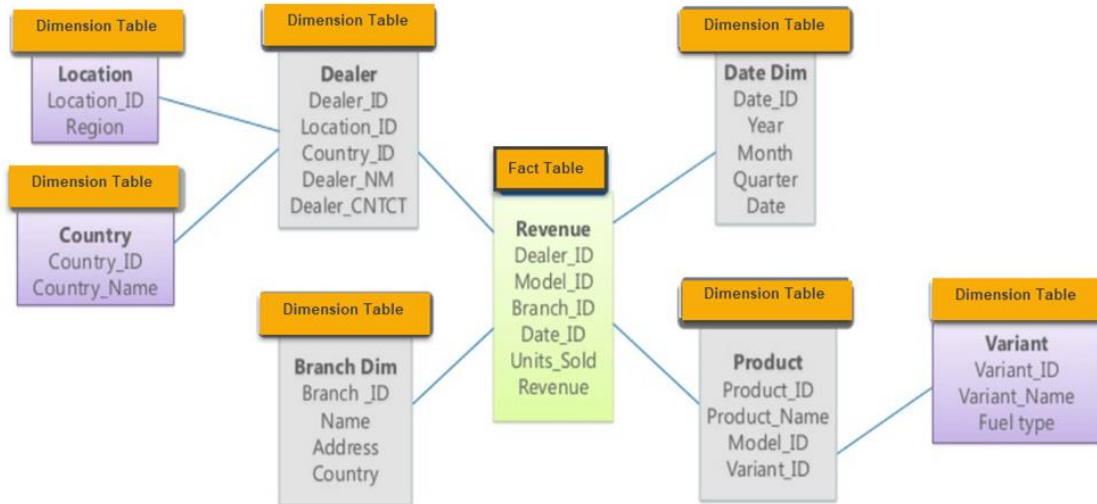


Figure 34 - Example of a snowflake schema (Guru99, 2020a)

A Fact Constellation schema also referred to as Galaxy schema, is a collection of multiple fact tables that share some dimension tables. Therefore, it can be viewed as multiple star schemas with some dimension tables in common. It is one of the most widely used schemas for data warehouse designing and it is much more complex than the star and snowflake schema (GeeksforGeeks, 2020a). In Figure 35 we can see an example of a Fact Constellation schema, where two fact tables: Revenue and Product, shares the dimension tables Branch Dim and Date Dim.

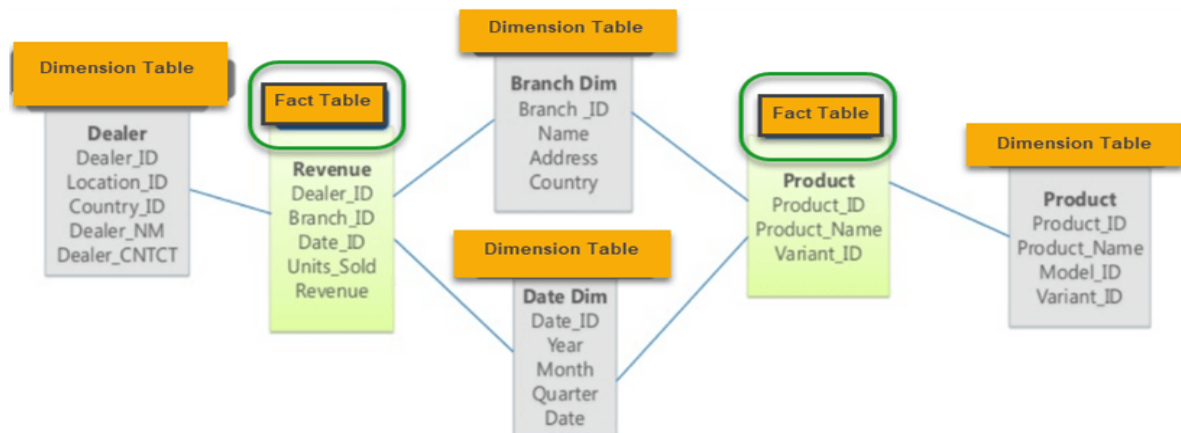


Figure 35 - Example of a Constellation schema (Guru99, 2020a)

The data gathered from the different sources to fill the tables on the data warehouse can be structured, semi-structured, and/or unstructured data. Thus, the data warehousing process includes cleaning, integration, and data consolidation. This data is loaded, processed, and used to assist in decision-making. The process of cleaning, integrate, and performing all the changes to transform the data to a uniformized format in the data warehouse is called ETL (Extract, Transform, and Load) (Almeida, 2017).

2.4.3 ETL (Extract, Transform and Load)

ETL is the abbreviation for Extract, Transform, Load. It corresponds to the process responsible for the extraction of data from different data sources, their conversion, cleaning, normalization, and their loading into the target system (Trujillo & Luján-Mora, 2003).

As the name indicates, the process is divided into three phases:

- **Extract** - It is the process of reading all data necessary to the data warehouse. The data can be gathered from multiple and different types of sources or a single source.
- **Transform** - It is the process of converting the extracted data into the required format. In this process is performed all the conversions, cleaning, normalization, and the changes needed to transform the gathered data into a unified format in the data warehouse.
- **Load** - It is the process of populating the transformed data into the target database, data warehouse, or another system.

Figure 36 shows the general schema of an ETL process, where the data is extracted from different sources, it is transformed into the required format, and then loaded into the target system.

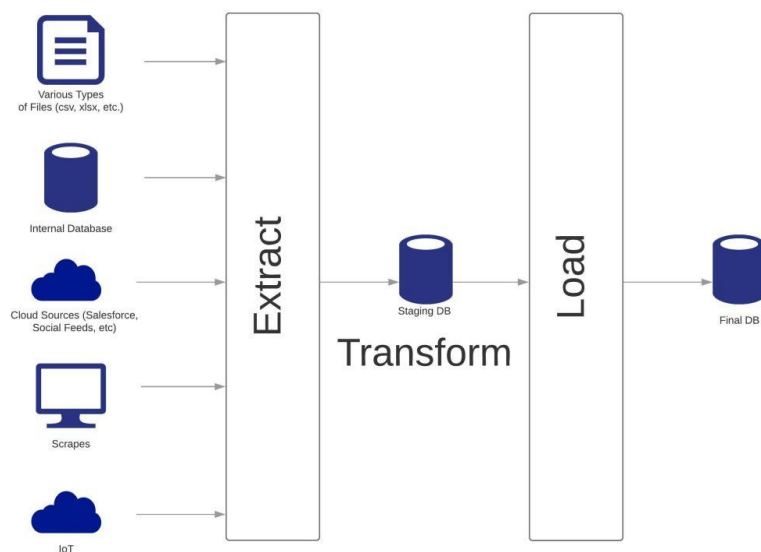


Figure 36 - Schema of an ETL process (Folly, 2018)

An important step of an ETL process is to ensure that changing business requirements will be incorporated in the data processing (Ahmed, 2020). This step is called Periodic ETL, it usually involves an initial loading of all data, followed by a periodic loading of incremental data changes (IBM Cloud Education, 2020).

2.4.4 Open-source Data warehousing Tools

Open-source tools play an important role in the data warehousing field. Open-source communities have helped to improve and accelerate the development of tools. They can be a

low-cost alternative to commercial solutions. The use of open-source software brings many benefits to users, namely (Alley, 2018; Congdon, 2015):

- Lower software costs, saving on licensing, maintenance fees, and other costs;
- Huge support, since open-source solutions generally have broad support from the scientific community;
- Escapes of constant frustrations such as lack of portability and inability to customize the software among other compliances.

The benefits provided by open-source solutions contributed to the researcher's belief that open-source data and analytic solutions are the future. Thus, considering the many benefits of using open-source solutions, all the tools to build, store, and analyse our data warehouse will be open-source.

The process of building a data warehouse usually includes choosing the ETL tool, the database engine, and the analytics tools to help organisations understand and visualise their key strengths and weaknesses. To ensure interoperability, it is crucial to ensure compatibility between the adopted ETL tool and the database engine, as well as the reporting tools.

2.4.4.1 ETL Tools

ETL tools, also known as Data Integration Tools, perform the data integration process on a data warehouse. The process includes the extraction of data from different systems, its transformation and polishing (e.g., cleaning, normalization, and mapping), and finally, the loading to the data warehouse.

After a study among the well-known and prestigious open-source ETL tools and their data integration capabilities (Alley, 2018; King, 2019; Techroba, 2015), we analysed the following ETL tools:

1. Pentaho Data Integration (Kettle);
2. Talend.

Pentaho Data Integration and Talend are two of the most popular and impeccable open-source ETL tools (Bruce, 2020).

➤ Pentaho Data Integration (Kettle)

Pentaho Data Integration (PDI), also referred to as Kettle, which stands for "Kettle Extraction Transformation Transport Load Environment" (Hitachi Vantara LLC., 2020b), is a component of Pentaho, a subsidiary of Hitachi Vantara (Hitachi Vantara LLC., 2020a). Kettle is an open-source project that offers ETL capabilities using a metadata-driven approach.

The tool is easy to use and simple to learn and understand. It has a user-friendly graphical design with drag-and-drop features that allow users to manipulate the data without executing a single line of code. However, this does not reduce the ability to implement complex ETL procedures.

In addition, Kettle provides a shared repository that enables remote ETL execution (Techroba, 2015).

The software comes in a free community edition and a paid enterprise edition, that offers more capabilities than the free version, such as automated scheduler, shared repository, team collaboration, versioning and locking, and secure data integration.

➤ **Talend Open Studio for Data Integration**

Talend Open Studio for Data Integration (Talend, 2020) is an open-source (Apache license) segment of Talend, a platform that owns various software and services for data integration, data management, cloud storage, and Big Data.

It works as an ETL tool, helping users to turn disorganised data into business insights. It offers a clear and easy-to-understand Eclipse-based interface (with drag-and-drop design flow), which allows the development and deployment of data integration jobs faster than hand-coding (Guru99, 2020b). The tool offers more than 400 connectors for various RDBMS (Relational Database Management System), SaaS, packaged apps, and other technologies. It also enables the exporting and execution of standalone jobs in runtime environments (Techroba, 2015).

Talend Open Studio for Data Integration was the first open-source data integration software, making Talend company the first commercial vendor to present this solution to the market (Chand, 2020). Talend Product Suite has three main products: Talend Big Data, Data Integration, and Integration Cloud.

➤ **Comparison**

Choosing the tool to perform the ETL process of this work, implies a tool that is open-source, easy to use, and that supports the characteristics of the data. Therefore, to choose the right ETL tool for this project we evaluated the following factors: Budget, Data Needs, Interface.

- **Budget:** The choice of tool is influenced by the available budget. In this study, to avoid costs, we prefer tools that are open-source and that are available for free. Based on the study carried out we observed that the analysed tools are both open-source and the commercial ones offered community editions, with ETL features, for free.
- **Data Needs:** Within this factor, it is important to answer questions, such as whether the volume of data currently processed and expected in the future, the number of sources accessed, and the data sources and data transformations supported. Concerning the volume of data, both of the tools support millions of records. Both tools support more than two data sources - the number of data sources required for this project. The analysed tools support equivalent data transformations features.
- **Interface:** Within this factor, it is necessary to analyse if the tool is easy to use, visually appealing, and provides a user-friendly interface. Both of the tools are user-friendly and easy to use, with drag-and-drop capabilities that allow users to perform data transformations without having expertise knowledge or execute a single line of code. However, Pentaho Kettle's GUI is easier to run when compared to Talend's GUI.

- **Familiarity:** This factor analyses how comfortable the team is with the tool and the compatibility of the tool with the technologies used in the hosted company. In terms of technologies, one of the requirements was the adoption of an ETL tool that can be installed in a Linux distribution - the main operating system used by the hosted company. All the analysed tools support distributions for Linux. In terms of how comfortable the team is with the tool, the intern has experience with Pentaho Data Integration.

As we can see, both Talend and Pentaho Data Integration carry similar characteristics. All the solutions are open-source, present almost the same architecture and data transformation features, and have a user-friendly interface, with drag-and-drop capabilities. However, Pentaho Data Integration has the advantage of the Interface factor for presenting an easier to run GUI compared to Talend's GUI and on factor Familiarity, in which the intern presents experience with the tool.

Among the factors analysed, Pentaho Data Integration seems to present some slight advantages compared to Talend Open Studio for Data Integration. Thus, to analyse if the former is the best choice for this project, bellow we present some of the other advantages of Pentaho Data Integration over Talend, which are not crucial to this work but would add value to it (Bruce, 2020; Educba, 2020):

- Pentaho Data Integration is twice faster than Talend;
- Pentaho offers a wide range of connectivity to extensive databases, while Talend offers limited connectivity to concurrent databases;
- Talend depends on Java drivers to connect to the data sources;
- In Pentaho, users have the option to store the files in their personal systems or in a centralized database repository, while Talend works at the filesystem level, where the data can only be stored in the personal system;
- In collaboration with Hitachi Vantara portal, Pentaho offers its users customer support, while in Talend, to get support, users need to register for a technical support account.

Taking into account the factors analysed and the advantages of Pentaho Data Integration over Talend Open Studio for Data Integration, the former is the tool chosen to perform the ETL processes of this work.

2.4.4.2 Database Engines

Traditionally, databases have been proprietary tools provided by companies like Oracle, IBM, and Microsoft. However, over de past years, open-source databases have grown steadily in maturity and importance, providing similar capabilities to non-open-source database engines (TrustRadius, 2020). In addition to the obvious cost savings, the richness of data available through the community, and the ease of finding out how to adjust the database to function as intended.

After a study among the most used and well-known open-source database engines, and the ones supported by Pentaho Data Integration (Miller, 2019; Slant, 2020; Tamang, 2019; TrustRadius, 2020), we reached the following database engines:

1. PostgreSQL;
2. MySQL.

➤ **PostgreSQL**

The University of California at Berkeley developed the PostgreSQL database system in 1986 (as part of the POSTGRES project) (The PostgreSQL Global Development Group, 2020). It has more than 30 years of active development and it is one of the most advanced and powerful open-source object-relational database systems.

PostgreSQL aims to help developers and administrators to build applications and fault-tolerant environments, as well as protect data integrity and facilitate the management of the data regardless of how big or small is the dataset. It supports many of the standard SQL features, although sometimes with a slightly different syntax or function.

The database system under review is free and open-source, and it runs on all the major operating systems. It has been ACID-compliant since 2001 and has one of the best existing community supports (The PostgreSQL Global Development Group, 2020). Furthermore, it is highly extensible and flexible, allowing the definition of custom data types and functions and even writes code from programming languages like Python, Perl, Java, Ruby, C, and R.

Due to PostgreSQL's proven architecture, reliability, data integrity, robust feature set, extensibility, and the dedication of its open-source community, it has earned a strong reputation and has long been the preferred solution for businesses of all sizes (Bocetta, 2019).

➤ **MySQL**

MySQL (Oracle Corporation, 2020a) is an open-source SQL relational database management system, written in C and C++, acquired by Oracle Corporation in 2010 (Oracle Corporation, 2020c). The name MySQL is a concatenation of "My", the name of the daughter of Monty Widenius, one of the co-founders of MySQL, and "SQL", an acronym for Structured Query Language.

MySQL Database Software is a client/server system that supports multiple back-end technologies, many different client programs and libraries, management tools, and a wide range of Application Programming Interfaces (API). It also provides an integrated multithreaded library (MySQL Server), a fast and easy to use solution that can be linked to applications for smaller, faster, and easier to manage standalone products.

MySQL uses a very fast thread-based memory allocation system and has independent modules within a multi-layered server. It provides transactional and non-transactional storage engines, executes very fast joins using an optimized nested-loop join, and it implements SQL functions using a highly optimized class library (Oracle Corporation, 2020b).

On the official website, MySQL is presented as “the most popular open-source SQL database management system”. It is under the GNU General Public License and available under some private licenses. A huge and active community of open-source developers supports it.

➤ **Comparison**

The choice of the database to be used in this project depended on the ones supported by the ETL tool chosen and among these the security factor will be differential. Considering that we are working with data from critical systems, the chosen database needs to have a set of measures, processes, and methodologies to protect and secure the data from illegitimate use, malicious threats, and attacks.

In terms of database engines supported by the chosen ETL tool, both PostgreSQL and MySQL are supported by Pentaho Data Integration. In terms of security, PostgreSQL is one of the database engines most concerned with security issues, it claims to always choose security over anything else, while other databases engines mainly try to show better benchmark results (Slant, 2020). This does not mean that it is the most secure database engine or that it does not present any vulnerability but it indicates that PostgreSQL takes security issues seriously and that these issues are always resolved as quickly as possible. For this reason, PostgreSQL will be the chosen database engine to use in this work.

2.4.4.3 Analytics Tools

Analytic tools stand for any application designed to automate the process of building and visualise reports, queries, graphs, and other methods of data analysis to infer conclusions. Data visualization can be a valuable support in the decision-making process.

In (Oliveira & Bernardino, 2011) the authors did an exhaustive search to find open-source analytics solutions that allow users to analyse data without an IT background. Curiously, that research revealed a reduced number of open-source solutions and many of them were not well rated or recommended by distinguished research companies. Considering the reduced number of tools found, the authors built a top five that includes two open-source solutions: Metabase and Pentaho Business Analytics, and three of the most popular free solutions, which are not open-source: Power BI Free, QlikView, Tableau Public. Due to the decision of using only open-source solutions on this work, we will analyse Metabase and Pentaho Business Analytics.

➤ **Metabase**

Metabase (Metabase, 2020) is a web-based, open-source visual query and business intelligence tool, released in 2015, intended for companies of all sizes. It is built and packaged as a Java jar file and can be run anywhere where Java is available.

It allows filter and/or group data according to user needs, without resorting to SQL, and also provides an SQL interface for users to build SQL queries if needed. This tool allows users to ask and monitor questions, as well as learn from available data. These questions can be visualised through graphs and charts, which can be organised and saved into great-looking

Dashboards (Santos et al., 2019), to be seen later. It is also easy to share questions and dashboards with the rest of the team. Metabase supports the following databases: BigQuery, Druid, Google Analytics, H2, MongoDB, MySQL, MariaDB, PostgreSQL, Presto, Amazon Redshift, Snowflake, Spark SQL, SQLite, and Microsoft SQL Server.

Metabase is available under three types of licenses: Affero General Public License (AGPL), which is free of costs, and other two versions: Premium Embedding License Metabase and Commercial License with acquisition costs (Oliveira & Bernardino, 2011).

➤ **Pentaho Business Analytics**

Pentaho Business Analytics is an open-source component of Pentaho, responsible for the data analysis processes, released in 2009, and it is frequently updated. The tool was one of the first open-source solutions released for data analysis, and it aims to “Empower data consumers with interactive, real-time visual data analysis and predictive modelling, with minimal IT support” (Hitachi Vantara LLC., 2020a).

It runs on almost every platform: Android, iPhone, iPad, Mac, Web-based, and Windows, and supports the following databases: JDBC, IBM DB2, Microsoft SQL Server, MySQL, Oracle, PostgreSQL, NCR Teradata, and Firebird. The type of graphs and layouts proposed for some diagrams and charts are limited and partly outdated.

Pentaho Business Analytics provides two different editions: The Community Edition that is the open-source version, with restricts access to some more advanced functionalities regarding to data visualization, and the Enterprise Edition that is the commercial version. The latter includes all the features of the Community Edition and some more advanced ones.

➤ **Comparison**

To choose the appropriate analytic tool to perform the data analysis, we listed some of the features that we want the tool to support for free. Table 5 presents this list.

Table 5 - Features supported by Pentaho Business Analytics and Metabase for free (Hitachi Vantara LLC., 2020a; Metabase, 2020)

Functionalities	Pentaho Business Analytics	Metabase
Ad-hoc reporting		✓
Ad-hoc query		✓
Data Visualization variety	✓	✓
Dashboard Designer		✓
Interactive Visualization		✓
Predictive Analytics	✓	
Report customization		✓

Although Pentaho is a consolidated open-source product on the market, it lacks several features regarding to data visualization when compared to Metabase, which turned out to be a good option as it presents nearly all basic data analytic functionalities compared with Pentaho.

It is worth highlighting that the fact that we have chosen a Pentaho tool for the ETL process does not imply that we also need to choose a Pentaho Analytic tool. Therefore, considering the variety of functionalities that we have for free, we chose Metabase as the Analytic tool to be used in this work.

2.5 Security Principles

System security principles are principles that govern the security of information systems within an organisation, minimising the chances of an attack on its systems. Taking into account that one of the objectives of this work is the identification and analysis of the vulnerabilities found on the ICSs, this section presents the fundamental principles of security used to measure the risk that a vulnerability represents to the ICSs.

2.5.1 Key principles of security

There are several principles of information security, however, all of them try to address at least one of the three core principles that provide the basis for the major security standards (Merkow & Breithaupt, 2014). These principles are Confidentiality, Integrity, and Availability, also known as the CIA Triad (Keung, 2013; Rountree, 2011).

2.5.1.1 Confidentiality

The principle of confidentiality prevents access of information to unauthorized individuals or systems. It specifies that only the sender and receiver will be able to access the information shared between them. The objective of the confidentiality principle is to ensure that private information remains private and that it can only be viewed or accessed by authorized individuals or systems. The degree of confidentiality determines the secrecy of the information (Keung, 2013).

Confidentiality is compromised if an unauthorized person or system can access the information. For example, sender A wants to share some confidential information with receiver B. It is a breach of confidentiality if the information gets intercepted by an unauthorized attacker C. The interception causes loss of message confidentiality (Rountree, 2011).

2.5.1.2 Integrity

The principle of integrity prevents the modification of data by unauthorized individuals or systems. It specifies that changes need to be done only by authorized entities and through authorized mechanisms. It guarantees protection against unauthorized modifications of data (Keung, 2013).

Integrity is compromised when a person or system can make changes to data without authorization. For example, when: an employee, accidentally or with malicious intent, deletes important data files; an employee can modify the database without authorization; the content

of a message is changed after the sender sends it and before it reaches the intended receiver. The modification causes the loss of message integrity. The violation of integrity is not necessarily the result of a malicious act.

2.5.1.3 Availability

The principle of availability ensures that data is fully available at the time it is needed by authorized entities. It is compromised if authorized entities are unable to have access to the information when it is needed (Keung, 2013).

The principle of availability states that the information created and stored is useless if it cannot be accessed when it is needed. For example, there is a violation of the availability principle when bank customers cannot access their accounts for transactions. The interruption puts the availability of resources in danger.

Figure 37 illustrates the balance between the basic principles of security mentioned above.

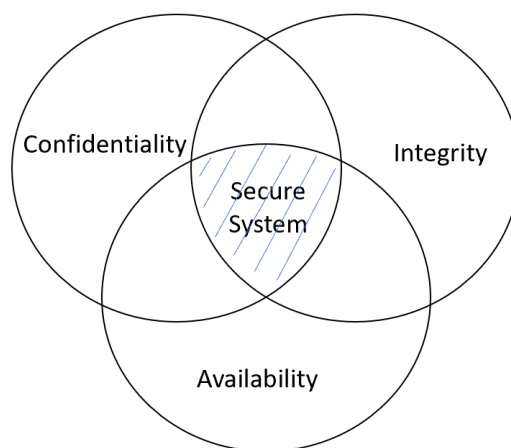


Figure 37 - Balance between the basic principles of security

The right balance between the three principles of security is needed to build a secure system. If the principles are not balanced, then a small gap is created for attackers to nullify the other security principles (Keung, 2013). Having a highly confidential system but low integrity, or vice versa, implies that the system is not secure. On the other hand, a system can protect confidentiality and integrity but if the information is not available when needed, the other two principles are useless (Merkow & Breithaupt, 2014).

In addition to the basic principles of security presented above, there are some other common-sense security principles employed to achieve confidentiality, integrity and availability of systems. Among them are the principles of Defence in Depth, Minimum Privilege, and the Minimization of Attack Surface.

The principle of Defence in Depth defends layered security. It states that several defence mechanisms must be layered to protect valuable data and information. Thus, if a mechanism fails, further steps are taken immediately to prevent an attack. The intentional redundancies caused by the multi-layered approach increases the overall security of the system. Defence in

depth also seeks to compensate the weaknesses of one layer of security with the strengths of two or more layers (Coole et al., 2012; Lee, 2008). The Defence in Depth was a military strategy, which was based on the assumption that an attack would lose momentum over a period of time, and this would allow an appropriate response from the parties being attacked (Coole et al., 2012).

The principle of Minimum Privilege states that a subject should receive only the privileges necessary to accomplish the task. The minimum necessary privilege should be granted to a subject and should be of the shortest duration necessary. Users are limited to access only the files and resources needed to perform their work. It also restricts the access for systems, processes and devices to only the necessary permissions to perform authorised activities. In summary, it is the concept and practice of restricting the access rights of users, accounts and information technology processes only to those resources that are absolutely necessary to perform legitimate routine activities for the necessary time (Gegick & Barnum, 2013; Schneider, 2003).

The principle of Minimization of Attack Surface defends that a secure system has the minimum number of entry points necessary for its function, and a system with more entry points tends to be less secure than one with less. It also states that unused functionalities should be deactivated. Every system has functionalities that serve as entry points for security breaches, and for every new functionality added, the attack surface area is expanded, making the system more vulnerable to attacks. One way to minimise the attack surface area is to identify existing vulnerabilities and implement techniques to minimise the risk of being exploited. One of the possible solutions to minimize the attack surface of a system involves the deactivation of components that are not used. An example would be to close unnecessary services and ports to limit the possibilities of remote interactions with the system (Atighetchi et al., 2015; Majeed & Quadri, 2016; Steinegger et al., 2014).

The principles presented above, were very relevant in the identification of the risk that each vulnerability presents to the identified ICSs.

2.6 Related Work

In this work, we intend to identify ICSs in Portugal connected to the Internet and evaluate their associated risk level in terms of security. Below, we present some existing public studies related to the identification of ICSs and the analysis of their security, according to different parameters.

In (Samtani et al., 2018), the authors assessed the vulnerabilities of SCADA devices found in the Shodan database, by using text mining, data mining, and vulnerability assessment tools. They used Nessus as the framework component to assess the vulnerabilities and the algorithm Random Forest achieved the highest result at 99.4 percentage - identifying 587,158 out of the 627 million devices as SCADA devices. Nessus found that, out of the 587,158 devices identified, 7.77 % (2,942) presented critical issues, 11.21 % (4,241) had high vulnerabilities, 62.55 % (23,673) had medium vulnerabilities, and 18.47 % (6,989) had low vulnerabilities. The study concluded that Shodan was able to found more than 500,000 devices from the most

common SCADA vendors and that many of these devices presented critical vulnerabilities that ranged from default credential issues to outdated software.

In (Vavra & Hromada, 2016), the authors also evaluate the cybersecurity of SCADA systems based on data from Shodan and ICS-CERT databases. They found that 974 SCADA devices can be accessed on the public Internet, almost 94% of these devices use the DNP3 communication protocol and almost 50% of all devices were affected by the ICSA-16-026-02 vulnerability. They also found 188 SCADA systems operating with Windows XP, considered as an untrustworthy operating system. The most affected country was the USA with 50% (491) of all affected devices.

In (Ceron et al., 2019), the authors also used Shodan to discover vulnerabilities of ICSs in the Netherlands. They found about 989 devices using Shodan and Tridium was the manufacturer of 557 of them. They were not able to identify the organisations to which the ICSs belongs, once the organisations hide behind their Internet Service Providers (ISPs).

By using the Shodan and Censys database, in (Andreeva et al., 2016), the authors researched the availability of ICSs over the Internet. The work developed showed that most of the ICSs components were in the United States and that in Europe, most of them were located in Germany. Like the previous studies, Tridium was the manufacturer of most of the systems found.

In (Wojciech, 2019), the author developed a project named *λamerka*, which works as an interface of Shodan for ICSs and IoT (Internet of Things) devices for users with Shodan and Binary Edge Enterprise Accounts or users that have Shodan/Binary Edge credits. The project claimed to allow users to search for ICSs and IoT devices, and together with Binary Edge (Binaryedge, 2017), it shows potential attack vectors that the device may be exposed to. By July 23, 2020, the project was discontinued and its repository on GitHub was archived, but it seems that the project has been resumed by September 2020.

The presented studies have influenced the direction of this work, serving as a guide that helped in the decisions to achieve the objectives of this work. To the best of our knowledge, there is no public work that identifies the Industrial Control System exposed on the Internet in Portugal and evaluates their level of risk.

Conclusions

This chapter presented the entire study of the literature performed to support the work. It addressed the topics related to ICSs, Port Scan, Data Warehouse, the key principles of security of information, and other basic concepts and terminologies used throughout this work. Finally, it presented the most relevant public studies related to the identification of the ICSs and the analysis of their vulnerabilities.

Within the section of ICS, it introduced the concept of ICSs, including an overview of the most common types, the main components of an ICS environment, and the standard protocols used in these systems. Within this section, we discovered that ICSs use their own protocols,

considered industry standard, to communicate. We considered this, one of the main findings of this section as it will allow us to distinguish ICSs from other systems, based on the used protocols.

In the section of Port Scan, it presented the concept of port scan, the main tools used to perform this technique, and some of the existing projects that perform port scan techniques on systems connected to the Internet and store the information obtained. Within this section, we find out that there is an ethical discussion regarding the execution of port scan techniques against systems. Thus, one of the main findings of this section was the identification of projects/search engines that can be used as alternatives to obtain information about systems connected to the Internet, avoiding performing port scan techniques again on a system whose information has already been stored.

Since more than one source of data, with different structures, has been identified to be used in this work, one of the decisions was to build a data warehouse to gather and organise the data, allowing easier analysis of it. Based on this, the section of Data Warehouse presented the concept of the data warehouse, including the ETL process and an overview and analysis of the most used and well-known open-source tools used in the context of data warehousing. This section has enabled us to identify the best data warehouse tools to be used in this work.

One of the objectives of this work is the identification of the risk level of the systems found. Thus, this chapter presented a section with the principles of information security. The presented principles were very relevant in the identification of the risk that each vulnerability presents to the identified ICSs.

This chapter also presented a brief explanation of some of the main concepts and terminologies related to the identification and analysis of systems connected to the Internet. This will allow the reader to better understand this work.

Finally, it presented some existing and most relevant public studies related to the identification of the ICSs and the analysis of their vulnerabilities. This section was one of the most relevant of this chapter as it had strongly influenced the direction of this work, serving as a guide that helped in the decisions to achieve the objectives of this work.

3 PROPOSED METHODOLOGY

One of the core objectives of this work is the development of a methodology to identify ICSs exposed on the Internet in Portugal and their associated risk level. The proposed methodology is composed of 6 phases. During those phases, we identified sources containing data about ICSs connected to the public Internet in Portugal, and the characteristics that may represent vulnerabilities to them. Next, we identified the level of risk that each of the possible vulnerabilities identified presents to these systems and based on that, we proposed a formula to calculate their overall level of risk. Also, within the methodology, a case study was performed using the data on the ICSs found. The data were analysed using the proposal of calculation of the level of risk made.

This chapter presents the proposed methodology. It describes the phases of the methodology and the main decisions made on each one. Figure 38 presents the overall structure of the proposed methodology.

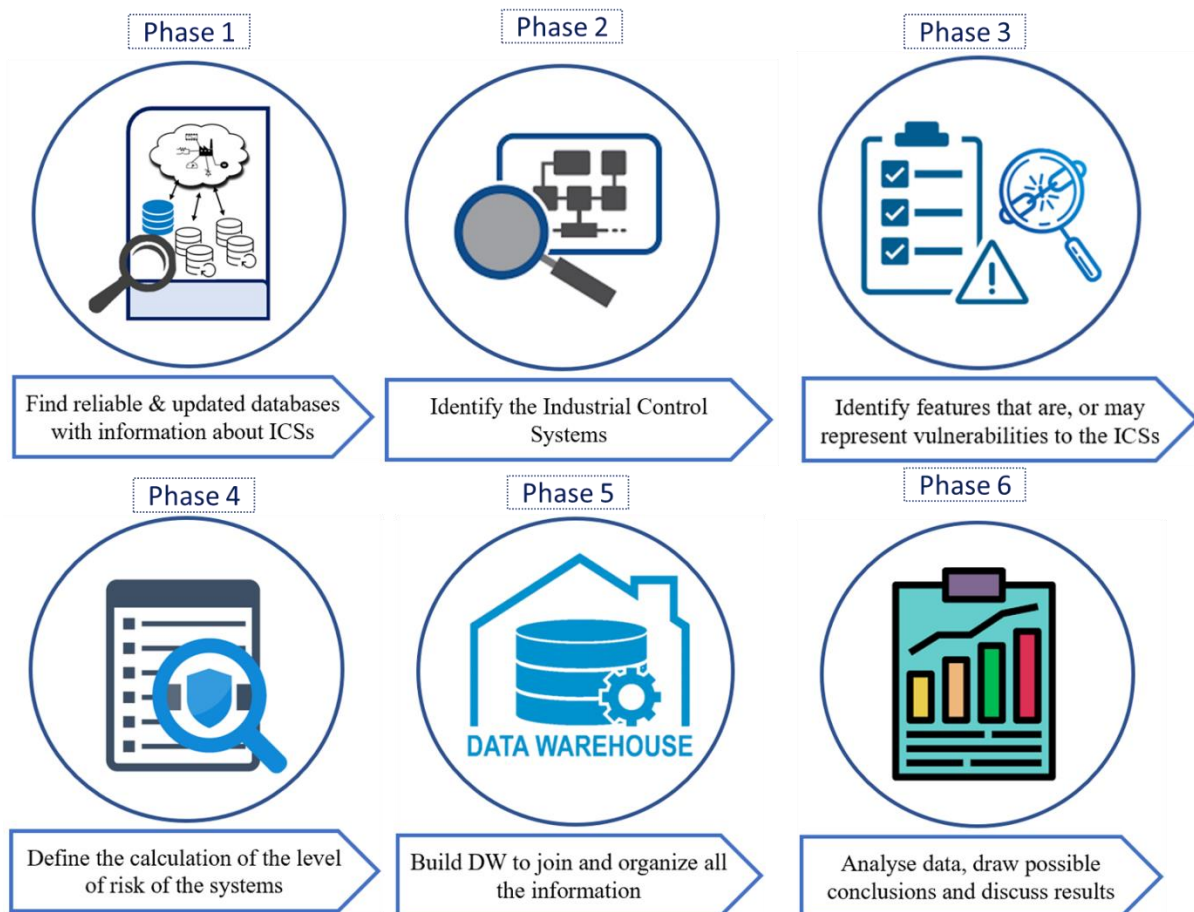


Figure 38 - Overall structure of the proposed methodology

3.1 Phases of the methodology

In order to achieve the objectives of this work, 6 phases were defined for the proposed methodology, namely:

- Phase 1: Find reliable and updated databases with information about ICSs;
- Phase 2: Identify the Industrial Control Systems;
- Phase 3: Identify the features that are, or may represent vulnerabilities to the ICSs;
- Phase 4: Define the calculation of the level of risk of the systems;
- Phase 5: Build a data warehouse to join and organise all the data;
- Phase 6: Analyse the data, draw possible conclusions, and discuss the results.

The definition of each phase of the methodology was based on the objectives of this study and the available data. Some of the phases of the methodology were also inspired by the methodology proposed in (Ceron et al., 2019), with adaptations to the available data and the current context. Each phase of the proposed methodology is explained and discussed in the following sections.

3.1.1 Phase 1: Find reliable and updated databases with information about ICSs

In this phase, it is necessary to perform a literature study to find reliable projects that perform port scanning techniques and maintain up-to-date databases with information about ICSs connected to the Internet in Portugal.

The decision to use data that already exists on databases was conditioned by the need to avoid causing potential problems that, as explained before, may occur when it is performed port scanning techniques against vulnerable systems. We also avoid generating unnecessary network traffic against systems whose information has already been stored. Besides, performing port scanning techniques implies an ethical discussion, since, at the same time that it is used by cybersecurity professionals to identify vulnerabilities, it is also used by malicious users to obtain sensitive information that may allow them to improperly access and interact with systems (Ceron et al., 2019; Partridge & Allman, 2016).

To the best of our knowledge, there is no open project that provides information about systems connected to the Internet via IP address version 6 (IPv6), and manual scanning of the IPv6 address space is not feasible, considering the impact of the volume of requests generated, as it has almost 2^{128} (340 282 366 920 938 463 463 374 607 431 768 211 456) valid addresses (Ceron et al., 2019). Thus, within this methodology, we do not consider systems connected to the Internet via IPv6.

Through the literature study, we observed that the most well-known and used port scan projects are Shodan and Censys (Ceron et al., 2019). These projects perform port scan techniques to a comprehensive set of systems (using IP version 4 addresses - IPv4) and store the information about them. Therefore, we studied these projects (section 2.3.2) and were able to verify that they provide data about ICSs including the information about their geolocation, in which we verified many systems located in Portugal, which may allow us to achieve the objective of the

work. Besides, we were granted access to their Academic account to access the data for academic purposes.

Based on the reasons presented aforementioned, we decided to use these two projects as sources of data for this work. From these two sources, we gathered information about the IP of the systems, the ISP that provides internet services and the AS that regulates the routing policies of them, the organisation to which the systems belong, the city where they are located, the active ports and protocols running on them, the information about the algorithm of encryption available for encrypting their communication, and the CVEs identified on the systems. The data gathered were in CSV format separated by a comma delimiter.

3.1.2 Phase 2: Identify the Industrial Control Systems

In this phase, it is necessary to study the data from the chosen sources and identify what will allow the distinction between the ICSs and the other systems.

Through the literature study about ICSs, we found that due to several problems caused by the use of proprietary protocols by the ICSs manufacturers, they started to use industry-accepted standards protocols on these systems. These protocols are known as ICS standard protocols.

By analysing the data, we noticed that the protocol is one of the information available on the gathered data. Thus, we used this information to distinguish ICSs from the other systems. It is important to note that a system with a port running an ICS standard protocol may be running, at the same time in different ports, other ICS non-standard protocols. For example, it is possible to have an ICS communicating with another ICS through port 502 (ICS standard protocol - Modbus) and at the same time receiving an HTTPS connection through port 443 (ICS non-standard protocol).

3.1.3 Phase 3: Identify the features that are or may represent vulnerabilities to the ICSs

Based on the data gathered from the sources, in this phase, it is necessary to identify what features are or may contain information that may threaten or compromise the proper functioning of the systems.

From the chosen sources, Shodan and Censys, we had information about the IP of the systems, the ISP that provides internet services and the AS that regulates the routing policies of them, the organisation to which the systems belong, the city where they are located, the active ports and protocols running on them, the information about the algorithm of encryption available for encrypting their communication, and the CVEs identified on the systems.

Based on these information, the features that we considered that are or may represent vulnerabilities to the systems are:

- The exposed ports of the systems;
- The cipher suite used to encrypt and/or decrypt the communication;
- The SSL/TLS version available to be used on the encryption;

- The CVEs identified in the systems.

Below we present the reasons for choosing these features.

We considered the exposed ports of the system, as one of the things that may represent vulnerability for the system since there are many ports where it has been proven that they are no longer secure and their use is no longer recommended (Beaver, 2020; Muncaster, 2019; Pascucci, 2014). These ports, in general, are ports running unnecessary services (default services not disabled or old services no longer used), using insecure protocols (protocols without encryption or with encryption considered unsecured), or are misconfigured ports. A system running these ports becomes automatically vulnerable, and even if protected by a Firewall, a compromised system on the same network could easily access and attempt brute force attacks against any insecure ports that it finds. On the other hand, even secure ports are susceptible to be vulnerable when providing information about the system. Malicious and experienced users can use the provided information to attack the system.

On these bases, in the case study, the fact that we have information about the ports used by ICSs through the Internet, taking into account that these systems are always potential targets to cyber-attacks, this information represents a threat to the system. The information about these ports should not be exposed to everyone on the public Internet, and thus, in our opinion, the exposed ports is one of the features that may represent vulnerabilities for the systems.

We considered the cipher suite and the version of the SSL/TLS used on the system to encrypt and/or decrypt the communication as one of the features that may represent vulnerabilities to the system because there are many cipher suites and version of SSL/TLS that are no longer considered secure (Barracuda Networks, 2020; Beyondsecurity, 2020; CipherSuite, 2019; Hebert & GlobalSign, 2020; Nohe, 2019; Packetlabs, 2020). Old or outdated cipher suites are often vulnerable to attacks, and many common SSL/TLS misconfigurations are caused by choosing the insecure cipher suites (Muscat, 2019). The use of cipher suites or versions of SSL/TLS that are no longer considered secure cause a false sense of security. This false sense of security also happens when it is used a secure cipher suite in conjunction with insecure versions of SSL/TLS or vice versa.

As we can see, only by using a secure cipher suite or a secure version of SSL/TLS is not enough. Both the cipher suite and the SSL/TLS version must be secure to ensure the security of system. On these bases, we considered the cipher suite and the version of SSL/TLS as features that have information that may represent vulnerabilities for the systems.

The feature CVE contains information about known vulnerabilities already identified in the system. Based on that, we also considered the feature CVEs as one of the features that contain information that represents vulnerabilities for the systems.

3.1.4 Phase 4: Define the calculation of the level of risk of the systems

In this phase, it is necessary to define the level of risk that the content of each feature, selected in the previous phase, presents for the system. It is also necessary to define how it will be calculated the total level of risk of the system.

Before start, let us define the meaning of “risk” and “level of risk” in this work. Generally, in popular language, risk is the possibility of something bad happening. In (Gantz & Philpott, 2013), the author defines risk in the context of information systems security as “the impact to an organisation and its stakeholders that could occur due to the threats and vulnerabilities associated with the operation and use of information systems and the environments in which those systems operate”.

Based on the definition presented above, in this work, “risk” is the combination of the possibility of exploiting vulnerabilities identified in a system and the potential for damage that this exploitation has on it. The “level of risk” is a quantitative value that reflects the risk. It is a quantitative value that represents the combination between the possibility of exploiting vulnerabilities identified in a system and the potential for damage that this exploitation has on it.

To calculate the level of risk of the systems, we made a proposal of calculation that translates the possibility of exploiting the vulnerabilities identified on the systems and its potential for damage into risk scores. The risk score will then make it possible to identify the level of risk that each vulnerability presents for the system, and the calculation of the overall level of risk of the system. Below we present the proposal of calculation of the level of risk.

3.1.4.1 Proposal for calculation of the level of risk of a system

The proposal of calculation of the level of risk of a system is based on Security Risk Assessment proposals such as (Carnegie Mellon University, 2015; Freund, 2015; Isaca journal archives, 2014; M.Talabis & Martin, 2013; Sultan et al., 2008; White, 2014). In these proposals, the authors choose a set of threats and vulnerabilities that may be associated to the systems and assign a number, based on the characteristics of vulnerabilities, that classifies them in terms of risk. Finally, the risk of the systems is the combination between the probability of exploitation of the vulnerabilities found and the impact that the exploitation would bring to the organisation to which the system belongs. To calculate this combination, each one uses a specific formula, based on the purpose of the risk calculation and the type of the systems that they were classifying. However, regardless of scale, they all have in common the fact that the systems with the most severe vulnerabilities (whose exploitation causes serious damage) or the systems that have vulnerabilities more likely to be exploited have a higher level of risk, compared to those with less severe vulnerabilities or with vulnerabilities less likely to be successfully exploited.

For our proposal, we follow the same line of reasoning, using more precisely the proposed risk calculation presented in (M.Talabis & Martin, 2013). The choice of this proposal in particular,

was based on the fact that it considers the criticality of systems in the calculation of risk. And given that we are classifying the risk of ICSs that are considered critical systems, this proposal is the most appropriate one. In the book, the authors describe the methodology used to develop an Information Security Risk Assessment Toolkit, which they claim that “adopts the best parts of some established frameworks and teaches you how to use the information that is available (or not) to pull together an IT Security Risk Assessment that will allow you to identify High Risk areas”. The authors defend that:

- The risk of a system is directly proportional to the impact of its exploitation (e.g., the exploitation of a system that is business-critical presents a high risk to the organisation);
- The risk of the system is directly proportional to the likelihood of the exploitation of its vulnerabilities (e.g., the risk of a system is higher when it presents vulnerabilities that are more likely to be exploited, for example, known vulnerabilities, etc.);

As both impact and likelihood are directly proportional to the risk, the authors used the product of the two as the basis to calculate the risk score, as presented in Equation 1. Therefore, the risk score is the product of the Likelihood and the Impact of exploiting a vulnerability on a given system.

$$Risk\ Score = Impact \times Likelihood \quad (1)$$

Equation 1 - Formula to calculate the Risk score

The Impact is an estimate of the damage that exploiting a vulnerability could cause. For example, an exploitation of a severe vulnerability on an ICS could have a catastrophic impact. The Likelihood estimates the probability of a weakness in a system being successfully exploited. The Likelihood will try to answer the question: What is the probability of an attacker to gain improper access to a system by exploiting a vulnerability?

There are many different approaches to determining the Impact and the Likelihood score and there is no single correct method for determining it. In the study, the impact score is obtained by assigning scores for the potential impact that the exploitation of vulnerabilities can cause to the system. The higher the Impact score, the more severe is the consequences of exploiting the vulnerability on the system. The Likelihood score is obtained by assigning scores for the probability of exploiting a vulnerability on a system. As the Impact score, the higher the Likelihood score is, the higher is the probability of a vulnerability to be successfully exploited.

Both Impact and Likelihood scores are categorized ranging from 1 to 5, meaning Very Low, Low, Medium, High, and Very High respectively. The range from 1 to 5 was adopted based on ISO 27005, however, it is an increasing index that could assume any other values, as long as the vulnerabilities whose exploitation is more severe present higher Impact score than the less severe vulnerabilities, and the vulnerabilities more likely to be successfully exploited present higher Likelihood score than the vulnerabilities less likely to be successfully exploited. For the Impact score, the assignment has the following meanings:

- 1 - Very Low Impact - characteristics of the system that do not affect, any of the key security principles (the CIA Triad - Confidentiality, Integrity and Availability);

- 2 - Low Impact - characteristics of the system that do not affect so far, any of the key security principles, but it is no longer recommended for systems to have such characteristics;
- 3 - Medium Impact - characteristics of the system that, when exploited, affects at least one of the key security principles;
- 4 - High Impact - characteristics of the system that, when exploited, may affect at least two of the key security principles;
- 5 - Very High Impact - characteristics of the system that, when exploited, may affect all the key security principles.

For the Likelihood score, the assignment has the following meanings:

- 1 - Very Low Likelihood - characteristics of the system that are considered secure and have a minimum probability of being successfully exploited;
- 2 - Low Likelihood - characteristics of the system that are considered secure, but are no longer recommended for use. They may have some chance of being successfully exploited;
- 3 - Medium Likelihood - characteristics of the system that are no longer considered secure, and have a high chance of being successfully exploited by more advanced hackers;
- 4 - High Likelihood - characteristics of the system that are considered insecure and have a high probability of being successfully exploited, however, there are other characteristics that are even more likely to be successfully exploited.
- 5 - Very High Likelihood - characteristics of the system that are considered highly insecure and have a high probability of being successfully exploited. Systems with these characteristics are susceptible to being attacked at any time.

The risk score is then calculated based on this characterization and the formula presented in Equation 1. To classify the risk score into levels of risk, the authors divided the risk score into three ranges: High Risk, containing the risk scores within the black area, Medium Risk, containing the risk scores within the grey area, and Low Risk, containing the risk scores within the white area. Figure 39 presents the matrix classification based on Impact \times Likelihood proposed by (M.Talabis & Martin, 2013) .

		Likelihood				
		1	2	3	4	5
Impact	1	1	2	3	4	5
	2	2	4	6	8	10
	3	3	6	9	12	15
	4	4	8	12	16	20
	5	5	10	15	20	25

Area	Risk Classification
Black	High Risk
Grey	Medium Risk
White	High Risk

Figure 39 - Risk matrix classification based on Impact \times Likelihood (M.Talabis & Martin, 2013)

As presented in Figure 39 the risk is classified in 3 levels: High Risk, Medium Risk, and Low Risk, based on the risk score calculated by multiplying the Likelihood and the Impact of exploiting a vulnerability of the system:

- The Low Risk includes the risk score ranging from 1 to 4. They are largely acceptable, however, they should be reviewed and monitored periodically.
- The Medium Risk includes the risk score ranging from 4 to 12. They should only be tolerated for the short-term and as soon as possible measures should be taken to mitigate them.
- The High Risk includes the risk score ranging from 15 to 25. They should be mitigated as soon as possible with the maximum urgency. In some cases, it is recommended to deactivate the use of the systems until these risks are mitigated.

Finally, the overall risk score of a system is given by the aggregate risk scores (sum of all risk scores) of all the threat and vulnerability identified on the system. This way, the systems with more vulnerabilities or with vulnerabilities with highest risk scores present highest overall risk score associated, deserving more attention on further analysis. In theory, the higher the aggregate risk score is, the higher is the risk of the system, however, this does not always mean that the system has High Risk vulnerabilities. This is because the systems with several Low Risk vulnerabilities associated when aggregated can present a higher total risk than systems with few High Risk vulnerabilities. These systems also deserve attention because, although they have Low Risk vulnerabilities associated, they have many vulnerabilities, thus increasing the attack surface and the chances of being exploited.

In (M.Talabis & Martin, 2013), to help the better visualization of the overall risk score resulted from the aggregation of risk scores, the authors present the following example of the analysis of the risk of applications used in a health system. In the example, present in Figure 40, we can see the final risk score of each application resulted from the aggregation of the risk score of the identified vulnerabilities (sum of all risk scores of their identified vulnerabilities).

Risk Rank	Application	Aggregate Risk Score
1	HIS	60
2	HR Payroll	50
3	Cardio Research DB	47
4	Email	46
5	Imaging	45

Figure 40 - Aggregate risk scores of applications used in a health system (M.Talabis & Martin, 2013)

In order to use the risk calculation proposal explained above in the case study, some decisions have been made. In the proposal, the author suggests as the first step the identification of the information (available or not) that allows the identification of risk in a system. In the previous phase of the methodology, we identified, based on the information available, the features that could threaten or compromise the proper functioning of the systems. Thus, the information on the selected features will be used to calculate the risk of the systems.

As a means of simplification, instead of working with each potential vulnerabilities of the features separately, we divided the content of the features, selected in the previous phase, into

4 classes: Secure, Partially Secure, Insecure, and Totally Insecure, based on the characteristics of the potential vulnerabilities. Each class represents a set of vulnerabilities with the same characteristics. In order to know the vulnerabilities that the classes will include, we defined the meaning of the class for each feature.

➤ **Definition of the classes**

The definition of the classes resulted from the combination between the basic security principles presented in the [State of the Art](#) chapter: Confidentiality, Integrity, and Availability, and the following ones: Defence in-depth, Least privilege and Minimization of Attack Surface. Thus, based on the presented principles of security and the literature studies about the security of systems connected to the Internet (Beyondsecurity, 2020; CipherSuite, 2019; Comodo Group, 2020; GlobalSign, 2020; Mathew et al., 2014; McKay & Cooper, 2019; MITRE Corporation, 2020c; Ocunerix, 2020; SANS Institute, 2015; Villanueva, 2016), below we present the definition of each class for each of the features.

For the feature Port, we defined the following definition:

- **Secure** - a port is considered Secure if, so far, the protocol most commonly associated with it does not present any vulnerability and its use is highly recommended.
- **Partially Secure** - it is considered Partially Secure the set of secure ports running services that normally should not be running on ICSs, using here the principle of Minimization of Attack Surface. For example, e-mail ports should not be running on ICSs, since the latter should not be used as mail servers. It is also considered Partially Secure the set of secure ports running non-standard service, which is eye-catching for attackers.
- **Insecure** - a port is considered Insecure if the protocol most commonly associated with it had known flaws and security problems over the years, yet, with the information available, it is not possible to know if these flaws represent vulnerabilities for the specific ICSs. It is also considered Insecure every open database port since these ports are extremely vulnerable to attacks and must have associated security mechanisms (e.g. Firewalls) so that they are not directly visible on the public Internet. In this same logic, all ports running ICSs standard protocols are considered Insecure, as these ports should not be accessible to anyone on the Internet.
- **Totally Insecure** - a port is considered Totally Insecure if the protocol most commonly associated with it is known to be Insecure and its use is not recommended, much less in an ICS environment.

[Appendix B](#) presents the assignment of the classes to the ports of our dataset.

For the feature SSL version, we associate the following definition for the classes:

- **Secure** - a version of SSL/TLS is considered Secure, if so far, it does not present any vulnerability and its use is highly recommended (e.g., TLSv1.2);

- **Partially Secure** - a version of SSL/TLS is Partially Secure if it did not present so far, any vulnerability, however, there are better alternatives. Within this class is TLSv1.1. There is no security issue in TLSv1.1 that TLSv1.2 fixes, however, the latter have changes and improvements, namely more flexibility, clean-ups, etc. as explained in section 1.2 of [RFC 5246](#);
- **Insecure** - a version of SSL/TLS is Insecure if it has already presented some flaws, but it is not proven whether it represents a vulnerability in the specific ICS;
- **Totally Insecure** - a version of SSL/TLS is Totally Insecure if it has been proven to be Insecure and its use is no longer recommended, much less in an ICS environment.

Table 6 presents the assignment of the classes to the SSL/TLS version.

Table 6 - Classes of the SSL/TLS versions

SSL/TLS Version	Class
SSLv2	Totally Insecure
SSLv3	Totally Insecure
TLSv1	Totally Insecure
TLSv1.1	Partially Secure
TLSv1.2	Secure

For the feature Cipher, we associate the following definition for the classes:

- **Secure** - a cipher suite is Secure if so far, none of its components has presented any vulnerability and its use is highly recommended. They support Perfect Forward Secrecy (PFS) - property of the key-agreement that gives assurances that session keys used to encrypt the data will not be compromised even if a long-term private key is compromised in the future (Patel, 2013).
- **Partially Secure** - a cipher suite is Partially Secure if none of its components present, so far, any vulnerability, however, there are better alternatives. In (CipherSuite, 2019), the authors present a list of Recommended, Secure, Weak and Insecure cipher suites. The Secure cipher suites presented in (CipherSuite, 2019) fits in this class because they do not present vulnerabilities, however, there are other alternatives recommended.
- **Insecure** - it is considered Insecure, all the old and outdated cipher suites, even if no specific vulnerabilities are associated with them. These ciphers should only be used in special cases (e.g., where support for older operating systems, browsers or applications is required). Otherwise, they should be disabled and only be used in special use cases, for example.
- **Totally Insecure** - a cipher suite is Totally Insecure if it has broken protection and its use is no longer recommended in any circumstances. These ciphers are insecure and their protection can be broken, at any time, with minimal effort. The system is totally vulnerable using these cipher suites.

[Appendix C](#) presents the assignment of the classes to the cipher suites of our data set.

Unlike the other features, the feature CVE contains vulnerabilities already identified on the ICSs, which means that it already presents insecurities for the specific system. As seen in the literature study, each CVE has an associated qualitative severity ranking of "Low", "Medium", and "High" that is based on its score which assesses the impact of the CVE on a system. In this way, the CVEs are already divided into classes according to the impact that they have on the system. Thus, instead of defining new classes, we will use the existing classes:

- **CVEs with “Low severity”** - the set of CVEs with a score from 0.0 to 3,9;
- **CVEs with “Medium severity”** - the set of CVEs with a score from 4.0 to 6,9;
- **CVEs with “High severity”** - the set of CVEs with a score from 6,9 to 10.0.

After the definition of the meaning of each of the classes for each of the features, the next step is the assignment of the Impact and the Likelihood score for each of the classes of each feature. This assignment will allow the calculation of the risk score given by the formula presented in Equation 1, which will automatically give us the risk level classification based on what is presented in Figure 39.

➤ **Assignment of the Impact and Likelihood scores to each class of each feature**

Based on what we explained previously, both of the Impact and the Likelihood score was assigned from 1 to 5, taking into account the severity and the probability of exploitation of the vulnerabilities. According to the Impact and the Likelihood score attributed, for each of the features, we present below where its classes fit into the risk classification matrix. In the risk classification matrix of each feature, as presented in Figure 39, the black cells represent vulnerabilities with High Risk classification, the grey cells represent vulnerabilities with Medium Risk classification and the white cells represent vulnerabilities with Low Risk classification.

○ **Feature Port**

For the feature port, according to each of the classes we assigned the following Impact and the Likelihood scores:

- For ports considered Secure we attributed 1 for the Impact score and also 1 for the Likelihood score. This attribution was based on the fact that these ports are considered secure and are highly recommended, thus they fit into the Low Risk classification quadrant. However, they should be reviewed and monitored periodically.
- For ports considered Partially Secure, we attributed 2 for the Impact score and also 2 for the Likelihood score, as they are secure but not recommended to use in ICSs. With this assignment they also fit into the Low Risk classification quadrant, but with the highest score of the quadrant, in which case they should be reviewed as soon as the Medium and the High Risks are mitigated.
- For ports considered Insecure, we attributed 4 for the Impact score and 3 for the Likelihood score. The Impact score of the exploitation of these ports is high because these ports include ports running ICSs protocols, but at the same time this impact is lower than the Totally Insecure ports, and the Likelihood is medium because no study

proves that they represent vulnerabilities to the ICSs. With this assignment, these ports fit into the Medium Risk classification, with the highest associated risk of that quadrant, which means that, measures should be taken to mitigate them as soon as possible.

- For ports considered Totally Insecure, we assigned for both Impact and Likelihood the highest score. The impact of exploit those ports are severe as they can give direct access to the ICSs and the probability of been exploited successfully is high. With this assignment, they fit into the High Risk classification quadrant and they should be mitigated as soon as possible with the utmost urgency. If the use of the port is not crucial to the function of the system, it is recommended to close them.

Table 7 presents the risk classification matrix that reflects the attribution explained above.

Table 7 - Risk matrix classification for the feature Port

Port	Likelihood				
	1	2	3	4	5
Impact	1	Secure			
	2		Partially Secure		
	3				
	4			Insecure	
	5				Totally Insecure

○ **Feature SSL version**

For the feature SSL version, according to each of the classes we assigned the following Impact and the Likelihood scores:

- For the versions of SSL/TLS considered Secure we attributed 1 for the Impact score and also 1 for the Likelihood score, as they are secure and highly recommended, thus they fit into the Low Risk classification quadrant. However, they should be periodically reviewed and monitored, following the update of more secure versions.
- For the SSL/TLS version considered Partially Secure, we assigned 2 for the Impact score and also 2 for the Likelihood score, as they are secure but there is a better recommendation. With this assignment also they fit into the Low Risk quadrant but with the highest score of the quadrant. They should be upgraded, as soon as possible, to more recommended versions.
- For the SSL/TLS versions considered Insecure we attributed 4 for the Impact score and 3 for the Likelihood score. The Impact score is high because these versions have already presented some flaws but at the same time the score is lower than the impact score of Totally Insecure versions. The Likelihood is medium because there is no study that proves that the flaws represent vulnerabilities for the ICSs. With this assignment, these SSLT/TLS versions fit into the Medium Risk classification, which means that they should only be tolerated for the short-term and as soon as possible they should be updated to more secure versions.
- For the SSL/TLS versions considered Totally Insecure, we attributed the higher Impact score (5), because the consequences of exploiting them are severe, as they are versions

proven to be Insecure and are not recommended. We attributed 4 for the Likelihood score, as although the probability of exploiting systems using these versions are high, they are still less than the probability of exploiting systems without any type of encryption. In this sense, for the when the SSL/TLS version is not available (NA), meaning that the system does not have any encryption associated, we attributed the highest Impact score (5), and the highest Likelihood score, based on the assumption that an unencrypted technology always has more disadvantages in terms of security than an encrypted technology. Even the worst encryption with known vulnerabilities makes several attacks more difficult (Hytrust Cloud Under Control, 2015; Lantronix, 2009; K. Matthews, 2019). Thus, both NA and Totally Insecure versions of SSL/TLS fit into the High Risk classification quadrant, and they should be mitigated as soon as possible with the maximum urgency. However, the system without any encrypted algorithm associated should receive more attention.

Table 8 presents the risk classification matrix that reflects the assignments explained above and shows each of the classes fit.

Table 8 - Risk matrix classification for the feature SSL version

SSL version		Likelihood				
		1	2	3	4	5
Impact	1	Secure				
	2		Partially Secure			
	3				Insecure	
	4					
	5				Totally Insecure	NA

○ **Feature Cipher**

For the feature Cipher, according to each of the classes we assigned the following Impact and the Likelihood scores. The attribution of the Impact and Likelihood scores for this feature were similar to the attribution made for the SSL/TLS versions:

- For the cipher suites considered Secure, we attributed 1 for the Impact score and also 1 for the Likelihood score, as they are secure and highly recommended. With this attribution, they fit into the Low Risk classification quadrant with the lowest risk score. However, they should be reviewed and monitored periodically, following the update of more secure cipher suites.
- For the cipher suites considered Partially Secure, we attributed 2 for the Impact score and also 2 for the Likelihood score, as they are secure but there is a better recommendation. With this assignment they also fit into the Low Risk quadrant, with the highest score of this quadrant. They should be upgraded, as soon as possible, to more recommended versions.
- For the cipher suites considered Insecure, we assigned 4 for the Impact score and 3 for the Likelihood score. The Impact score is high because these versions have already

presented some flaws but at the same time lower than the impact score of Totally Insecure cipher suites. The Likelihood is medium because there is no study that proves that the flaws represent vulnerabilities to the ICSs. With this assignment, these SSL/TLS versions fit into the Medium Risk classification, which means that they should only be tolerated for the short-term and as soon as possible they should be updated to more secure cipher suites.

- For the cipher suites considered Totally Insecure, we assigned the higher Impact score (5), because the consequences of exploiting them are severe, as they have been proven to be Insecure and are not recommended. We attributed 4 for the Likelihood score, as although the probability of exploiting systems using these ciphers suites are high, they are still less than the probability of exploiting systems without any type of encryption. In the case of unavailable ciphers (NA), we took into account that ciphers are one of the sets of cryptographic algorithms chosen to implement SSL. Taking this into account, we already penalized empty ciphers in the NA SSL versions and, for this reason, the NA cipher suites are not considered in the calculation of the risk.

Table 9 presents the risk classification matrix that reflects the assignments explained above and shows where each of the classes fit.

Table 9 - Risk matrix classification for the feature Cipher

Cipher		Likelihood				
		1	2	3	4	5
Impact	1	Secure				
	2		Partially Secure			
	3				Insecure	
	4					
	5				Totally Insecure	

○ **Feature CVE**

For the feature CVE, the assignment of the level of risk is different from the other features since the classes are different, as they represent insecurities already identified for the specific system and should be corrected as soon as possible according to their severity. Regardless of the severity of the CVE, there is a huge probability a CVE to be successfully exploited. Thus, we assigned a Likelihood score of 5 for all the classes of the CVEs. For the Impact score, we attributed the scores 3, 4, 5 for CVEs with Low severity, CVEs with Medium severity and CVEs with High severity respectively. With this assignment, all the CVEs fit into the High Risk classification quadrant and they should be corrected, with the maximum urgency according to their severity, however the ones the highest severity has higher scores and they should receive more attention. Table 10 presents the risk classification matrix that reflects the assignments explained above and shows where each of the classes fit.

Table 10 - Risk matrix classification for the feature CVE

CVE		Likelihood				
		1	2	3	4	5
Impact	1					
	2					
	3					Low Severity
	4					Medium Severity
	5					High Severity

After finishing this phase, we would like to highlight some notes. It is important to keep in mind that the level of risk provide clues about the current state of the systems and consequently it helps to filter those that present many vulnerabilities. Based on the calculated level of risk of the systems, cybersecurity professionals should analyse them in more detail, in order to verify whether they really represent a threat to the organisations to which it belong, and/or alert organisations that they should give more attention to these systems. We cannot compare and/or state that one system is actually more Insecure than another just based on its total level of risk. The level of risk should not be automatically interpreted as being synonymous with the insecurity of a system. For instance, we can consider a case where we have two systems: one has several vulnerabilities with low level of risk and the other has few vulnerabilities with high level of risk. These two devices will appear with a high total level of risk, regardless of which is truly more Insecure. This will force both to deserve attention in terms of security since they are attractive targets for cybercriminals. The first one because, although it does not present any vulnerability with high level of risk, it presents many paths of exploitation. The second because, although it does not present many vulnerabilities, those that it presents have high level of risk. On the other hand, we can have two systems that present the same level of risk, and when analysed in more detail, it can be seen that one is much more Insecure and vulnerable than the other. Furthermore, a system can present a very high level of risk and when studied in more detail it can be verified that it is a honeypot - a system with vulnerabilities left purposely, which uses intrusion attempts to obtain information about cyber criminals or to distract them from other targets.

3.1.5 Phase 5: Build a Data Warehouse to join and organise all the information

In order to join the information about the ICSs found and facilitate their analysis, in this phase we propose the developing of a data warehouse to combine and organise the data. The data warehouse will allow the gathering and transformation of the data obtained about the ICSs from the different sources and will facilitate the analysis of it.

To build the data warehouse, it is necessary to establish and align its requirements with the objectives of the study. This phase presents an overview on the main characteristics of the data in order to better understand it, as well as the requirements analysis of the data warehouse, including the questions that must be answered and the data model that translates the requirements. Finally, it presents the main decisions made in the ETL process. In the data

warehouse context, the identified systems may also be addressed as network addresses identified by their IP.

3.1.5.1 Data Understanding

To better understand the data, it is necessary to take into account the following:

- A network address can have, at the same time, many (exposed) ports running.
- Different protocols can run by default on the same port number (e.g., ICCP and Siemens S7 operate by default on port 102) and there are also protocols that run by default on multiple port numbers (e.g., EtherNet/IP runs on port 2222 and port 44818; Niagara Tridium Fox runs on port 1911 and port 4911).
- The ports can be filtered, and if so, some of their information may be hidden.
- Many CVEs can be identified on a port running on a network address at a certain time.
- Some of the ports use encrypted algorithms to secure the connection. For those ports, there is information about the cipher suite and the list of SSL/TLS versions available to use on the connection. The level of risk of the version of the SSL/TLS considered is the level of risk of the worst (less secure) version available.
- The date corresponds to when the network address was port scanned. Thus, it is possible to have different records of the same network address with the same characteristics on different dates, which means that the network address characteristics remained the same through time.

3.1.5.2 Requirement analysis of the DW

Based on the results that we intend to obtain with the analysis of the data warehouse and on the studies and assignments made in the previous phases, the data warehouse should be able to answer the following questions:

- Q1. Number of network addresses running High/Medium/Low Risk ports;
- Q2. Number of network addresses using High/Medium/Low Risk SSL/TLS versions;
- Q3. Number of network addresses using High/Medium/Low Risk cipher suites;
- Q4. Number of network addresses with High/Medium/Low Risk CVEs;
- Q5. Number of exposed ports per network address/Most exposed network addresses;
- Q6. Level of risk per network address/Most vulnerable network addresses;
- Q7. Number of exposed ports running standard and non-standard ICS protocols;
- Q8. Most exposed protocols;
- Q9. Number of CVE per network address;
- Q10. Most common CVEs identified on the systems.

The list of questions presented above will help to better understand the main requirements of the data warehouse, its data model, and the ETL process. The first 6 questions are the most important since they will allow the identification of the systems with the highest risk characteristics associated and the identification of the most vulnerable systems.

3.1.5.3 Data Model

Considering all the characteristics of the data presented above, the fact that a network address can have several ports running at the same time and the ports can have several CVEs can identified, we concluded that there is different granularity between the network addresses, ports and CVES. Taking this into account, the ideal scheme for the data warehouse to deal with the different granularities would be a model with 3 fact tables.

Considering that the total risk level is given by the sum of the risk level of the ports, the risk level of the SSL/TLS version, the risk level of cipher suites and the risk level of CVEs identified on the system, if we decided to build the data warehouse with 3 fact tables, we have to somehow join the information from the various fact tables. To do this, instead of joining fact tables in a single query, it is recommended to use "drill across", which, in a simplified way, is the process of querying each fact table separately and merge the results (Adamson, 2011; Kimball, 2003).

The option of building a data warehouse model with 3 fact tables would make it easier for the end-users to analyse the data if they were using a reporting/analytics tool, but otherwise, it would imply that they had to build queries to search each fact table separately and merge the results (Adamson, 2011). On the other hand, the choice to build a model with 3 fact tables would increase the complexity of the data warehouse in terms of perception, making it more difficult to understand.

Considering the complexity of the cybersecurity area itself and the intricacy of data we are working with, we wanted to conceive a more intuitive and easy to understand data warehouse model.

Based on the considerations discussed above, we decided to build a model with only one fact table. This model makes some searches more complex due to the need to use some grouping functions to answer some questions, but makes the data warehouse very easy to understand and very intuitive. The fact that we have all the questions that we want the DW to answer already defined, made it easier to choose the complexity of the queries over the complexity of the data warehouse model. Furthermore, Dognædis is more comfortable with having some complexity in the queries (use of grouping functions), than with the complexity of understanding the data warehouse. Thus, we designed the following data warehouse structure presented in Figure 41. It is constituted by one fact table: Fact_risk, and 6 dimensions tables: Dim_Port, Dim_NetworkAddress, Dim_Time, Dim_CVE, Dim_ssl_version, and Dim_Cipher.

The fact table (Fact_risk) contains the events related to the vulnerabilities found on the ports and network addresses in a certain time. It also has a metric, level of risk, that is a combination of the level of risk of the vulnerabilities found in the system. The dimension tables contain the information that describes the events recorded on the fact table.

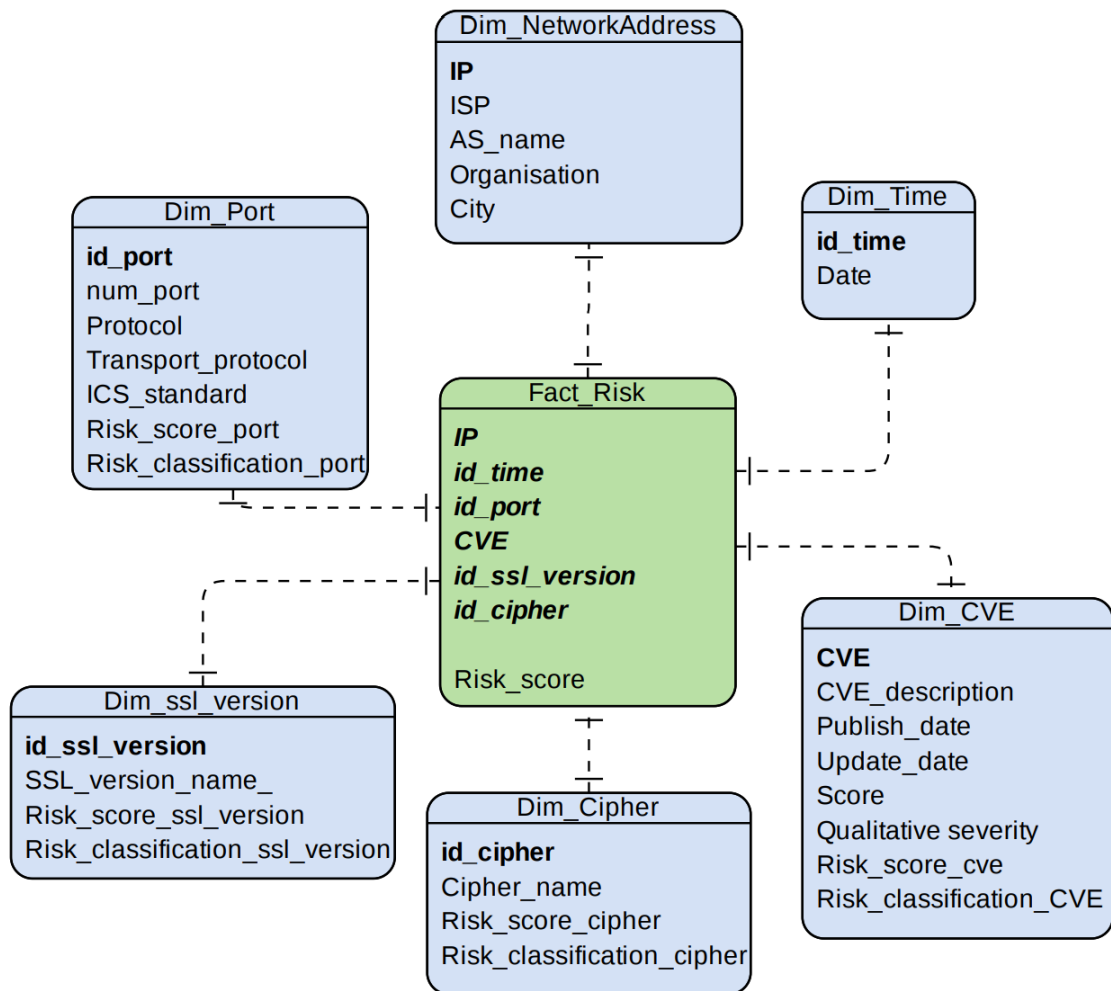


Figure 41 - Diagram of the data warehouse

The network addresses are the exposed ICSs about which we obtained information. The dimension, *Dim_NetworkAddress*, contains information about all network addresses retrieved from Shodan and Censys, between February 2020 and July 2020. The dimension is constituted by:

- **IP** - IPv4 address that uniquely identifies the network address, and distinguishes every record of this dimension;
- **Organisation** - the organisation to which the network address belongs;
- **City** - where the network address is located;
- **AS_name** - the name of Autonomous System that regulates the routing policies of the network address;
- **ISP** - the Internet Service Provider of the network address.

The dimension *Dim_Port*, represents the data about all the scanned ports on the retrieved network address. It is constituted by:

- **id_port** - a number that identifies and distinguishes the records of this dimension;
- **num_port** - the number of the port;
- **Protocol** - the protocol running in the port;

- Transport_protocol - the transport protocol used to communicate;
- ICS_Standard - identifies (yes or no) if the port is running a standard ICSs protocol;
- Risk_score_port - the risk score resulting from the multiplication between the Impact score and the Likelihood score of the port;
- Risk_classification_port - the qualitative risk classification based on the risk score of the port.

The dimension Dim_ssl_version has information about the SSL/TLS version available to use on the encryption of the systems. It is constituted by:

- id_ssl_version - a number that identifies the SSL/TLS version;
- SSL_version_name - the name of the SSL/TLS version;
- Risk_score_ssl_version - the risk score resulting from the multiplication between the Impact score and the Likelihood score of the SSL/TLS version;
- Risk_classification_port - the qualitative risk classification based on the risk score of the SSL/TLS version.

The dimension Dim_Cipher has information about the algorithms used to encrypt the communication of the systems. It is constituted by:

- cipher_id - a number that identifies and distinguishes the records of this dimension;
- Cipher_name - the name of the cipher suite;
- Risk_score_cipher - the risk score resulting from the multiplication between the Impact score and the Likelihood score of the cipher suite;
- Risk_classification_cipher - the qualitative risk classification based on the risk score of the cipher suite.

The dimension Dim_CVE has data about the CVEs retrieved on (MITRE Corporation, 2020b). It is constituted by:

- CVE - the identification number assigned to a security flaw. It identifies and distinguishes the records of this dimension;
- CVE_description - a brief description of the CVE;
- Publish_date - the date when the CVE was published;
- Update_date - the date when the CVE was last updated;
- Score - the score of the CVE;
- Qualitative_severity - the qualitative severity that the CVE presents to the systems;
- Risk_score_CVE - the risk score resulting from the multiplication between the Impact score and the Likelihood score of the CVE;
- Risk_classification_CVE - the qualitative risk classification based on the risk score of the CVE.

The dimension, Dim_Time, contains information about the time. It will allow to filter the information by the date when the network address was port scanned. It is constituted by:

- id_time - a number that identifies every record of this dimension;

- Date - contains all the dates, starting from November 2019. The used database engine allows us to unfold this date in days, weeks, months, quarters, and so on. Therefore, this single field enable us to do all the researches by the required time interval.

The fact table, Fact_Risk, contains the data concerning the results of port scan techniques retrieved from Shodan and Censys. It is constituted by a the primary key which is a combination of the foreign keys of the dimension tables and a metric named Risk_score that is the sum of the risk scores of the dimensions Dim_Port, Dim_ssl_version, Dim_Cipher, and Dim_CVE.

3.1.5.4 ETL Process

Bellow, we present some of the main decisions made to perform the extraction, transformation and loading of data to all the dimensions and fact tables of the data warehouse.

➤ Dimension Dim_NetworkAddress

For the data extraction process to fill in the dimension Dim_NetworkAddress, we get the data from Shodan and Censys and it was filtered to only gather systems located in Portugal. The information that comes from Shodan corresponds to the systems running ICS standard protocols. Since we had a download limit on the Shodan Academic Account, we choose to download only data from systems that are running ICS standard protocols. The information gathered from Censys corresponds to all the systems that the search engine port scanned.

Through Shodan, we obtained data about the IP, the ISP, the ASN, the port scan time, the city where the network address is located, and the organisation to which the system belongs. In the Censys data, we had information about the IP, the ASN, the name of the AS, the port scan time, and the city where the system is located. Finally, we merged the information from the two sources, complementing Shodan data with information from the Censys data. Then, we performed some transformations to uniformize the data.

In the process of transformation, we performed the following transformations:

- We replaced every null and empty field with the string “NA” (Not Available);
- We capitalized all the strings (on the fields: ISP, AS name, Organisation and City) and removed the space characters before and after the data on these fields, in other to uniformize them;
- Some of the network addresses had "Lisbon" in the field City, while others had "Lisboa", we uniformize them, turning "Lisbon" into "Lisboa". Still, on the field City, we also performed the following transformations on some cities that were badly written:
 - ?bidos to Óbidos;
 - ?gueda to Águeda;
 - Vila Nova de Famalic to Vila Nova de Famalicão;
- For the field Organisation, we carried out some transformations to correct and uniformize some designations. Figure 42 presents the transformations made in this filed. In the left side (Source value) we have the misspelt name of the organisations and on the right side (Target value) we have their correct name.

Source value	Target value
DIRECAO GERAL DE ESTATISTICAS DA EDUCACAO E CIENCI	DIRECAO GERAL DE ESTATISTICAS DA EDUCACAO E CIENCIA
FOUNDATION FOR SCIENCE AND TECHNOLOGY, PORTUGAL	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.
MESP, MOTA-ENGIL SERVICOS PARTILHADOS ADMINISTRATI	MESP-MOTA-ENGIL - SERVIÇOS PARTILHADOS ADMINISTRATIVOS E DE GESTÃO, S.A.
MINHOCOM, GESTAO DE INFRAESTUTURAS DE TELECOMUNICA	MINHOCOM, GESTAO DE INFRAESTUTURAS DE TELECOMUNICACAO
SECRETARIA-GERAL MINISTERIO DA ADMINISTRACAO INTER	SECRETARIA-GERAL MINISTERIO DA ADMINISTRACAO INTERNA
NET SOLUTIONS - CONSULTORIA EM TECNOLOGIAS DE INFO	NET SOLUTIONS - CONSULTORIA EM TECNOLOGIAS DE INFORMACAO
VIA NET.WORKS PORTUGAL - TECNOLOGIAS DE INFORMA,CA	VIA NET.WORKS PORTUGAL - TECNOLOGIAS DE INFORMACAO, SA
DOMINIOS, S.A.	DMNS - DOMINIOS, S.A.
ONI	ONITELECOM - INFOCOMUNICACOES, S.A.
ONI TELECOM	ONITELECOM - INFOCOMUNICACOES, S.A.
ONI CLOUD	ONITELECOM - INFOCOMUNICACOES, S.A.
OVH HOSTING	OVH HOSTING LDA
OVH SAS	OVH HOSTING LDA
OVH GMBH	OVH HOSTING LDA
CLARANET LTD	CLARANET PORTUGAL S.A

Figure 42 - Transformation of the field Organisation for the dimension Dim_NetworkAddress

After the transformations mentioned above, we loaded the data to the table Dim_NetworkAddress.

➤ Dimension Dim_Port

The data obtained through Shodan and Censys also allowed to fill in the dimension Dim_Port. In both, we had data about the number and the protocol running of the port, and the transport protocol used. We also were able to identify the ports running ICSs standard protocols, and the information about their level of risk. Then, we merged the information from the two sources, complementing the missing information in one source with information from the other source. As mentioned before, Shodan provides information about the systems running ICSs Standard protocols. Therefore, for every ICS in the Shodan data, we get the information about other ports running on that system from Censys data.

To standardize the data to load in the Dim_Port dimension, we performed the following transformations:

- We capitalized the fields Protocol and Transport_Protocol, in order to uniformize them;
- In the field Protocol, we performed the following transformations:
 - Some of the rows had “s7” and some “Siemens S7”. Both represent the same protocol (Siemens S7). We transform “s7” into “Siemens S7”.
 - The protocol “Modbus” in some of the rows had the name “Modbus/TCP”, we transform the latter into “Modbus”;
- The protocol ISAKMP in some of the rows came as “Fatek FB Series”, we transform the latter into “ISAKMP”;
- We used a script to add the risk level of ports based on their security, according to what was explained in phase 4 of the methodology.

After all the transformations mentioned above, we loaded the data to the table Dim_Port.

➤ **Dimension Dim_CVE**

To fill in the dimension Dim_CVE, we retrieved the information about all the published CVE from (MITRE Corporation, 2020c) and load it into a file. The file had the information about the identification of the CVE, its description, its publish date, the date of its latest update and its score. In the process of transformations, we performed the following tasks:

- We created the field qualitative severity, that were filled according to the Score of the CVE. For CVEs with a score between 0 and 4 the qualitative severity is Low, for CVEs with score between 4 and 7, the qualitative severity is Medium, and for CVEs with a score between 7 and 10, the qualitative severity is High.
- According to the qualitative severity, we used a script to add the risk level of each CVE, based on what was explained in phase 4 of the methodology.

After the transformations mentioned above, we loaded the data to the table Dim_CVE.

➤ **Dimension Dim_Cipher**

We retrieved the information about the cipher suites on (Rudolph & Grundmann, 2019) and loaded it into a file. On the file, we had the information about the name of the cipher suite and the security associated with it.

On the process of transformation, we performed a script to add the level of risk of the cipher suite, based on the risk score associated to it, according to on what was explained on phase 4 of the methodology. After the transformations, we loaded the data to the dimension Dim_Cipher.

➤ **Dimension Dim_ssl_version**

In a file, we had information about all the SSL/TLS versions and the security associated to each one. On the process of transformation, we performed a script to add the level of risk of the SSL/TLS version, based on the security associated to it. After the transformations, we loaded the data to the dimension Dim_ssl_version.

➤ **Dimension Dim_Time**

To fill in the dimension Dim_Time, we generated a timestamp with the format yyyy/MM/dd, starting from November 2019. This decision was based on the fact that although the data used in this study has begun to be downloaded in February 2020, by analysing it we observed that the oldest date when the Shodan/Censys performed the port scan techniques on the systems was December 2019. Thus, we gave an interval of 1 month before this date, to reduce the margin of error.

It was not made any transformation in the generated date. Next, we loaded the data to the dimension Dim_Time.

➤ **Fact table Fact_Risk**

The data obtained from Shodan and Censys through the port scan techniques allowed us to fill in the table Fact_Risk. Then, we merged the two sources, complementing the missing data in one source with data from the other source. On the transformation process, we performed the following actions:

- We capitalized all the strings;
- For each encrypted port, we had the information about the list of SSL/TLS versions available to be used in the connection. The SSL/TLS version considered for the risk calculation of the system is the worst version available since it represents the highest risk for the system. To get this version, we developed a script that runs through all the available SSL/TLS versions and returns the version with the highest risk score;
- The information about the CVEs identified in each system comes in a single field separated by commas (e.g., “CVE-2017-18911, CVE-2018-18911, CVE-2014-18911”). Thus, we performed a parse to separate the CVEs using the delimiter “,”;
- We replaced every null and empty field with the string “NA” (Not Available);
- Some of the ports were filtered, thus it was not possible to have some information about them. For these ports, the fields with missing information were filled with the string “filtered”;
- We performed a lookup on the dimension tables to link the events on the fact table to the information on the dimension tables;
- The metric “level of risk” in the fact tables is the sum of the level of risk of the used port, the level of risk of the used cipher, the level of risk of the worst version of SSL/TLS available for the connection, and the level of risk of the identified CVE.

After the transformations, we loaded the data to the fact table Fact_Risk.

Figure 43 presents an example of one of the process, build in Pentaho Data Integration, to perform the initial ETL of the dimension Dim_NetworkAddress.

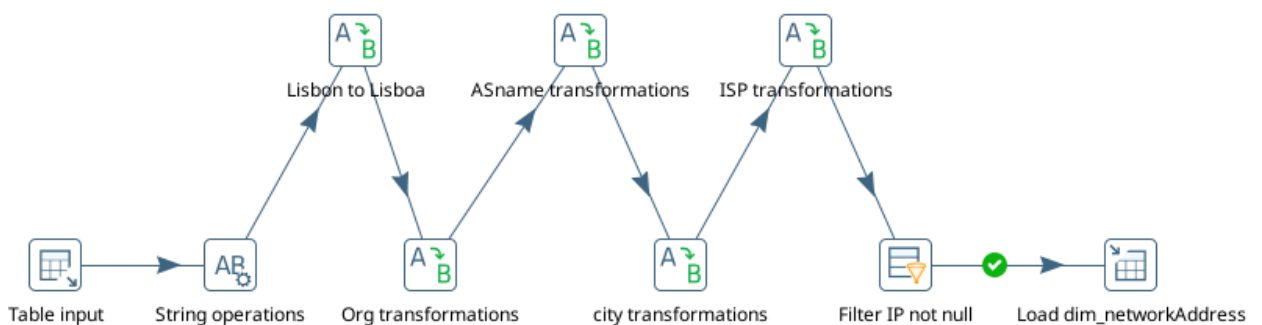


Figure 43 - ETL process of the dimension Dim_NetworkAddress

3.1.5.5 Periodic ETL

The periodic ETL is necessary to keep the data warehouse updated. It is performed after the initial load of all the dimensions and fact tables. In this process, it is decided what to do with

new incoming data, namely: update attributes or inserts new records on the data warehouse. The periodicity of the incremental loads depends on the availability of data on the data sources. It can be on a daily, weekly, monthly, quarterly, yearly basis, and so on.

Considering that the data was downloaded from the sources every month, as our Shodan account renewed the credits to be used for downloading the data monthly, the periodicity of the incremental loads of the dimension and the fact tables on our data warehouse were made on a monthly basis.

For the periodic ETL of the dimension Dim_NetworkAddress, we performed a database lookup for each network address (identified by its IP) to verify whether it was on the data warehouse. If so, it was verified whether the information of the fields ISP, AS_name, Organisation, City, or Scan_date was changed. In case of change, the network address would be updated. If the network address was not on the data warehouse a new record would be created with the new information.

For the periodic ETL of the dimension Dim_Port, we performed a database lookup for each port (identified by the id_Port) to verify if it was on the data warehouse. If so, it was verified whether the information of the fields: ports, protocol, transport protocol, the flag that identifies if the port is running an ICS standard protocol, or the level of risk of the port were changed. In case of change, the port would be updated with the new information. If the port was not on the data warehouse a new record would be created with the new information.

For the periodic ETL of the dimension Dim_ssl_version, we performed a database lookup for each SSL/TLS version (identified by the id_ssl_version) to verify whether it was on the data warehouse. If it already existed, it was verified whether the name or the risk level of the cipher suite were changed, if so, the cipher suite would be updated with the new information. If the SSL/TLS version was not on the data warehouse a new record would be created with the new information.

For the periodic ETL of the dimension Dim_Cipher, we performed a database lookup for each cipher suite (identified by the id_cipher) to verify whether the cipher suite was on the data warehouse. If it already existed, it was verified whether the name or the risk level of the cipher suite were changed, if so, the cipher suite would be updated with the new information. If the cipher suite did not exist on the data warehouse a new record would be created with the new information.

For the Periodic ETL of the dimension Dim_CVE, we performed a database lookup for each CVE (identified by the CVE ID) to verify whether the CVE was on the data warehouse. If so, it was verified whether the publish date, update date, score, qualitative severity, description, or the risk level of the CVE were changed. In case of change, the CVE would be updated with the new information. If the CVE did not exist on the data warehouse a new record would be created with the new information.

For the Periodic ETL of the dimension Dim_Time, we made a database lookup for each date to verify whether the date already existed, if it did not exist, the data warehouse would be updated with the new date.

For the Periodic ETL of the fact table Fact_risk, we performed a database lookup that for each new event verified whether it already existed. If it did not exist, the data warehouse would be updated with the new event.

Figure 44 presents an example of one of the process, build in Pentaho Data Integration, to perform the period ETL of the dimensions, in this case, the period ETL of the dimension Dim_NetworkAddress.

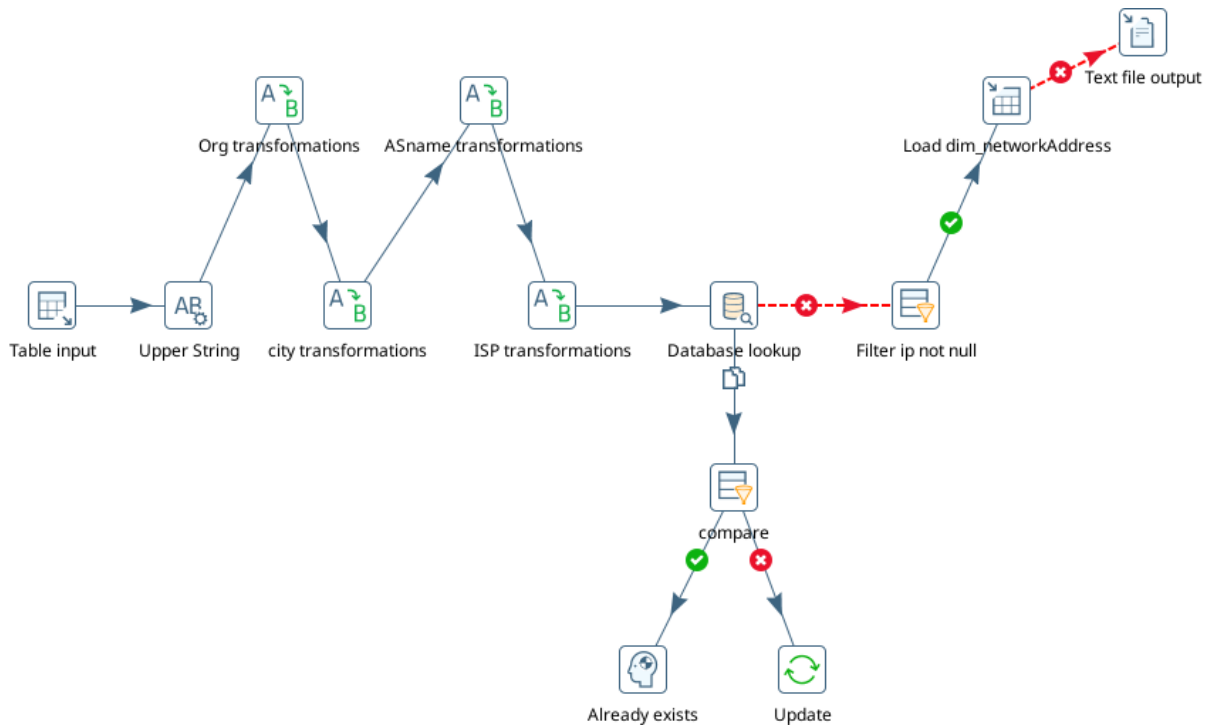


Figure 44 - Process to perform the periodic ETL of the dimension Dim_NetworkAddress

After finishing this phase, we would like to highlight that even though the construction of the data warehouse should be done with care, it was not one of the objectives of this work. The decision to build a data warehouse was motivated by the fact of having a repository with data from the various sources and the data warehouse would facilitate the analysis of it and visualisation of the results. Any other technology which allows the achievement of these goals would be interesting.

3.1.6 Phase 6: Analyse the data, draw possible conclusions and discuss the results

In this phase, it is presented the analysis and discussion of the results. It is answered the defined questions and, to help in the visualization of the results, it is presented some of the reports generated in Metabase. This phase also includes the conclusions and discussion of the results. The outcome of it is presented in the next chapter.

Conclusions

This chapter presented the methodology proposed to perform this work. The methodology is composed of 6 phases. Phase 1 - Find reliable and updated databases with information about ICSs. Through the literature study, we observed that the most well-known and used port scan projects are Shodan and Censys. By analyse them, we notice that they provide data about ICSs located in Portugal that would allow us to achieve the objectives of the work. Thus, we chose to use Shodan and Censys as the source of data of this work.

Phase 2 - Identify the Industrial Control Systems. Through the literature review, we notice that ICSs use specific protocols known as ICS standard protocols. By analysing the data, we found that the protocol used on the systems is one of the information available on the gathered data. Thus, we used this information to distinguish the ICSs from the other systems.

Phase 3 - Identify the features that are or may represent vulnerabilities to the ICSs. From the chosen sources, Shodan and Censys, we had information about the IP of the system, the ISP, the name of the AS that regulates the routing policies of the system, the organisation to which the system belongs, the city where the system is located, the active ports and protocols running on the system, the information about the SSL/TLS versions and cipher suite available to encrypt the communication, and the CVEs identified on the system. Based on the gathered data, the features that we considered that are or may represent vulnerabilities to the ICSs are the exposed ports, the cipher suite used to encrypt the communication, the version of SSL/TLS available to be used on the encryption, and the CVEs identified on the system.

Phase 4 - Define the calculation of the level of risk of the systems. It was defined the proposal of calculation of the risk level of the systems. The risk score was defined as the product of the Likelihood and the Impact of exploiting a vulnerability on a given system. As a mean of simplification, instead of working with each of the potential vulnerabilities separately, we divided the content of the features into different classes, based on the characteristics of the potential vulnerabilities. Then, we defined the meaning of each of the classes of each feature, and the Impact and the Likelihood score of each one.

Phase 5 - Build a data warehouse to join and organise all the data. It was built a data warehouse to combine and organise the data from the different sources into one comprehensive database in order to better analyse it. It was presented an overview of the data, the requirements analysis of the data warehouse, the data model that translate the requirements, and then it was presented the main decisions made on the initial and periodic ETL processes.

Finally, Phase 6 of the methodology - Analyse the data, draw possible conclusions and discuss the results. This phase includes the analysis and discussion of the results. The outcome of this phase is presented in the next chapter.

4 ANALYSIS OF THE DATA AND DISCUSSION OF THE RESULTS

This chapter presents the analysis and discussion of the results of this work. For the visualization and better understanding of the results we used reports generated in Metabase. The presented analysis corresponds to phase 6 of the methodology proposed in Chapter 3.

The analysed data was gathered from Censys and Shodan between February 2020 and July 2020, and all the data was filtered in order to only get the systems located in Portugal. The fact table contains 273803 records, with 167252 different network addresses, running 80 different ports. The network addresses were in 1072 different locations and belong to 92 different organisations. There are 78 ISPs providing Internet services to them and 81 different ASs that maintaining their network routing policies. It was also been identified 57 different cipher suites and 361 different CVEs on the network addresses.

The analysis of the data was based on the questions previously established to be answered by the data warehouse. Below, we present the results for each of these questions.

Q1. Number of network addresses running High/Medium/Low Risk ports

Figure 45 presents the number of exposed network addresses per risk classification of their exposed ports. From the 167252 network addresses identified, 12301 were running High Risk ports, 118225 were running Medium Risk ports, and 36696 were running Low Risk ports.

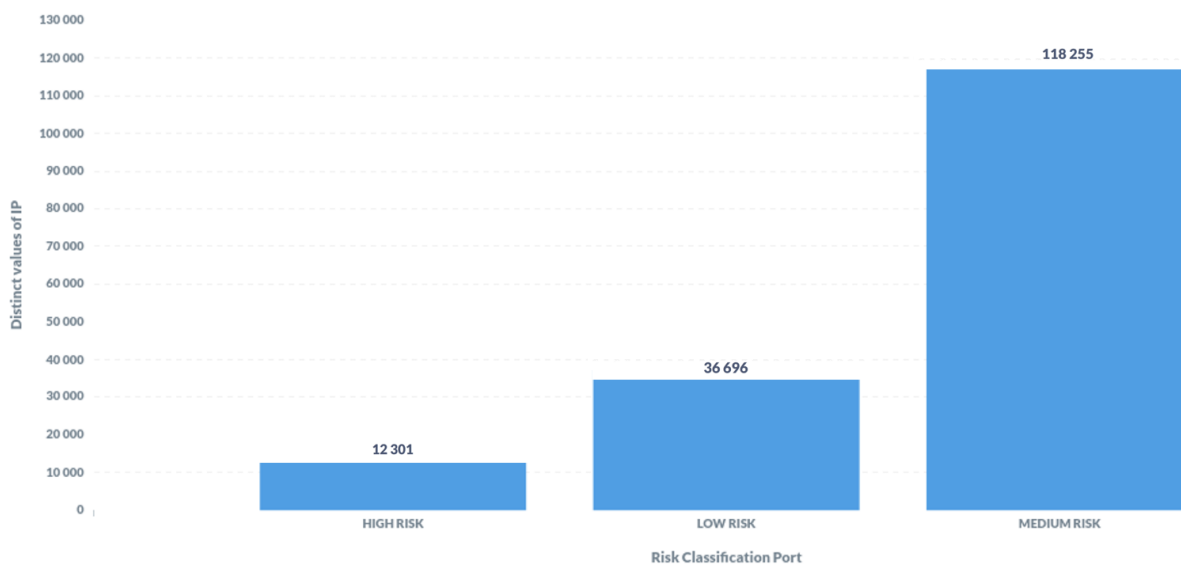


Figure 45 - Number of exposed network addresses per risk level of ports

By analysing the exposed High Risk ports, we found that most of them are running the protocol FTP, which is a concern, since FTP was not built to be secure, as data sent via FTP are vulnerable to sniffing (theft or interception of network traffic data), spoofing (when a person or program successfully identifies as another by falsifying data, to gain an illegitimate advantage), brute force attacks, and among other basic attack methods. This result shows that, despite all

the security recommendations, a large number of organisations still have their systems running protocols already known as insecure and not recommended.

Q2. Number of network addresses using High/Medium/Low Risk SSL/TLS versions

Figure 46 presents the number of exposed network addresses per risk classification of their SSL/TLS versions. As we can see, 140263 network addresses have High Risk SSL/TLS versions associated and 26989 has Low Risk SSL/TLS versions associated. The rest of the systems have the ports filtered, and thus we cannot determinate the SSL/TLS used.

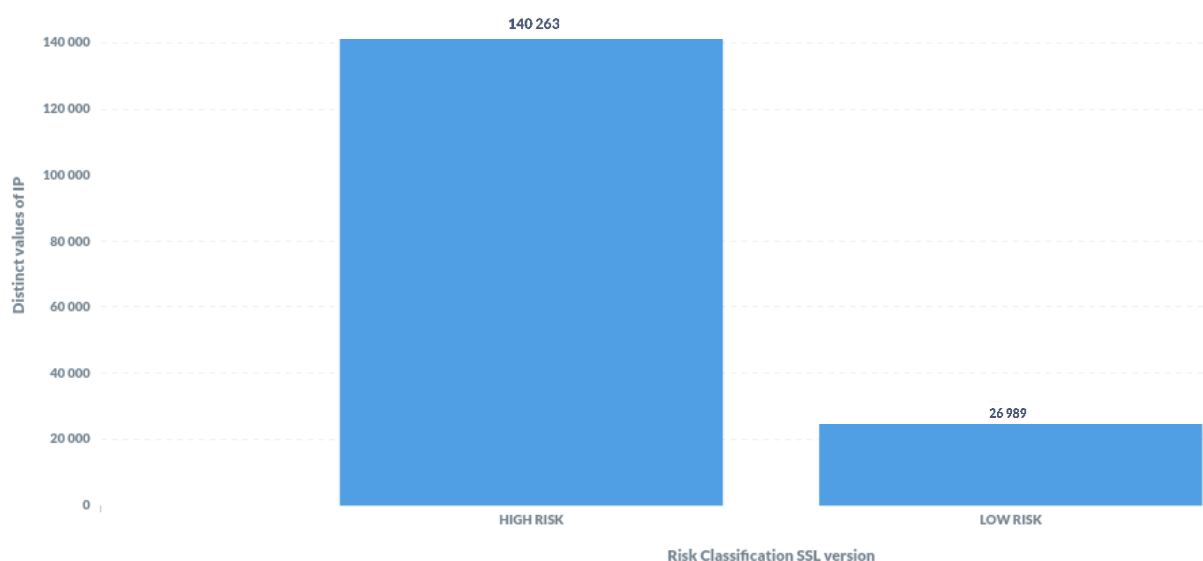


Figure 46 - Number of exposed network addresses per SSL/TLS version risk classification

As we can see most of the systems are using High Risk SSL/TLS versions. However, it is important to take into account that systems without any type of encryption associated are considered in this group. By analysing the systems with High Risk SSL/TLS version, we observe that most of them (152374) do not have any type of encryption associated. From those systems that have an encrypted connection, 786 of them are using SSLv2, 1157 of them are using SSLv3, and 3184 of them are using TLSv1.

The high number of systems without any type of associated encryption is worrying because as mentioned above, even a system with weak encryption is more secure than a system without any type of encryption. In addition, some systems present encrypted connection but they are using versions of SSL/TLS that are no longer considered secure, which gives the organisations a false sense of security.

Q3. Number of network addresses using High/Medium/Low Risk cipher suites

Figure 47 presents the number of exposed network addresses per risk classification of the used cipher suite. As we can see, there are 161016 network addresses using cipher suites with High

Risk classification associated, 5256 network addresses using cipher suites with Low Risk classification associated, and 980 network addresses using cipher suites with Medium Risk classification associated.

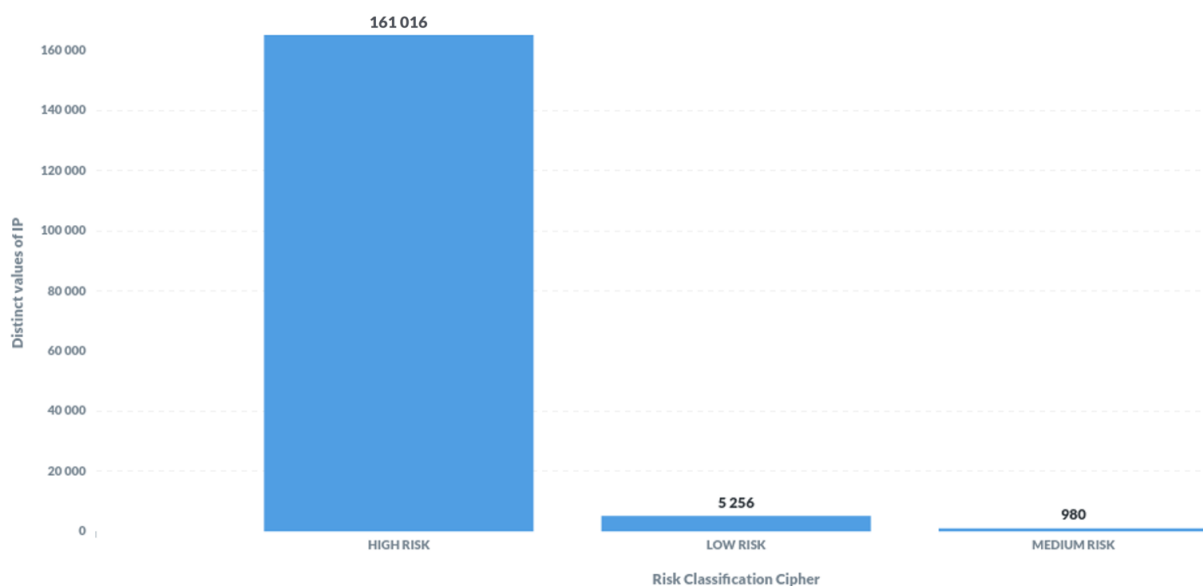


Figure 47 - Number of network addresses per risk classification of the used cipher suite

By analysing the systems using High Risk cipher suites, as the SSL version, we observe that most of the systems do not have any algorithm of encryption associated. From those that have, most of them are using the cipher suite TLS_RSA_WITH_RC4_128_SHA, followed by RC4-SHA and TLS_RSA_WITH_RC4_128_MD5. All cipher suites that are considered insecure and their use are not recommended.

As we can see, most of the systems with encrypted connections, are using cipher suites that are no longer considered secure. This highlights the observation made in the previous question that, there are organisations that are concerned about encrypting the connection of their system, however, they use insecure algorithms of encryption on them.

Q4. Number of network addresses with High/Medium/Low Risk CVEs

As seen before, all CVEs present High Risk for the systems where they have been identified. However, among the CVEs, some present a higher risk than others based on their severity. This can be seen in Figure 48, which presents the number of systems per qualitative severity of the CVEs identified on them. As we can see, from the systems where has been identified CVEs, most of them (830) have CVEs with Medium severity associated, 585 have CVEs with Low severity associated and 344 have CVEs with High severity associated.

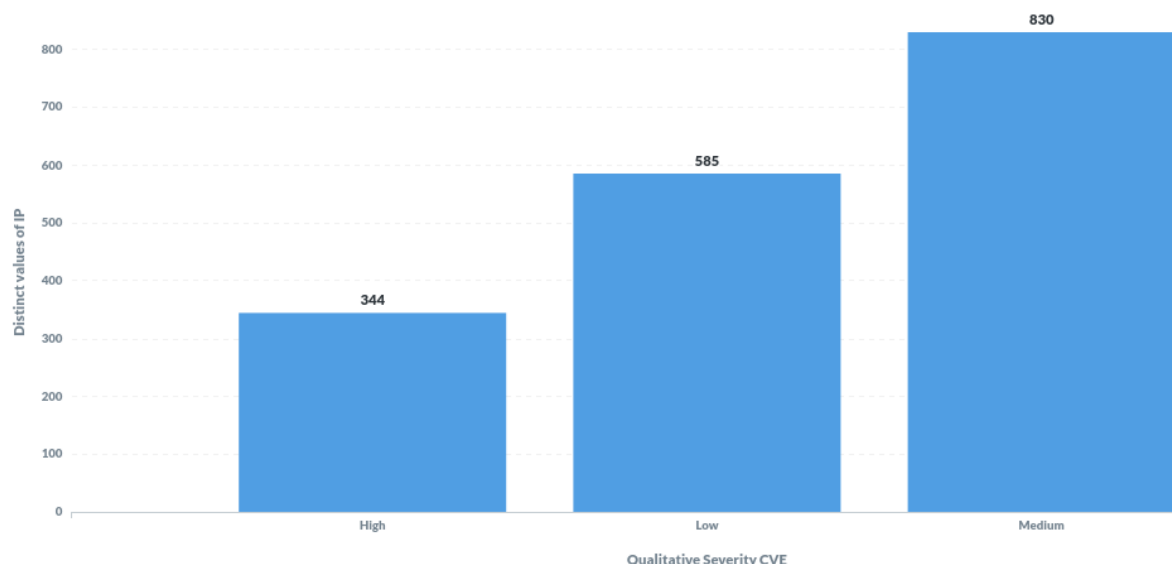


Figure 48 - Number of network addresses per qualitative severity of the CVEs identified

As we can see, there are many systems where CVEs have been identified, including CVEs with high severity, making the system very insecure. As CVEs are known vulnerabilities on the system, the organisations should make the effort to mitigate them, prioritizing those with higher severity.

Q5. Level of risk per network address/most vulnerable network addresses

This question is one of the most important in this study, as it aims to determine which systems present the highest level of risk. Figure 49 presents the systems with the highest level of risk of each month. As we can see, April is the month that presents the system with the highest level of risk (IP: 94.61.122.246) with the aggregation of the risk score equal to 7256. Although the system presented a single open port, running the protocol Yaskawa MP Series Ethernet, it was identified 123 CVEs in that port. The system does not present an encrypted algorithm to secure the connection. It was located in Coimbra and it is under the Vodafone Portugal network.

Month	IP	Sum of Level Of Risk
December, 2019	109.48.0.227	37
January, 2020	109.48.192.166	53
February, 2020	188.251.17.15	6 994
March, 2020	188.81.197.137	1 492
April, 2020	94.61.122.246	7 256
May, 2020	193.136.19.70	3 963
June, 2020	188.82.152.10	1 492
July, 2020	193.136.19.70	3 963

Figure 49 - Systems with the highest level of risk per month

In addition to the analysis of the systems with the highest level of risk per month, we made a general analysis to find out, in general, which are the systems with the highest level of risk, by summing up the risk that each one presented in each month. Figure 50 presents the top 10 of the systems with the highest level of risk. The list contains the IP of the systems, the organisations to which they belong, the name of their AS and ISP, the City where they are located, and the sum of the level of risk of the system. The sum of the level of risk is the aggregation of the risk score of all the vulnerabilities that the system contains.

IP	Organization	As Name	Isp	City	Sum of Level Of Risk
188.251.17.15	MEO	MEO-RESIDENCIAL	MEO	RIO TINTO	13 951
193.136.19.70	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	BRAGA	11 926
94.61.122.246	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	COIMBRA	7 256
213.58.234.184	ONITELECOM - INFOCOMUNICACOES, S.A.	ONI INTERNET SERVICE PROVIDER	ONITELECOM - INFOCOMUNICACOES, S.A.	PORTO	3 732
193.137.201.176	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	COIMBRA	3 341
62.28.221.86	MEO	MEO-EMPRESAS	MEO	PORTELA	3 234
195.23.53.194	NOS COMUNICACOES, S.A.	NOS_COMUNICACOES	NOS COMUNICACOES, S.A.	NA	2 996
93.108.249.125	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	ESTORIL	2 937
2.81.97.211	MEO	MEO-RESIDENCIAL	MEO	CONSTANCIA	2 840
213.13.123.92	MEO	MEO-RESIDENCIAL	MEO	CASTRO DAIRE	2 751

Figure 50 - Top 10 network addresses with the highest level of risk

The network address with the highest level of risk is the one identified by the IP: 188.251.17.15. It has only 2 ports exposed, however, one of the ports that is running the protocol SAIA S-BUS (standard ICS protocol) has 121 CVEs and the other, running the protocol MySQL, has 1 CVE. In total, 122 CVEs were identified in this system. None of the ports has an encrypted algorithm associated.

It is possible to observe that the name of the organisation and the ISP are the same. By analysing all the data with more details, we could observe that most of the IP addresses are pointing to the ASs of their internet providers, which means that most of the organisations do not have their own infrastructures to maintain the network routing policies of their systems. In this situation, it is not possible to identify the organisations because they were hidden behind their ISPs. This can be advantageous in the sense that the attacker cannot identify immediately the organisation to which the system belongs, but in the other hand, the organisation becomes dependent on the ISP, in such a way that if the ISP suffer an attack and goes offline, all the organisations behind it will be severely affected.

Q6. Number of exposed ports per network address

This question aims to know the number of different ports running on the systems, in order to identify the most exposed ones. Figure 51 presents the list of the top 10 most exposed network addresses, identified by their IP. The list presents the IP of the network addresses, the organisations to which they belong, the name of their AS, and ISP, the City where they are located, and the number of their exposed ports.

IP	Organization	As Name	Isp	City	Nr of exposed ports
193.137.201.176	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.R	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.R	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.R	COIMBRA	25
148.63.201.154	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	VILA DO CONDE	24
148.71.172.44	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	BRAGA	21
149.90.238.76	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	PORTO	21
149.90.87.114	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	NA	21
194.38.136.242	ONITELECOM - INFOCOMUNICACOES, S.A.	ONI INTERNET SERVICE PROVIDER	ONITELECOM - INFOCOMUNICACOES, S.A.	VILAMOURA	21
149.90.165.59	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	NA	20
178.166.44.158	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	LISBOA	17
46.189.154.131	VODAFONE PORTUGAL	VODAFONE-PT VODAFONE PORTUGAL	VODAFONE PORTUGAL	LISBOA	17
5.196.38.179	OVH HOSTING LDA	OVH	OVH HOSTING LDA	NA	17

Figure 51 - Top 10 of the most exposed systems

In the list presented above, the most exposed network address is the one identified by the IP: 193.137.201.176, with 25 distinct ports exposed. The network address belongs to Fundação para Ciencia e a Tecnologia, its routing policies are managed by Fundação para Ciencia e a Tecnologia, and its Internet is also provided by Fundação para Ciencia e a Tecnologia. The network address is located in Coimbra. Figure 52 presents the evolution over time of the number of exposed ports in this network address.

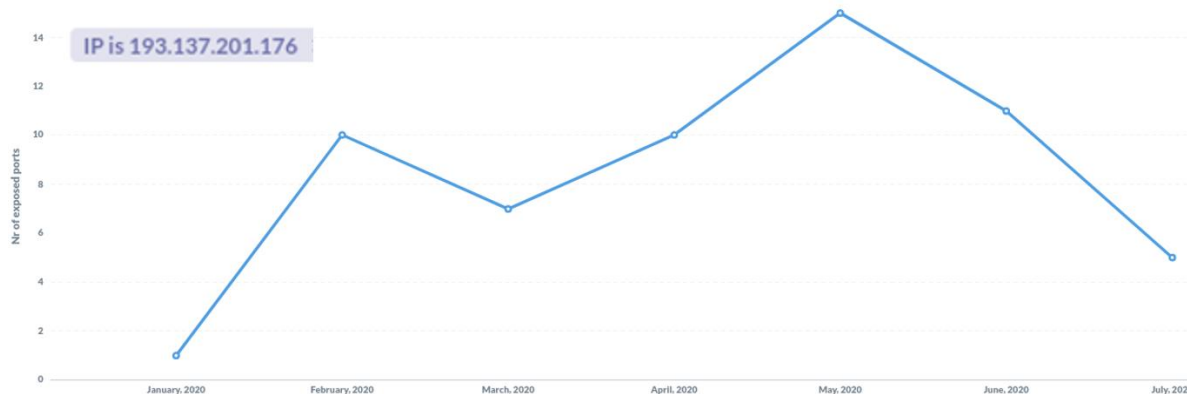


Figure 52 - Evolution of the number of exposed ports of the most exposed network address

The system was first port scanned in January where it presented 1 open port running the protocol SSH. It did not present any algorithm of encryption to secure the connection and it was not identified any CVE on the open port.

In February the number of exposed ports augmented to 10. The exposed ports were running the following protocols: Crimson 3, DNP3, FTP, HTTP, HTTPS, IMAP, POP3, SMTP, Telnet, and Yaskawa MP Series Ethernet. The CVE-2019-12815 was identified on 3 of the exposed ports (the ports running Yaskawa MP Series Ethernet, Crimson 3, and DNP3). Only the port running the protocol HTTPS was using an algorithm of encryption to secure the connection. However, the port was filtered and thus, it was not possible to identify the SSL/TLS version used.

In March the number of exposed ports decreased (from 10 to 6) when compared to February. The exposed ports were running the following protocols: Ethernet/IP, Codesys, Omron Fins, Schleicher XCX 300, Fox, DNP3, and Yaskawa MP Series Ethernet. There were 2 ports running the protocol Codesys. None of the ports was using an algorithm of encryption to secure the connection. The CVE-2019-12815 was identified in all of the exposed ports.

In April the number of exposed ports increased again (from 6 to 9) when compared to March. The exposed ports were running the following protocols: Unitronics Socket1, MELSEC Q, GE-SRTP, IEC-60870-5-104, PCWORX, Panasonic FP (Ethernet), Yaskawa MP2300Siec, Crimson 3, and Fox. None of the ports was using an algorithm of encryption to secure the connection. The CVE-2019-12815 was identified in all of the exposed ports.

In May the number of exposed ports of the network address increased (from 9 to 15) when compared to April. The exposed ports were running the following protocols: MELSEC Q, Omron Fins, Fox, Unitronics Socket1, Yaskawa MP Series Ethernet, Ethernet/IP, Siemens S7, Panasonic FP2 (Ethernet), DNP3, PCWORX, IEC-60870-5-104, Codesys, SCP Config, Schleicher XCX 300, and GE-SRTP. None of the ports was using an algorithm of encryption to secure the connection. The CVE-2019-12815 was identified in all of the exposed ports.

In June the number of exposed ports decreased (from 15 to 11) when compared to May. The exposed ports were running the following protocols: GE-SRTP, Fox, Yaskawa MP Series Ethernet, Siemens S7, Crimson 3, Codesys, DNP3, Schleicher XCX 300, PCWORX, Ethernet/IP, and Yaskawa MP2300Siec. None of the ports was using an algorithm of encryption to secure the connection. Except the ports running the protocols GE-SRTP, Fox, and Yaskawa MP Series Ethernet, the CVE-2019-12815 was identified in all of the exposed ports.

In July the number of exposed ports of the network address decreased again (from 11 to 5) when compared to June. The exposed ports were running the following protocols: Panasonic FP2 (Ethernet), Omron Fins, IEC-60870-5-104, MELSEC Q, and Unitronics Socket1. None of the ports was using an algorithm of encryption to secure the connection. The CVE-2019-12815 was identified in all of the exposed ports.

As we could observe, the CVE-2019-12815 were identified in almost every exposed ports of the most exposed network address (IP: 193.137.201.176). The CVE presents the following description: “An arbitrary file copy vulnerability in mod_copy in ProFTPD up to 1.3.5b allows for remote code execution and information disclosure without authentication”. In other words, the vulnerability represents an arbitrary file copy vulnerability in the module mod_copy of ProFTPD versions up to 1.3.5b, due to improper access control. According to (ProFTPD, 2019), the module mod_copy does not honor the “<Limit READ>” and “<Limit WRITE>” configuration settings in the proftpd.conf. This allows remote attackers without write permissions to copy files on the FTP server. The CVE is classified as representing High Risk to the system.

By comparing the list of the systems with more ports exposed (Figure 51) and the list of the systems with the highest level of risk (Figure 50), we could observe that the systems with the most exposed ports were not always the ones with the highest level of risk. The latter is dependent on the number of all the vulnerabilities identified on the system and the risk score of each one.

Taking into account the information about the geolocation of the exposed systems, we also analysed the cities with the highest number of exposed systems. Lisboa appears as the most

exposed city with the highest number of exposed network addresses, followed by Porto with the second-highest and Quinta do Conde.

Q7. Number of exposed ports running standard and non-standard ICS

This question aims to know from the different ports running on the identified systems, which ones are running standard and non-standard ICSs protocols. Figure 53 presents the number of ports running ICS standard and non-standard protocols. As we can see, from the 80 different ports running on the systems, almost 70% of the ports are running non-standard ICS protocols (green area) and only approximately 30% are running ICS standard protocols (violet area). This means that, on average, for each ICS standard protocol running on a system there are 2,3 non-standard ICS protocols running on the same system.

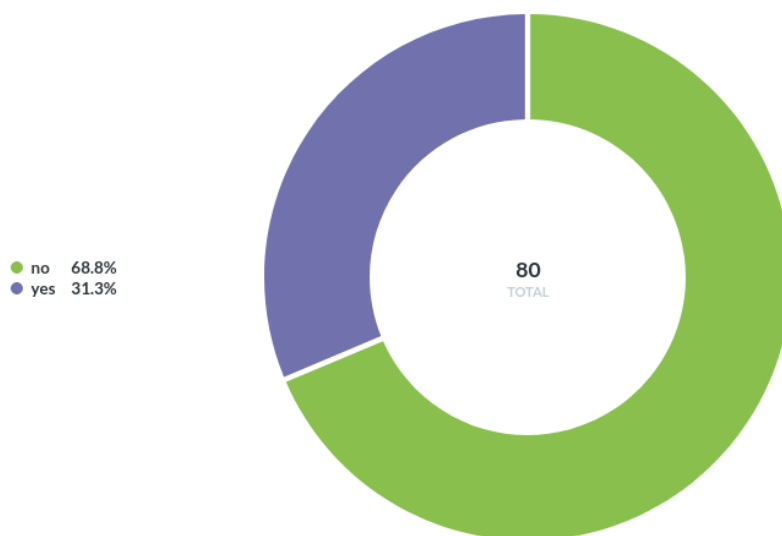


Figure 53 - Number of ports running ICS standard and non-standard protocols

The organisations should pay attention to the ports running non-standard protocols and monitor them, because they are often ports that should not be running on an ICS or unnecessary ports that only increase the attack surface of the system.

Q8. Most common protocols exposed

This question intendeds to identify the most exposed protocols. In other words, the number of different systems running a given protocol, as we can see in Figure 54. As we can see, HTTP is the most exposed protocol, running in more than 120000 of the 167252 exposed network addresses. This means that the majority of the identified systems have a port running an HTTP protocol. This may represent a problem because HTTP are not considered secure. It is then followed by HTTPS and ISAKMP protocol.

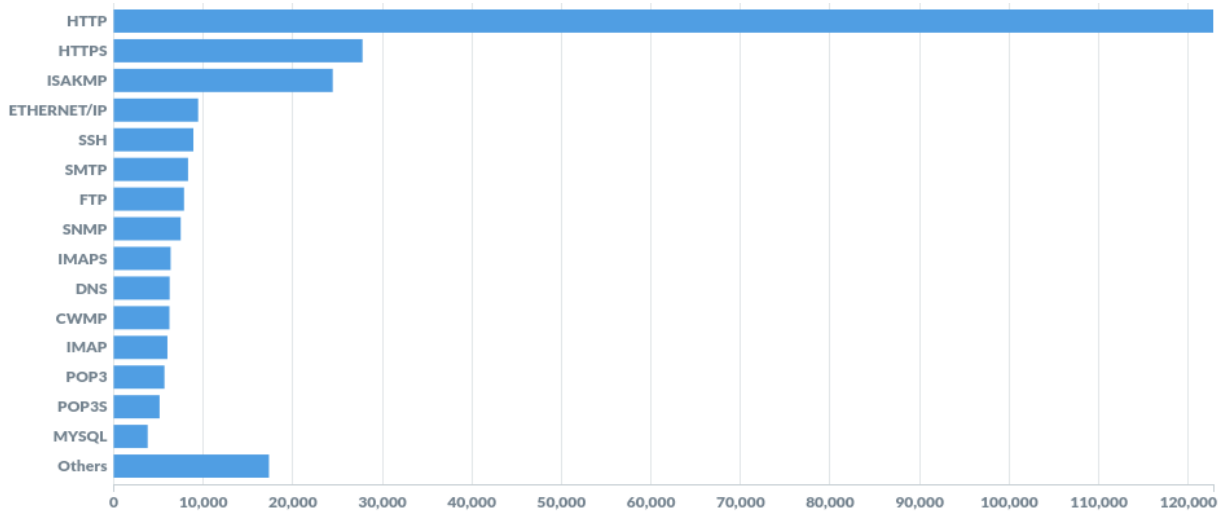


Figure 54 - Most Exposed protocols.

As we can see from the graph presented above, among the most exposed protocols there are some database protocols. Database protocols typically never have reason to be open and visible, as they are usually accessed from specific applications and therefore it is always possible to configure the firewall restricting access to only the intended sources. Even if these ports are protected through authentication, it should be assumed that authentication can fail, and therefore the exposition of these ports should be avoided. Figure 55 presents the number of distinct ICSs running database protocols.

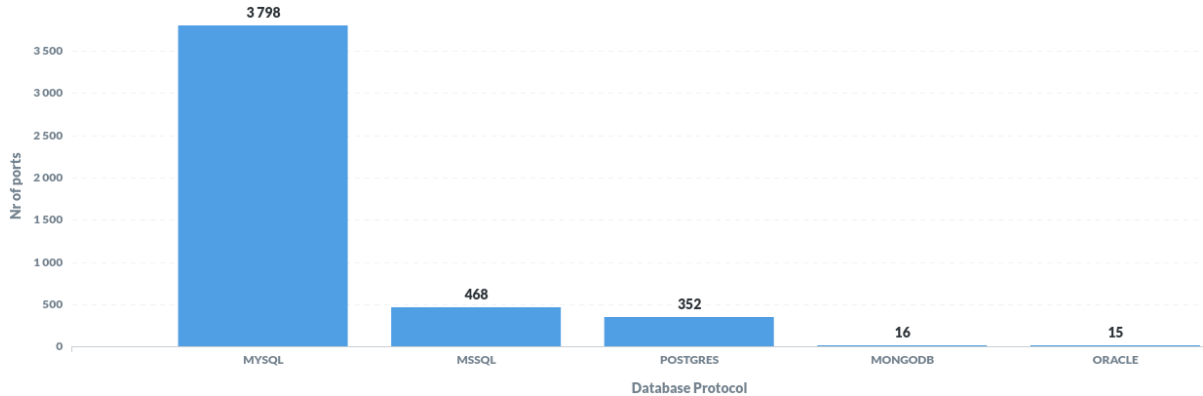


Figure 55 - Exposed database protocols

There are 3798 systems running the protocol MySQL, 468 running the protocol MSSQL, 352 running the protocol Postgres, 16 running the protocol MongoDB and 15 running Oracle. MySQL is the most exposed database protocol.

Q9. Number of CVEs per network address

Figure 56 presents the top 10 of the systems with more CVEs identified. It presents the system, identified by the IP and the number of distinct CVEs identified on each one. The system with more CVEs identified is the one identified by the IP: 94.61.122.246, with 123 CVEs. Then, it is followed by the system identified by the IP: 188.251.17.15, with 122 CVEs. After, it is possible to observe the system identified by the IP: 193.136.19.70, with 70 CVEs identified. As

expectable, the systems with the highest number of CVEs identified appear on the list of the most vulnerable systems.

IP	Distinct values of ID Cve
94.61.122.246	123
188.251.17.15	122
193.136.19.70	70
188.81.197.137	27
188.80.200.119	26
188.80.79.141	26
188.82.152.10	26
188.83.162.197	26
188.83.224.29	26
2.82.144.171	26

Figure 56 - Number of distinct CVEs per network address

Q10. Most common CVEs identified on the systems

Figure 57 presents the top 10 of the CVEs that have been most identified on the exposed systems, their qualitative severity, their published year and the number of systems that they are identified (Count column).

Cve	Qualitative Severity	Publish Date: Year	Count
CVE-2017-15906	Medium	2017	661
CVE-2018-20685	Low	2019	569
CVE-2019-6109	Medium	2019	569
CVE-2019-6110	Medium	2019	569
CVE-2019-6111	Medium	2019	569
CVE-2018-15919	Medium	2018	406
CVE-2016-10708	Medium	2018	381
CVE-2014-1692	High	2014	342
CVE-2010-5107	Medium	2013	334
CVE-2016-0777	Medium	2016	315

Figure 57 - Top 10 of the most identified CVEs

The most common CVE is CVE-2017-15906. It presents a medium severity to the systems. It was published in 2017 and it was identified on 661 network addresses. On the description of the CVE, it says: “The process_open function in sftp-server.c in OpenSSH before 7.6 does not properly prevent write operations in read-only mode, which allows attackers to create zero-length files”. In other words, the system is vulnerable to a denial of service (an interruption in

a machine or network, making them inaccessible to the intended users), caused by an error in the function `process_open()`, when it is on read-only mode (IBM, 2018).

On the top 10 of the most exposed CVEs, we only have one CVE with High severity, the CVE-2014-1692. This CVE was published in 2014 and it was identified on 342 network addresses. The description of the CVE says: “The `hash_buffer` function in `schnorr.c` in OpenSSH through 6.4, when `Makefile.inc` is modified to enable the J-PAKE protocol, does not initialize certain data structures, which might allow remote attackers to cause a denial of service (memory corruption) or have unspecified other impacts via vectors that trigger an error condition”. In other words, the vulnerability affects the function `hash_buffer` of the file `schnorr.c` of the component J-PAKE Protocol Handler, and its exploitation leads to a denial-of-service vulnerability (Dowd, 2020).

The older the CVE, the more risk it presents to the system. This is because with time more forms of exploitation of the CVE are found. Figure 58 presents the top 10 of the organisations with the oldest CVE.

Organization	Publish Date
UNIVERSIDADE DE LISBOA	December, 2003
VODAFONE PORTUGAL	November, 2004
UNIVERSIDADE DE LISBOA	December, 2004
VODAFONE PORTUGAL	July, 2005
MEO	August, 2005
UNIVERSIDADE DE LISBOA	August, 2005
VODAFONE PORTUGAL	August, 2005
MEO	September, 2005
UNIVERSIDADE DE LISBOA	September, 2005
MEO	January, 2006

Figure 58 - Top 10 organisations with the oldest CVEs

In general, the Universidade de Lisboa is the organisation that presents the oldest CVEs, having appeared 4 times in the list of top 10 organisations with the oldest CVE, with CVEs published in December 2003, December 2004, August 2005, and September 2005. Then we have some organisation hiding behind Vodafone Portugal and MEO.

Further analysis

In addition to the questions analysed above, we verified that there are some organisations found on the data warehouse, whose systems do not appear on the list of the most exposed or vulnerable systems. However, has caught our attention for the organisation itself, because they are crucial organisations for the functioning of the country and an attack on them could bring

much damage to the country. On this basis, we analysed risk of their systems. Those organisations are:

- Administração central do Sistema de Saúde, I.P. - a public institute, integrated into the indirect administration of the State, responsible for ensuring the integrated management of the resources of the National Health System.
- Aeroportos de Portugal - the authority responsible for the administration of airports in Portugal. It integrates several companies in the aviation sector.
- Comissão do Mercado de Valores Mobiliários (CMVM) - the organisation in Portugal responsible for supervising and regulating the operation of markets in financial instruments and the activity of all agents acting in them.
- Instituição Financeira de Crédito - Credibom - financial company for credit acquisitions, which currently belongs to the Sofinco group.
- Instituto Nacional de Estatística, I.P. (INE) - the official agency in Portugal responsible for producing and publishing, in an effective, efficient and impartial way, high-quality official statistical information.
- Redes Energéticas Nacionais (REN) - the Portuguese organisation responsible for the transportation and management of the national electric and natural gas.
- Secretaria Geral- Ministério da Administração Interna - the agency responsible for providing technical and administrative support to the offices of the members of the Government within the Ministry of Internal Administration.
- Transportes Aéreos Portugueses (TAP) – Oficial portuguese Airlines.

Figure 59 presents the level of risk of the organisations presented above. As we can see, The Instituição Financeira de Crédito - Credibom is the organisation that presents the highest level of risk (sum of the level of risk equal to 450). It is then followed by Aeroportos de Portugal with the sum of the level of risk equal to 315 and then TAP with the sum of the level of risk equal to 296. From the selected organisations, the CMVM is the one that presents the lower sum of the level of risk, equal to 82.

Organization	Sum of Level Of Risk
INSTITUICAO FINANCEIRA DE CREDITO SA	450
AEROPORTOS DE PORTUGAL, SA	315
TRANSPORTES AEREOS PORTUGUESES SA	296
INSTITUTO NACIONAL DE ESTATISTICA, I.P.	243
REDE ENERGETICAS NACIONAIS, S.A.	185
ADMINISTRACAO CENTRAL DO SISTEMA DE SAUDE, I.P.	119
SECRETARIA-GERAL MINISTERIO DA ADMINISTRACAO INTERNA	111
COMISSAO DO MERCADO DE VALORES MOBILIARIOS	82

Figure 59 - Sum of the level of risk of some critical organisations

Bellow, we will analyse each one of these organisations:

- **Instituição Financeira de Crédito, S.A - Credibom**

Figure 60 presents the exposed network addresses of Credibom. On the Figure, the name of the organisation appears as Instituição Financeira de Crédito, S.A, which may be the institution that provides Internet services to this and other credit institutions. Through the AS name we discovered that the network addresses belong to Credibom.

The organisation presents 3 exposed network addresses, all of them running the protocol SNMP, all with the same High Risk vulnerabilities, and apparently with the same characteristics. It was not identified any CVE on the systems.

The fact that all the exposed network addresses present a High Risk level, they should be analysed with more details by the organisation.

IP	Organization	Name Ssl Version	Protocol	Cve	As Name	Isp	Level Of Risk
193.36.190.128	INSTITUICAO FINANCEIRA DE CREDITO SA	NA	SNMP	NA	CREDIBOM	INSTITUICAO FINANCEIRA DE CREDITO SA	150
193.36.190.129	INSTITUICAO FINANCEIRA DE CREDITO SA	NA	SNMP	NA	CREDIBOM	INSTITUICAO FINANCEIRA DE CREDITO SA	150
193.36.190.255	INSTITUICAO FINANCEIRA DE CREDITO SA	NA	SNMP	NA	CREDIBOM	INSTITUICAO FINANCEIRA DE CREDITO SA	150

Figure 60 - Exposed network addresses of Credibom

- **Aeroportos de Portugal**

Figure 61 presents the list of the exposed network addresses of Aeroportos de Portugal identified. As we can see, there are 3 exposed network addresses. The first one identified by the IP: 104.244.9.87 has one open port running the protocol SNMP. The second one, identified by the IP: 104.244.9.129, has one open port also running the protocol ISAKMP. The third exposed network address is identified by the IP: 104.244.9.127. It has 3 open ports, running the protocols HTTP, ISAKMP and HTTPS respectively. It is the only one that presented an encrypted communication, using the most secure SSL version available: the TLSv1.2 on the port running the protocol HTTPS. It is also the network address that presented the highest aggregate level of risk, which results from the sum of the level of risk of each of its open ports. In none of the network, addresses were identified CVEs. The Aeroportos de Portugal, S.A. have its own infrastructures to maintain the network routing policies of its systems.

IP	Organization	Name Ssl Version	Protocol	Cve	As Name	Isp	Level Of Risk
104.244.9.87	AEROPORTOS DE PORTUGAL, SA	NA	SNMP	NA	ANA-AS	AEROPORTOS DE PORTUGAL, SA	50
104.244.9.129	AEROPORTOS DE PORTUGAL, SA	NA	ISAKMP	NA	ANA-AS	AEROPORTOS DE PORTUGAL, SA	111
104.244.9.127	AEROPORTOS DE PORTUGAL, SA	NA	HTTP	NA	ANA-AS	AEROPORTOS DE PORTUGAL, SA	37
104.244.9.127	AEROPORTOS DE PORTUGAL, SA	NA	ISAKMP	NA	ANA-AS	AEROPORTOS DE PORTUGAL, SA	111
104.244.9.127	AEROPORTOS DE PORTUGAL, SA	TLSv1.2	HTTPS	NA	ANA-AS	AEROPORTOS DE PORTUGAL, SA	6

Figure 61 - Exposed network addresses of Aeroportos de Portugal

- **Transportes Aéreos Portugueses (TAP)**

Figure 62 presents the exposed network addresses of Transportes Aéreos Portugueses (TAP). The organisation presents 3 exposed network addresses identified by the IPs: 91.198.90.11,

91.198.90.211, and 91.198.90.252. All of the network addresses had one open port. The network addresses identified by the IPs: 91.198.90.11 and 91.198.90.211 was running the protocol ISAKMP and the network address identified by the IP: 91.198.90.252 was running the protocol Ethernet/IP. All the network addresses identified, presented high risk levels and should be analysed by the organisation. None of the ports had algorithms of encryption available. It was not identified any CVE on the ports.

TAP is using their own infrastructures to maintain the network routing policies of their systems. Although the name of the AS was not available, by performing some searching we found that it [AS43643](#).

IP	Organization	Name Ssl Version	Protocol	Cve	As Name	Isp	Level Of Risk
91.198.90.11	TRANSPORTES AEREOS PORTUGUESES SA	NA	ISAKMP	NA	NA	TRANSPORTES AEREOS PORTUGUESES SA	111
91.198.90.211	TRANSPORTES AEREOS PORTUGUESES SA	NA	ETHERNET/IP	NA	NA	TRANSPORTES AEREOS PORTUGUESES SA	37
91.198.90.252	TRANSPORTES AEREOS PORTUGUESES SA	NA	ISAKMP	NA	NA	TRANSPORTES AEREOS PORTUGUESES SA	148

Figure 62 - Exposed network addresses of TAP

- **Instituto Nacional de Estatística**

Figure 63 presents the exposed network addresses of Instituto Nacional de Estatística (INE). The organisation presents two exposed network addresses identified by the IP: 193.192.11.1 and IP: 193.192.8.178 respectively. The network address identified by the IP: 193.192.11.1 has 2 open ports, running the protocols HTTPS and SNMP. The network address identified by the IP: 193.192.8.178 has 4 open ports, running the protocols HTTPS, SSH, HTTP and Yaskawa MP Series Ethernet, respectively. The ports running the protocols HTTPS from both network addresses and the port running the protocol SSH of the network address with IP: 193.192.11.1 are filtered and thus it is not possible to determinate the version of the SSL/TLS used. The other exposed ports do not have any encrypted algorithm available. The network address, identified by the IP: 193.192.11.1, has a port with a level of risk of 100, that should be analysed by the organisation. The network address identified by the IP: 193.192.8.178 should also be analysed with attention because it presents 4 open ports where at least 2 of them have characteristics that present High Risk to the system. It was not identified any CVEs on the exposed ports. The INE has its own infrastructures to maintain the network routing policies.

IP	Organization	Name Ssl Version	Protocol	Cve	As Name	Isp	Level Of Risk
193.192.11.1	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	filtered	HTTPS	NA	INE-AS	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	8
193.192.11.1	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	NA	SNMP	NA	INE-AS	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	100
193.192.8.178	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	filtered	HTTPS	NA	INE-AS	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	8
193.192.8.178	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	filtered	SSH	NA	INE-AS	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	16
193.192.8.178	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	NA	HTTP	NA	INE-AS	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	37
193.192.8.178	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	NA	YASKAWA MP SERIES ETHERNET	NA	INE-AS	INSTITUTO NACIONAL DE ESTATISTICA, I.P.	74

Figure 63 - Exposed network addresses of INE

- **Redes Energéticas Nacionais (REN)**

Figure 64 presents the exposed network addresses of Redes Energéticas Nacionais (REN). It presents one exposed network address identified by the IP: 185.165.104.250, with two ports open. Both of the open ports are running the protocol ISAKMP. Both of the ports have a high-risk level associated, and thus should be analysed by the organisation with more details. None

of the ports has an encrypted algorithm available. It was not identified any CVE on the system. None of the ports has an encrypted algorithm available. The REN has its own infrastructures to maintain the network routing policies of its systems.

IP	Organization	Name Ssl Version	Protocol	Cve	As Name	Isp	Level Of Risk
185.165.104.250	REDE ENERGETICAS NACIONAIS, S.A.	NA	ISAKMP	NA	REN	REDE ENERGETICAS NACIONAIS, S.A.	111
185.165.106.250	REDE ENERGETICAS NACIONAIS, S.A.	NA	ISAKMP	NA	REN	REDE ENERGETICAS NACIONAIS, S.A.	74

Figure 64 - Exposed network addresses of REN

- **Administração Central de Sistema de Saúde, I.P**

Figure 65 presents the list of exposed network addresses of the Administração Central de Sistema de Saúde, I.P. The organisation presents one exposed network address, identified by the IP: 193164.0.79, with 3 ports open. Two of the ports are running the protocol HTTPS and HTTP respectively. The port running the protocol HTTPS is filtered and thus it is not possible to determinate the version of the SSL/TLS used. The third port is running the protocol Yaskawa MP Series Ethernet, is the one with the highest level of risk and thus deserving more attention and analysis by the organisation. It was not identified any CVE in the identified network address. The organisation has its own infrastructures to maintain the network routing policies of its systems.

IP	Organization	Name Ssl Version	Protocol	Cve	As Name	Isp	Level Of Risk
193.164.0.79	ADMINISTRACAO CENTRAL DO SISTEMA DE SAUDE, I.P.	NA	YASKAWA MP SERIES ETHERNET	NA	IGIF-AS	ADMINISTRACAO CENTRAL DO SISTEMA DE SAUDE, I.P.	74
193.164.0.79	ADMINISTRACAO CENTRAL DO SISTEMA DE SAUDE, I.P.	NA	HTTP	NA	IGIF-AS	ADMINISTRACAO CENTRAL DO SISTEMA DE SAUDE, I.P.	37
193.164.0.79	ADMINISTRACAO CENTRAL DO SISTEMA DE SAUDE, I.P.	filtered	HTTPS	NA	IGIF-AS	ADMINISTRACAO CENTRAL DO SISTEMA DE SAUDE, I.P.	8

Figure 65 - Exposed network addresses of Administração Central de Sistema de Saude, I.P.

- **Secretaria Geral do Ministério da Administração Interna**

Figure 66 presents the exposed network addresses of Secretaria Geral do Ministério da Administração Interna. It presents one exposed network address identified by the IP: 91.227.100.249, with one open port, running the protocol ISAKMP. The open port presents a very high level of risk and thus should be analysed with more detail by the organisation. It was not identified any CVE on none of the ports.

IP	Organization	Name Ssl Version	Protocol	Cve	As Name	Sum of Level Of Risk
91.227.100.249	SECRETARIA-GERAL MINISTERIO DA ADMINISTRACAO INTERNA	NA	ISAKMP	NA	UTIS-AS	111

Figure 66 - Exposed network address of Secretaria Geral do Ministério da Administração Interna

- **Comissão de Mercado de Valores Mobiliários**

Figure 67 presents the exposed network addresses of Comissão de Mercado de Valores Mobiliários (CMVM). The organisation presents one exposed network address identified by IP: 193.16.103.193, with two ports open. The open ports are running the protocols ISAKMP and HTTPS respectively. The port running the protocol HTTPS is filtered and thus it is not possible to determinate the SSL/TLS version used. The port running the protocol ISAKMP presents the highest level of risk, thus it should be analysed by the organisation with more attention. It was

not identified any CVE on none of the open ports. The CMVM has its own infrastructures to maintain the network routing policies of its systems.

IP	Organization	Name Ssl Version	Protocol	Cve	Isp	As Name	Level Of Risk
193.16.103.193	COMISSAO DO MERCADO DE VALORES MOBILIARIOS	NA	ISAKMP	NA	COMISSAO DO MERCADO DE VALORES MOBILIARIOS	CMVM-PT-AS	74
193.16.103.193	COMISSAO DO MERCADO DE VALORES MOBILIARIOS	filtered	HTTPS	NA	COMISSAO DO MERCADO DE VALORES MOBILIARIOS	CMVM-PT-AS	8

Figure 67 - Exposed network addresses of CMVM

When analysing the organisations presented above, we verified that most of them have their own infrastructures to maintain the network routing policies of their systems, which is good as they are less dependent on external entities. In addition, we also observed that there are many exposed systems running ICSs standard protocols that belong to educational institutions. Figure 68 shows the most exposed educational institutions. The network routing policies of the majority of the systems from the educational institutions are managed by Fundação Para a Ciência e a Tecnologia, I.P. (the Portuguese public agency that supports science, technology and innovation, in all scientific domains, under the responsibility of the Ministry for Science, Technology and Higher Education (FCT, 2020)), with the exception of INESC Porto and INESC Lisboa that has its own infrastructures to maintain the network routing policies of its systems.

Organization	As Name	Isp	Nr of exposed IP
FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	172
UNIVERSIDADE DE LISBOA	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	15
INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES	INESC LISBOA, PORTUGAL	INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES	5
INESC PORTO	INESCTEC INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES	INESC PORTO	3
UNIVERSIDADE CATOLICA PORTUGUESA	UNIVERSIDADE CATOLICA PORTUGUESA	UNIVERSIDADE CATOLICA PORTUGUESA	3
INSTITUTO POLITECNICO DO PORTO	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	2
UNIVERSIDADE DE TRAS OS MONTES E ALTO DOURO	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	2
UNIVERSIDADE DO PORTO	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	2
UNIVERSIDADE DE COIMBRA	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	1
UNIVERSIDADE LUSIADA DO PORTO	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	1
UNIVERSIDADE LUSOFONA	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	1
UNIVERSIDADE NOVA DE LISBOA	RCCN FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	FUNDACAO PARA A CIENCIA E A TECNOLOGIA, I.P.	1

Figure 68 - Most exposed educational institutions

As we can see there are 172 exposed network addresses that belong to Fundação Para a Ciência e a Tecnologia, I.P., which means that there are many institutions that hide behind the ISP Fundação Para a Ciência e a Tecnologia, I.P. Next, we have the Universidade de Lisboa with the second highest number of exposed network addresses.

Figure 69 presents the level of risk of the educational institutions. The Fundação Para a Ciência e a Tecnologia, I.P. appears with the highest level of risk, but as we have seen before, there are a set of educational institutions that hide behind it. Next, we have the Universidade de Lisboa with the second highest level of risk.

Organization	Sum of Level Of Risk
FUNDAÇÃO PARA A CIÊNCIA E A TECNOLOGIA, I.P.	7 382
UNIVERSIDADE DE LISBOA	666
INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES	114
UNIVERSIDADE CATOLICA PORTUGUESA	57
INSTITUTO POLITECNICO DO PORTO	51
INESC PORTO	45
UNIVERSIDADE DO PORTO	37
UNIVERSIDADE DE TRAS OS MONTES E ALTO DOURO	29
UNIVERSIDADE DE COIMBRA	18
UNIVERSIDADE LUSIADA DO PORTO	14
UNIVERSIDADE LUSOFONA	6
UNIVERSIDADE NOVA DE LISBOA	6

Figure 69 - Level of risk of the educational institutions

Conclusions

This Chapter presented the analysis and discussion of the main results.

The analysed data was downloaded from the sources between February 2020 and July 2020. The fact table contains 273803 records, with 167252 different network addresses, running 80 different ports. The network addresses were in 1072 different locations, under the control of 81 different ASs and there are 78 ISPs providing Internet services to them. We identified 57 cipher suites and 361 different CVEs on the network addresses. During the analysis it was possible to infer the following conclusions:

- Most of the systems are running Medium Risk ports. From those running High Risk ports, most of them are running the protocol FTP, which is a concern, since data sent via FTP are vulnerable to several attacks.
- Most of the systems do not have an algorithm of encryption available. From those that have, a huge percentage are in ports that are filtered, and so it is not possible to determinate the SSL/TLS version available, but in the other hand it is advantageous because there is less information about these ports available.
- Most of the encrypted systems are using cipher suites that present High Risk to the system, giving the organisations a false sense of security. The most common cipher suite used is TLS_RSA_WITH_RC4_128_SHA, which presents a High Risk to the system.
- Most of the identified CVEs on the systems presents medium severity for them. The most common CVE identified was CVE-2017-15906. The systems that present this CVE are vulnerable to a denial of service (an interruption in a machine or network, making them inaccessible to the intended users), caused by an error in the function process_open(), when it is on read-only mode.

- Many of the organisations do not have their own infrastructure to maintain the routing policies of their systems. In this situation, it is not possible to identify the organisations because they are hidden behind their ISPs. This can be advantageous in the sense that the attacker cannot identify immediately the organisation to which the IP belongs, but on the other hand, the organisation becomes dependent on the ISP, in such a way that if the ISP suffers an attack and goes offline, all the organisations behind it would be severely affected.
- Lisboa is the city with the most exposed systems, followed by Porto and Quinta do Conde.
- Almost 70% of the ports are running non-standard ICS protocols and approximately 30% are running ICS standard protocols. Which means that, on average, for each 1 ICS standard protocol running on a system there are 2,3 non-standard ICS protocols running on the same system.
- HTTP is the most exposed protocol, running in more than 120000 of the 167252 different systems identified. This means that the majority of the identified systems have a port running a HTTP protocol. This is a concern because HTTP connections are not considered secure. We also identified ICSs with exposed database ports. Such ports typically never have reason to be open, and it should be assumed that authentication can fail, and therefore the exposition of these ports should be avoided. Among them, MySQL is the most exposed database protocol.
- The Universidade de Lisboa is the organisation with the oldest CVE identified. It has CVEs published in December 2003, December 2004, August 2005, and September 2005.
- Among the most exposed critical organisations, Credibom is the organisation that presents the higher level of risk, followed by Aeroportos de Portugal, and TAP. Most of these critical organisations, have their own infrastructure to maintain the routing policies of their systems, which is advantageous as they are less dependent on external entities.
- The network routing policies of the majority of the educational institutions systems, except INESC Porto and INESC Lisboa, are managed by Fundação Para a Ciência e a Tecnologia, I.P., therefore hiding themselves behind the latter. Universidade de Lisboa is the educational institution with the highest number of exposed network addresses and the highest level of risk.

5 CONCLUSION AND FUTURE WORK

The evolution of the Internet in conjunction with technological advances have contributed to the interconnectivity between ICSs and the outside world. Thus, nowadays, the majority of ICSs are connected directly, or indirectly, to the Internet. This connectivity brought many advantages to the industries, but it also increased the level of exposure and vulnerability of the ICSs, placing those infrastructures under the direct visibility of various malicious agents. However, despite all the negative consequences that an attack on ICSs can cause, there is no public work that allows the identification of the exposed ICSs on the Internet in Portugal and their associated cyber security risks. Based on the ideas laid out above, this work had as the main objectives the investigation and identification of ICSs exposed on the Internet in Portugal and their associated risks in terms of security, in order to allow the organisations to know how resilient they are in terms of cyber security, and devote more attention to the most exposed and vulnerable systems.

The literature review presented in Chapter 2 allowed us to identify and understand the main topics addressed to the scope of this work. The study of ICSs enabled us to identify the standard ICSs protocols, which we later used to distinguish ICSs from other systems. It allowed us to identify data sources about ICSs exposed on the Internet in Portugal, preventing us from performing port scan techniques against ICSs, once we noticed that vulnerable systems used to fail when performed port scanning techniques against them. Still, within the literature review, we performed a study on the principles of information security, which helped us in the identification and classification of the risk that each vulnerability presents to the ICSs. The fact that in the literature review we were able to find more than one source of data that can complement each other, we took the decision to build a data warehouse to facilitate the organisation and the analysis of the data and the visualisation of the results. Thus, we also performed a study that allowed us to identify the best data warehouse tools to be used in this work. Finally, we performed a study on some of the existing and most relevant public studies related to the identification of the ICSs and the analysis of their vulnerabilities. This section had strongly influenced the direction of this work, serving as a guide that helped in the decisions to achieve the main objectives of this work.

To achieve the objectives of the work, we have developed a methodology composed of 6 phases that allowed the identification and the analysis of the exposed ICSs in Portugal and their associated cyber security risks. In the first phase of the methodology, we decided to use Shodan and Censys as the source of data of this work. The second phase was to identify the ICSs and distinguish them from other systems - the protocol used by the systems was used as a differential factor. In the third phase of the methodology, we identified the features: Port, Cipher Suite, SSL/TLS version, and CVE, as the ones that have information that may represent vulnerabilities to the systems. In the fourth phase of the methodology, we defined the calculation of the level of risk that the data on the selected feature represents for the systems. The fifth phase proposed the development of the data warehouse to store and organise the data about the ICSs identified and to allow a better analysis of it. Finally, the sixth phase addresses the analysis and discussion of the results.

Within the analysis of the data, we reached the following findings. There are at least 167252 ICSs exposed and easily found on the public Internet in Portugal. The majority of them are located in Lisboa. The organisations to which these systems belong should be aware of them since information on the ICSs should not be exposed on the Internet to be accessed by anyone.

Most of the systems have at least one feature that presents a High Risk to their security. Furthermore, a huge number of the systems do not have an algorithm of encryption available to secure the connection. From those that have, a huge percentage are using algorithms not considered secure, which gives the organisations a false sense of security. The majority of the identified systems (more than 120000 of the 167252 systems found) have a port running an HTTP protocol, even though HTTP connections are not considered secure. From the systems running High Risk ports, most of them are running the protocol FTP, a protocol not built to be secure. These results show that, despite all the security recommendations, a large number of critical organisations have characteristics on their systems that are already known as insecure and that are not recommended. This indicates some negligence on the part of organisations concerning security issues, even though they have systems considered critical.

In average, for each ICS standard protocol running on the identified ICSs, there are 2,3 non-standard ICS protocols running on the same system. The organisations should be aware and monitor these ports because they are often ports that should not be running on ICSs or unnecessary ports that only increase the attack surface of the system.

Many of the organisations to which the systems belong do not have their own infrastructures to maintain the network routing policies. In this situation, it is not possible to identify the organisations because they hide behind their ISPs. This result is similar to the one found in this study (Ceron et al., 2019) made in the Netherlands. This appears to be a commonly used approach and it can be advantageous in the sense that the attacker cannot identify immediately the organisation to which the system belongs. However, in the other hand, the organisation becomes dependent on the ISP, in such a way that if the ISP suffer an attack and goes offline, all the organisations behind it will be severely affected.

The main limitations of this work are related to the used data. This study is limited to only IPv4 addresses. This limitation occurs because there is no open project that provides information on IPv6 addresses space and brute-force scanning of it is not feasible. It is also limited to the use of data from Shodan and Censys, not considering the results of other scanning projects or the results of a direct port scan to the systems. Still within the Shodan data, we only had access to 100 credits per month, which limits the amount of data that we had access to. The limitations presented, may have influenced the number of devices found. Thus, the results should be seen as a lower limit of what actually exists. It is also possible that at this time, the ICSs found are no longer exposed on the Internet in the same way, or do not present the same vulnerabilities. Furthermore, as we do not investigate what for what the systems are being used for, when studied in more detail, some of these systems can be recognized as honeypots.

In this study, we associated a level of risk to the systems, related to the vulnerabilities they present, with the aim to give a sense of their current state of security. It is important to note that this level of risk, serves to alert organisations about systems that may be vulnerable. It should

be used in the process of filtering and prioritizing systems with High Risk features to be analysed with more detail. However, this does not mean that only High Risk vulnerabilities should be mitigated. The level of risk should not be automatically interpreted as being synonymous with the insecurity of a system but should be seen as a warning factor to provide additional attention to some systems. In this sense, we cannot compare and/or state that one system is *de facto* more insecure than another, considering only the level of risk. It is also important to note that, whether a system is proven to be insecure or not, just because we can have information about it on the Internet, this exposure makes it more vulnerable than a system that it is not possible to have any information about. The former is more likely to be exploited. Thus, all potential risk factors must be analysed, however, some organisations may choose to analyse only High Risk factors first.

In general, considering all the conditions, all the objectives of this internship have been achieved. The trainee has acquired strong foundations which will contribute to her professional growth in the future, provided by the knowledge acquired in a professional environment and working with real data.

As future work, it would be interesting to perform port scans techniques using Nmap, through some of the systems found, with the authorisations of the affected organisations, to compare if the results would be similar. It would also be interesting to investigate other port scanning projects, in order to compare the outcome of the results obtained with Shodan and Censys. Still, as future work, it would be interesting the development of a data warehouse with a different structure in order to evaluate the performance of the searches.

REFERENCES

- Adamson, C. (2011). Three ways to drill across. Retrieved November 29, 2020, from <http://blog.chrisadamson.com/2011/12/three-ways-to-drill-across.html>
- Ahmed, I. (2020). ETL: What It Means and Why Is It Important? Retrieved September 4, 2020, from <https://www.astera.com/type/blog/etl-what-it-means-and-why-is-it-important/>
- Alley, G. (2018). Open Source ETL Tools Comparison. Retrieved July 8, 2020, from <https://www.alooma.com/blog/open-source-etl-tools-comparison>
- Almeida, F. (2017). Concepts and Fundamentals of Data Warehousing and OLAP. *INESC TEC and University of Porto, 1*(September), 1–40.
- Andreeva, O., Gordeychik, S., Gritsai, G., & Kochetova, O. (2016). *Industrial Control Systems and Their Online Availability*. 1–16.
- Atighetchi, M., Simidchieva, B., Soule, N., Yaman, F., Loyall, J., Last, D., ... Flatley, C. B. (2015). *Automatic Quantification and Minimization of Attack Surfaces*.
- Automated Results. (2019). Data Historian Overview. Retrieved December 11, 2020, from <http://www.automatedresults.com/PI/data-historian-overview.aspx>
- Barracuda Networks, I. (2020). TLS with Insecure Ciphers and SSLv2/SSLv3 No Longer Supported. Retrieved November 2, 2020, from <https://campus.barracuda.com/product/essentials/doc/91981918/tls-with-insecure-ciphers-and-sslv2-sslv3-no-longer-supported/>
- Beaver, K. (2020). Commonly Hacked Ports. Retrieved October 27, 2020, from <https://www.dummies.com/programming/networking/commonly-hacked-ports/>
- Belmonte, F., Boulanger, J. L., Schön, W., & Berkani, K. (2006). Role of supervision systems in railway safety. *WIT Transactions on the Built Environment*, 88(June), 129–138. <https://doi.org/10.2495/CR060131>
- Besen, S. M., & Israel, M. A. (2013). The evolution of internet interconnection from hierarchy to “mesh”: Implications for government regulation. *Information Economics and Policy*, 25(4), 235–245. <https://doi.org/10.1016/j.infoecopol.2013.07.003>
- Beyondsecurity. (2020). Finding and Fixing SSL RC4 Cipher Suites Supported Vulnerability. Retrieved May 6, 2020, from <https://beyondsecurity.com/scan-pentest-network-vulnerabilities-ssl-rc4-cipher-suites-supported.html?cn-reloaded=1&cn-reloaded=1>
- Binaryedge. (2017). How does BinaryEdge compute the score of an IP address? Retrieved May 5, 2020, from <https://github.com/balغان/ratemyip-openframework/blob/master/ip-score.md>
- Bocetta, S. (2019). Comparing 3 open source databases: PostgreSQL, MariaDB, and SQLite. Retrieved July 8, 2020, from <https://opensource.com/article/19/1/open-source-databases>
- Bodenheim, R., Butts, J., Dunlap, S., & Mullins, B. (2014). Evaluation of the ability of the Shodan search engine to identify Internet-facing industrial control devices. *International Journal of Critical Infrastructure Protection*, 7(2), 114–123. <https://doi.org/10.1016/j.ijcip.2014.03.001>
- Bruce, D. (2020). Pentaho vs. Talend: How the Two Data Integration Tools Compare?

Retrieved August 31, 2020, from <https://www.knowledgenile.com/blogs/pentaho-vs-talend/#DataIntegration>

Carnegie Mellon University. (2015). *Threats and Risk Calculation*. 1–22.

Censys. (2019a). Report Builder. Retrieved December 11, 2020, from https://censys.io/ipv4/report?q=*&field=tags.raw&max_buckets=1000

Censys. (2019b). See Your Entire Attack Surface in Real Time. Retrieved December 11, 2020, from <https://censys.io/>

Ceron, J., Chromik, J., Santanna, J., & Pras, A. (2019). *Online Discoverability and Vulnerabilities of ICS / SCADA Devices in the Netherlands: University Twente*.

Chand, S. (2020). What Is Talend? – An Unified Platform For Data Integration. Retrieved July 8, 2020, from <https://www.edureka.co/blog/what-is-talend-tool/#WhatIsTalend>

Chapin, L., & Owens, C. (2005). Interconnection and peering among Internet service providers. *An Interisle White Paper*. Retrieved from <http://www.interisle.net/sub/ISPInterconnection.pdf>

Chokalingam, A. (2019). SCADA Network Security Monitoring. Retrieved December 11, 2020, from <https://logrhythm.com/blog/scada-network-security-monitoring/>

Choudhary, V. (2019). Censys - Find and analyze any server and device on the Internet. Retrieved December 11, 2020, from <https://developerinsider.co/censys-find-and-analyze-any-server-and-device-on-the-internet/>

CipherSuite. (2019). 337 Cipher Suites. Retrieved May 6, 2020, from <https://ciphersuite.info/cs/?singlepage=true&page=8&software=openssl>

Clarke, G., Reynders, D., & Wright, E. (2004). *Practical Modern SCADA Protocols: DNP3, 60870.5 and Related Systems* (V. Mehra, Ed.). Mumbai, India. <https://doi.org/https://doi.org/10.1016/B978-0-7506-5799-0.X5015-3>

Cloudflare. (2020). What is an autonomous system? | What are ASNs? Retrieved July 13, 2020, from <https://www.cloudflare.com/learning/network-layer/what-is-an-autonomous-system/>

Collantes, M. H., & Padilla, A. L. (2015). *Protocols and network security in ICS infrastructures*.

Comodo Group, I. (2020). Deprecation of TLS 1.0. Retrieved May 6, 2020, from <https://www.comodo.com/e-commerce/ssl-certificates/tls-1-deprecation.php>

Congdon, L. (2015). 8 advantages of using open source in the enterprise. Retrieved July 8, 2020, from <https://enterpriseproject.com/article/2015/1/top-advantages-open-source-offers-over-proprietary-solutions%0D>

Coole, M., Corkill, J., & Woodward, A. (2012). Defence in Depth , Protection in Depth and Security in Depth : a Comparative Analysis Towards a Common Usage Language. *Australian Security and Intelligence Conference*, 1(1), 21–23. <https://doi.org/10.4225/75/57a034ccac5cd>

Coolfire. (2019). What Is The Difference Between IT And OT? Retrieved December 11, 2020, from <https://www.coolfiresolutions.com/blog/difference-between-it-ot/>

Csanyi, E. (2019). IED (Intelligent Electronic Device) advanced functions that make our life

- better. Retrieved December 11, 2020, from <https://electrical-engineering-portal.com/ied-intelligent-electronic-device-advanced-functions>
- Datapath.io. (2016). Understanding Last Mile Internet Access. Retrieved November 2, 2020, from https://medium.com/@datapath_io/understanding-last-mile-internet-access-a62ee96c0a00
- Department of Homeland Security. (2020). Department of Homeland Security. Retrieved July 8, 2020, from <https://www.dhs.gov/%0D>
- Digicert. (2020a). Behind the Scenes of SSL Cryptography. Retrieved July 8, 2020, from <https://www.digicert.com/ssl-cryptography.htm>
- Digicert. (2020b). What is SSL, TLS and HTTPS? Retrieved July 8, 2020, from <https://www.websecurity.digicert.com/security-topics/what-is-ssl-tls-https%0D>
- Dognaedis. (2015). Software Security Democratization. Retrieved December 11, 2020, from <https://codev.dognaedis.com/>
- Dognaedis. (2020). DOGNÆDIS, A Prosegur Company. Retrieved April 19, 2020, from <https://www.dognaedis.com/>
- Dowd, M. (2020). OPENSSSH 6.4 J-PAKE PROTOCOL SCHNORR.C HASH_BUFFER DENIAL OF SERVICE. Retrieved August 5, 2020, from <https://vuldb.com/?id.12124>
- Doyle, S. (2018). TCP vs. UDP: Understanding the Difference. Retrieved July 8, 2020, from <https://www.privateinternetaccess.com/blog/tcp-vs-udp-understanding-the-difference/>
- Durumeric, Z., Wustrow, E., & Halderman, J. A. (2013). ZMap: Fast Internet-wide Scanning and Its Security Applications. *Proceedings of the 22nd USENIX Security Symposium*, (August), 605–619. Retrieved from <https://zmap.io/paper.pdf>
- Educba. (2020). Talend vs Pentaho. Retrieved August 31, 2020, from <https://www.educba.com/talend-vs-pentaho/>
- FCT. (2020). About FCT. Retrieved July 29, 2020, from <https://www.fct.pt/fct/fct.phtml.en>
- FIRST. (2020). Common Vulnerability Scoring System SIG. Retrieved December 8, 2020, from <https://www.first.org/cvss/>
- Folly, A. (2018). Control Your ETL. Retrieved August 25, 2020, from <https://andrewsfolly.com/2018/08/08/control-your-etl/>
- Freund, J. (2015). *Measuring and Managing Information Risk: A FAIR Approach* (1st Editio). Butterworth-Heinemann. <https://doi.org/https://doi.org/10.1016/C2013-0-09966-5>
- Gantz, S. D., & Philpott, D. R. (2013). Thinking About Risk. In *FISMA and the Risk Management Framework* (pp. 53–78). <https://doi.org/10.1016/b978-1-59-749641-4.00003-5>
- GeeksforGeeks. (2020a). Fact Constellation in Data Warehouse modelling.
- GeeksforGeeks. (2020b). Internet Service Provider (ISP) hierarchy. Retrieved November 2, 2020, from <https://www.geeksforgeeks.org/internet-service-provider-isp-hierarchy/>
- Gegick, M., & Barnum, S. (2013). Least Privilege. Retrieved December 11, 2020, from <https://us-cert.cisa.gov/bsi/articles/knowledge/principles/least-privilege>
- GlobalSign. (2020). It's Time to Disable TLS 1.0 (and All SSL Versions) If You Haven't

- Already. Retrieved May 6, 2020, from <https://www.globalsign.com/en/blog/disable-tls-10-and-all-ssl-versions%0D>
- Gosain, A., & Singh, J. (2020). Comprehensive complexity metric for data warehouse multidimensional model understandability. *IET Software*, 14(3), 275–282. <https://doi.org/10.1049/iet-sen.2019.0150>
- Guru99. (2020a). Star and SnowFlake Schema in Data Warehouse. Retrieved July 8, 2020, from Star and SnowFlake Schema in Data Warehouse
- Guru99. (2020b). What is Talend? Retrieved July 8, 2020, from <https://www.guru99.com/talend-tutorial.html#6>
- Hebert, T., & GlobalSign. (2020). It's Time to Disable TLS 1.0 (and All SSL Versions) If You Haven't Already. Retrieved May 6, 2020, from <https://www.globalsign.com/en/blog/disable-tls-10-and-all-ssl-versions>
- Hitachi Vantara LLC. (2020a). Easy, Flexible Consumption. Retrieved July 8, 2020, from <https://www.hitachivantara.com/en-us/home.html>
- Hitachi Vantara LLC. (2020b). Pentaho Data Integration. Retrieved July 8, 2020, from https://help.pentaho.com/Documentation/9.0/Products/Pentaho_Data_Integration
- Hudak, P. (2018). Censys.io Guide: Discover SCADA and Phishing Sites. Retrieved November 25, 2019, from <https://0xpatrik.com/censys-guide/>
- Hytrust Cloud Under Control. (2015). *Top 10 Reasons You Need Encryption*.
- IAN. (2016). What does ACID mean in Database Systems? Retrieved July 13, 2020, from <https://database.guide/what-is-acid-in-databases/>
- IBM. (2018). Security Bulletin: OpenSSH vulnerability affects IBM Spectrum Protect Plus (CVE-2017-15906). Retrieved August 5, 2020, from <https://www.ibm.com/support/pages/security-bulletin-openssh-vulnerability-affects-ibm-spectrum-protect-plus-cve-2017-15906>
- IBM Cloud Education. (2020). ETL (Extract, Transform, Load) Analytics Integration. Retrieved September 4, 2020, from <https://www.ibm.com/cloud/learn/etl>
- IETF. (2020). The Internet Engineering Task Force (IETF). Retrieved April 19, 2020, from <https://ietf.org/about/>
- Inst Tools. (2019). What is Foundation Fieldbus (FF)? Retrieved December 11, 2020, from <https://instrumentationtools.com/foundation-fieldbus/#.WuIC8C5ubIU>
- Instituto Pedro Nunes. (2020). Instituto Pedro Nunes. Retrieved July 13, 2020, from <https://www.ipn.pt/ipn>
- Isaca journal archives. (2014). An Enhanced Risk Formula for Software Security Vulnerabilities. Retrieved November 7, 2020, from <https://www.isaca.org/resources/isaca-journal/past-issues/2014/an-enhanced-risk-formula-for-software-security-vulnerabilities>
- ISEC. (2018). ISEC. Retrieved June 4, 2018, from <https://www.isec.pt>
- Jiang, R. (2013). *Robust Group Key Management with Revocation and Collusion Resistance for SCADA in Smart Grid*. Retrieved from <https://www.slideserve.com/cosima/robust-group-key-management-with-revocation-and-collusion-resistance-for-scada-in-smart->

grid

- Keung, Y. A. U. H. (2013). Basic Principle of Information Security. *Advances in Robotics & Automation*, 03(03), 9695. <https://doi.org/10.4172/2168-9695.1000e120>
- Kimball, R. (2003). The Soul of the Data Warehouse, Part One: Drilling Down. *Intelligent Enterprise*, 6(5), 16-47. 3p. Retrieved from kimballgroup.com/2003/04/the-soul-of-the-data-warehouse-part-two-drilling-across/
- King, T. (2019). Top 12 Free and Open Source ETL Tools for Data Integration. Retrieved August 7, 2020, from <https://solutionsreview.com/data-integration/top-free-and-open-source-etl-tools-for-data-integration/>
- Lantronix. (2009). *Encryption and Its Importance to Device Networking* (Vol. 1, pp. 1–13). Vol. 1, pp. 1–13. Irvine, CA.
- Lee, A. (2008). Defence in depth. *Engineer*, 293(7687), 28–29.
- Lyon, G. F. (2008). *Nmap network scanning : official Nmap project guide to network discovery and security scanning*. 434. Retrieved from <https://nmap.org/book/>
- M.Talabis, M. R., & Martin, J. L. (2013). *Information Security Risk Assessment Toolkit - Practical Assessments through Data Collection and Data Analysis*. 225 Wyman Street, Waltham, MA 02451, USA.
- Majeed, M., & Quadri, S. M. K. (2016). *Analysis and Evaluation of “ Reducing the Attack Surfaces ” to improve the security of the software at Design Level*. 7(3), 1471–1475.
- Malinowski, E., & Zimányi, E. (2008). Advanced Data Warehouse Design From Conventional to Spatial and Temporal Applications. *Data Vault 2.0*, 1–15. <https://doi.org/10.1016/b978-0-12-802510-9.00001-5>
- Margea, R., & Margea, C. (2011). Open source approach to project management tools. *Informatica Economică*, 15(1), 196–206.
- Mathew, K., Tabassum, M., & Lu Ai Siok, M. V. (2014). A study of open ports as security vulnerabilities in common user computers. *2014 International Conference on Computational Science and Technology, ICCST 2014*, (August). <https://doi.org/10.1109/ICCST.2014.7045193>
- Matthews, K. (2019). 7 Advantages of Using Encryption Technology for Data Protection. Retrieved November 9, 2020, from <https://www.smartdatacollective.com/5-advantages-using-encryption-technology-data-protection/>
- Matthews, W. (2017). New roles for process historians. Retrieved December 11, 2020, from <https://www.isa.org/intech/20171202/>
- McKay, K. A., & Cooper, D. A. (2019). NIST Special Publication 800-52 Revision 2 - Guidelines for the selection, configuration, and use of Transport Layer Security (TLS) implementations. *NIST Special Publication*, (August). <https://doi.org/10.6028/NIST.SP.800-52r2>
- Merkow, M. S., & Breithaupt, J. (2014). Information Security: Principles and Practices Second Edition Warning and Disclaimer. In *Library of Congress Control Number*.
- Metabase. (2020). Metabase Documentation. Retrieved July 22, 2020, from <https://www.metabase.com/docs/latest/users-guide/01-what-is-metabase.html>

- Microsoft. (2018). Cipher Suites in TLS/SSL (Schannel SSP). Retrieved July 8, 2020, from <https://docs.microsoft.com/en-us/windows/win32/secauthn/cipher-suites-in-schannel>
- Miller, S. R. (2019). Top 10 Free and Open-Source Database Management Software Solutions. Retrieved July 8, 2020, from <https://www.goodfirms.co/blog/top-10-free-and-open-source-database-management-software-solutions%0D>
- Mirian, A., Ma, Z., Adrian, D., Tischer, M., Chuenchujit, T., Yardley, T., ... Bailey, M. (2016). An Internet-wide view of ICS devices. *2016 14th Annual Conference on Privacy, Security and Trust, PST 2016*, 96–103. <https://doi.org/10.1109/PST.2016.7906943>
- MITRE Corporation. (2020a). Common Vulnerabilities and Exposures. Retrieved July 8, 2020, from <https://cve.mitre.org/%0D>
- MITRE Corporation. (2020b). Current CVSS Score Distribution For All Vulnerabilities. Retrieved July 8, 2020, from <https://www.cvedetails.com/%0D>
- MITRE Corporation. (2020c). CVE Details. Retrieved April 19, 2020, from <https://www.cvedetails.com/>
- Mocan, J. (2018). Why is the term “SSL” still being used when it has been replaced by TLS for some time now? Retrieved July 8, 2020, from <https://www.quora.com/Why-is-the-term-SSL-still-being-used-when-it-has-been-replaced-by-TLS-for-some-time-now>
- Morris, T., & Pavurapu, K. (2010). A retrofit network transaction data logger and intrusion detection system for transmission and distribution substations. *PECon2010 - 2010 IEEE International Conference on Power and Energy*, (April), 958–963. <https://doi.org/10.1109/PECON.2010.5697717>
- MuleSoft LLC. (2020). What is an API? (Application Programming Interface). Retrieved July 13, 2020, from <https://www.mulesoft.com/resources/api/what-is-an-api>
- Muncaster, P. (2019). Most Port Vulnerabilities Are Found in Three Ports. Retrieved October 27, 2020, from <https://www.infosecurity-magazine.com/news/most-port-vulnerabilities-are/>
- Muscat, I. (2019). Recommendations for TLS/SSL Cipher Hardening. Retrieved November 3, 2020, from <https://www.acunetix.com/blog/articles/tls-ssl-cipher-hardening/>
- NIST. (2020a). National Institute of Standards and Technology. Retrieved July 13, 2020, from <https://www.nist.gov/>
- NIST. (2020b). National Vulnerability Database. Retrieved July 8, 2020, from <https://nvd.nist.gov/%0D>
- Nmap. (2019a). Introduction to Port Scanning. Retrieved December 11, 2020, from <https://nmap.org/book/port-scanning.html>
- Nmap. (2019b). Windows: Obtaining, Compiling, Installing, and Removing Nmap. Retrieved December 11, 2020, from <https://nmap.org/book/inst-windows.html>
- Nohe, P. (2019). SSL/TLS Cipher suites. Retrieved July 7, 2019, from <https://www.thesslstore.com/blog/cipher-suites-algorithms-security-settings/>
- Ocunerix. (2020). TLS Security 6: Examples of TLS Vulnerabilities and Attacks. Retrieved May 6, 2020, from <https://www.acunetix.com/blog/articles/tls-vulnerabilities-attacks-final-part/%0D>

- Oliveira, A., & Bernardino, J. (2011). *Evaluating Self-Service BI and Analytics Tools for SMEs*.
- Oracle Corporation. (2020a). MySQL Community Edition. Retrieved July 8, 2020, from <https://www.mysql.com/products/community/>
- Oracle Corporation. (2020b). The Main Features of MySQL. Retrieved July 8, 2020, from https://docs.oracle.com/cd/E17952_01/mysql-5.6-en/features.html
- Oracle Corporation. (2020c). What is MySQL? Retrieved July 8, 2020, from <https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html%0D>
- Outspoken Media Inc. (2020). Cipher Suites. Retrieved July 8, 2020, from <https://outspokenmedia.com/https/cipher-suites/>
- Packetlabs. (2020). TLS 1.0 and TLS 1.1 Are No Longer Secure. Retrieved November 2, 2020, from <https://www.packetlabs.net/tls-1-1-no-longer-secure/>
- Partridge, C., & Allman, M. (2016). Ethical considerations in network measurement papers. *Communications of the ACM*, 59(10), 58–64. <https://doi.org/10.1145/2896816>
- Pascucci, M. (2014). Let's Put Down Insecure Protocols For Good. Retrieved October 27, 2020, from <https://www.algosec.com/blog/lets-put-insecure-protocols-good/>
- Patel, A. (2013). *Perfect Forward Secrecy (PFS)*. Retrieved from <http://theconversation.com/explainer-what-is-perfect-forward-secrecy-20863>
- ProFTPD. (2019). Bug 4372 - SITE CPCR/CPTO do not honor <Limit> configurations. Retrieved December 13, 2020, from http://bugs.proftpd.org/show_bug.cgi?id=4372
- RealPars. (2019). WHAT IS RTU? Retrieved December 11, 2020, from <https://realpars.com/rtu/>
- Red Hat. (2020a). What is an API? Retrieved July 13, 2020, from <https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces>
- Red Hat, I. (2020b). O que é CVE? Retrieved July 8, 2020, from <https://www.redhat.com/pt-br/topics/security/what-is-cve%0D>
- Rountree, D. (2011). *Introduction to General Security Concepts* (D. Rountree, Ed.). <https://doi.org/https://doi.org/10.1016/B978-1-59749-594-3.00001-6>.
- Rouse, M. (2007). cipher. Retrieved July 7, 2020, from <https://searchsecurity.techtarget.com/definition/cipher%0D>
- Rudolph, C., & Grundmann, N. (2019). 372 Ciphers. Retrieved September 2, 2020, from <https://ciphersuite.info/cs/?singlepage=true>
- Samtani, S., Yu, S., Zhu, H., Patton, M., Matherly, J., & Chen, H. (2018). Identifying SCADA systems and their vulnerabilities on the internet of things: A text-mining approach. *IEEE Intelligent Systems*, 33(2), 63–73. <https://doi.org/10.1109/MIS.2018.111145022>
- SANS Institute. (2015). *Firewall Checklist*. (Security 401).
- Santos, B., Serio, F., Abrantes, S., Sa, F., Loureiro, J., Wanzeler, C., & Martins, P. (2019). Open source business intelligence tools: Metabase and redash. *IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 1(Ic3k), 467–474. <https://doi.org/10.5220/0008351704670474>

- Schneider, F. B. (2003). Least privilege and more. *IEEE Security and Privacy*, 1(5), 55–59. <https://doi.org/10.1109/MSECP.2003.1236236>
- SecurityTrails. (2018). Top 5 Best Port Scanners. Retrieved December 11, 2020, from <https://securitytrails.com/blog/best-port-scanners>
- Serbanescu, A. V., Obermeier, S., & Yu, D.-Y. (2015). *ICS Threat Analysis Using a Large-Scale Honeynet*. 20–30. <https://doi.org/10.14236/ewic/ics2015.3>
- Shodan. (2019). The search engine for the Web. Retrieved April 19, 2020, from <https://www.shodan.io/>
- Slant. (2020). What are the best open-source Relational Databases? Retrieved July 8, 2020, from <https://www.slant.co/topics/1273/~best-open-source-relational-databases>
- Speedguide. (2020). What are ports and protocols? Retrieved July 8, 2020, from <https://www.speedguide.net/faq/what-are-ports-and-protocols-75>
- Steinegger, R., Schäfer, J., Vogler, M., & Abeck, S. (2014). Attack surface reduction for web services based on authorization patterns. *SECURWARE 2014 - 8th International Conference on Emerging Security Information, Systems and Technologies*, (January), 194–201.
- Stouffer, K., Pillitteri, V., Lightman, S., Abrams, M., & Hahn, A. (2015). Guide to Industrial Control Systems (ICS) Security NIST Special Publication 800-82 Revision 2. *NIST Special Publication 800-82 Rev 2*, 1–157. <https://doi.org/http://dx.doi.org/10.6028/NIST.SP.800-82r1>
- Sultan, K., En-Nouaary, A., & Hamou-Lhadj, A. (2008). Catalog of metrics for assessing security risks of software throughout the software development life cycle. *Proceedings of the 2nd International Conference on Information Security and Assurance, ISA 2008*, (October 2014), 461–465. <https://doi.org/10.1109/ISA.2008.104>
- Talend. (2020). Clean, compliant, and accessible data for everyone. Retrieved July 8, 2020, from <https://www.talend.com/>
- Tamang, P. (2019). 6 Best Free and Open Source Database Software Options. Retrieved July 8, 2020, from <https://blog.capterra.com/free-database-software/>
- Techroba. (2015). 10 Open Source ETL Tools. Retrieved July 8, 2020, from <https://www.datasciencecentral.com/profiles/blogs/10-open-source-etl-tools%0D>
- TechTerms. (2020). SSL. Retrieved July 8, 2020, from <https://techterms.com/definition/ssl>
- Teja, D. (2011). Master Terminal Units (MTU) in SCADA systems. Retrieved December 11, 2020, from <http://electricalquestionsguide.blogspot.com/2011/06/master-terminal-units-mtu-in-scada.html>
- The PostgreSQL Global Development Group. (2020). PostgreSQL: The World's Most Advanced Open Source Relational Database. Retrieved July 8, 2020, from <https://www.postgresql.org/>
- Thousandeyes. (2020). What are Autonomous System Numbers (ASN)? Retrieved September 1, 2020, from <https://www.thousandeyes.com/learning/glossary/as-autonomous-system>
- Trend Micro. (2019). Industrial Control System. Retrieved November 12, 2019, from https://www.trendmicro.com/vinfo/us/security/definition/industrial-control-system#What_is_an_ICS_System

- Trujillo, J., & Luján-Mora, S. (2003). A UML based approach for modeling ETL processes in data warehouses. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2813, 307–320. https://doi.org/10.1007/978-3-540-39648-2_25
- TrustRadius. (2020). Open-Source Database Software. Retrieved July 8, 2020, from <https://www.trustradius.com/open-source-database>
- Tsang, R. (2009). Cyberthreats , Vulnerabilities and Attacks on SCADA Networks. *Control*, 1–23. Retrieved from <http://www.mendeley.com/research/cyberthreats-vulnerabilities-attacks-scada-networks/>
- Tutorialspoint. (2019). DWH data warehousing. *Tutorials Point (I) Pvt. Ltd.*, 1–13.
- Tutorialspoint. (2020). DBMS - Transaction. Retrieved July 13, 2020, from https://www.tutorialspoint.com/dbms/dbms_transaction.htm
- Unitronics. (2019). What is the definition of “PLC”? Retrieved December 11, 2020, from <https://unitronicsplc.com/what-is-plc-programmable-logic-controller/>
- Vavra, J., & Hromada, M. (2016). Possibilities of the Search Engine Shodan in Relation to SCADA. *SECURWARE 2016 : The Tenth International Conference on Emerging Security Information, Systems and Technologies*, (c), 130–135. Retrieved from <https://pdfs.semanticscholar.org/2e3a/f284355ae9922d3c3b405d722ed3aff89da0.pdf>
- Villanueva, J. C. (2016). Port Confusion - Is Security Through Obscurity Bad? Retrieved November 7, 2020, from <https://www.jscape.com/blog/using-nonstandard-ports-is-security-through-obscurity-really-bad>
- Wang, Y., Qiao, P., Chen, H., Luo, Z., & Sun, G. (2019). The reliability assessment of ics based on evidential reasoning and semi-quantitative information. *Ingenierie Des Systemes d'Information*, 24(2), 147–154. <https://doi.org/10.18280/isi.240203>
- Webranded. (2019). Information Technologies (IT) Vs Operational Technologies (OT). Retrieved December 11, 2020, from <https://randed.com/information-technologies-it-vs-operational-technologies-ot/?lang=en>
- WhatIs MyIPAddress. (2018). What is a port. Retrieved July 8, 2020, from <https://whatismyipaddress.com/port%0D>
- WhatIsMyIPAddress. (2019). Facts About Port Scanning. Retrieved December 11, 2020, from <https://whatismyipaddress.com/port-scan>
- White, J. M. (2014). *Security Risk Assessment: Managing Physical and Operational Security* (1st editio). Butterworth-Heinemann.
- Wojciech. (2019). Ultimate Internet of Things/Industrial Control Systems reconnaissance tool. Retrieved May 5, 2020, from <https://github.com/woj-ciech/Kamerka-GUI>
- Zmap. (2019). <https://zmap.io/>. Retrieved December 11, 2020, from <https://zmap.io/>

APPENDIX

Appendix A - Industrial Control Systems Standard Protocols and their port numbers

Protocol	Port number
ANSI C12.22	1153
BACNet	47808
Beckhoff-ADS communication	48898
CANopen	7234
CodeSys	2455
Crimson 3	789
DNP3	20000
Danfoss ECL apex	5050
EtherCAT	34980
EtherNet/IP	44818
EtherNet/IP	2222
FATEK FB Series	500
GE-SRTP	18245
GE-SRTP	18246
HART-IP	5094
HITACHI EHV Series	3004
ICCP	102
IEC 60870-5-104	2404
IEC 61850 / MMS	102
KEYENCE KV-5000	8501
KOYO Ethernet	28784
LS Fenet	2005
LS Fenet	2004
MELSEC Q	5006
MELSEC Q	5007
Modbus/TCP	502
Moxa	4800
Niagara Tridium Fox	1911
Niagara Tridium Fox	4911
OMRON FINS	9600
OPC	135
Panasonic FP (Ethernet)	9094
Panasonic FP2 (Ethernet)	8500
ProConOS	20547
Quick Panel GE	57176
SAIA S-BUS (Ethernet)	5050
Schleicher XCX 300	20547
Siemens S7	102
Simatic	161
Unitronics Socket1	20256
Unitronics Socket1	20257
YASKAWA MP Series Ethernet	10000
YASKAWA MP2300Siec	44818
Yokogawa FA-M3 (Ethernet)	12289

Appendix B - Assignment of the classes to the exposed ports

Port number	Protocol	Class
21	FTP	Totally Insecure
22	SSH	Insecure
23	TELNET	Totally Insecure
25	SMTTP	Totally Insecure
53	DNS	Partially Secure
80	HTTP	Insecure
81	HTTP	Insecure
102	ICCP	Insecure
102	IEC 61850 / MMS	Insecure
102	SIEMENS S7	Insecure
110	POP3	Totally Insecure
135	OPC	Totally Insecure
137	Windows NetBIOS	Totally Insecure
138	Windows NetBIOS	Totally Insecure
139	Windows NetBIOS	Totally Insecure
143	IMAP	Insecure
161	SIMATIC	Totally Insecure
161	SNMP	Totally Insecure
443	HTTPS	Partially Secure
445	SMB	Totally Insecure
465	SMTTP	Insecure
500	FATEK FB SERIES	Insecure
502	MODBUS	Insecure
587	SMTTP	Insecure
623	IPMI	secure
631	IPP	Partially Secure
789	CRIMSON 3	Insecure
993	IMAPS	Partially Secure
995	POP3S	Partially Secure
1153	ANSI	Insecure
1433	MSSQL	Insecure
1434	MSSQL	Insecure
1521	ORACLE	Insecure
1883	MQTT	Insecure
1911	FOX	Insecure
1911	NIAGARA TRIDIUM FOX	Insecure
1962	PCWORX	Insecure
2004	LS Fenet	Insecure
2005	LS Fenet	Insecure
2222	ETHERNET/IP	Insecure
2323	TELNET	Totally Insecure
2404	IEC 60870-5-104	Insecure
2455	CODESYS	Insecure
3004	HITACHI EHV Series	Insecure
3306	MYSQL	Insecure
3389	RDP	Totally Insecure
4800	MOXA	Insecure
4911	NIAGARA TRIDIUM FOX	Insecure
5006	MELSEC Q	Insecure
5007	MELSEC Q	Insecure
5050	Danfoss ECL apex	Insecure
5050	SAIA S-BUS (ETHERNET)	Insecure
5094	HART-IP	Insecure
5432	POSTGRES	Insecure
5632	PCA	Partially Secure

Port number	Protocol	Class
5672	AMQP	Insecure
5900	VNC	Insecure
5901	VNC	Insecure
5902	VNC	Insecure
5903	VNC	Insecure
6379	REDIS	Insecure
6443	KUBERNETES	Partially Secure
7234	CANopen	Insecure
7547	CWMP	Insecure
8500	PANASONIC FP2 (ETHERNET)	Insecure
8501	KEYENCE KV-5000	Insecure
8883	MQTT	Partially Secure
9090	PROMETHEUS	Insecure
9094	PANASONIC FP (ETHERNET)	Insecure
9200	ELASTICSEARCH	Insecure
9600	OMRON FINS	Insecure
10000	YASKAWA MP SERIES ETHERNET	Insecure
10001	SCP CONFIG	Insecure
11211	MEMCACHED	Insecure
12289	Yokogawa FA-M3 (Ethernet)	Insecure
16993	HTTPS	Partially Secure
18245	GE-SRTP	Insecure
18246	GE-SRTP	Insecure
20000	DNP3	Insecure
20256	UNITRONICS SOCKET1	Insecure
20257	UNITRONICS SOCKET1	Insecure
20547	SCHLEICHER XCX 300	Insecure
27017	MONGODB	Insecure
28784	KOYO Ethernet	Insecure
30718	LANTRONIX DISCOVERY	Insecure
34980	EtherCAT	Insecure
44818	ETHERNET/IP	Insecure
44818	YASKAWA MP2300SIEC	Insecure
47808	BACNET	Insecure
48898	Beckhoff-ADS communication	Insecure
57176	Quick Panel GE	Insecure

Appendix C - Assignment of the classes to the cipher suites


Cipher Suite	Class
ADH-AES128-GCM-SHA256	Totally Insecure
ADH-AES128-SHA	Totally Insecure
ADH-AES128-SHA256	Totally Insecure
ADH-AES256-GCM-SHA384	Totally Insecure
ADH-AES256-SHA	Totally Insecure
ADH-AES256-SHA256	Totally Insecure
ADH-CAMELLIA128-SHA	Totally Insecure
ADH-CAMELLIA128-SHA256	Totally Insecure
ADH-CAMELLIA256-SHA	Totally Insecure
ADH-CAMELLIA256-SHA256	Totally Insecure
ADH-DES-CBC3-SHA	Totally Insecure
ADH-SEED-SHA	Totally Insecure
AECDH-AES128-SHA	Totally Insecure
AECDH-AES256-SHA	Totally Insecure
AECDH-DES-CBC3-SHA	Totally Insecure
AECDH-NULL-SHA	Totally Insecure
AES128-CCM	Partially Secure
AES128-CCM8	Partially Secure
AES128-GCM-SHA256	Partially Secure
AES128-SHA	Partially Secure
AES128-SHA256	Partially Secure
AES256-CCM	Partially Secure
AES256-CCM8	Partially Secure
AES256-GCM-SHA384	Partially Secure
AES256-SHA	Partially Secure
AES256-SHA256	Partially Secure
CAMELLIA128-SHA	Partially Secure
CAMELLIA128-SHA256	Partially Secure
CAMELLIA256-SHA	Partially Secure
CAMELLIA256-SHA256	Partially Secure
DES-CBC3-SHA	Insecure
DHE-DSS-AES128-GCM-SHA256	Secure
DHE-DSS-AES128-SHA	Partially Secure
DHE-DSS-AES128-SHA256	Partially Secure
DHE-DSS-AES256-GCM-SHA384	Secure
DHE-DSS-AES256-SHA	Partially Secure
DHE-DSS-AES256-SHA256	Partially Secure
DHE-DSS-CAMELLIA128-SHA	Partially Secure
DHE-DSS-CAMELLIA128-SHA256	Partially Secure
DHE-DSS-CAMELLIA256-SHA	Partially Secure
DHE-DSS-CAMELLIA256-SHA256	Partially Secure
DHE-DSS-DES-CBC3-SHA	Insecure
DHE-DSS-SEED-SHA	Partially Secure
DHE-PSK-3DES-EDE-CBC-SHA	Insecure
DHE-PSK-AES128-CBC-SHA	Partially Secure
DHE-PSK-AES128-CBC-SHA256	Partially Secure
DHE-PSK-AES128-CCM	Partially Secure
DHE-PSK-AES128-CCM8	Partially Secure
DHE-PSK-AES128-GCM-SHA256	Secure
DHE-PSK-AES256-CBC-SHA	Partially Secure
DHE-PSK-AES256-CBC-SHA384	Partially Secure
DHE-PSK-AES256-CCM	Partially Secure

Cipher Suite	Class
DHE-PSK-AES256-CCM8	Partially Secure
DHE-PSK-AES256-GCM-SHA384	Secure
DHE-PSK-CAMELLIA128-SHA256	Partially Secure
DHE-PSK-CAMELLIA256-SHA384	Partially Secure
DHE-PSK-CHACHA20-POLY1305	Secure
DHE-PSK-NULL-SHA	Totally Insecure
DHE-PSK-NULL-SHA256	Totally Insecure
DHE-PSK-NULL-SHA384	Totally Insecure
DHE-RSA-AES128-CCM	Partially Secure
DHE-RSA-AES128-CCM8	Secure
DHE-RSA-AES128-GCM-SHA256	Secure
DHE-RSA-AES128-SHA	Partially Secure
DHE-RSA-AES128-SHA256	Partially Secure
DHE-RSA-AES256-CCM	Partially Secure
DHE-RSA-AES256-CCM8	Secure
DHE-RSA-AES256-GCM-SHA384	Secure
DHE-RSA-AES256-SHA	Partially Secure
DHE-RSA-AES256-SHA256	Partially Secure
DHE-RSA-CAMELLIA128-SHA	Partially Secure
DHE-RSA-CAMELLIA128-SHA256	Partially Secure
DHE-RSA-CAMELLIA256-SHA	Partially Secure
DHE-RSA-CAMELLIA256-SHA256	Partially Secure
DHE-RSA-CHACHA20-POLY1305	Secure
DHE-RSA-DES-CBC3-SHA	Insecure
DHE-RSA-SEED-SHA	Partially Secure
ECDHE-ECDSA-AES128-CCM	Partially Secure
ECDHE-ECDSA-AES128-CCM8	Secure
ECDHE-ECDSA-AES128-GCM-SHA256	Secure
ECDHE-ECDSA-AES128-SHA	Partially Secure
ECDHE-ECDSA-AES128-SHA256	Partially Secure
ECDHE-ECDSA-AES256-CCM	Partially Secure
ECDHE-ECDSA-AES256-CCM8	Secure
ECDHE-ECDSA-AES256-GCM-SHA384	Secure
ECDHE-ECDSA-AES256-SHA	Partially Secure
ECDHE-ECDSA-AES256-SHA384	Partially Secure
ECDHE-ECDSA-CAMELLIA128-SHA256	Partially Secure
ECDHE-ECDSA-CAMELLIA256-SHA384	Partially Secure
ECDHE-ECDSA-CHACHA20-POLY1305	Secure
ECDHE-ECDSA-DES-CBC3-SHA	Insecure
ECDHE-ECDSA-NULL-SHA	Totally Insecure
ECDHE-PSK-3DES-EDE-CBC-SHA	Insecure
ECDHE-PSK-AES128-CBC-SHA	Partially Secure
ECDHE-PSK-AES128-CBC-SHA256	Partially Secure
ECDHE-PSK-AES256-CBC-SHA	Partially Secure
ECDHE-PSK-AES256-CBC-SHA384	Partially Secure
ECDHE-PSK-CAMELLIA128-SHA256	Partially Secure
ECDHE-PSK-CAMELLIA256-SHA384	Partially Secure
ECDHE-PSK-CHACHA20-POLY1305	Secure
ECDHE-PSK-NULL-SHA	Totally Insecure
ECDHE-PSK-NULL-SHA256	Totally Insecure
ECDHE-PSK-NULL-SHA384	Totally Insecure
ECDHE-RSA-AES128-GCM-SHA256	Secure
ECDHE-RSA-AES128-SHA	Partially Secure
ECDHE-RSA-AES128-SHA256	Partially Secure
ECDHE-RSA-AES256-GCM-SHA384	Secure
ECDHE-RSA-AES256-SHA	Partially Secure
ECDHE-RSA-AES256-SHA384	Partially Secure
ECDHE-RSA-CAMELLIA128-SHA256	Partially Secure

Cipher Suite	Class
ECDHE-RSA-CAMELLIA256-SHA384	Partially Secure
ECDHE-RSA-CHACHA20-POLY1305	Secure
ECDHE-RSA-DES-CBC3-SHA	Insecure
ECDHE-RSA-NULS-SHA	Totally Insecure
IDEA-CBC-SHA	Partially Secure
NULL-MD5	Totally Insecure
NULL-SHA	Totally Insecure
NULL-SHA256	Totally Insecure
PSK-3DES-EDE-CBC-SHA	Insecure
PSK-AES128-CBC-SHA	Partially Secure
PSK-AES128-CBC-SHA256	Partially Secure
PSK-AES128-CCM	Partially Secure
PSK-AES128-CCM8	Partially Secure
PSK-AES128-GCM-SHA256	Partially Secure
PSK-AES256-CBC-SHA	Partially Secure
PSK-AES256-CBC-SHA384	Partially Secure
PSK-AES256-CCM	Partially Secure
PSK-AES256-CCM8	Partially Secure
PSK-AES256-GCM-SHA384	Partially Secure
PSK-CAMELLIA128-SHA256	Partially Secure
PSK-CAMELLIA256-SHA384	Partially Secure
PSK-CHACHA20-POLY1305	Partially Secure
PSK-NULS-SHA	Totally Insecure
PSK-NULS-SHA256	Totally Insecure
PSK-NULS-SHA384	Totally Insecure
RC4-SHA	Totally Insecure
RSA-PSK-3DES-EDE-CBC-SHA	Insecure
RSA-PSK-AES128-CBC-SHA	Partially Secure
RSA-PSK-AES128-CBC-SHA256	Partially Secure
RSA-PSK-AES128-GCM-SHA256	Partially Secure
RSA-PSK-AES256-CBC-SHA	Partially Secure
RSA-PSK-AES256-CBC-SHA384	Partially Secure
RSA-PSK-AES256-GCM-SHA384	Partially Secure
RSA-PSK-CAMELLIA128-SHA256	Partially Secure
RSA-PSK-CAMELLIA256-SHA384	Partially Secure
RSA-PSK-CHACHA20-POLY1305	Partially Secure
RSA-PSK-NULS-SHA	Totally Insecure
RSA-PSK-NULS-SHA256	Totally Insecure
RSA-PSK-NULS-SHA384	Totally Insecure
SEED-SHA	Partially Secure
SRP-3DES-EDE-CBC-SHA	Insecure
SRP-AES-128-CBC-SHA	Insecure
SRP-AES-256-CBC-SHA	Insecure
SRP-DSS-3DES-EDE-CBC-SHA	Insecure
SRP-DSS-AES-128-CBC-SHA	Insecure
SRP-DSS-AES-256-CBC-SHA	Insecure
SRP-RSA-3DES-EDE-CBC-SHA	Insecure
SRP-RSA-AES-128-CBC-SHA	Insecure
SRP-RSA-AES-256-CBC-SHA	Insecure
TLS_AES_128_CCM_SHA256	Secure
TLS_AES_128_GCM_SHA256	Secure
TLS_AES_256_GCM_SHA384	Secure
TLS_DH_anon_WITH_3DES_EDE_CBC_SHA	Totally Insecure
TLS_DH_anon_WITH_AES_256_CBC_SHA	Totally Insecure
TLS_DH_anon_WITH_AES_256_CBC_SHA256	Totally Insecure
TLS_DH_anon_WITH_CAMELLIA_128_CBC_SHA	Totally Insecure
TLS_DH_anon_WITH_CAMELLIA_128_CBC_SHA256	Totally Insecure
TLS_DH_anon_WITH_RC4_128_MD5	Totally Insecure

Cipher Suite	Class
TLS_DHE_PSK_WITH_AES_256_CBC_SHA	Partially Secure
TLS_DHE_PSK_WITH_AES_256_CBC_SHA384	Partially Secure
TLS_DHE_RSA_EXPORT_WITH_DES40_CBC_SHA	Totally Insecure
TLS_DHE_RSA_WITH_3DES_EDE_CBC_SHA	Insecure
TLS_DHE_RSA_WITH_AES_128_CBC_SHA	Partially Secure
TLS_DHE_RSA_WITH_AES_128_CBC_SHA256	Partially Secure
TLS_DHE_RSA_WITH_AES_128_GCM_SHA256	Secure
TLS_DHE_RSA_WITH_AES_256_CBC_SHA	Partially Secure
TLS_DHE_RSA_WITH_AES_256_CBC_SHA256	Partially Secure
TLS_DHE_RSA_WITH_AES_256_GCM_SHA384	Secure
TLS_DHE_RSA_WITH_CAMELLIA_256_CBC_SHA	Partially Secure
TLS_DHE_RSA_WITH_CAMELLIA_256_CBC_SHA256	Partially Secure
TLS_DHE_RSA_WITH_DES_CBC_SHA	Totally Insecure
TLS_DHE_RSA_WITH_SEED_CBC_SHA	Partially Secure
TLS_ECDH_anon_WITH_3DES_EDE_CBC_SHA	Totally Insecure
TLS_ECDHE_ECDSA_WITH_3DES_EDE_CBC_SHA	Insecure
TLS_ECDHE_ECDSA_WITH_AES_128_CBC_SHA	Partially Secure
TLS_ECDHE_ECDSA_WITH_AES_128_CBC_SHA256	Partially Secure
TLS_ECDHE_ECDSA_WITH_AES_128_GCM_SHA256	Secure
TLS_ECDHE_ECDSA_WITH_AES_256_CBC_SHA	Partially Secure
TLS_ECDHE_ECDSA_WITH_AES_256_CBC_SHA384	Partially Secure
TLS_ECDHE_ECDSA_WITH_CHACHA20_POLY1305_SHA256	Secure
TLS_ECDHE_RSA_WITH_3DES_EDE_CBC_SHA	Insecure
TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA	Partially Secure
TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256	Partially Secure
TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256	Secure
TLS_ECDHE_RSA_WITH_AES_256_CBC_SHA	Partially Secure
TLS_ECDHE_RSA_WITH_AES_256_CBC_SHA384	Partially Secure
TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384	Secure
TLS_ECDHE_RSA_WITH_CHACHA20_POLY1305_SHA256	Secure
TLS_ECDHE_RSA_WITH_RC4_128_SHA	Totally Insecure
TLS_RSA_EXPORT_WITH_DES40_CBC_SHA	Totally Insecure
TLS_RSA_EXPORT1024_WITH_DES_CBC_SHA	Totally Insecure
TLS_RSA_EXPORT1024_WITH_RC2_CBC_56_MD5	Totally Insecure
TLS_RSA_WITH_3DES_EDE_CBC_SHA	Insecure
TLS_RSA_WITH_AES_128_CBC_SHA	Partially Secure
TLS_RSA_WITH_AES_128_CBC_SHA256	Partially Secure
TLS_RSA_WITH_AES_128_GCM_SHA256	Partially Secure
TLS_RSA_WITH_AES_256_CBC_SHA	Partially Secure
TLS_RSA_WITH_AES_256_CBC_SHA256	Partially Secure
TLS_RSA_WITH_AES_256_GCM_SHA384	Partially Secure
TLS_RSA_WITH_DES_CBC_SHA	Totally Insecure
TLS_RSA_WITH_RC4_128_MD5	Totally Insecure
TLS_RSA_WITH_RC4_128_SHA	Totally Insecure

Appendix D - Internship Proposal



DEPARTAMENTO DE ENGENHARIA INFORMÁTICA E DE SISTEMAS

PROPOSTA DE ESTÁGIO

Ano Lectivo de 2019/2020

em [Mestrado em Informática e Sistemas](#) (Tecnologias da Informação e do Conhecimento)

TEMA

Resiliência a Ciberataques nas Infraestruturas Críticas de Portugal

SUMÁRIO

Atualmente, as infraestruturas críticas dos países estão cada vez mais subordinadas à progressiva informatização e, conseqüentemente, a interconectividade provida pela Internet. Infraestruturas críticas (internacionalmente conhecidos como Infrastructure Critical Systems - ICS - e Supervisory Control And Data Acquisition - SCADA - Systems) são sistemas que controlam e permitem acesso aos grandes serviços de infraestrutura de um país. Exemplos de sistemas críticos são os sistemas de aviação, energéticos, de fornecimento de água, das forças policiais e militares, de resposta a emergências, entre outros. Ataques bem sucedidos a estes sistemas podem ter repercussões catastróficas.

A interconetividade progressiva dos sistemas críticos coloca a possibilidade dos mesmos estarem visíveis e desprotegidos contra simples utilizadores da Internet, e, conseqüentemente, a todos os potenciais hackers existentes no mundo. O grau de vulnerabilidade e a facilidade da sua descoberta podem ser a diferença entre a vida normal de um país e uma catástrofe nacional iniciada por uma entidade estrangeira (ou mesmo nacional). A avaliação do grau de segurança destes sistemas face aos utilizadores maliciosos na Internet é uma atividade que todos os países deveriam executar regularmente. Apesar da importância, e da existência clara da preocupação na segurança nacional por parte do Governo Português e correspondentes organismos técnicos, a verdade é que não existem medidas públicas do grau de exposição do Estado Português a estes ataques. Isto é particularmente crítico, pois sem saber a situação atual dos sistemas é difícil enfatizar e exigir a sua melhoria.

O objetivo deste projeto é o de propor uma metodologia e desenvolver uma investigação sobre o estado atual de resiliência das infraestruturas críticas de Portugal. Entre os diversos desafios da investigação estão a investigação do estado atual das informações disponíveis sobre o assunto (incluindo trabalhos públicos e académicos na área), a investigação de ferramentas e bases de dados que possam auxiliar na

1/5



DEPARTAMENTO DE ENGENHARIA INFORMÁTICA E DE SISTEMAS

atividade, a proposta de uma metodologia que permita diferenciar endereços que fazem parte de infraestruturas críticas e outros não relevantes e uma potencial classificação de grau de risco associado, e a execução de um levantamento da situação atual, completo ou em regime de prova de conceito.

1. ÂMBITO

As infraestruturas críticas de um país constituem um alvo potencial de ciberataques, e as repercussões de um ataque bem sucedido podem ser catastróficas. Ao mesmo tempo a conectividade que as infraestruturas tem apresentado ao longo do tempo pode colocá-las sob visibilidade direta de todos os agente maliciosos que se encontram na Internet. Identificar o grau de visibilidade e vulnerabilidade das infraestruturas críticas de um país é de importância vital. Atualmente, não é de conhecimento público o estado de visibilidade e vulnerabilidade das infraestruturas críticas de Portugal.

As técnicas de reconhecimento e enumeração tipicamente utilizadas na avaliação de segurança de sistemas informáticos não são suficientes para a avaliação, pois a identificação dos equipamentos e *endpoints* que representam infraestruturas críticas em um país é uma tarefa ainda sem uma estratégia de solução evidente. Um dos resultados principais deste trabalho será a investigação e proposição de técnicas que permitam a identificação de dispositivos que fazem parte da infraestrutura crítica e a avaliação do grau de risco associado. Por mais que existam trabalhos similares em âmbito sigiloso, o objetivo é determinar o grau de informação disponível à qualquer pessoa com acesso a informações públicas e à Internet, e portanto, a todos os agentes maliciosos que existem.

2. OBJETIVOS

O presente projecto/estágio pretende atingir os seguintes objectivos genéricos:

- Investigação do estado da arte na identificação e avaliação de segurança de infraestruturas críticas;
- A investigação e proposta de ferramentas e bases de dados públicas que auxiliem nos objetivos propostos;
- O desenvolvimento de uma metodologia que permita identificar dispositivos que fazem parte de uma rede de serviços de infraestrutura crítica e estimar seu grau de risco e visibilidade;
- A aplicação da metodologia proposta na prática, através de uma investigação completa ou prova de conceito no âmbito das infraestruturas críticas de Portugal;
- A análise e avaliação dos resultados.



DEPARTAMENTO DE ENGENHARIA INFORMÁTICA E DE SISTEMAS

3. PROGRAMA DE TRABALHOS

O estágio consistirá nas seguintes atividades e respetivas tarefas:

- *T1 - Levantamento do Estado da Arte* - Investigação do estado da arte na identificação e avaliação de segurança de infraestruturas críticas.
- *T2 - Levantamento de Ferramentas e Bases de Dados* - Investigação e proposta de ferramentas e bases de dados públicas que auxiliem nos objetivos propostos.
- *T3 - Desenvolvimento da Metodologia* - Desenvolvimento de uma metodologia que permita identificar dispositivos que fazem parte de uma rede de serviços de infraestrutura crítica e estimar seu grau de risco e visibilidade.
- *T4 - Aplicação prática da Metodologia Desenvolvida* - Aplicação da metodologia proposta, através de uma investigação completa ou prova de conceito no âmbito das infraestruturas críticas de Portugal.
- *T5 - Análise de Resultados* - Análise e avaliação dos resultados.
- *T6 - Relatório de Estágio.*

4. CALENDARIZAÇÃO DAS TAREFAS

As Tarefas acima descritas, incluindo os testes de validação de cada módulo, serão executadas de acordo com a seguinte calendarização:

O plano de escalonamento dos trabalhos é apresentado em seguida, sendo que está planeada a execução de 1200 horas de trabalho, o que equivale a 30 semanas:

Tarefas	Meses					
	N	N+1	N+2	N+3	N+4	N+5
T1	■	■				
T2			■	■		
T3				■	■	■
T4					■	■
T5						■
Metas	INI	M1	M2	M3		M4 M5

- INI Início dos trabalhos
M1 (INI + 5 Semanas) Tarefa T1 terminada
M2 (INI + 5 Semanas) Tarefa T2 terminada
M3 (INI + 5 Semanas) Tarefa T3 terminada
M4 (INI + 5 Semanas) Tarefa T4 terminada
M5 (INI + 10 Semanas) Tarefa T5 + T6 terminadas



DEPARTAMENTO DE ENGENHARIA INFORMÁTICA E DE SISTEMAS

5. RESULTADOS

Os resultados a apresentar pelo estagiário serão consubstanciados nas várias metas estabelecidas para o projeto, sendo que serão espetáveis 2 períodos de avaliação fundamentais:

Dezembro de 2019

- Avaliação das tarefas 1 e 2: Avaliação das proposta de ferramentas e bases de dados públicas que serão utilizadas. Neste momento já deverá haver um draft da proposta da escolha de metodologia utilizada.

Julho de 2020

- Avaliação dos resultados: Será avaliada a metodologia proposta e os resultados obtidos com a utilização prática da mesma;
- Será também avaliada a qualidade da documentação elaborada pelo estagiário, materializada no seu relatório de estágio.

6. LOCAL DE TRABALHO

O trabalho será realizado a partir dos escritórios da Dognaedis em Coimbra.

Horário de Trabalho: 9h00 às 18h00, com uma hora para almoço.

7. METODOLOGIA

Embora estejam salvaguardadas 10 semanas no final do projeto, o dossier de projeto deverá ser elaborado pelo estagiário ao longo da execução do projeto, com o apoio do orientador.

Prevê-se que o desenvolvimento seja assente numa metodologia ágil, assegurando a realização de reuniões semanais de acompanhamento ao projeto, sobre as quais será verificada a conformidade com o planeamento estabelecido neste documento enquanto macro-tarefas e com o planeamento delineado pelo estagiário para execução da fase de implementação.



DEPARTAMENTO DE ENGENHARIA INFORMÁTICA E DE
SISTEMAS

8. ORIENTAÇÃO

ISEC:

Nome (nome@isec.pt)
Categoria

Entidade de Acolhimento:

Afonso Neto (hr@dognaedis.com)
Senior Cybersecurity Consultant

9. CARACTERIZAÇÃO E REMUNERAÇÃO

- Horário de Trabalho: 9h00 às 18h00, com uma hora para almoço.
- Será fornecido ao estagiário todo o material necessário para a realização do trabalho, incluindo workstation.
- O estagiário será remunerado mediante um modelo de mérito, com base em avaliações e semestrais. Sendo que ao cumprir todos os objetivos (apresentados no início do estágio), poderá atingir uma remuneração total de 3.000€. Adicionalmente o estagiário terá direito a subsídio de alimentação de acordo com a política da empresa.
- O estagiário terá acesso a todas formações proporcionadas internamente pela Dognaedis.

