Universidade do Minho
Escola de Engenharia

Ana Regina Coelho de Sousa

Big Data and Real-Time Knowledge Discovery in Healthcare Institutions

December, 2023

Big Data and Real-Time Knowledge Discovery in Healthcare Institutions

Regina Sousa

UMinho | 2023

Universidade do Minho
Escola de Engenharia

Ana Regina Coelho de Sousa

Big Data and Real-Time Knowledge Discovery in Healthcare Institutions

Doctorate Thesis

Doctorate in Doctoral Programme in Biomedical Engineering

Work developed under the supervision of:
José Manuel Ferreira Machado
António Carlos da Silva Abelha

December, 2023

COPYRIGHT AND TERMS OF USE OF THIS WORK BY A THIRD PARTY

Acknowledgements

One more journey comes to an end. After a long road full of hope, nervousness and insecurity I reached another milestone in my life, and I owe thanks to so many people who supported and encouraged this project in the most varied ways.

To Professor José Machado I thank all the trust and hope deposited in my work. More than an advisor he was a good counselor that I always cherish. I thank him above all for believing in my potential and providing me with incredible opportunities that made me grow as a person and as a researcher.

To Professor António Abelha I thank for all the support, thank you very much for the calm and tranquility transmitted during the course of these years. I also thank him for his advice, opportunities and words of encouragement

To my friends who shared with me the experience and adventure (KEG Lab) thank you very much for sharing, encouraging and most of all for the friendship created.

To my family and close friends thank you for all the support you have always offered me throughout this long journey full of good moments (and some bad ones) that is life. To my parents thank you for all the love, education and spirit of sacrifice that you always transmitted to me. And to my sister, thank you for unconditionally supporting the peculiar human being that sometimes I can be. I also want to thank my grandparents for all the protection they have over the whole family, for showing the meaning of perseverance and caring for others. Thank you all, *"Os de Casa"* for giving meaning to the words family and love, each in their own way. Thank you all for being with open arms for my dramas.

To my boyfriend, Tiago, I express only gratitude and unconditional love, thank you especially for your patience. Without you I would have given up in the first 6 months. You have been incredibly understanding and I can't thank you enough for all the support.

To all of you, thank you so much.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the Universidade do Minho.

________________________, _______________________________________

(Place) (Date)

____

(Ana Regina Coelho de Sousa)

“

‘ Advancing healthcare through the power of information and technology. ’ (Dr. Donald Lindberg)

”

Resumo

Big Data e Descoberta de Conhecimento, em Tempo Real, nas Instituições de Saúde

Nas instituições de saúde, a quantidade e complexidade dos dados gerados tornam a recolha, armazenamento, processamento e disponibilização de informações um processo desafiador. Com a crescente adoção de tecnologias, como registos eletrónicos de saúde, as instituições de saúde têm acesso a vastas quantidades de informações. Estes dados são provenientes de vários dispositivos que, por vezes, são incapazes de trocar informação entre si o que eleva a complexidade dos sistemas de informação.

O uso de Big Data, Interoperabilidade e *Cloud-Computing* surgem como soluções promissoras para melhorar a eficiência e a eficácia dos sistemas de informação na área da saúde. Ajudam a fornecer informações em tempo real, melhorar a tomada de decisões clínicas e providenciar atendimento personalizado aos pacientes. Contudo, a falta de recursos sentida de forma transversal, aliada à complexidade do problema bem como da solução dificulta a aceitação e investimento por parte das instituições.

Esta tese pretende mostrar que, sistemas desenvolvidos na *cloud* podem providenciar acesso a recursos computacionais poderosos e escaláveis sem a necessidade do assustador investimento inicial. Mais ainda, mostra-se que, a combinação do paradigma *cloud* com ferramentas de Big Data possibilita a informatização de sistemas a instituições de saúde de qualquer dimensão revelando um grande avanço para a partilha de dados e interoperabilidade de dados.

O objetivo desta investigação é o desenvolvimento de um *Software as a Service (SaaS)* que, com a implementação de padrões de dados conhecidos na área da saúde, consiga interoperar com as fontes de dados das instituições de saúde. Este *software* deve ser adaptável aos novos modelos de trabalho (remoto e/ou híbrido), possibilitando a diminuição significativa de gastos em recursos humanos e materiais. Os resultados são extremamente promissores consistindo num *software*, adaptativo, escalável e modular que permite a customização a qualquer instituição de saúde. Apesar dos casos de estudos se encontrarem em diferentes estados de maturidade foram amplamente aceites pelos utilizadores.

Palavras-chave: Big Data, Clinical Data Standards, Cloud-Computing, Health Information Systems,Interoperability

Abstract

Big Data and Real-Time Knowledge Discovery in Healthcare Institutions

In healthcare institutions, the amount and complexity of data generated makes it difficult to capture, store, process, and distribute information. With the increasing adoption of digital technologies such as electronic health records and other data sources, healthcare institutions have access to voluminous quantities of data. This data originates from a variety of devices that are sometimes unable to communicate with one another, thereby increasing the complexity of information systems.

Consequently, Big Data and Interoperability are emerging as promising solutions for enhancing the efficacy and efficiency of healthcare information systems. They contribute to the provision of vital information in real time, the improvement of clinical and managerial decision-making, and the delivery of personalized, high-quality care to patients. However, the widespread lack of resources and the complexity of the issue and its solution make it challenging for institutions to accept and invest in them.

This thesis aims to show that cloud-based systems can provide access to powerful and scalable computing resources without the need for a daunting initial investment. Furthermore, it is shown that the combination of the cloud paradigm with Big Data tools, such as Spark, enables the computerization of systems for healthcare institutions of any size, revealing a significant advance in data sharing and interoperability.

Thus, the main objective of this investigation is the development of a generic SaaS that, with the implementation of widely known data standards in the healthcare field (Health Level Seven (HL7)), can interoperate with any source of data from healthcare institutions. It is also intended that this software be adaptable to new models of work (remote and/or hybrid), enabling a significant reduction in spending on human and material resources. The results are extremely promising, consisting of a generic, adaptive, scalable and modular software that allows adaptation to any healthcare institution, its professionals, patients and equipment. Although the case studies are in different stages of maturity, they were widely accepted by users.

Keywords: Big Data, Clinical Data Standards, Cloud-Computing, Health Information Systems,Interoperability

Contents

List of Figures

List of Tables

Acronyms

ACID	Atomicity, Consistency, Isolation, Durability *(pp. 23, 24)*
AI	Artificial Intelligence *(pp. 20, 46, 102, 104, 105)*
API	Application Programming Interface *(pp. xvii, 39, 49, 50, 52–54, 58, 59, 63, 64, 81, 88, 99, 100, 102, 104, 110)*
AWS	Amazon Web Services *(pp. 39, 41–43, 59, 62–66, 81, 100, 111)*
BI	Business Intelligence *(pp. 13, 25, 26, 46, 54, 108)*
CDA	Clinical Document Architecture *(pp. 35, 36)*
CDSS	Clinical Decision Support Systems *(p. 2)*
CEO	Chief Executive Officer *(p. 26)*
CMS	Centers for Medicare and Medicaid Services *(p. 20)*
CRM	Customer Relationship Management *(p. 38)*
CSP	Cloud Service Provider *(pp. xvii, 11, 42–44, 59, 100)*
CSV	Comma-Separated Values *(p. 87)*
DCMI	Dublin Core Metadata Initiative *(p. 29)*
DICOM	Digital Imaging and Communications in Medicine *(pp. 31–33)*
DL	Deep Learning *(p. 54)*
DSR	Design Science Research *(pp. 5–7, 99)*
EEA	European Economic Area *(p. 35)*
EHR	Electronic Health Records *(pp. 2, 14, 20, 22, 29, 36, 46, 48)*
ETL	Extract, Transform, Load *(pp. 88, 96, 101)*
EU	European Union *(p. 35)*

FAIR	Findability, Accessibility, Interoperability, Reusability *(pp. 35, 47, 100)*
FHIR	Fast Healthcare Interoperability Resources *(pp. 30–32, 35–37, 54)*
FSN	Fully Specified Name *(p. 34)*
GCP	Google Cloud Platform *(pp. 42, 43)*
GDH	Grupos de Diagnósticos Homogéneos *(p. 51)*
GDP	Gross Domestic Product *(p. 103)*
GDPR	General Data Protection Regulation *(pp. 35, 48)*
HDFS	Hadoop Distributed File System *(pp. 16, 17, 27)*
HDI	Human Development Index *(p. 89)*
HIMSS	Healthcare Information and Management Systems Society *(pp. xiv, 29, 30)*
HIPAA	Health Insurance Portability and Accountability Act *(p. 35)*
HL7	Health Level Seven *(pp. vii, ix, 30–32, 34–37, 47–50, 59, 62, 63, 99, 100, 104–106, 108)*
HTTP	HyperText Transfer Protocol *(pp. 37, 59, 63)*
HTTPS	HyperText Transfer Protocol Secure *(pp. 44, 59, 63)*
HUE	Hadoop User Experience *(p. 18)*
IaaS	Infrastructure as a Service *(pp. 38–41)*
IBM	International Business Machines Corporation *(p. 13)*
ICD	International Classification of Diseases *(pp. 32, 33)*
ICD-10	International Statistical Classification of Diseases and Related Health Problems 10th Revision *(pp. 33, 47, 49–51)*
INE	Instituto Nacional de Estatística *(p. 55)*
IoT	Internet of Things *(pp. vi, 20, 42)*
IP	Internet Protocol *(p. 49)*
IS	Information Systems *(p. 14)*
ISO	International Standardization Organization *(pp. 85, 87, 101)*
IT	Information Technologies *(pp. vii, 5, 30, 38, 107)*
JPEG	Joint Photographic Experts Group *(p. 33)*
JSON	JavaScript Object Notation *(pp. 25, 52, 53, 81, 87, 112)*
KFF	Kaiser Family Foundation *(p. 103)*
LAN	Local Area Network *(p. 39)*
LIS	Laboratory Information System *(p. 48)*

LOINC	Logical Observation Identifiers Names and Codes *(pp. 32, 34, 47, 49, 51)*
ML	Machine Learning *(pp. 18, 25, 42, 46, 54)*
NHANES	National Health and Nutrition Examination Survey *(p. 102)*
NHS	National Health Service *(pp. 20, 28)*
NIST	National Institute of Standard and Technology *(p. 39)*
NLM	National Library of Medicine *(p. 21)*
NoSQL	Not only SQL *(pp. xiv, 17, 22, 24, 25, 87)*
OECD	Organization for Economic Cooperation and Development *(p. 103)*
OSI	Open Systems Interconnection *(p. 63)*
PaaS	Platform as a Service *(pp. 38–41)*
PACS	Picture Archiving and Communication System *(p. 33)*
PDF	Portable Document Format *(pp. 62, 63, 65, 70, 71, 81, 106)*
PHI	Personal Health Information *(p. 35)*
PoC	Proof of Concept *(pp. 8, 99, 107)*
RDBMS	Relational Database Management System *(pp. 18, 22, 23)*
REST	Representational State Transfer *(pp. 52, 63)*
SaaS	Software as a Service *(pp. vii, ix, xiii, xiv, xvii, 6, 38–41, 56–63, 100–102, 105, 108, 110–112, 130)*
SDK	Software Development Kit *(p. 59)*
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms *(pp. 31, 32, 34)*
SPMS	Serviços Partilhados do Ministério da Saúde *(p. 37)*
SQL	Structured Query Language *(pp. 18, 23, 24, 27, 42, 49, 99)*
SWOT	Strenghs,Weeknesses,Oportunities,Threats *(pp. xiv, 8, 9, 107)*
UK	United Kingdom *(pp. 20, 28, 34)*
URL	Uniform Resource Locators *(p. 81)*
USA	United States of America *(pp. xv, 35, 89, 92, 93, 103)*
VPN	Virtual Private Network *(pp. 39, 44)*
WAN	Wide Area Network *(p. 39)*

WBC	White Blood Cells *(p. 51)*
WHO	World Health Organization *(pp. 33, 55, 82, 101, 102)*
XML	Extensible Markup Language *(pp. 25, 42)*
YARN	Yet Another Resource Negotiator *(p. 17)*

Part I

Introduction

1

Introduction

This chapter introduces the scope and motivation for this doctoral thesis (Section 1.1). Then it presents the research problem, opportunity and objectives (Sections 1.2 and 1.3) as well as the expected outcomes. Finally, the design for conducting the research process, along with the reason for its adoption, are described in Section 1.4. In addition, the summary of the research process (Section 3) and the structure of this document (Section 1.6) are clarified.

1.1 Scope and Motivation

This investigation focuses on data-driven knowledge discovery in healthcare, particularly when vast amounts of data are involved. These data, which include Electronic Health Records (EHR), laboratory test results, and patient monitoring data, can potentially improve patient care and stimulate innovation in the health industry. So, different data types are identified, including structured, semi-structured, and unstructured data from multiple sources [128, 139]. Thus, collecting, storing, processing, and accessing this large amount of data is particularly challenging since it requires the employment of sophisticated tools and technologies to accommodate their heterogeneity [50]. This is when the concept of Big Data emerges.

Big Data is one of the most popular terms worldwide, yet its definition is controversial. The term was created from an effort to describe the huge amounts of data generated at any given time [91]. However, it is challenging to determine when data becomes ”Big”. Therefore, the concept is now defined by its characteristics (e.g., variety, velocity, volume) to eliminate ambiguity. These characteristics make Big Data challenging to manage and analyze and makes it a popular research topic in healthcare [91]. As healthcare organizations rely increasingly on data-driven approaches, it is crucial to have rigorously supported proof that developing methodologies and technology, like Clinical Decision Support Systems (CDSS), may assist them in advancing in data-driven healthcare contexts [73].

Big Data as a research topic offers as many opportunities as challenges. Various sectors are experiencing several difficulties due to the topic [33]. From the general dilemmas that have already been presented, some of them are highly important such as the lack of consensus and rigour in the definition, models and architectures, to those related to ethics, the ones related to privacy, security, and discrimination, and even those related to organizational changes, such as the need for new skills, changes in workflows, and resistance to change, among many others [102, 45]. Thus, research results are highly relevant, allowing companies like healthcare institutions to provide robust evidence based on contemporary techniques and technologies [88].

Using Big Data Analysis to improve healthcare is one of the primary drivers of research in this field. By analyzing vast amounts of data, it may be feasible to find patterns and trends that can be used to forecast patient outcomes and optimize treatment. This can aid healthcare professionals in providing more effective and personalized care to patients [125]. In addition, the sector's potential for cost reduction is underlined. By identifying and fixing healthcare system inefficiencies, it may be possible to lower costs and enhance the system's overall efficiency [28].

However, the lack of resources, including computing power and storage capacity, can make it challenging for healthcare organizations to manage and analyze this data effectively. Traditional on-premise systems may not have the capacity or the capability to handle such large datasets, leading to the need for more scalable and flexible solutions [5]. In addition, traditional systems demand expensive hardware and software license investments that may result in a financial drain for the healthcare organization.

Cloud-based systems offer a solution to this problem by providing access to unlimited computing and storage resources on a pay-as-you-go basis. This allows healthcare organizations to scale up their resources as needed to handle the growing data volume without investing in expensive infrastructures [118]. In addition to the scalability and flexibility of cloud-based systems, they also offer other benefits, such as improved data security and privacy and the ability to access data from anywhere with an internet connection. These features make cloud-based systems particularly well-suited for the healthcare sector, where data security and accessibility are critical considerations [63].

Combining Big Data and the lack of resources has driven the adoption of cloud-based systems in the healthcare sector. They provide a cost-effective and scalable solution for managing and analyzing data. Lastly, research in this area encourages innovation in the healthcare industry, new research and development opportunities will emerge, and new technologies and procedures that can enhance patient care will be created.

1.2 Research Problem

Research on *Real-Time Big Data Analysis* and even *Real-Time Knowledge Discovery in the Healthcare* was uncommon a few years ago due to the topic's novelty, complexity, and absence of a practical implementation guide or set of best practices. It has been possible to identify the optimum technology to fulfil the demands of the data volumes now being generated, using case-oriented strategies [24].

A literature search was performed to precisely define a pertinent research problem, as described in Table 1. Various search engines, keywords, and inclusion criteria are highlighted there.

Table 1: Literature Review Process

Search Engine	Keywords	Inclusion Criteria

Scopus	Big Data;	Open Access;
		Since 2015;
		Written in English;
		Subject area is Computer Science, Engineering, Decision Systems or Healthcare.

Science Direct	Healthcare Information systems;	
IEEE Xploree	Data Repositories;	
Web Of Science	Interoperability;	
Google Scholar	Clinical Data Standards;	
NCBI	Cloud Computing Paradigm.	

This search returned 1526 documents, the titles and abstracts of which were evaluated. After that, the introductions and conclusions of the revealing documents were read. The reading continued if they matched the theme for further discussion and analysis. Ultimately, a deeper examination was conducted, and the work was cited. Note: citations from works before 2015 may appear in transactions if the work is cited in another work under analysis or if the work is a reference in the study field.

In summary, the research identifies a considerable gap between distinguishing which data can be studied in real-time and which data is required for analysis as a problem, for demonstrating essential advancements in healthcare, and how this data may be examined in real-time, resulting in a use-case-driven strategy. Providing a standardized method for creating and implementing solutions with scalability and modification capabilities based on modern cloud-based architectures leads to new efforts for a new method of developing decision support systems in healthcare. This work intends to enhance large-scale data-driven techniques in which models and methodologies are as context-insensitive as possible.

Given the complexity of the environment, the requirements for real-time knowledge construction, and a previously studied set of requirements, the following research question was formulated:

”How can Big Data in healthcare institutions become relevant knowledge for real-time decision-making, with the ability to assist clinical management and generate clinical and performance indicators, improving processes and resources, especially in evidence-based medicine contexts?”

To address the deficiency mentioned above and in light of the growing demand for real-time knowledge extraction from large data sources in healthcare, Section 1.3 contains a comprehensive listing of all research objectives to be accomplished.

1.3 Objectives

This doctoral thesis seeks to develop a cloud-based health information system solution capable of real-time managing massive amounts of data. In addition, it was necessary to develop an interoperable data repository based on laboratory test results. These solutions address the research question presented in Section 1.2 and will be described in further detail later in this paper. As a result, this innovative method reveals how to rapidly design and execute comprehensive, scalable, and low-cost clinical software for an institution.

Based on this primary objective and given the identified gap, after seeing some health professionals in the field, some objectives for this work can be postulated:

- Investigation and analysis of the existing level of knowledge concerning the context of the PhD dissertation.
- Exploring the difficulties in Portuguese health institutions that can be addressed by employing techniques for real-time processing of large amounts of data.
- Determine the most effective research methodologies and technologies.
- Proposal systems' Architecture Design.
- Solutions development based on the aforementioned objectives.
- Prototype simulation, deployment and implementation.

In order to fulfil the objectives mentioned above, secondary research questions were outlined and will be described below:

- RQ1 - What is the importance of Big Data in Healthcare?
- RQ2 - How can cloud technologies address the resource deficit in healthcare institutions?
- RQ3 - How do cloud services accommodate Big Data volumes when required in real-time?
- RQ4 - Is the use of standards recommended in this type of solution?
- RQ5 - What are the most suitable methodologies to conduct this research?
- RQ6 - What are the most appropriate technologies to develop the solution?
- RQ7 - What will be the best architecture for the novel system?
- RQ8 - What are the major challenges encountered throughout the development process? How were they overcome?
- RQ9 - How would the established investigation build relevant knowledge for real-time decision-making processes in healthcare institutions?
- RQ10 - Why is the study developed relevant, significant, and original in comparison to similar solutions?

In order to address the main research question and the additional research questions, it was necessary to identify the research methodology to be followed. According to the preceding contextualization, it is evident that the research falls under the domain of applied sciences. Due to the extensive usage of the Design Science Research (DSR) methodology in the development of new solutions and approaches in the Information Technologies (IT) sector, this methodology was chosen. This methodology will be described in further detail in Section 1.4.

1.4 Research Design

This project will be supported by a set of primary methods and procedures so that there are guiding principles and an orderly path to follow. After identifying and analyzing the available possibilities, the decisions taken were deemed the most effective means of achieving the desired outcomes.

This scientific research process involves the constant acquisition of knowledge not just about the domain to which it applies but also about the subprocesses that comprise it (Bibliographical Research, Bibliographical Review, Research Methodologies, and Text Production, among others). To develop a grounded theory, the research must employ both quantitative and qualitative methodologies. Therefore it is based on a **Mixed** approach for the data analysis. Since the development will follow the analysis of data, which will then be linked to the literature, the **Inductive** approach has been selected. As a result, and in conjunction with the primary purpose of solving a problem, the ontological philosophy embraced in this study is **Pragmatism**, as no concepts will be used as the cause. Saunders, Lewis, and Thornhill believe that pragmatism is intuitively appealing because it precludes researchers from engaging in intricate discussions of notions such as truth and reality [126].

According to Brocke, Hevner, and Maedche, a DSR is a problem-solving paradigm that attempts to expand human knowledge by creating inventive artifacts; as such, it is in complete harmony with the aforementioned pragmatic approach. This methodology will be described in Section 1.4.1 [12].

1.4.1 Design Science Research

DSR is a relatively recent research methodology, but its roots are in engineering and the artificial sciences [132]. As such, it seeks to create something that is novel and solves a specific problem, as opposed to explaining an existing reality [120, 59].

The concept and meaning of this methodology have been interpreted differently by numerous authors. In each case, it is agreed that the technique has two primary goals: 1 - apply knowledge to problem-solving, and 2 - develop new knowledge, insights, and theoretical explanations. The methodology is built on a critical approach, allowing it to be applied iteratively until the optimal solution is achieved based on the examination of the generated solution [84]. Thus, according to the authors of *A Design Science Research Methodology for Information Systems Research*, the methodology is divided into six parts illustrated in Figure 1.

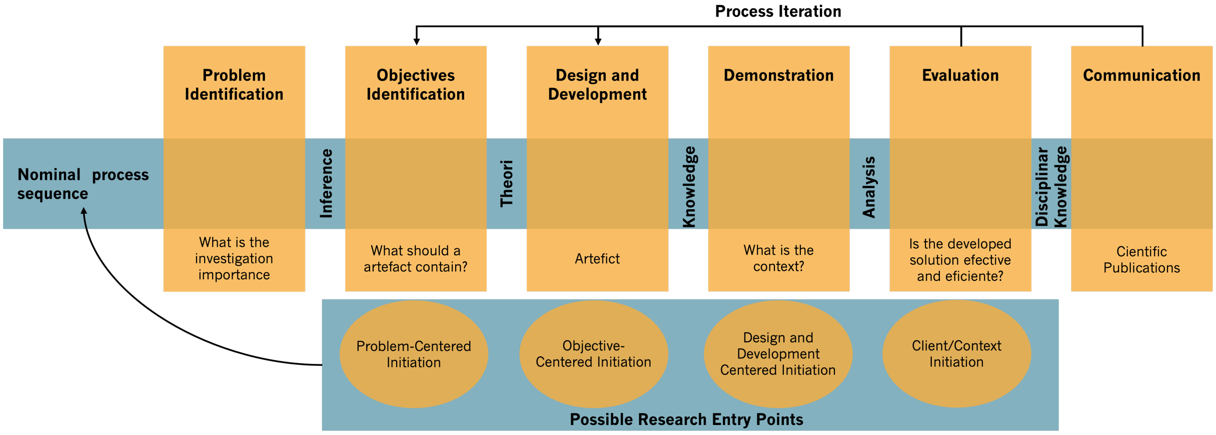

Process Iteration
Problem Identification
Objectives Identification
Design and Development
Demonstration
Evaluation
Communication
Nominal process sequence
Inference
Theori
Knowledge
Analysis
Disciplinar Knowledge
What is the investigation importance
What should a artefact contain?
Artefict
What is the context?
Is the developed solution efective and eficiente?
Cientific Publications
Problem-Centered Initiation
Objective-Centered Initiation
Design and Development Centered Initiation
Client/Context Initiation
Possible Research Entry Points

Figure 1: Design Science Research Process. Adapted from [110].

The procedure for the research is as follows:

1. Problem identification and motivation - It involves defining the investigation and justification of the potential value of an artifact. This stage is crucial since it will determine the success of all subsequent steps. The impression of the problem's complexity is essential to the success of the item.

2. Objectives Definition - At this stage, the viability and practicability of the idealized artifacts must be questioned. The more precisely and rigorously objectives are specified, the more achievable they become. Artifact is understood as the overall solution, which is composed of the data repository, the SaaS, and the case study about covid.

3. Artifact Design and Development - Creation of the artifact of any type (construct, method, model, among others). This step should determine the ideal functionality of the artifact and its architecture.

4. Demonstration - Developing Proof of Concept of the artifact to solve the identified problem(s). It may involve using the artifact, a simulation, a case study, or another appropriate activity.

5. Evaluation - Observing and measuring the effectiveness of the artifact for the problem. In this step, the proposed artifact objectives should be compared with the actual results observed when using the artifact.

6. Communication - Communication to the entire target community of the initial problem and its importance, as well as of the elaborated artifact and its usefulness, effectiveness and relevance.

In accordance with this methodology, the research is **Exploratory** and seeks a solution to the inefficiencies seen in a particular context. As a result, and taking into account the fact that a survey of the gaps present in the institution was conducted via interviews, analysis of existing systems, and other means, it is considered that the research is of a mixed nature, as both quantitative and qualitative methods were employed, both individually and in combination. Classifying the project in terms of its time horizon, it has a cross-sectional nature due to its instantaneous execution and concurrent occurrence in several healthcare institutions.

In addition to the DSR methodology, two additional strategies were employed. The first, outlined in Section 1.4.2, is intended to support the design and development phase of the artifact(s). In contrast, the second, described in Section 1.4.3, is intended to accompany the demonstration phase of the developed solution's viability.

1.4.2 Case Study Approach

The case study method is a research technique that entails an in-depth investigation of a single case or collection of cases. It is a technique used frequently in the social sciences, psychology, education, and business to examine a complicated subject or phenomena in a real-world setting [93]. In the context of the DSR research methodology, the case study approach entails collecting and analyzing data from a range of sources to get a comprehensive knowledge of a particular problem or issue and suggest viable answers or recommendations for its resolution. Usually resulting in the creation of comprehensive state-of-the-art research.

In order to conduct a case study while applying DSR methodology, the following steps are often considered:

Problem Identification : The first phase may involve reviewing the literature or interviewing experts or stakeholders to better understand the problem and its causes. In our work, both approaches were taken into consideration. Several interviews were conducted with clinic administrators together with health professionals. However, an extensive literature search was also conducted in order to find similar work that could shed light on the way forward.

Data acquisition : This can include interviews with individuals, observations, and written materials such as documents, reports, and other published literature. Of all these options, during the course of the research, professionals in action were observed, but also various reports and other software were analyzed and discussed in order to understand how to develop a better solution.

Data Analyses : When studying the results of the interviews and the selected software, a careful analysis was performed to identify patterns and trends and gain a deeper understanding of the problem.

artifact Development : Based on the insights gained from the data analysis, we moved on to the development that is detailed in each of the case studies, Chapters 6, 7 and 8.

Evaluate and refine solution : Finally, the case study approach was used to evaluate the effectiveness of the proposed solution through a proof of concept.

1.4.3 Proof of Concept

The Proof of Concept (PoC) methodology is a practical demonstration model, which can prove whether a concept, theory or solution is feasible for potential application in the real world [129]. Therefore, performing a PoC is often pointed out as one of the most important steps in the process of designing, developing, implementing and proposing a prototype. In other words, it verifies if the solution meets the requirements and objectives defined for which it was initially designed, identifying potential flaws or errors in the developed solution.

In short, a PoC demonstrates in practice the concepts, methodologies, and technologies involved in elaborating a given project and thus validates the proposed solution, proving its viability and usefulness.

In this PhD thesis, the defence of the feasibility and usefulness of the proposed system went through the application of SWOT research technique, in which a Strength, Weaknesses, Opportunities and Threats were analyzed in detail. The SWOT analysis is meticulously explained in Section 1.4.3.1.

In this investigation, SWOT analysis will be supported by data collected with self-administered and interviewer-administered questionnaires. In the future, it is intended to re-distribute the questionnaire and see if the users' responses change with continued use of the system.

1.4.3.1 SWOT Analysis

The origin of the acronym SWOT is comprised of four English words: *Strengths*, *Weaknesses*, *Opportunities*, and *Threats*. As the titles imply, this analysis enables us to analyze the developed solution's inherent strengths and shortcomings and its opportunities and risks [111].

During a study project between 1960 and 1970, Albert S. Humphrey is credited for developing this technique. The investors in this study were Fortune 500 firms, a yearly list of the five hundred largest companies in the United States, published to determine where business planning was going wrong and designing a new management and administration system [47, 53].

This technique is now a strategy-planning tool that examines both the internal and external environments of a product or business. The complete process is represented by a matrix, which facilitates the display of a solution's characteristics to further justify decisions, revealing its strengths and weaknesses, opportunities and even threats. Frequently, the matrix resembles the one represented in Figure 2. Analyzing Figure 2 succinctly, it can be stated that the positive factors are all those that contribute incrementally to achieving the initially proposed objectives, and these are positioned on the left side of the matrix. On the other hand, the negative factors are those that somehow hinder evolution and are therefore seen as adversities. These two types of factors are further divided into two others, those that only depend on the developer or the organization in question and are, therefore, internal factors, and those that are external to the organization and, therefore, beyond our control, external factors.

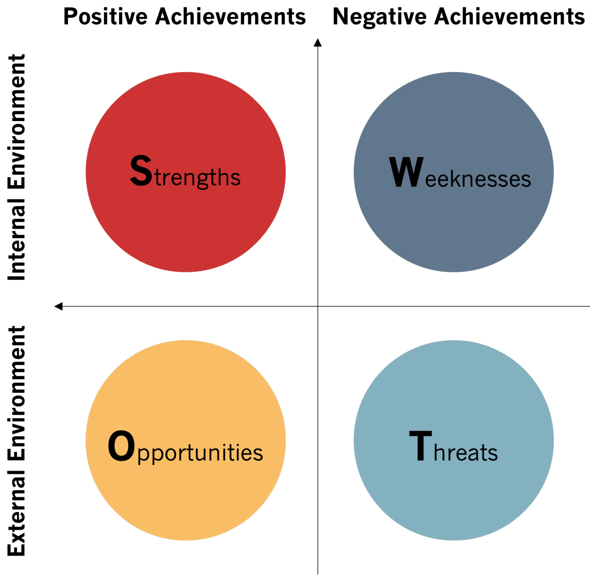
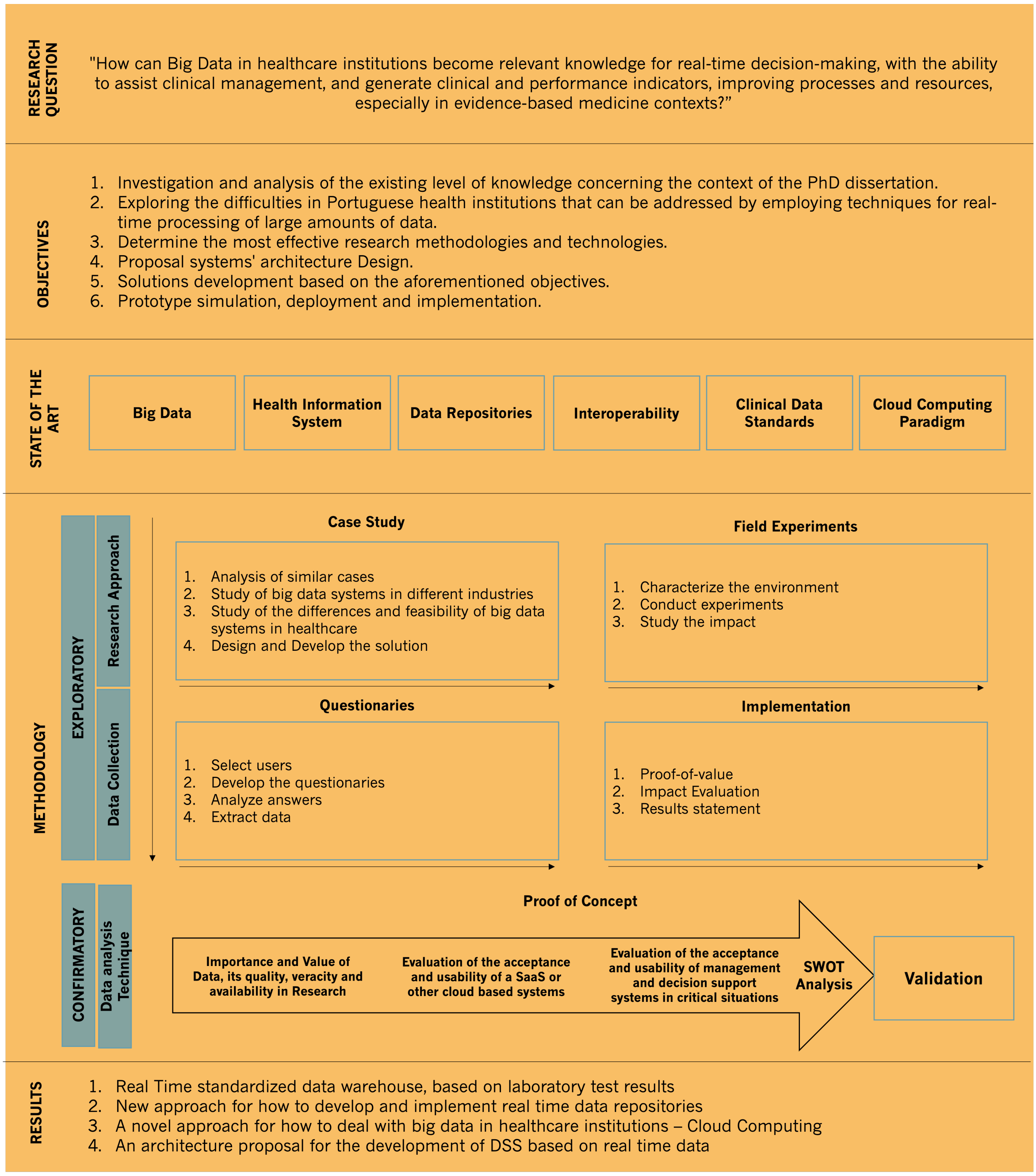


Figure 2: Illustration of the SWOT analysis matrix. Adapted from [47].

1.5 Research Process Overview

Figure 1.5 represents the entire research process through one illustration. It includes the core research question followed by the primary objectives. Next, the main keywords searched during the literature review research process and the various research strategies employed are addressed. Finally, the expected outcomes of this doctoral dissertation are mentioned.



RESEARCH QUESTION

"How can Big Data in healthcare institutions become relevant knowledge for real-time decision-making, with the ability to assist clinical management, and generate clinical and performance indicators, improving processes and resources, especially in evidence-based medicine contexts?"

OBJECTIVES

1. Investigation and analysis of the existing level of knowledge concerning the context of the PhD dissertation.
2. Exploring the difficulties in Portuguese health institutions that can be addressed by employing techniques for real-time processing of large amounts of data.
3. Determine the most effective research methodologies and technologies.
4. Proposal systems' architecture Design.
5. Solutions development based on the aforementioned objectives.
6. Prototype simulation, deployment and implementation.

STATE OF THE ART

Big Data | Health Information System | Data Repositories | Interoperability | Clinical Data Standards | Cloud Computing Paradigm

METHODOLOGY

EXPLORATORY

Research Approach

Case Study

1. Analysis of similar cases
2. Study of big data systems in different industries
3. Study of the differences and feasibility of big data systems in healthcare
4. Design and Develop the solution

Field Experiments

1. Characterize the environment
2. Conduct experiments
3. Study the impact

Data Collection

Questionaries

1. Select users
2. Develop the questionaries
3. Analyze answers
4. Extract data

Implementation

1. Proof-of-value
2. Impact Evaluation
3. Results statement

CONFIRMATORY

Data analysis Technique

Proof of Concept

Importance and Value of Data, its quality, veracity and availability in Research

Evaluation of the acceptance and usability of a SaaS or other cloud based systems

Evaluation of the acceptance and usability of management and decision support systems in critical situations

SWOT Analysis

Validation

RESULTS

1. Real Time standardized data warehouse, based on laboratory test results
2. New approach for how to develop and implement real time data repositories
3. A novel approach for how to deal with big data in healthcare institutions – Cloud Computing
4. An architecture proposal for the development of DSS based on real time data

Figure 3: Research Process Summary.

1.6 Document Structure

This manuscript is divided into four parts, subdivided into ten chapters. Part I has the introduction. The introduction chapter (Chapter 1) includes the scope and motivation for the research, the research problem being addressed, the research objectives of the study and the methodology used. To summarise, a research process overview is provided. Additionally, the structure of the document is outlined in this chapter.

Part II aims to contextualize the readers about the background and concepts discussed throughout the document. The second chapter will provide an overview of the concept of Big Data, its characteristics, the types of data it encompasses, the challenges it presents, the ethical considerations that must be considered when working with Big Data, and some case studies.

Chapter 3 will approach the concept of Data Repositories, the history and concept, the types of repositories that are available, the challenges they present

The next chapter (Chapter 4) will provide an overview of the Interoperability and Clinical Data Standards concept. Beginning with the concept and importance, it addresses the types of data standards. Then the challenges are presented, ending with some case studies.

Chapter 5 will provide an overview of the concept of the Cloud Computing Paradigm, the history and concept, implementation and service models, the existing CSP, the issue of data security in the cloud, and some practical examples.

Furthermore, a third part (Part III) containing three chapters presents the case studies. Chapter (Chapter 6) describes an integrated real-time Big Data repository based on laboratory test results. Chapter 7 describes a novel approach to developing software as a service for test results, and Chapter 8 the development of a study of the impact of covid-19 on the population's desire to stay at home.

Part IV of the document presents the final considerations. Chapter 9 contains a detailed discussion and summary of the projects and scientific contributions. In addition, Chapter 10 provides some conclusions and stages that follow the work presented.

Part II

Background and Concepts

2

Big Data

The purpose of this chapter is to define Big Data and its characteristics. Section 2.1 begins with a brief history of Big Data. Then, Section 2.2 presents the definition for the Big Data Characteristics. Section 2.3 covers the data types that can be processed. Next, the technologies normally associated with Big Data are briefly described (Section 2.4). Finally, it concludes with a discussion of the challenges (Section 2.5) as well as case studies in the healthcare industry (Section 2.6).

2.1 History and Concept

When considering the concept of Big Data, it is evident that although the word did not appear until 2005, the usage of big amounts of data and the necessity to comprehend and manage it dates back decades, if not centuries. John Graunt conducted the first reported statistical data analysis experiment in 1663. By collecting data on mortality, he theorized that it might be feasible to create an early warning system for the bubonic plague, ravaging Europe at the time [9].

Since then, several technological improvements have evolved, including internet, Business Intelligence (BI), data centers, and, in 1989, Big Data. Erik Larson penned an article for Harper's Magazine describing data usage for uninvited advertising. In the 20th century, in 1937, the government of Franklin D. Roosevelt in the United States of America initiated the first major data processing initiative [9, 80].

Six years later, during World War II, the first data processing computer arrived with the primary objective of deciphering Nazi codes. At the time, the government was required to assess the contributions of 26 million Americans. International Business Machines Corporation (IBM) was awarded the contract to develop the perforated card reading system for this massive accounting undertaking. The National Security Agency was created on November 4, 1952. During the Cold War, it faced tremendous information overload due to the automatic collecting and processing of intelligence signals. This technology, capable of deciphering 5,000 characters per second, identified patterns in intercepted messages [85, 112].

Already in the 1990s, the proliferation of Internet-connected gadgets spurred data production. Roger Mougalas of O'Reilly Media coined the term "Big Data" in 2005. As new social networks emerge, more data are generated every day. Larger corporations have developed in recent years than startups that engage in Big Data research and comprehension, according to [115, 85].

Regarding the boundary between what is and is not deemed Big Data, the concept of Big Data remains a relative word. Big Data is a vastly different notion and scale for a corporation like Google than for a medium-sized business [140].

The term 'Big Data' refers not only to the enormous volume of data, which can be measured in petabytes or exabytes, but also to its velocity (pace of data gathering) and its variety due to the numerous types of data, including structured, unstructured, and semi-structured data. However, there are multiple definitions of Big Data. The most widely recognized explanation was provided by Douglas Laney, who noted that Big Data increased in three distinct dimensions: volume, velocity, and variety (3 Vs) [78]. These three characteristics compose the mainstream model for 'Big Data'. Other authors, however, have combined these traits and added several more V's, including Value, Veracity, Visualization, Viscosity, and Virality. These characteristics will be meticulously described in Section 2.2.

2.2 Characteristics

Big data has become increasingly important today as organizations seek to extract value from the vast amounts produced and consumed data. In order to effectively utilize data analytics and incorporate them into product and process development, organizations need to understand and analyze relevant data flows and bring them closer to the core business [114, 161].

Big data has the potential to have a substantial effect on the healthcare industry. Healthcare organisations can acquire significant insights into the patterns and trends that influence patient outcomes by analyzing huge amounts of data from many sources, such as EHR, wearable devices, and population-level data. These insights can be used to enhance patient care, cut healthcare costs, and stimulate innovation in the healthcare business [145, 114].

To explain this new field of Information Systems (IS) in a more consensus way, the characteristics linked with the term Big Data arose. According to Gandomi and Haider, the most significant attribute of Big Data is its volume [35]. However, the definition now requires additional qualities. 3V's model was introduced in 2001, as illustrated in Figure 4 [67, 78]. As time passed, further features appeared. Initially, value and veracity emerged. Variability and complexity emerged subsequently, although they are not as remarkable according to the literature [35]. In addition to these qualities, three others emerged: ambiguity, viscosity, and virality [74]. Figure 5 summarises all these characteristics identified in the research.

These first characteristics are straightforward to define. The amount of data that is created and stored is denoted by volume. Variety refers to the various types of data that are generated, including semi-structured, unstructured, and structured data. Velocity refers to the rate at which data is generated and must be processed [140].

Variability is associated with the varying rates at which data flows and is hence strongly related to

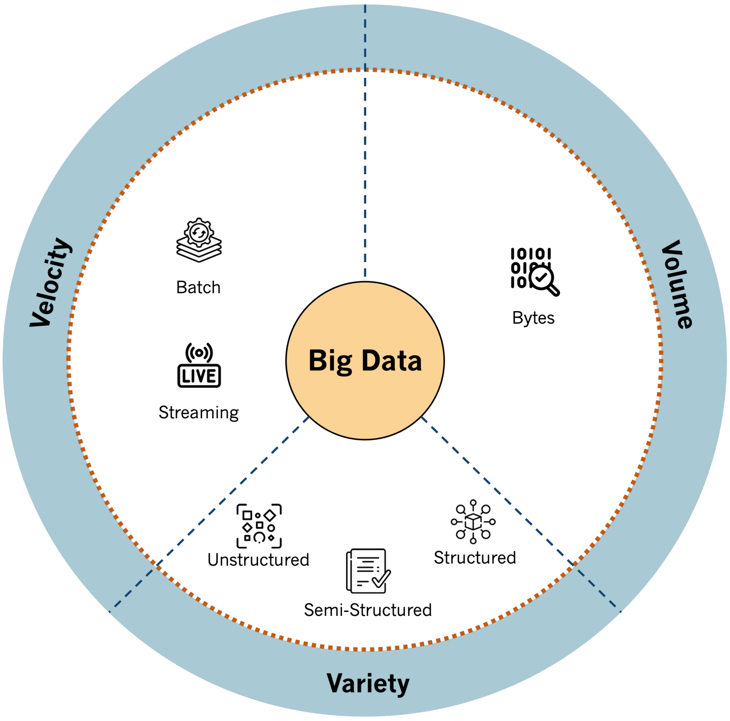


Figure 4: The 3 Vs as main Big Data Characteristics. Adapted from [161].

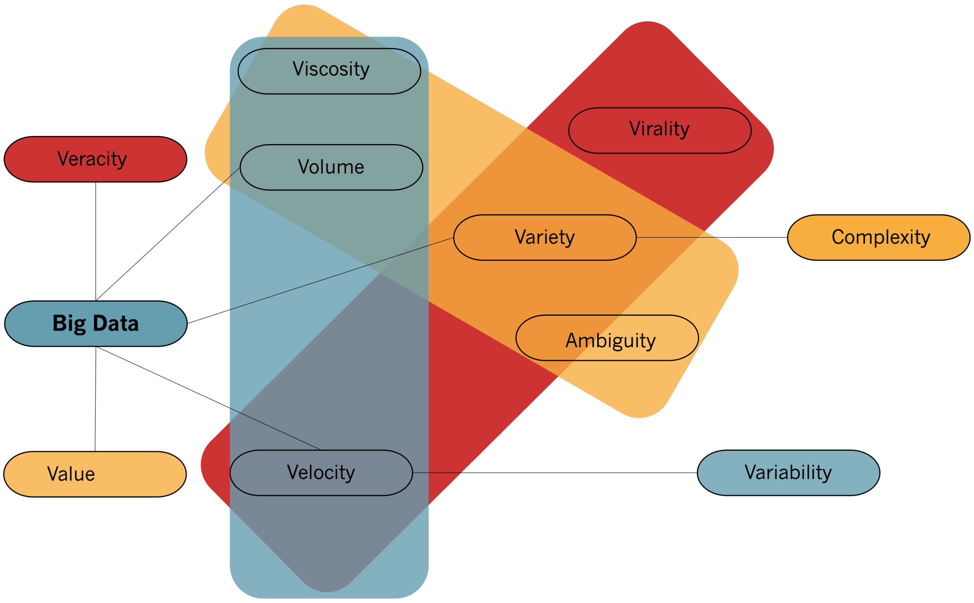


Figure 5: The 9 Vs as main Big Data Characteristics. Adapted from [22].

Velocity. The complexity of the others increases due to their dependence on other characteristics. Complexity emphasizes the difficulty of dealing with many data sources, for example, connecting, combining, cleansing, and transforming them, which depends on their forms and, therefore Variety. The absence of suitable metadata and the combination of volume and variety causes ambiguity. Viscosity is produced when the combination of volume and data velocity generates resistance in data flows and Virality, which measures the time it takes for data to spread between peers in a network [140].

The V's of Big Data are crucial considerations for companies seeking to effectively handle and analyze massive, complex data collections. At this point, attempting to quantify any of these traits appears difficult. Big data remains an abstract notion, and it must be acknowledged that it can comprise several features [162]. If approaches and technologies are insufficient to deal with the concept, it should be acknowledged as information that prompts a shift in our thinking about them.

2.3 Data Types

Big Data encompasses a wide range of data types, including structured data wich implies constant fields across all entries, that can be organized, for instance, in database tables, and can be easily processed using traditional tools; unstructured data, which is not organized in a predetermined manner and cannot be easily processed and stored using traditional tools; and semi-structured data, which is partially structured and partially unstructured [140]. Table 2 presents the distinguishing characteristics of the three data types.

Table 2: Differences between structured, semi-structured and unstructured data. Adapted from [87]

Characteristics	**Types**		
	Structured	Semi-Structured	Unstructured
Schema	Well Defined	Not Required	Not Defined
Sctructure	Regular	Unregular	Unregular
Data Structure	Independent	Embedded	Source-Dependent
Adaptability	No	Yes	Yes

2.4 Big Data Technologies

There are several technologies related to Big Data. The most common one is Apache Hadoop. This section presents the Hadoop ecosystem and other technologies for Big Data analytics. Hadoop is an open-source Apache project based on Hadoop Distributed File System (HDFS) and MapReduce [8].

Some advantages of Hadoop are:

- Higher speed and agility;

- Reduced administrative complexity;
- Integration with other services in the cloud;
- Improved availability and disaster recovery;
- Flexible capacity

In early versions of Hadoop, there were two major components: the HDFS, where files are distributed and replicated across nodes enabling the system for fault tolerance and availability, and a distributed processing framework called Hadoop MapReduce, where data present in the HDFS is processed on a *divide-and-conquer* basis [16, 137].

Over the years, Hadoop has evolved considerably, including transitioning from MapReduce to Yet Another Resource Negotiator (YARN) [52]. YARN takes a new approach to the JobTracker and TaskTracker components, replacing them with a ResourceManager, a NodeManager, and an ApplicationMaster, to solve some problems present in previous versions, such as scalability on large clusters or support for alternative programming paradigms [130].

Other related projects were emerging. Some of them and their main characteristics are illustrated in Figure 6.

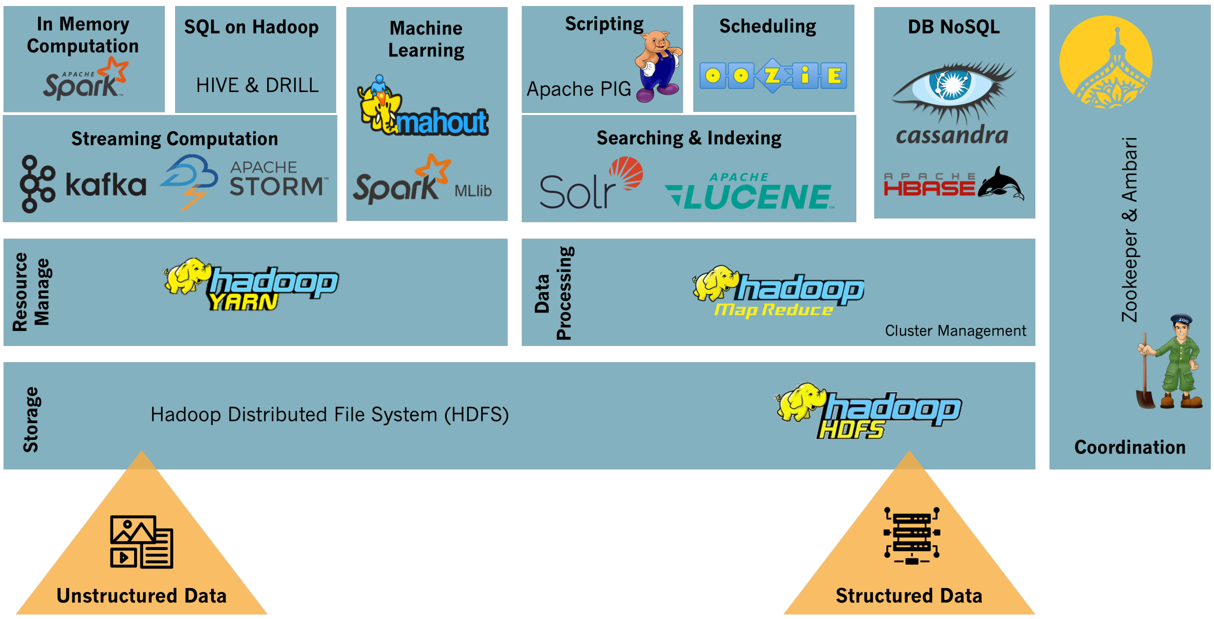


Figure 6: Hadoop Ecosystem. Adapted from [76].

Next, the Hadoop technologies are described according to [76, 137, 52].

1. Core Components

   a) HDFS – Distributed Storage System

   b) YARN – Distributed Resource Management Layer
   c) MapReduce – Distributed Data Processing Component

2. High Level Data Processing Components
   a) Pig – Top level data processing engine
   b) Hive – Data Warehousing on top of Hadoop, interface to query data.
3. NoSQL Components
   a) HBase – Column Oriented NoSQL database
4. Data Analysis Components
   a) Drill – Schema free Structured Query Language (SQL) query engine
   b) Hama – Framework for Big Data analysis
   c) Crunch – Framework to write, test and run MapReduce pipelines
   d) Mahout – Scalable Machine Learning (ML) library
   e) Lucene – High performance text search engine
5. Data Serialization Components
   a) Avro – Data serialization framework
   b) Thrift – Interface definition language and binary communication protocol
6. Data Transfer Components
   a) Sqoop – Tool designed for efficiently transferring bulk data between Hadoop and Relational Database Management System (RDBMS);
   b) Chukwa – Data collection system for monitoring large distributed systems
   c) Flume – Data Collection and aggregation system;
7. Management Components
   a) Oozie – Server-based workflow scheduling system;
   b) HCatalog – Table and storage management layer. Interface between Hive, Pig and MapReduce;
8. Monitoring Components
   a) Ambari – Hadoop deployment, management and monitoring tool
   b) ZooKeeper – Highly Reliable distributed coordination system
   c) Hadoop User Experience (HUE) – Open source Hadoop Web interface

2.5 Challenges

Working with Big Data presents many challenges, including the need for specialized skills and technologies, the need to handle large volumes of data, and the need to ensure the privacy and security of sensitive data. To effectively design and implement Big Data solutions, organizations must overcome these challenges and adopt techniques and technologies tailored to their specific needs [1, 6].

2.5.1 General Dilemmas

The absence of an agreement on the concept of ”Big Data” is the first of the general problems. The same applies to the definitions of the linked models and architectures. Even the complete use of Big Data remains an unexplored topic, such as its applications in research, engineering, health, finance, education, the government, retail, transportation, and telecommunications. Big Data problems include selecting the most relevant data from several sources and determining its worth. How Big Data can better reflect the population than limited data collection is a second often debated challenge. This varies by circumstance, but many authors caution against supposing that more data is always preferable.

2.5.2 Technical Issues

Technical issues include the collection, integration, cleansing, transformation, storage, processing, analysis, and management of Big Data. Working with Big Data involves several challenges, such as the need for specialized infrastructure and tools to process and analyze the data, professional staff to manage and analyze the data, and the need to protect data security and privacy.

Data Storage : Large amounts of data can be difficult to store, particularly in the healthcare industry, where data is frequently sensitive and regulated. Healthcare businesses must ensure the necessary infrastructure and systems are in place to securely store and handle data. Therefore, it is necessary to reconsider both storage devices and architectures.

Data Redundancy: Data redundancy is a crucial tool to ensure the reliability and availability of important data. More so with systems that rely on large amounts of data and synchronization between them. Therefore multiple redundancies must be ensured. Ideally, online backups should be implemented in addition to redundancy. Still, in Big Data systems, implementing redundancy can be a problem because a greater need for storage also means a significant cost increase.

Data Management : Scalability becomes essential for data storage and analysis. Growing data volumes demand redesigning databases and algorithms to extract their value. Distributed/parallel computing, especially when hosted in the cloud, is required to manage Big Data in order to ensure availability, cost-effectiveness, and flexibility.

Data Quality : Ensure the quality and accuracy of data, particularly in the healthcare industry, in order to make informed decisions.The higher the number of data sources, the higher the risk of quality concerns. Multiple sources of heterogeneity exacerbate these difficulties, as typical data analysis tools assume homogenous data. Heterogeneity has ramifications for data integration and repercussions for Big Data analytics since the unstructured nature of data sources provides a number of issues in terms of transformations required to allow the execution of relevant analytic activities.

Data Visualization : Visualizing Big Data requires rethinking old ways due to the volume of data. Therefore, it is essential to combine form and function. Visualization in the healthcare industry also demands particular skills and knowledge. Healthcare institutions must ensure they have the professionals necessary to evaluate and interpret data efficiently but even more so that the information is easily accessible for consultation.

Data Security : Data security is a concern for any project that relies on data. However, especially when Big Data specific tools are used some, special care is needed. There are several security projects that provide for the use of Hadoop, for example. Kerberos, Apache Knox, and Apache Ranger are three projects that aim to implement the five pillars of security: administration, authentication, authorization, auditing, and data protection.

2.5.3 Ethical Considerations

The use of Big Data raises a number of ethical considerations, including the need to safeguard the privacy of people whose data are being gathered and processed, the potential for bias in the data analysis, and the need to ensure that the data is being used ethically and responsibly. Organizations and individuals working with Big Data must take these ethical considerations into account in order to ensure that the data is used in a way that is fair and just.

Privacy and security/Ethical Considerations: Protecting the privacy and security of patient data is a major concern in the healthcare sector. Healthcare organizations need to ensure that they have the necessary safeguards to protect against unauthorized access or data misuse.

2.6 Practical Examples

There are many examples of how Big Data is being used worldwide in healthcare to improve patient outcomes and reduce costs. Next are presented a few examples of practical implementations.

In 2013, Dr. Chow Wai Leng and her team identified that public hospitals faced a recurring challenge : Patient wait times in the emergency department. After analyzing the trends, the team adjusted personnel to better fit patient arrival patterns [147]. After it, The Singapore Ministry of Health used Big Data to improve the efficiency of its healthcare system. By analyzing data from electronic health records, the ministry identified bottlenecks and inefficiencies in the system. This allowed them to make changes that reduced wait times and improved patient satisfaction [119].

In 2018, The United Kingdom (UK) National Health Service (NHS) employed Big Data techniques to reduce the time it takes to diagnose and treat patients with breast cancer. By analyzing data from mammograms, electronic health records, and other sources, the NHS could identify patterns indicating breast cancer's presence. This allowed physicians to initiate treatment earlier, which can enhance patient outcomes [81, 57]. With the establishment of NHS Artificial Intelligence (AI) Laboratory, the institution claimed in 2020, to have been at the vanguard of the AI revolution. Consequently, NHS patients will be among the first in the world to benefit from new AItechnologies, owning to the AI in Health and Care Award and a 50 million investment increase [149].

In 2021, Mayo Clinic launched two new companies to use patient data and AI to improve the accuracy of diagnoses and treatment plans for patients with rare and complex diseases. By analyzing EHR, laboratory results, and other data sources, the clinic was able to identify patterns and correlations that helped doctors make more informed decisions [77]. In Denmark, the same approach was taken by the NHS [29].

The Centers for Medicare and Medicaid Services (CMS) used Big Data to identify and reduce fraudulent billing practices in the United States. By analyzing claims data, CMS was able to identify patterns that indicated potential fraud and take action to prevent it [121].

A review of publications on Industry 4.0, Big Data, healthcare operations and future perspectives of these keywords provided by [70] suggests that wearable devices, the Internet of Things (IoT), and cloud-based technologies will serve as the foundation of future personalized healthcare. It also anticipates that as sensors and smart devices improve quality, functionality, and energy efficiency, the way we approach individualized healthcare will shift, resulting in an explosion of data. Undoubtedly Big Data, defined as the processing of diverse data in format, source, and qualities, will be a future buzzword in the healthcare industry.

3

Data Repositories

The following chapter will provide a comprehensive description of data repositories. Initially, the history of the term is outlined (Section 3.1). Yet, the various types of data repositories that can be adapted to the universe of Big Data are described in Section 3.2. In addition, some of the difficulties connected with developing data repositories are highlighted in Section 3.3. Finally, some case studies illustrating their application are presented in Section 3.4.

3.1 History and Concept

The term ”Data Repository” refers to a generic infrastructure that stores segmented data for analysis or reporting. These infrastructures are currently highly appreciated because they enable businesses to make decisions supported by information that is typically more reliable than gut feelings [151]. An appropriate data repository increases the speed of data access and sharing, as well as the preservation and archiving of sensitive data. This challenge increases when we are facing large amounts of data. In this case, we refer to our repositories as Big Data Repository. A Big Data Repository is designed for the same purpose but has the particularity to store and manage large and complex data. Therefore the design process needs to handle the main characteristics of Big Data: volume, velocity, and variety. These data can be structured, unstructured, or semi-structured and can be generated from a wide range of sources, including social media, websites, sensors, and other devices, at an unpredictable velocity. Big Data Repositories provide the necessary infrastructure and tools to process and analyze these datasets in order to extract valuable insights and knowledge.

In healthcare, data repositories are applied for a variety of purposes. They may be used to store and retain patient records, which healthcare professionals may subsequently use to guide treatment decisions and improve patient care. Eventually, it may also be used to store and manage research study data, which researchers can then use to aid their job. Finally, it may be used to store and manage administrative data, such as billing and scheduling information, hence speeding and enhancing administrative tasks, or even store and manage public health data, such as data on disease outbreaks, which can impact public health policies and activities.

One of the first data repositories in the healthcare field was the National Library of Medicine (NLM) MEDLINE database, which was created in the 1960s. MEDLINE is a bibliographic database that contains abstracts and citations for biomedical literature from around the world, including articles from scientific journals, proceedings, and other types of literature. It is widely used by healthcare professionals, researchers, and students to access information related to various aspects of medicine and healthcare. Later, the NLM created the National Cancer Institute's Cancer Data Standards Repository, which is used to store and manage data related to cancer research, the Genetic Testing Registry, which is used to store and manage information about genetic tests and the ClinicalTrials.gov database, which is used to store and manage information about clinical trials. Since 2000s, EHR has become more widely adopted, leading to the development of large, centralized data repositories used to store and manage patient records.

There are inherent preoccupations that must be addressed, such as the repository's potential behaviour as the data expands or the possibility of a system failure necessitating more frequent backups [7]. Therefore, these are legitimate challenges, but the data repository management team can be aware of and plan for them. More about this topic is detailed in Section 3.3.

There are several types of Big Data repositories, including data lakes, data warehouses, among others. Each type of Big Data repository has its own strengths and limitations, and the appropriate choice for a given organization will depend on its specific needs and requirements and will be described in Section 3.2.

3.2 Types of Data Repositories

Nowadays, institutions have several alternatives for storing the information they produce, gather, and use. Options for data repositories comprise relational databases, data lakes, data warehouses, data marts, NoSQL databases, among others. These repositories differ in terms of their architecture, data model, and use cases. Each repository type has its own strengths and weaknesses and is suited for a variety of data types and use cases.

3.2.1 Relational Database

Edgar Frank Codd was an Englishman who studied mathematics and chemistry at Oxford and then served as a Royal Air Force pilot in World War II. In 1970 he invented the relational database, which is nothing more than simply a set of relationships used to store and handle the data in the database. A connection consists just of a table of values. Each connection has columns and rows called attributes and tuples, respectively [142, 51]. To better specify and conceptualize the relational model, E.F Codd published in 1985, 12 rules that any relational model must follow [19, 142]:

Rule 0 - Foundation: Any RDBMS must be able to manage the data it stores, in its entirety, through its relational capabilities.

Rule 1 - Information: All data must be stored in table format only (columns and rows). Each row represents a single fact. Each column describes a single property of an object. Each value is defined by the intersection of a column and a row.

Rule 2 - Guaranteed Access: Access to the data must be ensured by the table name, the column name, and the primary key value.

Rule 3 - Null Value Support: Regardless of the domain of the column, null values are acceptable as a way to express information that is not available or inapplicable. The null value does not mean the absence of value.

Rule 4 - Dynamic Relation: In RDBMS, metadata are data that characterize the database or the data itself. Therefore, there are tables defined by the system and to which only authorized users to have access where all the metadata of the system is detailed.

Rule 5 - Comprehensive Data Sub-language: There must be a single language capable of defining integrity resets, views, and other operations. Although SQL is not the only data query language, it is the most common.

Rule 6 - Updating Views: Views must reflect possible updates of the source tables. The same applies to the reverse operation. So these help in the implementation of data abstraction and access control.

Rule 7 - Set level insertion, update and deletion: RDBMS must have the ability through an instruction to handle operations such as inserting, selecting, and updating data.

Rule 8 - Physical Data Independence: Physical operations must be separated from user operations with a physical storage layer. Changes in physical storage or access techniques must not cause logical damage to applications.

Rule 9 - Logical Data Independence: Users and applications are, to some extent, independent of the logical structure of a database. The logical structure can be changed without redeveloping the database and/or the application.

Rule 10 - Integrity Independence: RDBMS must implement data integrity internally. This is not the application's purpose. Data integrity ensures the consistency and precision of database data.

Rule 11 - Distribution Independence: The data manipulation language must be able to work with centralized and distributed databases. Views, for example, must be able to join data from tables on different servers.

Rule 12 - Non Subversion: Any row in any table must comply with the imposed security and integrity requirements. There are no unique privileges.

Consequently, the relational database model has a relatively rigorous schema om, meaning that a schema must be established in advance before the data is loaded and that all attributes in the schema are uniform for all components; in the case of missing values, null values are used. Changing the schema of databases is complex, especially when the database is a partitioned relational database that is distributed across numerous servers. If our requirements for data capture and administration are continually changing, a strict schema might soon become an impediment to change. The majority of relational databases use SQL to view and edit the database's contents. It is segmented into pieces such as clauses, predicates, queries, and statements and is originally based on relational calculus and relational algebra [48].

Four crucial properties define relational database transactions: atomicity, consistency, isolation, and durability - typically referred to as Atomicity, Consistency, Isolation, Durability (ACID) [51]:

Atomicity A transaction should have all its operations executed in case of success. In case of failure, no results of any operation are reflected in the database.

Consistency The execution of a transaction must take the database from one consistent state to another consistent state, A transaction must respect data integrity rules (such as key uniqueness, logical integrity constraints, etc.).

Isolation Isolation is a collection of strategies that prohibit concurrent transactions from interacting with one another as if they were done sequentially (one after the other). The primary purpose of concurrency control is isolation.

Durability The outcome of a successful transaction execution must persist in the database despite power outages, deadlocks, or other faults. They must ensure the data is ultimately accessible.

A relational model has as its main goals to provide a high degree of independence between data, simplify administration and maintenance work, allow the expansion of multiple-oriented data manipulation languages and eliminate redundancies, inconsistencies and errors [51].

3.2.2 Non Relational Database

Non-relational databases are a class of database management systems that differ significantly from relational systems in numerous aspects. Notably, neither relationships (tables) nor SQL are used as the storage structure or query language. Join operations cannot be performed, and ACID characteristics cannot be guaranteed. In addition, it is horizontally scalable. This means that when the volume of data, the number of users, or the number of requests increases, additional machines can be added to the system to manage the increased demand. [48, 98].

Non-relational databases are designed to manage vast volumes of data that are typically unstructured and do not conform to the usual tabular relationships utilized by SQL databases. With the development of NoSQL, cloud computing can now manage massive amounts of data and analyze it efficiently [15].

They are often used for Big Data and real-time web applications. Some popular NoSQL databases, Hadoop and Cassandra, are specifically designed for handling very large amounts of data and are used by many organizations as the primary repository for their Big Data. Other NoSQL databases, such as MongoDB and Redis, can also be used for Big Data applications, although they may not be as well-suited for extremely large datasets. NoSQL databases are a suitable option for Big Data applications that demand horizontal scalability and may benefit from the flexibility of many data models [4, 124].

NoSQL could be categorized in 4 types:

Key-Value database – Employ standalone tables to store data as basic identifiers (keys) and their associated values. These tables are also known as Hash Tables. The values might be of various types, ranging from simple text strings to the most complicated arrays or objects. However, data inquiries and modifications are only possible with the identifying keys. These databases are exceptionally well-sized, which makes the space occupied cost-effective and eliminates empty columns. As a result, they are extremely partitionable and permit horizontal scaling. The most prevalent are Redis, Riak, Amazon Dynamo DB, etc [10, 117].

Document-Oriented database – As an alternative to storing data in rows and columns, document databases, as their name suggests, utilize documents. They are the most prevalent alternative to relational and tabular databases. Documents are discrete records that store information about an object and its metadata in pairs of field values. The kinds and structures of values include strings, numbers, dates, arrays, and objects. Documents can be stored in JavaScript Object Notation (JSON), BSON, and Extensible Markup Language (XML) formats [10, 90]. The most popular is, unquestionably, MongoDB.

Column Oriented database – A relational database is built to store rows of data, but a columnar database is optimized for retrieving columns quickly. These databases are suited for analytical querying since column-based querying considerably reduces disk I/O requirements and the amount of data required to be loaded from the disk. Consequently, they are highly available, scalable, and designed for a distributed environment. The most prevalent include Apache Cassandra and HBase [90].

Graph database – Graph databases are specifically designed to store and navigate nodes and relationships. Relationships are unique elements that contribute the majority of value to graph databases. Nodes are used to store data entities, whereas edges are used to store the relationships between those entities. There is no restriction on the number or types of relationships a node can have. The structure of graphs permits data to be stored once and interpreted differently based on its relationships. The most prevalent are OrientDB and NEO4j. etc[122].

Other models, such as Column Oriented Databases, Object Oriented Databases, XML Databases, Multidimensional Databases, Multivalue Databases, and Multimodel Databases, are also possible. The

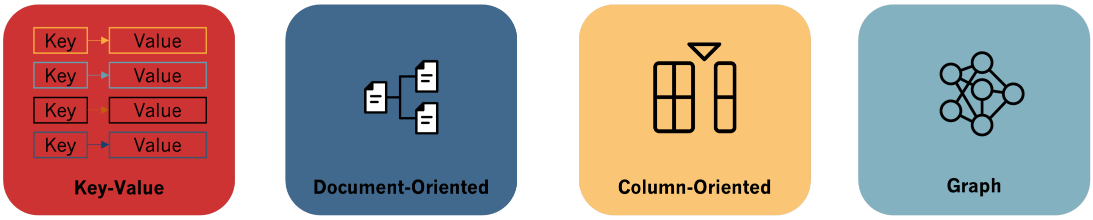


Figure 7: Most Common Types of NoSQL Databases.

optimal method for choosing between No Relational and Relational databases is to examine the data that will be stored. Consider a few factors, such as: Hierarchical structure will be used? How many users will access the database? How do you assess the scalability? What is the purpose of the database (BI, Historical Data,ML, Real-Time Data)? The most common types are illustrated in Figure 7.

In conclusion, the No Relational approach provides numerous advantages. But it is not the reason we should apply it in every circumstance. A relational model may and should be applied in numerous systems. The primary distinctions between relational and non-relational databases are outlined in the table 3.

Table 3: Comparison between Relational and Non-Relational Databases

Feature	Relational Database	Non Relational Database
Availability	Good	Good
Consistency	Good	Low
Data Storage	Medium	Good(for Big Data)
Performance	Low	Good
Reliability	Good	Low
Scalibility	Good (Expensive)	Good

Rather than these more common concepts (relational and non-relational databases), additional concepts are becoming increasingly popular due to the current interest in Big Data. Data Lake and Data Mart, to be precise. These will be briefly detailed in Sections 3.2.3, 3.2.4 and 3.2.5, respectively.

3.2.3 Data Lake

The concept of Data Lake has been around since the late 1990s. However, the term was coined by the Chief Executive Officer (CEO) of Pentaho, James Dixon, in 2010. Data lakes are large, centralized stores of raw data that can serve multiple purposes. Data lakes store a variety of data types, such as written documents, photographs, videos, and audio recordings. Data lakes are intended to hold both structured and unstructured data from various sources. Consequently, this type of repository has the ability to provide scalability and flexibility while being cost-effective. However, the data cannot be used immediately. Numerous applications, including data discovery, analytics, and machine learning, make substantial use of data lakes [94, 42]. Also, for really large datasets setting up and maintaining a data lake can be complex. A data lake can store data from many different sources and formats, which can make it difficult to enforce data governance policies and ensure the quality of the data. Also, it is important to implement robust security measures to protect this data [100].

3.2.4 Data Warehouse

Data Warehouse was defined in 1995, for Inmon as *"a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."* [60]. So it is interpreted that a data warehouse is a centralized repository, designed for fast querying and analysis, that stores large volumes of data from multiple sources in order to organize, analyze and report on it or even support decision-making activities [86]. In short, this structure stores structured data, mostly used by data visualization or BI reporting platforms, since the data present here is ready to be used. Typically the data comes initially from relational databases and internal or external systems or platforms. The most typical data warehouse architecture is the Star Schema, devised by Ralph Kimball in 1996. It consists of a central table called the fact table surrounded by several dimension tables [39, 99].

3.2.5 Data Mart

Data marts are a more specific version of data warehouses. It often focuses on a certain topic area or industry. For this reason, it is, only exceptionally well-structured and more detailed data is selected for this structure. They are meant to enable corporate intelligence and analytical applications and to give a quick query performance. Despite the fact that they are frequently "subsets" of a data warehouse, they can be independent and populated with data from a range of sources, including transactional systems, operational databases, and other data warehouses. They are typically developed to serve a particular business activity or department inside a major firm, such as sales, marketing, or finance [99].

Figure 8 highlights the significant architecture of data lakes and data warehouses.

3.3 Challenges

There are several challenges associated with the building and use of Big Data Repositories, including the need for specialized infrastructure and tools, the need for skilled personnel to manage and analyze the data, and the need to ensure the security and privacy of the data [66].

Managing Big Data often requires specialized hardware and software, such as distributed storage systems like Hadoop and HDFS which can scale to store very large amounts of data and parallel processing frameworks like MapReduce and Spark, allow data to be processed in parallel across a cluster of machines, which can significantly improve processing speed. This type of environment can be expensive to acquire and maintain. After being stored and processed, analyzing and deriving insights from Big Data requires

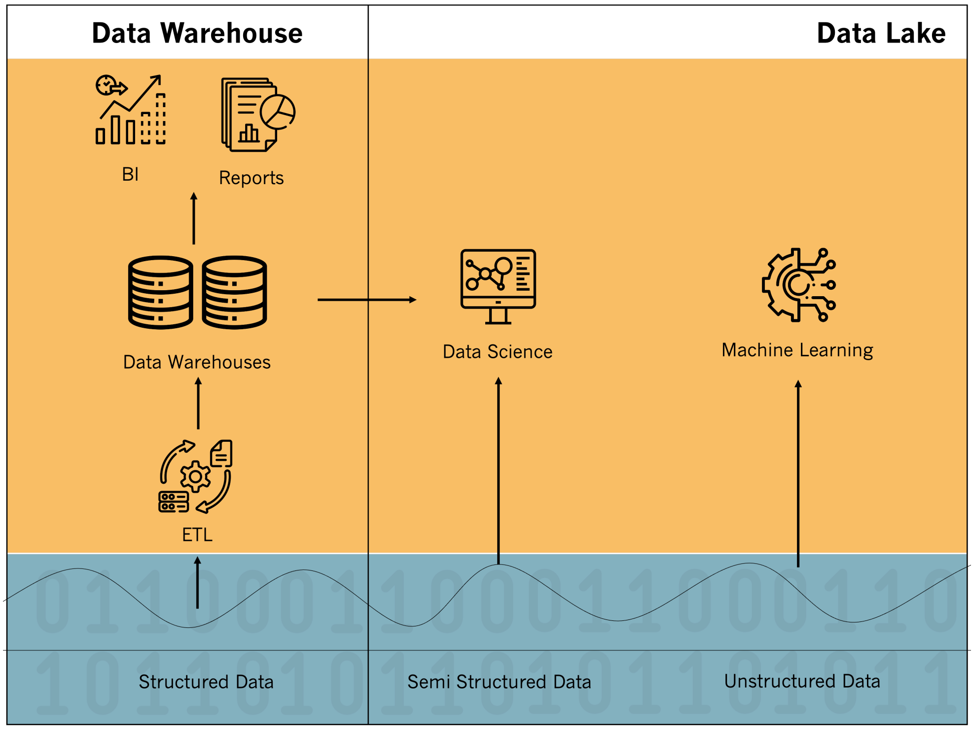


Figure 8: Data Warehouse and Data Lake Architecture.

specialized skills, such as knowledge of programming languages like Python and SQL, and experience with Big Data analysis tools mentioned above. Data scientists, data engineers, and other professionals with these talents are in high demand. Thus it might be difficult to find and retain individuals with them [17, 134].

Big data repositories often contain sensitive and confidential data, and it is important to ensure that this data is protected from unauthorized access and breaches. This can be challenging due to the volume and complexity of the data, and the need to balance security with the ability to access and analyze the data. It is important to implement measures such as encryption, access controls, and security monitoring to protect the data in a Big Data repository [144, 146].

3.4 Practical Examples

Numerous implemented case studies highlight the use of Big Data Repositories in various sectors and fields. A healthcare institution, for instance, could utilize a Big Data Repository to store and analyze patient data in order to enhance the quality of care and minimize expenses. A retailer might use a Big Data Repository to store and analyze customer data to enhance its marketing and sales operations. These are only a handful of the numerous methods in which Big Data Repositories can be utilized to promote value and innovation. In the real world, a number of healthcare organizations are focusing on building massive data repositories [114].

The UK's NHS has implemented a national electronic health record system called NHS Spine. NHS

Spine is the largest public healthcare platform in the world and a crucial element of the UK's national infrastructure. The system contains data on every patient in the country. Spine provides secure interoperability through national services such as the Electronic Prescribing Service, Summary Care Record and Electronic Referral Service. This technology is used to increase the efficiency of care and decrease the number of errors by making patient information more accessible to medical practitioners [103, 141].

The Swedish National Patient Registry, started in 1960, is a large data repository that contains information on all hospital admissions, outpatient visits, and causes of death in Sweden. It has for example, Statistics on disease and surgical treatment of patients in Sweden have been published for over 100 years. This registry is used for a wide range of purposes, including healthcare planning, quality improvement, and research [101].

More ambitious approaches are emerging in the genomic data arena. For example, the National Cancer Institute's Genomic Data Commons (GDC). This is a central database that provides access to both genomic and clinical data from patients with cancer registries. This allows researchers in the field to analyze thousands of patients in detail so that they can gain insight into the genetic basis of cancer and how it evolves. Ideally, these studies would be extremely effective in developing treatments for the symptoms or even the disease.

4

Interoperability and Clinical Data Standards

This chapter aims to provide information to the reader on the significance of interoperability and data standards. Consequently, the purpose of Section 4.1 is to define interoperability. The types of data standards utilized in healthcare are described in Section 4.2. In the final two chapters, the challenges of applying standards (Section 4.3) and instances of applicable case studies are described (Section 4.4).

4.1 Concept and Importance

Clinical data refers to patients, their symptoms, diseases, procedures, medications, and laboratory results, among other information. This may include demographic information, medical history, laboratory test results, imaging investigations, vital signs, and treatment plans. These data are collected in several information systems (EHR, illness registries, clinical trial documentation, mortality databases), which are varied, context-dependent, frequently insufficient, and occasionally erroneous [3].

Whenever statistical analysis or case-based reimbursement is necessary, the data must be structured with a trade-off regarding breadth and granularity. Therefore, it is essential to underline that openness and shared features are essential success criteria for e-Health solutions. Knowledge is power. Therefore, by implementing interoperability, institutions may maximize their data assets' value and their potential [138].

The concept of interoperability can be defined as the ability of one system to communicate and share information with another system that arises from the heterogeneity and distribution of several different sources of information. Interoperability is a concept that can be applied in several areas; therefore, each of them will have its own approach and definition. The Dublin Core Metadata Initiative (DCMI) describes interoperability as ”the ability of different computer networks, systems, components or applications, to work effectively to exchange information in a useful, meaningful and usable way.” This definition can be completed with the interaction and exchange of information between devices [58, 138].

However, its definition of interoperability has been updated notably by HIMSS which adds that this exchange should be integrated. Data should be used cooperatively in a coordinated manner, within and across organizational, regional and national boundaries to provide seamless and timely information portability and optimize the health of individuals and populations globally [54].

According to HIMSS, there are four levels of interoperability, presented in Figure 9.

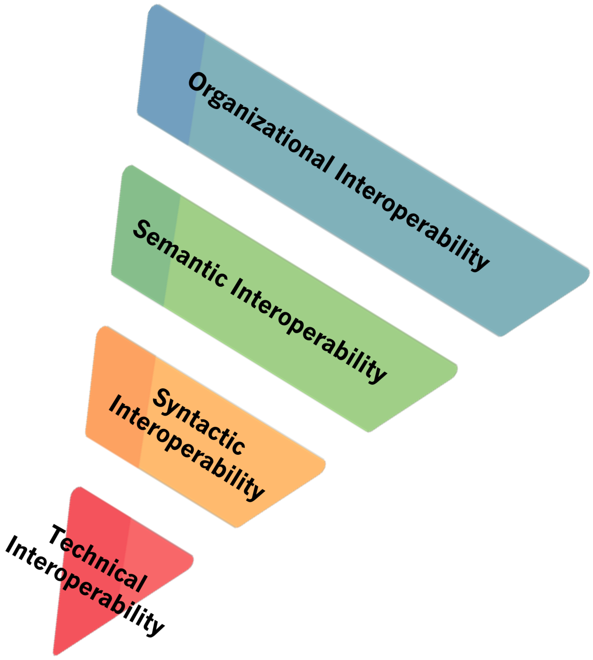

Organizational Interoperability
Semantic Interoperability
Syntactic Interoperability
Technical Interoperability

Figure 9: Interoperability Levels adopted by HIMSS. Adapted from[159].

The first layer is the Technical layer. This is based on the simple ability to exchange data and/or information between two or more systems. There are no interpretation requirements for the system receiving the data. Only the integration and compatibility requirements necessary to share and receive data are established. Thus, technical interoperability serves as the basis of the communication pyramid illustrated in the figure, offering the most basic data exchange services [160, 55].

Syntatic interoperability represents the ability of the receiving system to not only receive but also interpret information at the data field level [160, 95]. As an intermediate layer, syntatic interoperability defines the structure or format of medical information. Standards for electronic health information exchange, such as Fast Healthcare Interoperability Resources (FHIR) and HL7, allows the automatic detection and interpretation of predetermined data fields while preserving the purpose and clinical or operational meaning of the data.

Semantic interoperability resides in the second intermediate layer of the pyramid. HIMSS defines it as the ability of health IT systems to exchange, interpret information, and actively use the information exchanged. As such, data standardization and coding models are required. Semantic interoperability is key to bridging the terminology gap between divergent information systems and data sources in the healthcare sector. It allows organizations to share and interpret patient information, reducing duplicate tests, improving clinical decision-making, improving inter- and intra-hospital care coordination, reducing hospital readmissions, and ultimately saving hospitals money [13, 55].

At the top of the pyramid is organizational interoperability. While all organisations want this, it is also the most difficult to achieve. Most healthcare organizations are still working to establish technical and syntactic interoperability. According to HIMSS, organizational interoperability refers to the sharing and interpretation of patient data between multiple organizations with different goals, always taking into consideration non-technical constraints such as policy, legal, social and organizational aspects, since multiple stakeholders, organizations, and individuals are involved [55].

While current solutions and technologies can solve most of the basic and syntactic interoperability issues, semantic interoperability remains the most concerning. Making health information understandable to the computer increases the challenge of establishing a semantic level. Data standards may help in solving this problem [41]. A standard is a document that provides requirements, specifications, guidelines, or characteristics that can be used consistently to ensure that materials, products, processes, and services fit their purpose [62].

Clinical data standards are a set of rules, guidelines, or terminologies that are used to define structure, and format clinical data. They provide a common language for data exchange and interpretation, which is essential to ensure data accuracy and consistency and to enable the integration and analysis of data from different sources [31]. In the healthcare industry, they play a crucial role in improving patient care and outcomes. By enabling the exchange and analysis of clinical data, data standards can help healthcare providers make more informed decisions, identify trends and patterns, and develop new treatments. In addition, data standards can help reduce errors and improve the efficiency of healthcare processes, for example, by enabling the automation of certain tasks. Overall, using data standards is an essential part of modern healthcare and is likely to become increasingly important as the volume and complexity of healthcare data continue to grow. Overall, clinical data standards can help improve healthcare delivery quality, safety and efficiency.

There are many different types of clinical data standards, including standards for terminology, coding systems, data models, data exchange, and others. Some examples of clinical data standards include HL7, Digital Imaging and Communications in Medicine (DICOM), FHIR, Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), and others. Thus, standards are used in various settings, including hospitals, clinics, research laboratories, and other healthcare organizations. They help ensure that clinical data is collected, stored, and used consistently and reliably, essential for improving patient care and advancing medical knowledge. More specifics about the standards are given in Sections 4.2.1 to 4.2.5.

4.2 Data standards

Data standards can be divided into several macro categories. The most widely agreed upon in the literature are [127]:

Terminology Standards - These standards address a fundamental requirement for communication: the ability to represent concepts unambiguously between a sender and receiver of information. Most communication between health information systems relies on structured vocabularies, terminologies, code sets, and classification systems to represent health concepts. Some examples are International Classification of Diseases (ICD), SNOMED CT and Logical Observation Identifiers Names and Codes (LOINC) described in Section 4.2.2.

Content Standards - Content standards ensure that both the sender and receiver exchanging electronic messages or documents understand how the content is structured and/or what data sets it contains. Typically these standards require a communication standard, described below, to effect the exchange of information. Content standards can be, for example, HL7 described in Section 4.2.3.

Transport or Exchange Standards - Communication standards specify how information should flow between systems rather than how each system should organize its information internally. Therefore, these standards refer to the protocols used to exchange and transmit information between different systems and organizations. These standards should ensure that information is transmitted securely and accurately, enabling the interoperability of different healthcare systems and applications. Some of the best known standards are, for example, the FHIR and DICOM described in Sections 4.2.3 and 4.2.1, respectively.

Privacy and Security Standards - Privacy and data security standards are guidelines or rules that are used to protect the privacy and security of personal and confidential information. These standards define the steps that should be taken to protect information from unauthorized access, disclosure, or misuse and help ensure that personal and confidential information is treated responsibly and ethically.

Many other examples, not mentioned above, are illustrated in Figure 10.

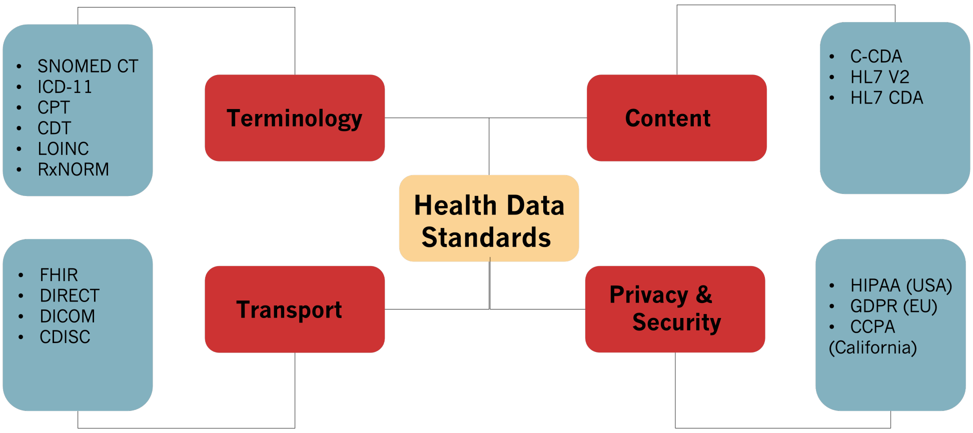


Figure 10: Summary of Existing Data Standards in healthcare. Adapted from [127].

4.2.1 DICOM and PACS

The DICOM standard began developing in the 1980s by the National Electrical Manufacturers Association (NEMA) to be a standard for storing and transmitting medical images and related information. This standard is widely used in hospitals and other healthcare organizations to manage and share medical images and data, such as X-rays, CT scans, and MRIs [49]. The first version of the DICOM standard, version 1.0, was released in 1993. Since then, the standard is continually being updated to meet the medical community's needs and to take advantage of technological advances, with the latest version released in 2021.

DICOM images can be quite large due to the high resolution and detailed nature of the images. In order to effectively manage the storage and transmission of these images, it is necessary to compress them. The DICOM standard defines a number of image compression methods that can be used, including Joint Photographic Experts Group (JPEG) and JPEG 2000. DICOM images are typically compressed and stored in a Picture Archiving and Communication System (PACS) system to manage the images' storage and transmission efficiency.

Medical images are stored electronically in a PACS system. They can be accessed from any device with the necessary permissions, eliminating the need for paper copies of the images. As illustrated in Figure 11, PACS is a medical imaging technology that uses digital images and DICOM standards to facilitate the exchange and management of images and medical data.

In short, DICOM is the standard used by PACS systems to store and transmit medical images and related data. PACS systems rely on DICOM to facilitate the exchange and management of medical images and data within a healthcare organization.

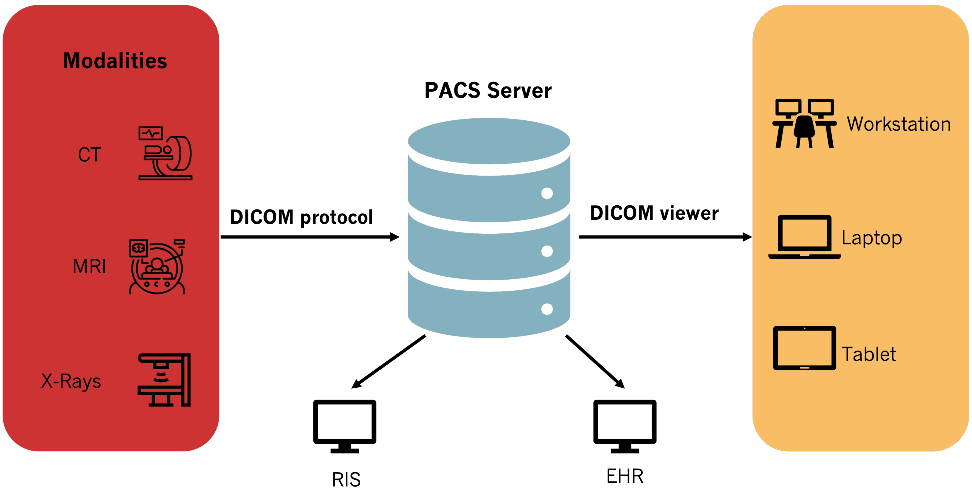


Figure 11: Arquitetura de comunicação entre PACS e DICOM. Adapted from [49].

4.2.2 ICD, LOINC and SNOMED CT

Some of the first attempts to classify diseases systematically were made between the 1600s and 1700s. Despite this advance, the resulting classifications were considered of little use, largely because of inconsistencies in nomenclature and poor statistical data [152].

In 1893 the ICD emerged as a standardized system for classifying diseases and health conditions. The main purpose of this standard is to code and classify morbidity and mortality data from death certificates and medical records. Most hospitals use these codes for billing and reimbursement purposes. The ICD is maintained by the World Health Organization (WHO) and is currently in its tenth revision - International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). The ICD-10 is used internationally and consists of 21 chapters divided into categories based on the type of disease or condition. Each category is assigned a unique code consisting of a letter followed by up to six digits. The first three elements represent a single category. The second three digits describe aetiology, anatomical location, severity, and other vital details. At the same time, the seventh character specifies an episode of care for injuries, poisoning and other conditions with external causes.

Later, in 1994, the LOINC standard was developed by the Regenstrief Institute, a non-profit organization based in Indianapolis, Indiana. The first version was released in 1996. This standard identifies laboratory and other clinical observations, such as vital signs and test results. It ensures that the same observations can be recorded and compared in different settings and healthcare systems [89].

The LOINC codes follow a semantic data model containing six major and up to four minor attributes to create fully specified names for concepts. The major attributes of the LOINC name are Component (What is being observed), a Property (How it is being measured), Time (The time interval during which an observation was made), System (The thing on which the observation was made), Scale (How the value of the observation is quantified or expressed: quantitative, ordinal, nominal), and Method (A high-level classification of how the observation was made. Only needed when the technique affects the clinical interpretation of the results). If a simple question had to be associated with each of these five mandatory points, it would be: What? (Component), How? (Property), When? (Time), Where? (System), Which? (Scale) [56].

SNOMED CT is the most comprehensive, multilingual standard for electronic health records worldwide. It encodes information about clinical concepts, such as diagnoses, procedures, and medications, in an accurate and human-readable form. It is owned, maintained and distributed by SNOMED International, an international non-profit organization in the UK, and used in over eight countries.

The SNOMED CT is composed of concepts (represents a unique clinical meaning) and descriptions (It can be a Fully Specified Name (FSN) or a Synonym. The FSN represents a unique and unambiguous description of the meaning of a concept. A synonym represents a term that can be used to display or select a concept) and relationships (represents an association between two concepts) accurately represent clinical information based on a logic organized in hierarchies for the entire health area. Its use provides interoperable coded data that improves the implementation of clinical practice by facilitating decision support systems [104, 38]. As of January 2020, the published SNOMED CT catalogue had 352,567 concepts on diagnosis and clinical findings, such as surgical, therapeutic, and other relevant information that supports clinical knowledge [61].

4.2.3 HL7 and FHIR

HL7 (Health Level Seven) was founded in 1987 by a non-profit organization (HL7 International) dedicated to developing standards for health information exchange. The organization was formed by a group of healthcare professionals, software developers, and representatives of healthcare technology companies who recognized the need for a standard way to exchange health data between different systems and organizations. The first version of the HL7 standard was released in 1988 and focused on exchanging patient admission and discharge information. Since then, HL7 has continued to evolve and expand its scope and now includes a series of standards for exchanging a wide variety of healthcare data, including clinical, administrative, and financial information [131].

In the late 1980s and early 1990s, HL7 released additional versions of its standard, including HL7v2 and HL7v3. These versions expanded the scope of the standard to include the exchange of a wider range of health data, including clinical, administrative, and financial information. HL7v2, in particular, became widely adopted and continues to be used today [95].

In the 2000s, HL7 released the Clinical Document Architecture (CDA) standard, which provided a way to structure and exchange clinical documents such as discharge summaries and reference letters. CDA was designed to be more flexible and easier to use than previous HL7 standards.

FHIR emerged in the 2010's as a modern web-based standard development for exchanging healthcare information. FHIR was designed to address some of the challenges and limitations of previous HL7 standards, which were seen as complex and difficult to implement. It was developed with a focus on simplicity, flexibility and ease of use, with a modular design that allows different parts of the standard to be used as needed. FHIR is based on a set of ”resources” representing different types of health data and a set of rules for exchanging and accessing these resources over the web [82].

HL7 has become a widely used standard in the healthcare industry, with many hospitals, clinics, and other healthcare organizations using it for electronic information exchange.

4.2.4 GDPR and HIPAA

The General Data Protection Regulation (GDPR) is a regulation in data protection and privacy legislation for all individuals in the European Union (EU) and European Economic Area (EEA). This regulation sets out a number of principles that organizations must follow when processing personal data, including the requirement to obtain the explicit consent of individuals to process their personal data, the need to inform individuals of their rights under the GDPR, and the obligation to protect personal data using appropriate security measures. It also addresses the export of personal data outside the EU and the EEA. Under the GDPR, non-compliant organizations may be subject to fines and/or other sanctions for non-compliance [148].

The Health Insurance Portability and Accountability Act (HIPAA) is a USA law that aims to protect the privacy and safety of Personal Health Information (PHI). It applies to healthcare providers, health plans, healthcare clearinghouses, and their business associates that deal with PHI. The law requires these entities to implement safeguards to protect and appropriately disclose data. It also gives individuals the right to access and request copies of their PHI and to request amendments to their data if it is inaccurate or incomplete [43].

While both address the protection of personal data, there are some differences between the two. The HIPAA is specific to the healthcare industry and applies only to health data only, whereas the GDPR applies to all personal data and to any organization that processes personal data of individuals in the EU, regardless of the location of the organization.

4.2.5 FAIR

It has been established that data sharing is crucial in areas such as healthcare when reacting to crises such as infectious disease outbreaks. While preparing data, some rules can be followed. The Findability, Accessibility, Interoperability, Reusability (FAIR) (Findable, Accessible, Interoperable, and Reusable) Data Principles were published in 2016 to facilitate the reuse of digital assets. This section discusses the principles of keeping research data in line with the FAIR principles [155, 154]. According to FAIR Principles, research data must be [64]:

- **Findable**: The information must be uniquely and persistently identifiable. Data should be readily available to other researchers;

- **Accessible**: Both people and machines must be able to comprehend the conditions under which the data may be used;

- **Interoperable**: Data should be machine-readable and interoperable through the use of industry-standard terminologies, vocabularies, and ontologies;
- **Reusable**: Data must correspond to the standards and be adequately characterized with metadata and provenance information to permit connection or integration with other data sources and correct citation.

4.3 Challenges

Implementing health data standards can be a challenging task due to a number of factors. One of the main challenges is the complexity of health data itself. Health data can have a lot of variety and come from a wide range of sources, making it difficult to develop a single set of standards that can be applied across all healthcare organizations and settings. Interoperability is another challenge because healthcare systems often use different data standards, making it difficult to exchange information. This can hinder the ability of healthcare providers to access and use critical patient information when making treatment decisions.

Another major challenge is ensuring the quality of health data. In order to make informed decisions about patient care, it is essential to ensure data quality. But, inconsistencies in collected data have a huge impact on it. In addition, implementing new data standards can be costly, as it often requires the purchase of new software and training of staff to use it. This can be a major barrier for smaller healthcare organizations that may not have the resources to make these investments.

Finally, resistance to change can also be a challenge in implementing health data standards. Some healthcare organizations may be resistant to adopting new data standards as they may require significant changes to current processes and systems. It can be difficult to overcome this resistance, but it is critical to the success of any health data standardization efforts.

4.4 Practical Examples

National health systems worldwide are increasingly adopting measures that promote interoperability safely and efficiently. In Utah and Idaho, the Intermountain health system is based on a clinical data repository that allows for the integration of data from various sources, such as EHR, laboratory systems, and radiology systems. This allows healthcare providers to access a patient's complete history, improving care and quality of care while reducing costs. Complementing this system is now the HL7 CDA standard to enable the exchange of clinical documents between healthcare providers [18].

Children's Hospital of Philadelphia (CHOP) has implemented an EHR system that allows for the integration of data from various sources, such as clinical data, administrative data, and research data. The EHR system also allows for the sharing of patient information with other healthcare providers through the use of theHL7 FHIR standard, which enables continuity of care and research [32].

Kaiser Permanente has fully implemented a system called HealthConnect that shares data across all of their facilities and makes it easier to use EHRs. A McKinsey report on Big Data healthcare states that "The integrated system has improved outcomes in cardiovascular disease and achieved an estimated $1 billion in savings from reduced office visits and lab tests." [37].

Moreover, new mandatory guidelines for the electronic submission of results of complementary diagnostic and therapeutic tests from health providers in Portugal's National Health Service were recently launched by the Serviços Partilhados do Ministério da Saúde (SPMS). FHIR Messaging version 3 has to be used as the data transfer mechanism. Among the Exchange Frameworks specified by HL7 FHIR, Messaging is an HyperText Transfer Protocol (HTTP)-based option. This technique, in brief, specifies a Bundle data structure for transporting messages with a MessageHeader data structure.

These examples illustrate how healthcare institutions are implementing interoperability and clinical data standards to improve patient care and reduce costs by allowing for the integration of data from multiple sources and enabling the sharing of patient information with other healthcare providers.

5

Cloud Computing Paradigm in Healthcare

This chapter highlights the concept of Cloud Computing when applied to the healthcare sector. Thus, after a brief history of the concept is presented (Section 5.1). Next, the Implementation Models (Section 5.2) as well as the Service Models (Section 5.3) are described. In Section 5.4, three Cloud Service Providers are present and a brief comparison between the three. Then, the data security challenge is discussed (Section 5.5). Finally, some case studies are presented in Section 5.6.

5.1 History and Concept

Cloud Computing is a novel computing paradigm made feasible by providing software, infrastructure, and all computing services for a platform as a service. In contrast to traditional web hosting, cloud computing offers paid services for a specific event, which is typically linked with a click. This means that customers/users only pay for the services they use [27].

As defined by ISO/IEC 17788, cloud computing is ”a paradigm that enables network access to a scalable and elastic set of shareable physical or virtual resources with self-service provisioning and administration on demand”. Cloud Computing can also be seen as a collection of concepts related to numerous fields of expertise, including distributed computing, grid computer and virtualization [157].

Despite the innovation associated with the term, the concept underlying cloud computing is not at all novel. In the 1960s, John McCarthy prophesied that the general people would be charged for access to computing services. However, the first recognizable cloud computing services did not emerge until the late 1990s [36].

Salesforce, which introduced its cloud-based Customer Relationship Management (CRM) service in 1999, and Amazon, which released Elastic Compute Cloud (EC2) in 2002, provided the first recognizable cloud computing services. These early cloud computing services focused primarily on delivering remote storage and computing capability to enterprises rather than individual consumers [143].

Several businesses, including Google and Microsoft, entered the cloud computing sector in the mid-2000s. These businesses focused on expanding their service offerings to include SaaS, Platform as a

Service (PaaS), and Infrastructure as a Service (IaaS). Since the beginning of the 2010s, a rising number of businesses and individuals have utilized cloud-based services for various purposes. Today, cloud computing is an integral component of the IT landscape, with many businesses relying on cloud-based services for storage, computing power, software, and infrastructure [143].

5.2 Implementation Models

When migrating to or adopting a cloud environment, there are several crucial considerations. The National Institute of Standard and Technology (NIST) presents three deployment models, described from Section 5.2 to Section 5.2, with objectives ranging from operational cost reduction to reliability [92]. A summary of the implementation models is provided in Figure 12 [44].

Private Cloud

Private clouds refer to a solution, commonly known as data centers, dedicated to a single enterprise and often operates behind a firewall. All clouds become private when the underlying computer infrastructure is committed to a single customer with entirely isolated access. The private cloud offers high performance and security independence, but there is an associated upfront investment cost. This architecture has no shared resources with other businesses, and local apps and users do not experience multi-tenancy or latency concerns [92, 44].

Public Cloud

A cloud provider typically sets up and manages the environment in the public cloud model. The absence of initial infrastructure investment and risk transfer are two of the most advantageous aspects of the public cloud for service providers. There is no control over data configurations, the network, or security, among other aspects. Nonetheless, a shared responsibility model compels enterprises that subscribe to these cloud services to ensure the security of their applications and networks, such as by analyzing packets for malware or encrypting data at rest and in transit. Amazon Web Services (AWS), Google Cloud, Microsoft Azure and IBM Cloud are among the most prominent public cloud providers [92, 44]..

Hybrid Cloud

Multiple environments connected via Local Area Network (LAN), Wide Area Network (WAN), Virtual Private Network (VPN), and/or API comprise a hybrid cloud. Hybrid clouds can use zero or more private and public clouds in numerous configurations. The properties of hybrid clouds are intricate, and their requirements can vary depending on who demands them [92, 44]..

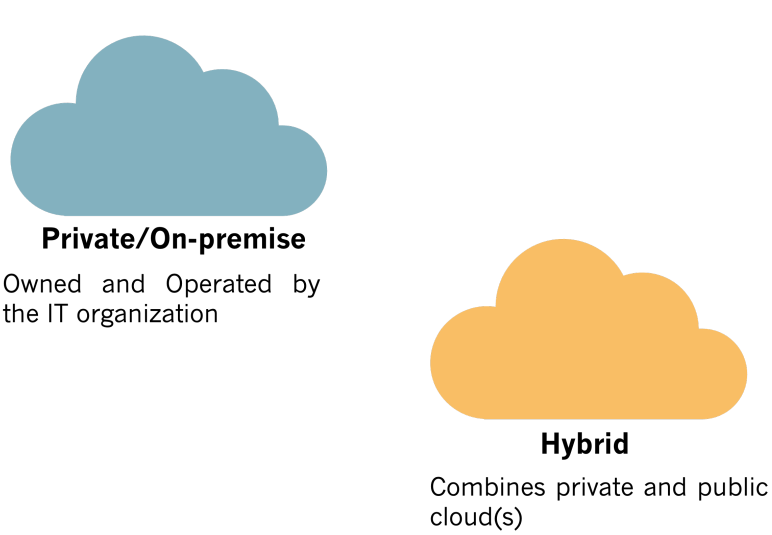

Private/On-premise
Owned and Operated by the IT organization
Public/Multicloud
Defined and Provided by the CSP, shared among multiple tenants
Hybrid
Combines private and public cloud(s)

Figure 12: Types of Implementation Models.

5.3 Service Models

According to the level of abstraction of the capabilities supplied and the service model of the providers, the model described by NIST defines three service models for services to be delivered: IaaS, PaaS, and SaaS. These models can be considered as a layered architecture, where services from a higher layer incorporate services from lower layers [72]. Depending on the user and desktop configuration, each model offers distinct functions. A comparison between what is controlled by the client and by the service provider is illustrated in Figure 13.

Infrastructure as a Service

Em IaaS o fornecedor de serviços fornece um conjunto de recursos informáticos virtualizados como CPU, Memória, SO, e Software Aplicativo, etc. na cloud. A IaaS utiliza tecnologia de virtualização para converter recursos físicos em recursos lógicos que podem ser provisionados e libertados dinamicamente pelos clientes, conforme necessário. Algumas das principais empresas que oferecem infra-estruturas como serviço incluem Rackspace Cloud Servers, Google, Amazon EC2, IBM, e Verizon [72, 83].

Platform as a Service

A service provider, in PaaS, administers and maintains system software and other computing resources. Design, development, and hosting of apps are also included. Other services include collaboration, database integration, security, online service integration, and escalation, among others. Users need not manage their own hardware and software resources or engage specialists to manage these resources. Scalability is another benefit of PaaS. This scheme allows for the flexible installation of software on the system. Suppliers' lack of interoperability and portability is a significant factor in the rise of PaaS [72, 97].

Software as a Service

The SaaS design presents to the client as a web-based application interface where services are accessed via a web browser over the Internet. In this model, service providers are responsible for operating and maintaining application software, the operating system, and additional resources. Unlike traditional software, SaaS offers the benefit of not requiring the consumer to purchase licenses, install, update, maintain, or run software on his computer [68]. In addition, it provides multitenant efficiency, configurable resources, and scalability that cannot be attained with an on-premise service [40]

Software as a Service (SaaS) is a secure choice if implemented appropriately. Numerous SaaS providers protect their clients' data via encryption, secure servers, and routine security updates. Before employing a SaaS service, it is essential to investigate and evaluate its security to ensure that it fulfils your organization's security standards. In addition, it is essential to adhere to best practices when using SaaS services, such as using secure passwords and two-factor authentication [72, 26].

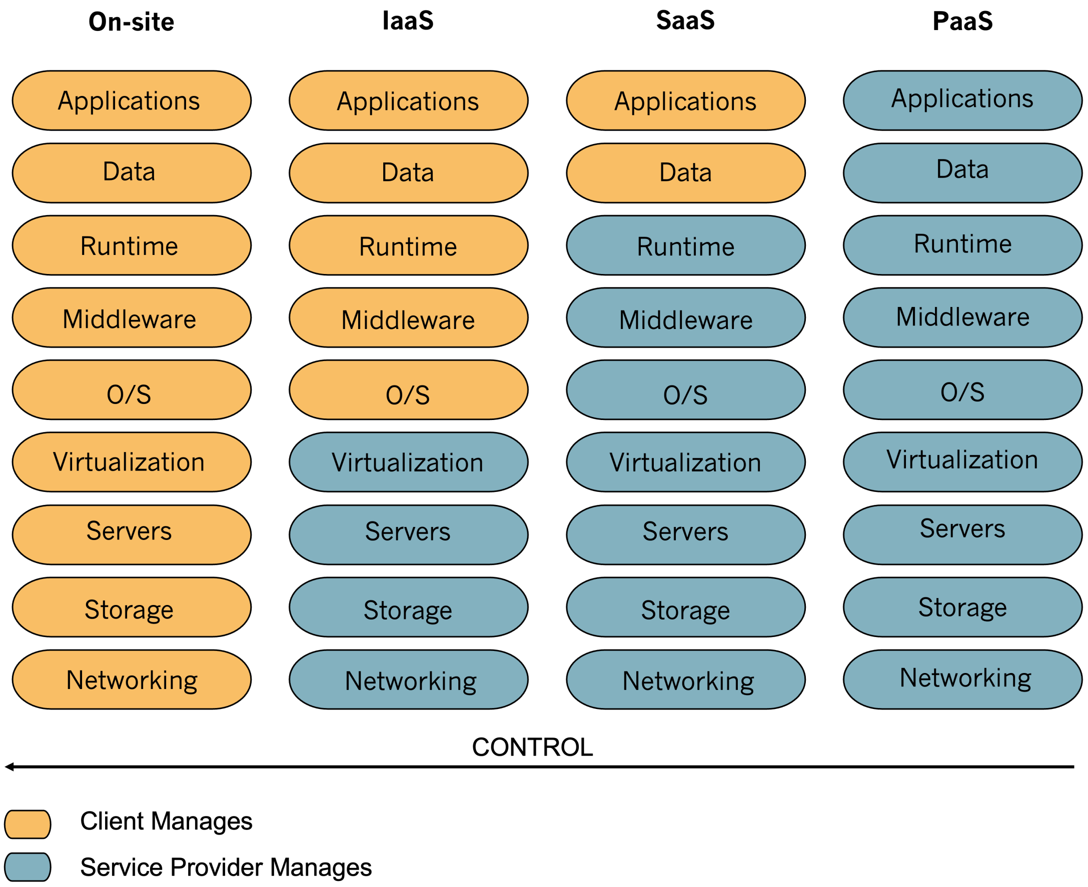


Figure 13: Comparison between different service models. Adapted from [72].

5.4 Cloud Service Providers

Amazom Web Services (AWS)

A substantial portion of Amazon's services is associated with the Cloud Computing paradigm. This enables the user to develop fully integrated and easily realizable business solutions. It offers a combination of IaaS, PaaS, and SaaS, including computational capacity, data storage and management, networking, data analytics, load balancing, and even autonomic scaling. In addition, Amazon offers systems that promote better productivity and efficiency, particularly in its service delivery. According to the book "Amazon Web Services in Action," over one hundred services organizedare currently into twenty categories. Amazon EC2 (computing capacity), Amazon RDS (storage and data management), Amazon S3 (scalable data storage), and Elastic Load Balancing (distribution of traffic amongst applications) are the most well-known [156, 46]. AWS also offers services for networking, security, analytics, XML, mobile applications, and the IoT. AWS is designed to be scalable, fault-tolerant, and highly available and provides a vast array of tools and services to assist customers in developing, deploying, and managing cloud-based applications.

Amazon's pricing models include "pay as you go," in which you only pay for what you use, "pay less on reserve," which allows you to save money on your margin by using reserved instances, and "pay even less per unit for 100% usage," in which storage and data transfer fees are taxed at pre-defined levels.

Microsoft Azure

Microsoft Azure, a key competitor to Amazon, provides a vast array of services with similar functions, including creating, deploying, and managing applications and services over a global network of Microsoft-managed data centers. It offers a variety of cloud services, such as computation, analytics, storage, and networking. Users can select and modify these services to match their particular requirements [46, 20]. Azure's flexibility to handle a variety of operating systems, programming languages, frameworks, and tools is one of its primary advantages. This renders it an adaptable platform for developing and deploying a vast array of applications. Additionally, many users choose Azure over AWS due to the connection with Microsoft tools and the availability of open-source support [20]. This CSP offers three solutions, "pay as you go". This service is very similar to AWS, but the machines are much more economical monetarily. In terms of capacity, namely SQL Server, there are savings of around 85% compared to Amazon RDS [150].

Google Cloud Platform

Google Cloud Platform is a CSP which enables customers to develop, test, and deploy applications using Google's infrastructure. It offers a variety of services, such as storage, computing, networking, and even ML, as well as management and monitoring tools for applications. Google Compute Engine, Google Kubernetes Engine, Google Cloud Storage, and Google BigQuery are the most popular services on the Google Cloud Platform. The Google Cloud Platform is designed to be scalable, adaptable, and secure, making it an excellent option for various applications and workloads. It cannot execute software in parallel on a wide scale. As the newest member of the CSP, it offers enticing pricing. Higher consumption results in a greater discount. Some tariffs correspond to the time unit second and usage-based reductions.

Comparison

AWS, Microsoft Azure, and Google Cloud Platform (GCP) are all cloud-computing platforms that provide similar services for developing and delivering applications. Each CSP has its advantages and is suited for various applications. Amazon Web Services is the oldest and most developed of the three platforms and offers the most services. It is renowned for its dependability and adaptability, making it an excellent option for various applications and workloads. Microsoft Azure is second only to Amazon Web Services in terms of the number of services it provides, and it is especially popular among enterprises that use Microsoft technologies. In addition, it excels in fields like artificial intelligence and machine learning. GCP is the most recent of the three platforms and is well-known for its emphasis on data analytics and machine learning. Additionally, it offers a wide variety of services and is generally regarded as user-friendly. The optimal option will rely on the application's and organization's specific requirements [11, 46].

Table 4 intends to summarize the qualitative and quantitative analysis of the three CSP. Understanding that these values may range based on the service, area, instance type, and consumed resources is important. Moreover, prices may vary based on usage frequency and volume.

Table 4: Comparison metris to evaluate CSP. Adapted from [150, 69, 71]

Evaluation Metrics	**AWS**	**Azure**	**GCP**
Starting Date	March 2006	February 2010	April 2008
Storage Price ($/hour)	0,05	0,007	0,04
Computing Power Price (GB/month)	0,0058	0,0107	0,007
Transfer Price ($/GB)	0,01	0,087	0,08
Discount 1-year comitment	40%	40%	60%
Performance (operation/sec)	20000	20000	15000
Latency (ms)	15-20	15-20	20-30
Scalability	5	4	4
Response Time (s)	0,5-1	0,5-1	1-1,5
Throughput (mbps)	5	4	4
Security	5	5	4
Total Score	9/11	4/11	2/11

Note that for scalability as well as security, a scale from 1 to 5 was used. For scalability, the maximum score was only assigned to AWS since the configuration for the other two CSPs is more complex. In the case of security, the maximum score was attributed to the AWS and Azure Microsoft since both have more monitoring and threat detection capabilities when compared to the GCP. Therefore, and analyzing the score obtained, according to the presented metrics, the CSP that reveals better characteristics is AWS. It should be mentioned that, apart from the score, the metrics where AWS was not classified as best were those associated with price, and even in those, the difference is not significant.

5.5 Data Security in Cloud

Cloud computing can offer numerous benefits to businesses but also present some issues. Security is probably one of the most troubling issues. Cloud-based data is susceptible to cyber attacks, and there have been data breaches in the past. If cloud services are not appropriately protected, an increasing number of users dispersed throughout the globe will be able to access huge quantities of client information. Among the security procedures that assure cloud security are the following:

Passwords: A unique password is required to access cloud-based services. In addition, two-factor authentication should be enabled for all accounts. Finally, CSP should ensure that user names and passwords kept in their database have no direct association.

Encryption: A good encryption technique protects client data. However, for example, homomorphic encryption techniques are not yet completely feasible in real-time scenarios.

Secure Connection: The use of secure connections such as HyperText Transfer Protocol Secure (HTTPS), VPN, Firewall to access cloud-based resources is encouraged.

Security Tools: Users should install effective anti-virus and anti-spyware software on their devices.

Limit Login Devices: Users should be cautious about where they log into the Cloud in order to access services. They should avoid using several personal gadgets, as some of these devices may have keyloggers.

Login Monitor Customers and CSP must check recent devices used to access cloud services. On the basis of this data, users can determine if someone has signed in using their credentials and reset their passwords in the event of a questionable login from an unexpected device or location.

Nowadays, cloud computing has progressed significantly and addresses many simple security concerns. Still, other unsolved issues must be resolved for the cloud computing business to expand.

5.6 Practical Examples

Several healthcare organizations have discussed the partial or complete use of cloud-based technologies.

There are a number of extremely inspiring examples of implemented applications throughout the world. Since 2016, Pfizer, a pharmaceutical and biotechnology business that was recently featured for its involvement with a COVID-19 vaccine, has utilized cloud services for its initiatives. Also, the company has begun cooperating with Amazon Web Services to develop cloud-based solutions for expediting and enhancing the creation, manufacture, and distribution of clinical trial tests [2].

Allscripts is a cloud-based platform that incorporates numerous applications and services for managing health paperwork, patient engagement, and analytics in a second, more extended region. In 2020, Microsoft announced a five-year extension of its partnership with Sunrise, an electronic health records program developed by Allscripts. The vice president of Microsoft's U.S. Health and Life Sciences stated that the partnership with Allscripts could be ”a disruptive force in the healthcare business.”

ClearDATA's HIPAA-compliant cloud is compatible with various public clouds and secures sensitive patient data with compliance measures, DevOps automation, and healthcare expertise. The platform also powers mission-critical apps and detects changes to cloud accounts automatically, allowing the company to respond to these changes in a variety of ways [96].

Meanwhile, Nintex is committed to eliminating paper paperwork. It streamlines laborious operations and extracts vital information from silos, hence enhancing the entire patient experience. The company's automation services are provided to a variety of healthcare industry professionals, including physicians, nurses, pharmaceutical and medical device makers.

The open CareCloud platform assists healthcare providers in enhancing their efficacy and efficiency. Also, it enables them to communicate directly with patients to give superior care. Applications include revenue cycle management, practice management, electronic health records, patient experience, app support for mobile devices, and healthcare analytics [14].

Part III

Case Studies

6

Case Study I: Building a Real-Time Data Repository based on Laboratory Test Results

6.1 Introduction and Problem Identification

The COVID-19 pandemic has highlighted the enormous difficulties that might develop when evidence-based medicine is inadequate. One of the major concerns is the lack of proof regarding occasionally appearing harmful microbes. This can lead to a lack of knowledge regarding preventing and treating infectious diseases efficiently. In addition, the lack of resources and effective therapeutic choices might make it difficult to give patients essential care, particularly during a pandemic when the demand for health services is at its peak. This issue has emphasized the urgent need for new health information systems based on BI, AI, and ML to replace conventional information processing and dissemination methods.

The development and implementation of AI and ML models require access to a large amount of data, including medical data such as EHR. Yet, rules established to safeguard patients' privacy can make it difficult for researchers to acquire and analyze this information. Synthetic data, derived from genuine data but respecting patient privacy, is one solution to this problem [30, 21]. It may be tempting to utilize synthetic data due to the time and resources required to gather and classify big real-world datasets. However, suppose the synthetic data is insufficiently accurate. In that case, it will not reflect the important patterns in the training or test data, and modelling attempts based on implausible data cannot provide valuable conclusions [123]. Hence, synthetic data cannot always resolve an issue [135].

The alternative is to use data processing techniques, such as anonymization, which anonymize genuine medical data and give important information on patient visits to healthcare facilities, yielding a time series dataset influenced by protected characteristics such as age, gender, and location. By entering the data universe of a patient's medical record, the quantity of data to be processed increases dramatically [79]. Every day, millions of individuals generate valuable medical data records fed into information systems, such as those used for prediction or decision support, to sequentially enhance their performance. They need data that has been preprocessed, and standardized [113, 136, 106].

The gap mentioned is the absence of real-time updated data repositories incorporating field-collected data. Thus, many EHR-based research papers lack follow-up data, making conducting additional studies and validation difficult. Providing a standardized, real-time data repository based on interoperability technologies and laboratory analysis results would greatly enhance healthcare delivery, particularly for assisting practitioners in tracking disease progression, assessing risk, and completing other tasks [107, 137].

This case study aims to build a repository of standardized real-time data within the scope of laboratory test results in order to use it in the solution presented in Chapter 7. Furthermore, the goal is for the development to be carried out abstractly so that the solution can be reused for comparable problems and inspire other research projects.

HL7, ICD-10, LOINC, and FAIR, all with distinct functions, were adopted to facilitate data collection and interoperability in order to meet the delineated goal. In the end, validating the information obtained by health professionals proved beneficial to ensure that users of the system employ safe and appropriate medical terms.

In conclusion, the developed data warehouse will allow the establishment of an organized and standardized data structure. Furthermore, the implementation of the principles has made the process of communication and reuse by potential stakeholders feasible. On the other hand, the construction of the repository encourages the construction of new knowledge from the stored data since the information is standardized.

6.2 Solution Objectives

Based on the problem outlined in the previous section, the development of this solution must achieve the following key objectives:

- Collect and store large amounts of laboratory test results in a centralized location for easy access and analysis.

- Use HL7 integration to automatically receive and update the repository with new test results in real-time, eliminating the need for manual data entry.

- Allow for efficient data retrieval and analysis to improve patient care and support research.

- Ensure data security and privacy through proper data management and compliance with relevant regulations.

- Improve communication and coordination between different healthcare providers and organizations by providing a shared, real-time view of patient laboratory test results.

- Enable population health management and public health surveillance through the analysis of large-scale laboratory test data over time.

- Provide support for research by making large, diverse datasets of laboratory test results available for analysis.
- Continuously improve the data repository by incorporating new data sources and functionalities, such as machine learning models and natural language processing, to extract insights from unstructured data.

6.3 Design and Development

It is also important to note that this system takes into consideration several functional requirements for development. Among all, the following stand out:

- Must be able to manage large amounts of laboratory test results, including structured and unstructured data.
- Be able to receive and process HL7 messages in real-time, from multiple sources.
- Have robust security measures to protect patient data and comply with relevant regulations, such as GDPR.
- Provide authorized users with the ability to search, retrieve, and analyze data while having proper access control.
- Have built-in quality control measures to ensure the accuracy and integrity of the data.
- Have a robust backup and disaster recovery plan in place to ensure the availability and integrity of data.
- Must be able to integrate with other systems, such as electronic health records and laboratory information systems, to improve data completeness.

To achieve these requirements, some tasks were outlined:

1. Select a platform for building the repository, such as a database management system, data warehousing solution, or cloud-based service.
2. Design the database schema, tables, and fields to support the data types and relationships required for the repository.
3. Integrate data from various sources, such as EHR, Laboratory Information System (LIS), and other clinical systems.
4. Implement security and privacy measures to protect the integrity, confidentiality, and availability of the data in the repository.

5. Build data pipelines to automate data collection and integration from various sources in real-time.

6. Test the repository and validate its performance, data quality, and compliance with regulations and standards.

7. Deploy the repository and establish a maintenance plan to ensure its ongoing performance, security, and compliance.

8. Provide Documentation.

9. Monitor the repository's usage, performance, and impact, and evaluate its effectiveness in meeting the goals and objectives previously defined.

6.3.1 General Architecture

A high-level architecture is shown in Figure 14.

As shown, the workflow begins at the healthcare facility. The number of institutions is limitless. Each healthcare institution has its own collecting sites, which may be positioned, among other places, in the emergency room or intensive care unit. After collecting the sample, it is sent to the laboratory along with the HL7 message. At the laboratory, analysis is conducted, and the results are sent to the interoperability engine. This HL7 message provides the sequence number of the patient, the episode associated with the analysis, the analysis codes, the analytical values discovered during the test, the units for that value, and the predicted limits. Each institution has its own port for sending and receiving results with this distributed interoperability engine. In the interoperability engine, the data undergoes a transformation procedure to prepare it for processing after the results have been collected. Moreover, the API permits any authorized client to query the repository.

MirthConnect is an integration engine in conjunction with the Javascript programming language for mapping the fields of the HL7 messages; pySpark for data preprocessing. SQL for constructing and putting data into the repository; Node.js for constructing the API; and Swagger for interactive API documentation.

The work accomplished yielded two solutions. The first one is a data repository, and the second is an API.

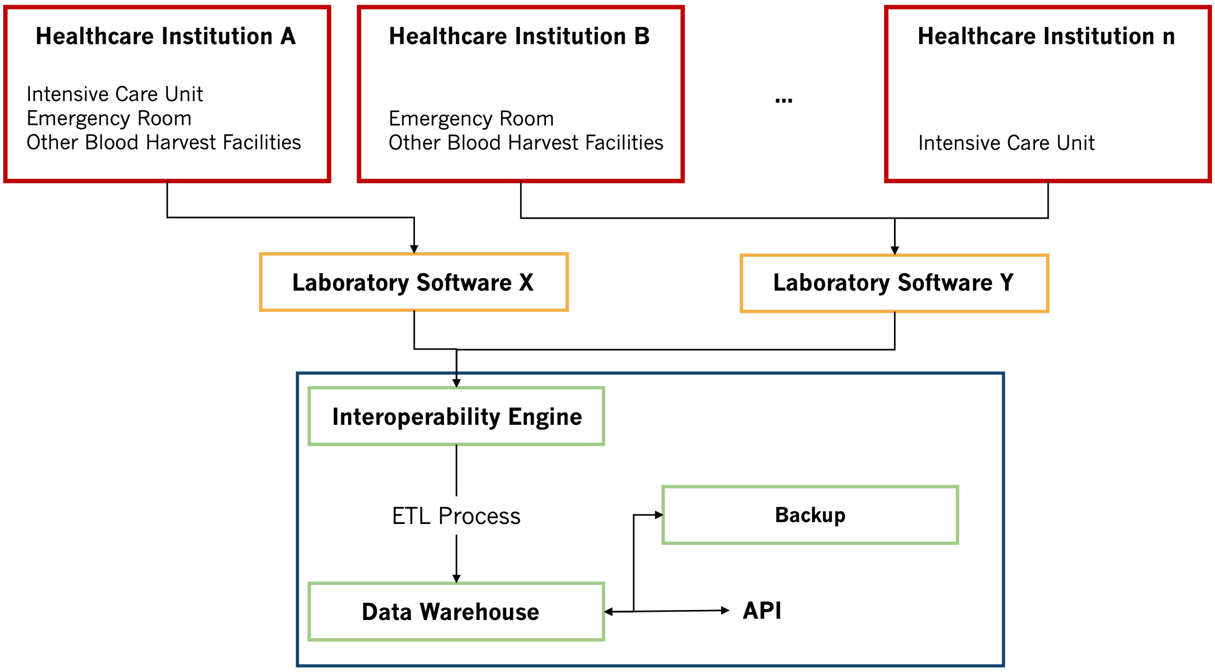


Figure 14: General architecture of the developed solution.

6.3.2 Development Stages

Three phases were involved in the development of the given solution. The initial step involved data gathering, analysis, preprocessing, and integration. This phase was implemented with the HL7 standard, Javascript programming language and the help of the Mirth Connect engine. This entire progression will be described in detail in Section 6.3.2.1. This section describes the creation of the data warehouse in detail. The development of the API followed the two phases described previously, as will be explained in Section 6.4.2.

6.3.2.1 Data Integration

HL7 data standard was used to implement the integration of data analysis and transformation. Using a connection (Internet Protocol (IP) and port), the Mirth Connect interoperability engine receives data. Each institution has a specified connection channel that receives messages containing the laboratory's analytical results. After receiving the HL7 message, it undergoes a processing procedure in which all fields of interest are mapped. Concurrently, numerous typical modifications are carried out (anonymization, all caps, date format standardization, etc.). This section concludes with a comprehensive list of the data cleansing procedure. In addition, in order to promote standardization and enable interoperability, it has been determined that diagnostic and analytic coding information must be included. The global standards ICD-10 and LOINC were employed for this purpose.

- Anonymization: All patient-related fields were anonymized. Only the sequential number concealed by randomly generated digits was taken into account.

- Uppercase: All words were capitalized and without accents.
- Date Format: dd/mm/yyyy hh:mm:ss
- Missing values: There are two possible cases. In case they are values that do not affect the quality of the repository (sequence number, for example), this number is generated randomly, and the record is kept. Otherwise, the record is ignored.
- Noisy Data: Since HL7 messages have mainly information depending on the health institution and the laboratory that publishes the result, each institution has its channel.
- Outliers: The records outside the cluster are older/inconsistent data. These tuples have been ignored.

The anonymized results are then transmitted to the data repository, described in the next section.

6.4 Results

This section presents two distinct sections obtained with the development of the case study. Section 6.4.1 describes in detail the Real Time Data Repository. Section 6.4.2 presents the API to access and query the stored data.

6.4.1 Real Time Data Repository

An estimated 2.5 million records are submitted annually to each healthcare facility's repository.Since only data from 2020 onwards were considered, there are around 7 million records for the project's pilot institution. In light of the quantity and classification of the data stored from multiple sources, the constructed data repository is classified as a data warehouse. The time at which the data is meant to be processed is an additional reason to use a data warehouse. In this situation, the raw data is processed and then stored, immediately saving substantial storage space. In addition, this transformation procedure before storage offers the data warehouse characteristics such as data quality and integrity.

The data warehouse structure comprises a single fact table, Results, and tables for three other dimensions, DemographicData, DiagnosisType, and ExamType.

The purpose of these tables is not to supplement the fact table. Thus, the ”DemographicData” table solely contains data about patients. The ExamType table contains all analytical codes, definitions, unit divisions, and normal ranges. It is possible to locate the code and description ICD-10 accompanying the DiagnoseType table.

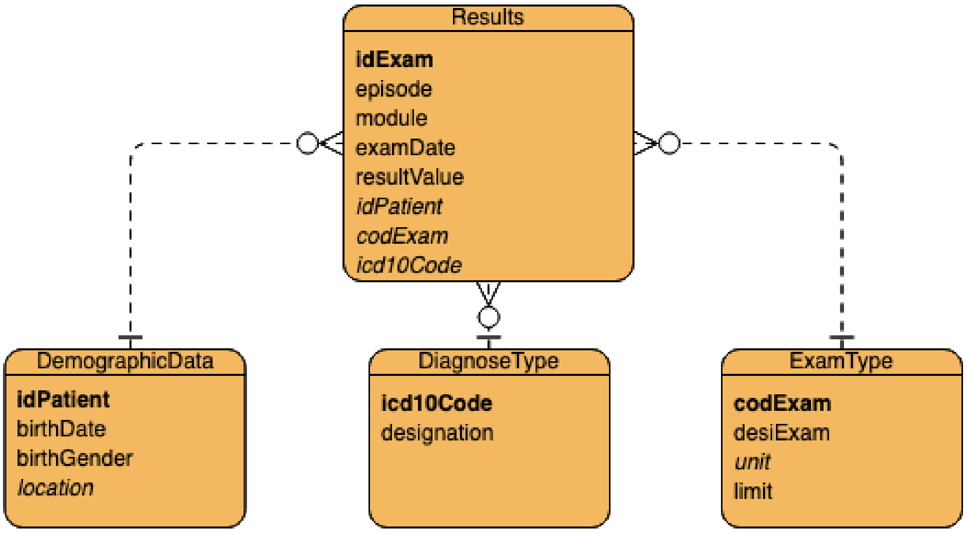


Figure 15: Entity Relationship Diagram of the Data Warehouse.

The Star Schema is the outcome of creating the fact table and the three-dimensional tables, as depicted in Figure 15.

- **Dimension Table 1:** DemographicData
  - **idPatient**: The sequential number is assigned, as the name implies, sequentially to the patient the first time he/she enters the healthcare institution. The same patient may be present in the database with different sequential numbers because he/she attends n institutions. To the original sequential number and in order to guarantee the anonymization of the dataset, a series of 4 numbers is added randomly;
  - **birthDate**: The date of birth also refers to the patient. To maintain the consistency of the data repository, all dates follow the format: dd-mm-yyyy hh:mm:ss;
  - **birthGender**: The birth gender is the attribute usually called administrative gender and refers to the identification, by a family member, midwife, nurse or doctor, of the external genital organs when the baby is born. In the developed repository, this gender can contain two values: F-Female, M-Male;
  - **location**: The location refers to where the patient has his or her fiscal address. In Portugal, this can be a city, a town or even a village.
- **Dimension Table 2:** DiagnoseType
  - **diagnoseCode**: The diagnosis used in work developed is that of the *Grupos de Diagnósticos Homogéneos (GDH)* since it is the one that is most reliable. This system, GDH, is the disease classification system used in Portugal. Because it is so specific, this code was transformed into an ICD-10 code, so other countries can contribute to the repository developed in the future.

  - **diagnoseDescription**: The description of the diagnosis corresponds to the ICD-10 code.

- **Dimension Table 3:** ExamType

  - **codExam**: The examination code is the LOINC code that will pair with the designation described below;
  - **desiExam**: The exam name is in English and is LOINC - Long Common Name;
  - **unit**: The unit corresponds to the dimensional unit associated with the value of the analysis performed;
  - **limit**: The limit refers to a standard of normality for the test code. For example, for the White Blood Cells (WBC) analysis the value is 6.0 - 16.0;

- **Fact Table:** Results

  - **idExam**: The exam number is, as its name indicates, the number that is assigned to the sample. In other words, there may be n results for the same test number because several parameters can be analyzed with the same sample. For example, patient ABC who has a test with the number 111111, has two results, one for the Leukocyte value and another for the Monocyte value;
  - **episode**: The episode corresponds to a numeric value referring to an hospital visit. An episode is opened when the patient enters the hospital and closes when the patient is discharged. In short, each patient may contain various episodes that may occur in the scope of the emergency room, inpatient stay and appointment, among others;
  - **module**: The module identifies the type of service in which the patient entered the hospital. For example, it could be CON - Appointment, INT - Internment, URG - Emergency, among others;
  - **examDate**: the date of examination is only identifying the time of sample collection. Follows the pattern defined for the repository dd-mm-yyyy hh:mm:ss;
  - **resultValue**: The value refers to the result obtained from the analysis of the sample. Typically it is a numeric value that will be completed by the two other columns that follow, the unit and the limit;
  - **Foreign Keys**: idPatient, codExam, icd10Code;

6.4.2 API

The need for an interface between the data repository and its users led to creating a API in which the web services are based on the Representational State Transfer (REST) architecture. The requests made available to the user are of the GET type since it is only possible to consult resources and not modify them.

The created API was built on the javascript programming language, specifically, Node.js with express.js framework, and Swagger to automatically generate interactive documentation. Swagger is an open-source framework to automate the generation of REST API documentation largely used in several projects.

The endpoints available for the users are described in table 5.

Table 5: Description of the endpoints available in the developed API

Endpoint	Type	Description
/login	POST	Generates JSON web token and is used for authorization on the remaining routes. This token is generated from the email and password pair.
/sample	GET	Returns an object with the last 100 records from the repository in descending order
/results/:idPatient/:idExam	GET	According to the exam number and the patient id as parameters, return a list of objects with all the results associated with that patient's exam sorted by descending date.
/results/:idPatient	GET	Receiving as parameter the patient id returns all results related to this patient sorted by exam and date pair.
/analysis/:idPatient/:idExam	GET	According to an analysis code and the patient's id returns the date of the last examination performed;
/analysis/:idExam	GET	According to an analysis type and a certain age range, it returns the minimum, average and maximum value;
/diagnosis	GET	According to the diagnosis code (ICD10), returns the most performed analyses;
/diagnosis/:icd10Code	GET	According to an analysis type and a certain diagnosis code returns the minimum, average and maximum value for each code of analysis;

Each of the aforementioned routes returns a JSON object with the most recent exam at the top. The schema documentation, necessary inputs, and API response are all available on Swagger.

6.5 Discussion and Conclusion

Creating a data repository centered on the outcomes of specific laboratory tests has many advantages.

As discussed throughout this chapter, it provides easier and faster access to information, improves data integrity, consistency, and security. The data are more accurate and reliable because health professionals have validated them, it reduces redundancy and data loss, and it enforces data format and terminology standards. In addition, it facilitates the extraction of knowledge because the information is standardized, i.e. using BI technology, which enables the production of clinical BI and performance indicators that offer users valuable insights and an intelligent decision-making process.

Improving and expanding the API knowledge extension's endpoints to additional areas of interest is one of the most pertinent issues. If the information that is not currently represented in the database has to be added in the future, the data model will have to be updated. Hence, users can later submit fresh feature requests on GitHub (routes). In addition, it is possible to create a web application that, for instance, enables real-time viewing of clinical test results based on variables such as patient, exam, location, and diagnosis.

Additionally, the integration should grow to embrace data integration with FHIR and ICD-11 (the most recent version of the classification system) for diagnostic standardization.

Including extra data or even segmentation into data marts can enhance the capabilities of the data warehouse and promote further study. In addition, there is significant potential for feeding ML, Deep Learning (DL) algorithms with data from these. Finally, algorithms could be devised for filling in missing values, such as missing units of a given result.

7

Case Study II: Novel Approach for a Software as a Service for Clinical Test Results Reporting

7.1 Introduction and Problem Identification

Statistics on population ageing are becoming increasingly obvious on a global scale. According to the Instituto Nacional de Estatística (INE), the number of aged people (65+ years old) in Portugal will increase from 2.2 million to 3 million by 2080, signifying an increase in the country's ageing index. This ratio will nearly double from 159 to 300 seniors per 100 youths [116]. According to the WHO, "Between 2015 and 2050, the proportion of the world's population over 60 years will nearly double from 12% to 22%", adding particular emphasis on the fact that "The number of people aged 80 years or older is expected to triple between 2020 and 2050 to reach 426 million." [108].

Consequently, an urgent dilemma arises for healthcare practitioners. Typically, older people have more complex medical conditions or exacerbated chronic illnesses. Therefore, it is extremely difficult to appropriately evaluate and respond to symptoms and test results. This causes a paradigm shift in medical practice. In the beginning, reactive medicine was practised, in which one acted based on what was observed, without clinical or laboratory evidence. Based on a clinical and laboratory data study, the physician diagnoses and forecasts potential disorders and then takes action to reverse the condition. The practice of Personalized Medicine now considers each individual's unique characteristics. General prescriptions are not offered. The paradigm of customized medicine enables more effective prevention and action, reducing the number of diseases or enhancing the treatment of ageing-related symptoms. However, tailored research of each leads to a significant increase in the number of prescribed and executed tests.

For personalized medicine to become more effective and less costly, information systems with the ability to interoperate are needed, with transparency and correspondence between the data from one or more institutions, eliminating duplication of studies, exams, etc. Unfortunately, these systems do not exist or are unprepared for the volume of data they encounter daily. The initial investment required to buy software, a license, or even hardware represents a considerable effort for small and medium-sized companies. As an alternative to investment, such companies typically become dependent on paper records or exchanging information by email. This alternative has proven to be very impractical. It can also pose a severe security threat, including patient data breaches, data loss, lack of compatibility with other systems, and high maintenance costs.

New approaches, such as cloud computing, could help bridge this gap. Allied to this new approach is a recent systems architecture, - "Software as a Service," that works as a pay-as-you-go service and allows for ideal scalability for small and medium-sized businesses. SaaS offers a cost-effective and secure solution for healthcare organizations, allowing them to access the necessary technology without the upfront investment in hardware and software. It also enables greater interoperability, as SaaS systems can easily communicate with each other and share data.

Overall, this type of software presents a great opportunity for the healthcare industry to improve patient care and streamline operations. By harnessing the power of the cloud, healthcare organizations can access the latest technology and services without breaking the bank, all while keeping patient data secure.

The developed solution is based on cloud-based software capable of integrating and displaying results to physicians so that reporting and access to results are as quick as possible. Fundamentally, it is characterized as "software as a service" with a multi-tenant architecture. The main objectives of this development are described in the following section.

7.2 Objectives

The main objectives to be fulfilled with the developed solution are:

- Automate the process of generating and distributing clinical results, reducing the need for manual reporting and facilitating efficient reporting.
- Reduce the need for initial investment, thus enabling small and medium-sized companies to computerize their healthcare system.
- Real-time access to clinical/exam results enables healthcare providers to diagnose and treat patients quickly, improving patient outcomes.
- Share clinical results with other healthcare providers, enabling better coordination of care and Enhancing communication and collaboration.
- Automate the reporting process to help ensure compliance with regulations and accreditation standards related to data management and reporting.
- Provide patients with access to their clinical results, helping empower them to take a more active role in their healthcare.

- Include built-in validation and quality checks to improve data accuracy, completeness, and consistency.
- Acquired data to be used to identify patterns and trends in patient populations, enabling healthcare providers to take a more proactive approach to population health management.
- Allow easy access for reports from any device, location, and time, improving the speed and quality of communication and decision-making.
- Development of a solution capable of fitting any type of exam or even analytic results.
- Ensure data security and privacy.

7.3 Design and Development

The intended audience for the designed system is fairly extensive. It comprises the Administrator, Physicians, and Technicians of any healthcare unit that conducts exams and/or generates reports. Other types of users are configurable. Consequently, a user type was established for each of these characters. Additionally, a user type was added for the SaaS administrator.

- User 0: Administrator
- User 1: Physician
- User 2: Technician
- User 3: SaaS Manager

It is also important to note that this system takes into consideration several functional requirements for development. Among all, the following stand out:

- Possibility to have several clients from different institutions.
- Ensure that clients can independently manage their clinics and their users.
- Ability to integrate with various systems that provide test results and analysis.
- Ensure separation of data by customer (security and privacy).
- Allow automatic integration from equipment.
- Have an alternative for integration when the automatic integration engine fails, or interoperability is not feasible.
- Allow the visualization of results, preparation and publication of reports and respective addenda.

- Allow the patient to access reports.
- Provide statistics, warnings, and monitoring indicators.

To fulfil these requirements the following tasks were outlined:

1. Select a technology stack to build the software, such as a programming language, framework, and database.
2. Design the software architecture and user interface to support the data types and relationships required for the software.
3. Implement a storage solution for the clinical/exam results data, such as a database or cloud storage.
4. Implement security and privacy measures to protect the integrity, confidentiality, and availability of the data in the software.
5. Build data pipelines to automate data collection and integration from various sources in real-time.
6. Implement reporting and visualization functions to allow users to access and analyze the data in the software.
7. Test the software and validate its performance, data quality, and compliance with regulations and standards.
8. Deploy the software and establish a maintenance plan to ensure its ongoing performance, security, and compliance.
9. Train users on how to access and use the software, including data entry, query, and reporting functions.
10. Monitor the software's usage, performance, and impact, and evaluate its effectiveness in meeting the goals and objectives defined in step 1.
11. Continuously gather feedback from users and stakeholders to improve the software, adding new features and fixing bugs.

7.3.1 General Architecture

The architecture of the solution, shown in Figure 16, encompasses several components in order to meet the requirements previously raised. This architecture has a multi-tenant typology since the resources will be partitioned so that several clients with several clinics can access the same instance in a particular way. This typology proved to be the most advantageous because, on the one hand, it saves resources and therefore reduces costs. On the other, it allows faster and more centralized management of all the software. Finally, since multi-tenancy allows a single instance of the software to serve multiple tenants, it can be more easily scaled to handle many users.

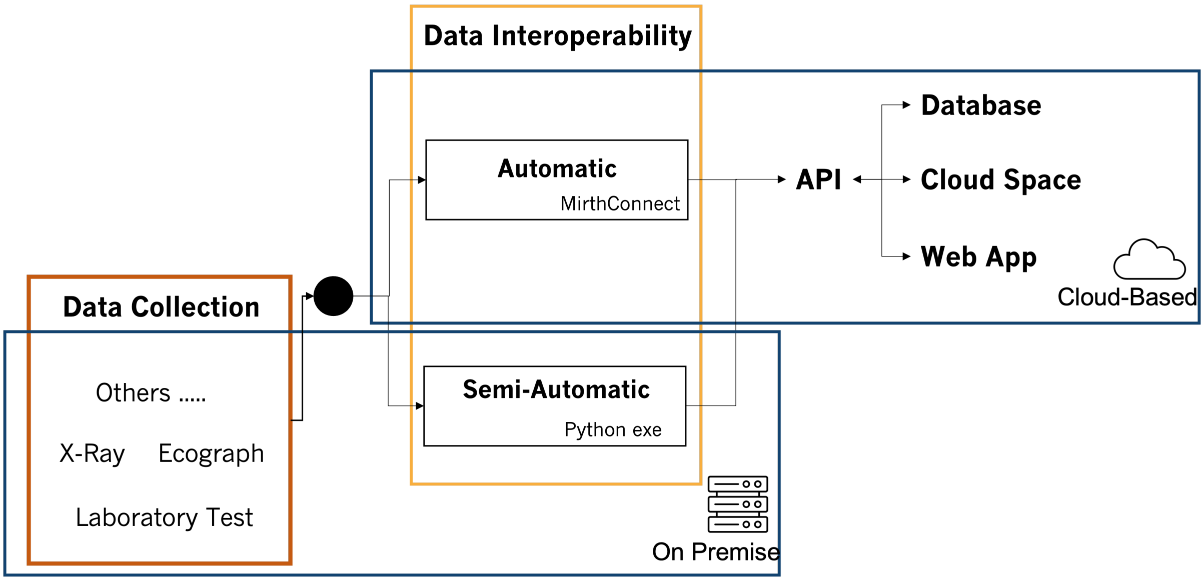


Figure 16: General Architecture of the SaaS.

As seen in the diagram, the architecture consists of five major components. On the cloud side, there are the database, document storage, API, automatic interoperability engine, and web app. On the on-premises side, only clinic-hosted equipment and a semi-automatic integrator are present.

Postgres was chosen as the data storage system. This is a popular option for database management in the development of SaaS due to its support for simultaneous connections, high performance, simple extension, and replication features.

To host this and other components, the CSP chosen was AWS. The advantages of using it have already been discussed in Section 5.4. However, in summary, AWS offers scalability, global coverage, security, compliance, flexibility, and cost-effectiveness. In addition, its robust ecosystem facilitates integrations with, for example, Amazon S3, which was the object storage selected for document storage. In this structure, each institution has its folder, protected by a private key. Within each institution, three other folders facilitate document management (processed, unprocessed, and reports).

Regarding the backend and document management, NodeJS was chosen as the programming language. JavaScript, one of the world's most extensively used programming languages, is the foundation of Node.js. This enables the usage of the same language on the frontend and backend, which facilitates the exchange of programming concepts. In addition, Node.js offers integration libraries for the chosen programming language, such as the AWS Software Development Kit (SDK), which greatly simplifies implementation. The built API will serve as the data processing and storage hub.

Real-time integration of exams and results is a crucial component of the system. Since certain equipment lacked the capacity for automated interoperability in HL7, it was clear from the outset that two solutions would be required. Python was chosen for the semi-automatic alternative because it is an easy-to-use scripting language, it integrates well with Amazon Web Services, it can be compiled for multiple operating systems, and it matches the specific problem. This agent is simple and has limited data processing responsibilities, restricting itself to the movement and synchronization of folders. Regarding automatic integration, the Mirth Connect integration engine was chosen due to its user-friendliness and ability to manage many connection channels.

The final product is visualized via a web app written in React because, while retaining the javascript programming language, it employs the reconciliation technique to provide outstanding efficiency when handling many updates. In addition, React may be simply combined with other libraries and frameworks to enhance its capabilities.

Nginx is the web server responsible for receiving and handling HTTPS requests. This was one of the options that showed excellent performance, the capacity to manage a huge number of simultaneous connections, and the capacity to operate as a reverse proxy and load balancer. Additionally, it is highly stable and safe when properly configured. Using Apache HTTP Server or Caddy was also possible, although the learning curve was significantly more complex.

All of the development described in the following section had to account for security measures relating to sensitive data encryption, route protection, document protection, web server configurations, HTTPS usage, update guarantee, minimizing the use of external packages, and monitoring their origin, among others.

7.3.2 Development Stages

Three phases were engaged in developing the solution. The first was the database modelling and design, Section 7.3.2.1. Section 7.3.2.2 explains in detail the subsequent creation of the two interoperability alternatives. Backend (Section 7.3.2.3) and frontend (Section 7.3.2.4) development of the platform was initiated simultaneously.

7.3.2.1 Database

The conception and design of the database were determined by numerous considerations. First, the requirements specification was considered, then the difficulty of supporting a multi-tenancy architecture, and lastly, normalization, redundancy, and data security criteria were considered. The resulting entity-relationship schema is depicted in the referenced diagram, Figure 17. A more comprehensive description is in the table 6. The fundamental flow begins with the client subscribing to a Plan. This customer may have multiple clinics with numerous users of various types. Users are allowed to conduct multiple actions on the Reports that are recorded in the Activity table. The user can extract information from the reports using regular expressions located in the Regex table, connected with the clinic and type of exam.

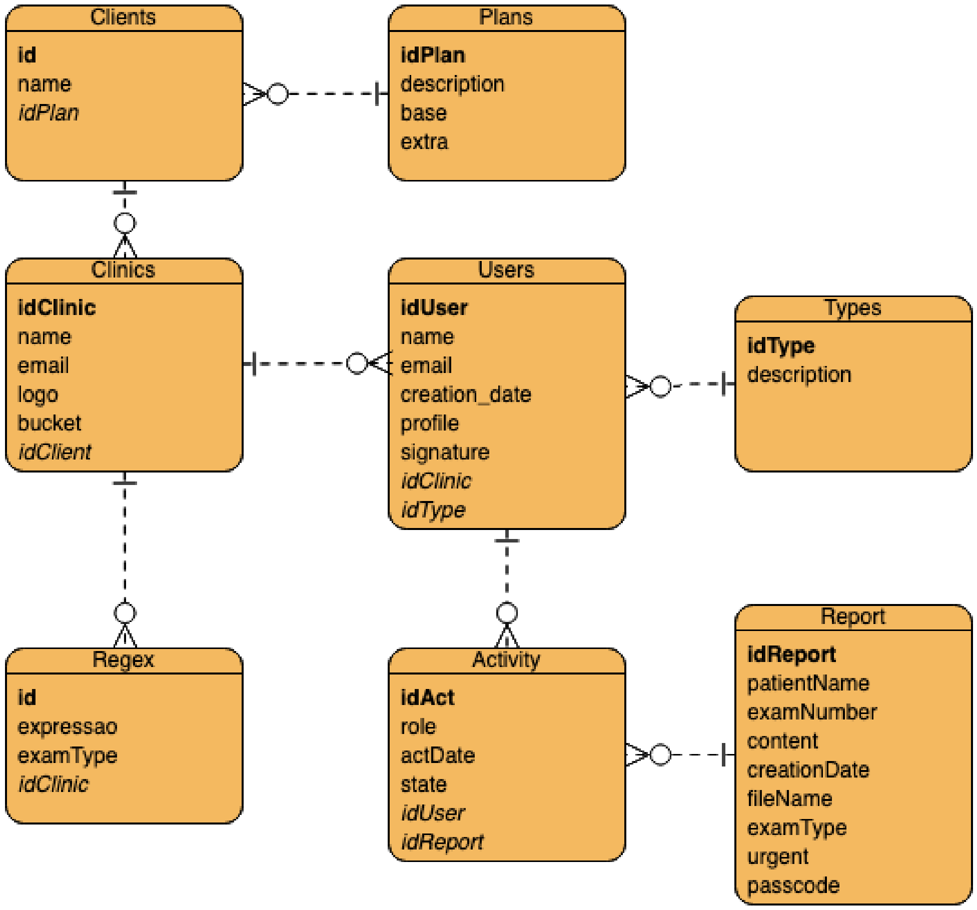


Figure 17: Entity Relationship Diagram of the Developed SaaS.

Table 6: Description of the tables modelled in the ER Diagram

Table Name	**Description**
Clients	The Clients table contains information on the legal entity or business entity that plans to use the produced SaaS. The client is characterized by the automatic increment (integer) identifier and the client's name. Additionally, the client must choose an appropriate subscription package in advance. Therefore the table contains two relationships, the plans table and the clinics table, which are discussed in the following section. A client can represent a clinic or a collection of clinics.

Table 6 continued from previous page

Table Name	**Description**
Plans	The available contract choices are listed in the Plans table. Each option contains a type, a description, a base value, and an additional value. The basic amount indicates the monthly subscription charge for 100 reports. The additional value will be charged for each unit in excess of 100. There may be multiple plans, such as Basic, Advanced, and Premium. For example, the Base Plan would cost 100 euros for 100 reports, and each additional unit would cost 0.75 euros. The only relationships of the Plans table is to the Clients table.
Clinics	The clinics, as previously mentioned, belong to one client. The Clinics table has the necessary data for its characteristics such as identifier, incrementally generated, name, email, logo and bucket. The logo is a link to where the image is hosted in the cloud, and the bucket is the name of the folder where all other sub-folders (unprocessed, processed, reports) are located. The clinic has three relationships: Clients, Regex and Users. In short, each Clinic belongs to one Client, has n Regex expressions and contains n Users.
Regex	The regular expressions that will later be used to extract information from the embedded PDFs are stored in the Regex table. Each instance of this table contains the identifier, the expression and the exam type to which it applies.
Users	The Users table holds all of the encrypted user information. These are distinguished by an automatically incrementing identifier, an email address, a password, the creation date, the name, a profile image, and a signature. This table's associations include Types, Activities, and Clinics. The Clinics' relationship has been discussed previously. As for the remaining associations, each user has just one type and can have up to n roles and date-related records in the activity table (report, validated, responsible, among others).
Types	The Types table is included in the user classification. There are now three types of users. However, because this information may be dynamic, it is saved in table format. Only an identity and a description are connected with user types.

Table 6 continued from previous page

Table Name	Description
Activity	The activity will serve as the entity connecting users to reports. There can be multiple users with various roles for each report. For instance, a report will include the user who generates the report, a user to whom the report is assigned for validation, and a user who validates the report, who may or may not be the same as the person who generated the report. It will also contain the status that will be utilized to manage the frontend lists. This table has the most actions because it undergoes many revisions throughout the procedure.
Report	The reports are characterized by an identifier, the name of the file associated with the exam, the name of the patient (if it is in the pdf), the content, the creation date, the exam number (if it exists in the pdf), the urgency (1 for urgent and 0 for not urgent) and a code to open the file safely. From the web app, this code is decrypted.

7.3.2.2 Interoperability

The primary purpose of the two integration options is the conversion of results to Portable Document Format (PDF) and subsequent upload to the cloud. Consequently, the next section discusses in detail the processing performed by both Semi-Automatic (Intelligent Agent) and Automatic (HL7 message exchange) systems, respectively.

Semi-Automatic The semi-automatic integrator was developed in Python. It works as an intelligent agent that adapts to its environment by perceiving the environment and reacting whenever it is stimulated. By validating and verifying the conditions, it acts with an intelligent purpose to facilitate the integration transparent to the end user. The integrator is available to clinics and agnostic to the operating system. The process starts by checking if the bucket for the clinic exists in the AWS cloud. If it does not exist, it will be created with the name of the clinic. Then, the same happens with the database instance. If it does not exist, an insertion is made through the POST endpoint (/clinics/) of the clinic and the X endpoint of the administrator user. Furthermore, the folders will be created on the device that is in the clinic (unprocessed and processed). The trigger of the intelligence agent is the appearance of a document in the unprocessed folder, which is where the new, non-integrated exams are placed. Each time a new document appears in that folder, it is uploaded to the cloud and moved to the processed folder associating the local machine timestamp to the file name. This integrator is imperceptible to the user, enabling an intuitive and simple experience.

Automatic The automated integration relies on an HL7 message exchange that is managed by the Mirth Connect interoperability engine. Mirth Connect is an extremely well-known open-source integration engine.

It is frequently used due to its flexibility to accommodate a wide range of data types and protocols, many channels, and both inbound and outbound messages. In addition, it provides a user-friendly interface for constructing and managing integrations, and it can be readily changed and extended using JavaScript and other computer languages. Additionally, Mirth Connect is renowned for its scalability, dependability, and security features. This engine can be hosted on-premises, in the cloud, or in a hybrid environment, and it can connect SaaS applications running on different platforms.

An HL7 message exchange normally takes place via channels. A channel consists of a source connector and a destination connector, which together define the protocol and settings for both incoming and outgoing connections. Using the appropriate protocol, a client system initiates the interoperability process by delivering an HL7 message to a Mirth Connect server (HTTPS). Consequently, the Mirth Connect server receives the message via a source connector set with the proper protocol, address, and port. The communication is then modified, filtered, or validated by a transformer before being transmitted to its destination. The field containing the scan results in base 64 is converted to pdf format. The created pdf is then delivered to the target system via a target connector (such as an API request) with the proper protocol, host, and port, and the remaining data is transmitted to the Database.

The developed automatic integration process begins upon receipt of the HL7 message via an HTTP request to a predetermined IP address and port. Each message is then subjected to a transformation and mapping of the message fields. Several verifications are performed on the data types and date format at this point. In addition, at this phase, the message field's base64 is retrieved and converted to a pdf. The channel destination is based on a call to the AWS API, where the PDF is sent.

Notably, depending on the volume and complexity of your SaaS integration requirements, other integration engines may be more appropriate, although they are not as well-suited to the subject of healthcare or HL7 messaging. In addition, when selecting an integration engine, Mirth was favoured due to its extensive selection of connections and certification by the Open Systems Interconnection (OSI) model.

7.3.2.3 API

The database system was designed and implemented using the Sequelize JavaScript package in Node.js Web services, which permits the definition of models and their associations, and, therefore, the formation of a database by setting the described database. Sequelize is one of the numerous libraries used in RESTful Web services to enable the sharing of data between the frontend, i.e. Web and mobile applications, and the database. Web services are built on REST technology, which is a communication style and method. The chosen strategy was monolithic and is prepared to be subdivided into microservices if required. Hence, four distinct HTTP requests were used to manipulate the data contained in the database through the system, namely:

- GET: to retrieve resources.
- POST: to create a resource.

- PUT: to change the state of or update a resource.
- DELETE: to remove a resource.

Client, clinics, users, examinations, and reports are the five core endpoints of the API. Table 7 displays the primary routes developed for each entity.

Table 7: Description of the endpoints available in the developed SaaS API

Endpoint	Route	Type	Description
/Clients	'/'	GET	Returns an object list with the records from the Clients table.
	'/:clientName'	GET	Returns an object corresponding to a record for a specific client.
	'/'	POST	Allows the creation of a client by inserting it in the Clients table of the database. This route, since it depends on the choice of plan, it is only available to manage users.
	'/:clientName'	PUT	It allows you to update a client's information, enabling its activation or deactivation.
/Clinics	'/'	GET	Returns an object list with the records from the clinics table.
	'/exe/:idClinic'	GET	Receiving the clinic id as a parameter accesses the database to fetch the access key, secret and the bucket name. After that, it accesses the AWS endpoint and returns the executable that belongs to a clinic.
	'/clients/:idClient'	GET	Returns an object list with all records resulting from the inner join between the Clinics table and the Clients table.
	'/:idClinic'	GET	Returns an object corresponding to a specific clinic, as well as the name of the client it belongs to.
	'/'	POST	Receiving as parameters the name, email, bucket name and the Client to which it belongs. It then creates a clinic by inserting the data in the Clinics table.

Table 7 continued from previous page

Endpoint	Route	Type	Description
	'/:idClinic'	PUT	Receiving as a parameter the id of the Clinic allows the name and email to be updated.
	':/idClinic'	DELETE	Receiving as parameter the id of the Clinic allows the removal of a clinic
	'/logo/:idClinic'	PUT	Receiving as a parameter the id of the Clinic allows the logo to be updated.
/Users	'/'	POST	Allows the creation of a user for a clinic-client pair by inserting an instance in the tableusers
	'/login'	POST	Allows user authentication by generating a json web token. In this route, encryption techniques are used to compare the user's password with the one encrypted in the database. The generated token expires every 2 hours or when the user leaves the platform.
	'/imgProfile/:id'	PUT	Receiving as a parameter the id of the user that is logged in allows updating a user's profile picture.
	'/signature/:id'	PUT	Receiving as a parameter the id of the user allows the user's signature to be updated.
	'/disable/:id'	POST	Receiving as a parameter the user's id, it allows the inactivation of the user through disassociation from all clinics.
	'/signature/:id'	GET	Receiving as a parameter the user's id returns the signature of a user.
	'/:id'	PUT	Receiving as a parameter the user's id allows updating a user's name, type, and email adress
	'/:id'	DELETE	Receiving as a parameter the user's id allows the removal of a user.
/Exams	'/delete/:idClinic/:id'	DELETE	Allows a specific exam to be archived by changing the status in the Activity table.

Table 7 continued from previous page

Endpoint	Route	Type	Description
	'/clinic/:idClinic'	GET	Returns a list of objects with all the unreported exam records for a given clinic. Receiving the id of the clinic as the parameter, it accesses the database to get the access key, the secret key and the bucket name. Then it accesses the AWS unprocessed folder. It furthermore applies to extracting information such as patient name, exam number, etc. The regex is found in the Regex table associated with a clinic, and an exam type since the PDF formats can be different.
	'/:idClinic/:id'	GET	Receiving as parameter the id of the clinic and the id of the exam returns the complete record of that exam
	'/processed/:idClinic/ :id'	GET	Receiving as parameter the id of the clinic and the id of the exam returns the data to render in the report area.
/Reports	'/:idClinic'	GET	Returns the object with all the records in the reports table, which published and which were assigned to the currently logged in physician in descending order of creation.
	'/toValidate/:idClinic'	GET	Returns a list of objects with the reports that are awaiting validation and are either assigned to the physician with login or have no assignment. The sorting of the list is according to urgency.
	'/pending/:idClinic/ :idDoc'	GET	Returns the number of reports that are pending validation by the logged in physician.
	'/published/:idClinic/:id Doc'	GET	Returns the number of reports published from the logged in physician.
	'/content/:id'	GET	Returns the object with the contents of a specific report

Table 7 continued from previous page

Endpoint	Route	Type	Description
	'/getReport/:filename/ :idClinic	GET	According to the parameters exam id and clinic id, it accesses the database o fetch the access key and secret key. After that, it accesses the clinic bucket and returns the pdf of a specific report.
	'/remove/:id'	POST	Removes the assignment of the physician who is responsible for the validation
	'/create/:idClinic'	PUT	Update the status of the report in the database so that it is in the list of reports to validate. Also, with AWS credentials, access the bucket and move the file from the unprocessed folder to the processed one.
	'/:idClinic/:idDoc'	POST	It generates the pdf of the report with the content that was validated and changed, or not, by the physician.

7.3.2.4 Web App

Authentication Module An authentication page safeguards login to the software. As depicted in Figure 18, the user must provide his credentials (email and password). If these are valid, a token is issued that grants access to the remainder of the application. If the user is an Administrator and logs into the software, he or she will have access to the Clinic, User, and Statistics Modules, as well as his or her user profile. If the user is a technician or physician, he or she has access to the Examination and Reports Modules as well as his or her user profile. When the token expires or when the user chooses to log out, the user quits the application.

Clinics Management Module The Clinics management module allows management by the user. This module, as shown in the workflow of Figure 19, has three main functionalities: Add a Clinic, Edit Information and Delete a Clinic.

Adding the clinic can be done in two ways. The first involves a request to the software manager to provide the instructions for automatic integration and the semi-automatic integrator that uploads the exams to the cloud. If the addition of the clinic is made through the software, after entering the information, the user submits the request, and then an email is sent with the instructions for the automatic integrator and the executable for the semi-automatic integrator. The necessary information for the addition of a clinic is the name, email and logo.

Editing information about the clinic is also the Administrator's responsibility. All clinics that belong to the Administrator with a session started have a row in the clinics table. To edit information, just search for the desired clinic. The user can use the search filters and sort functionalities available in the table. When you select the clinic, the user can then "Edit Information". The form with the current information will be presented. After making the desired changes, press "Save".

Deleting a clinic involves a process similar to editing the information. Search for the clinic you want to delete and click on "Delete Clinic". A confirmation message will open. After clicking on "Delete" the action is irreversible.

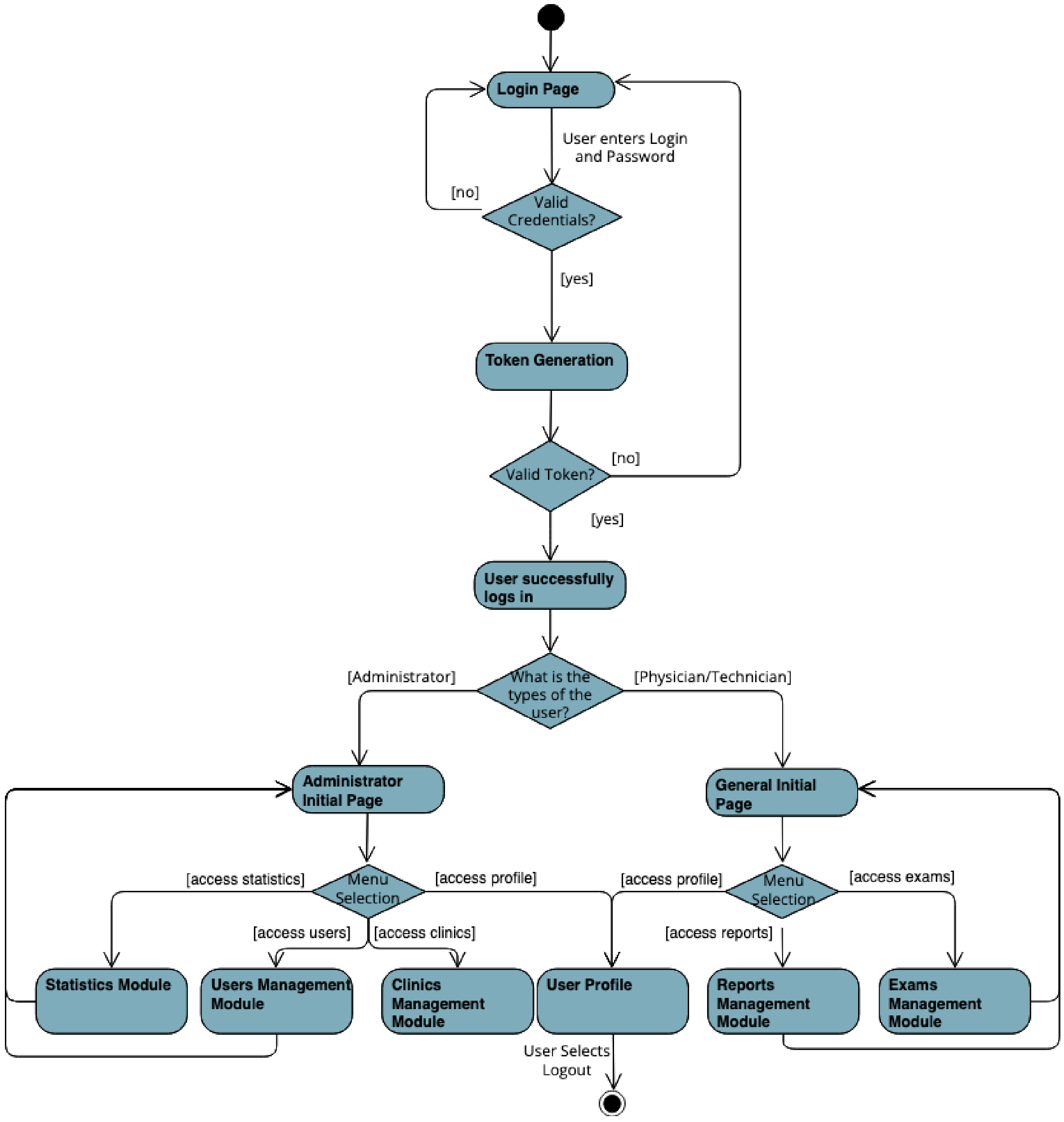


Figure 18: Workflow Diagram for the Authentication Module.

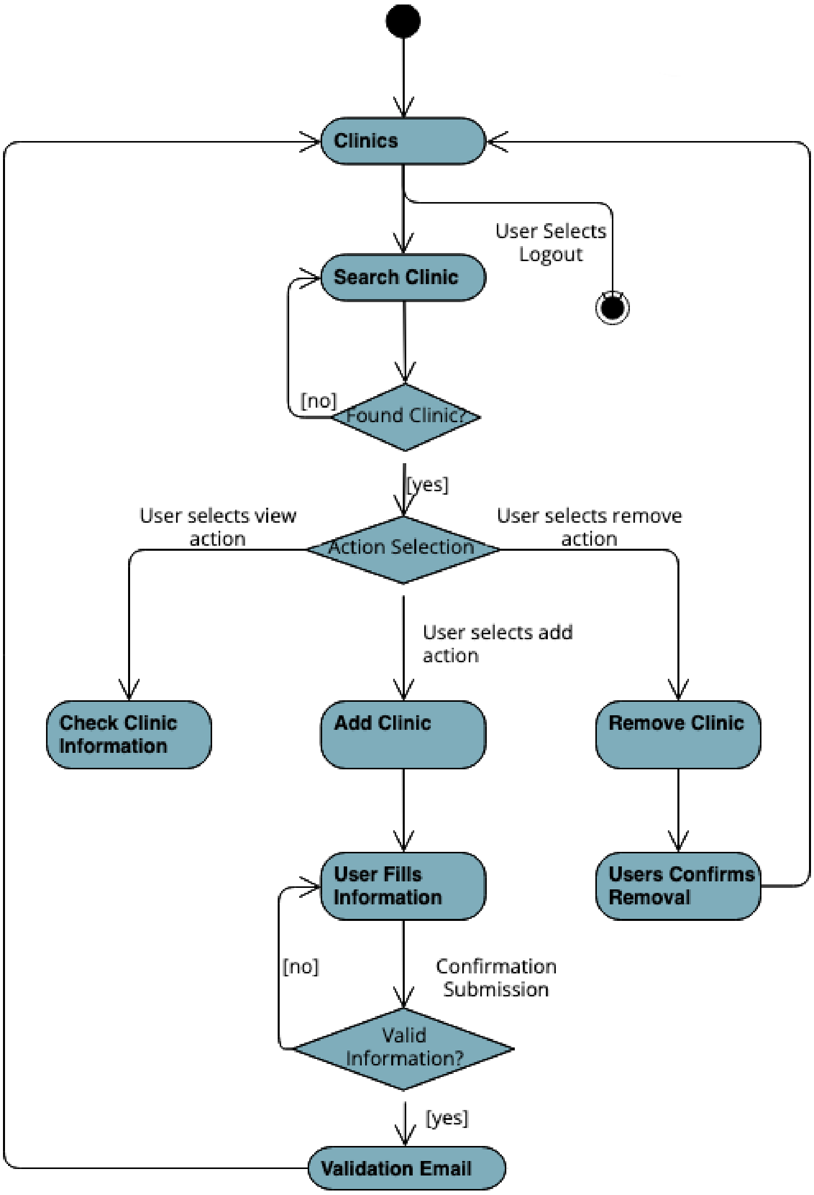


Figure 19: Workflow Diagram for the Clinics Management Module.

Users Management Module This module is similar to the one previously described. It allows the administrator to add users, edit user information and also delete users. To add a user, the administrator must first ensure that the user does not exist and that the clinic where you want to add the user is created. After that, he must fill in the four required fields (First Name, Professional Email, User Type and Clinic) and submit the request. After creating the user, you will receive an email with the password. All users of the medical type, who intend to elaborate and validate reports, must log into the platform and, in the profile, upload their signature. The image of the signature must be in .png or .jpeg format.

All users that are part of the Administrator's clinics and are logged in have an entry in the users' table. To edit information, just search for the desired user (the user can use the table filters) and click on "Edit Information". The form with the current information will be displayed. After making the desired changes, you should click ”Save”, and the changes will be validated.

To delete a user, the Administrator must search for the desired user and click ”Delete.” A confirmation message will serve to verify this activity. Once the confirmation is sent, the action is irreversible. The workflow of the complete user administration module is depicted in the diagram in Figure 20.

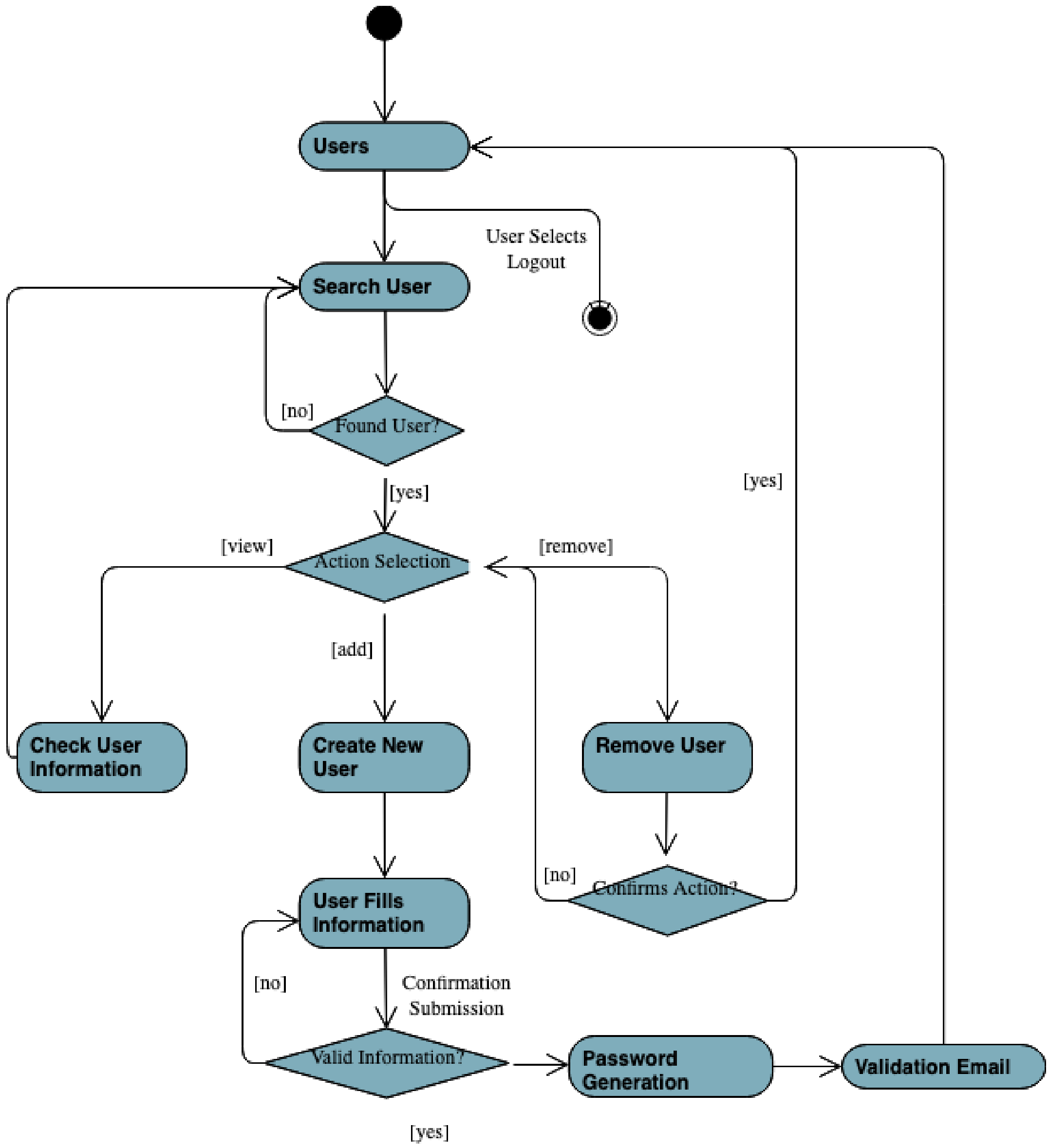


Figure 20: Workflow Diagram for the Users Management Module.

Statistics and Monitorization Module Administrators are able to access the statistics module. This page has several parameters, such as tests each month, tests per clinic, and the number of reports to validate, among others. This module is supplemented by user notification processes. For instance: When a technician submits a report for physician X to validate. Daily, the physician receives an email stating, ”There are X reports under your responsibility that are pending validation.”

Exams Management Module All unreported tests belonging to the clinic are in the ”Exams” view. The table contains the Patient Name (if available in PDF), the Test (PDF Name), the Date of Entry and the available actions. If you want to report an exam, the user must search for one of the available parameters and press the ”Report” button. The user will be forwarded to the page so that he can view the exam and the form where you fill in the User Name and the report. The first element that appears on the page to report is the PDF corresponding to the exam. The user can adjust the zoom and rotation of the PDF. Next, the fields to insert and/or correct the Name of the Patient, the Report and the physician to validate appear. If there is no physician preference, the selection must remain on ”None”. This way, the report will appear to all physicians. It is possible to set the indication ”Urgent” to make the test appear first in the list of physician reports to validate. Validations are made on the field of the name of the user, it must contain at least seven characters. Also, the report field cannot be empty. The submission is made through the ”Send to Validation” button. The action is confirmed through a confirmation window, where you must confirm the action. This action is irreversible. After confirming the action, the report is available in the list of reports to validate. It can appear only to the physician responsible or to all of them (if you have not assigned any).

Reports Management Module Under the ”Validate Reports” pane, all tests with reports are displayed. This view contains two sections: Reports assigned to the physician with the session begun and Reports without an assigned physician. In both tables, the Name of the User, the Exam (PDF Name), the name of the reporter (physician or technician), the Report Date, the Urgency (Green/Red), and the available actions are displayed.

Only the ”Validate Report” button is displayed if the user is a physician. The ”Validate Report” button exposes a page similar to the report preparation page but with the content already filled in. The physician must confirm all information before clicking ”Publish.” This button opens a confirmation box for the publication. When the ”Publish” button is clicked, a PDF is generated, which opens in a new window and is printable in the report management area.

The ”Remove Assignment” button is the only option displayed if the user is associated with the technician profile. The ”Remove Assignment” button provides a confirmation window for assignment removal. When the deletion is confirmed, the test is returned to the list of tests to report. Eliminating this assignment will require you to resubmit the exam-related report.

All users can access the published reports. Therefore, whenever you need to consult, you must access the ”Published Reports” page and search for one of the parameters available in the table. When you find the desired test, press ”Print” and you will be able to consult the respective PDF. Additionally, the user can send the report, by email, to the patient. The workflow of Figure 22 represents the entire flow of the reports module.

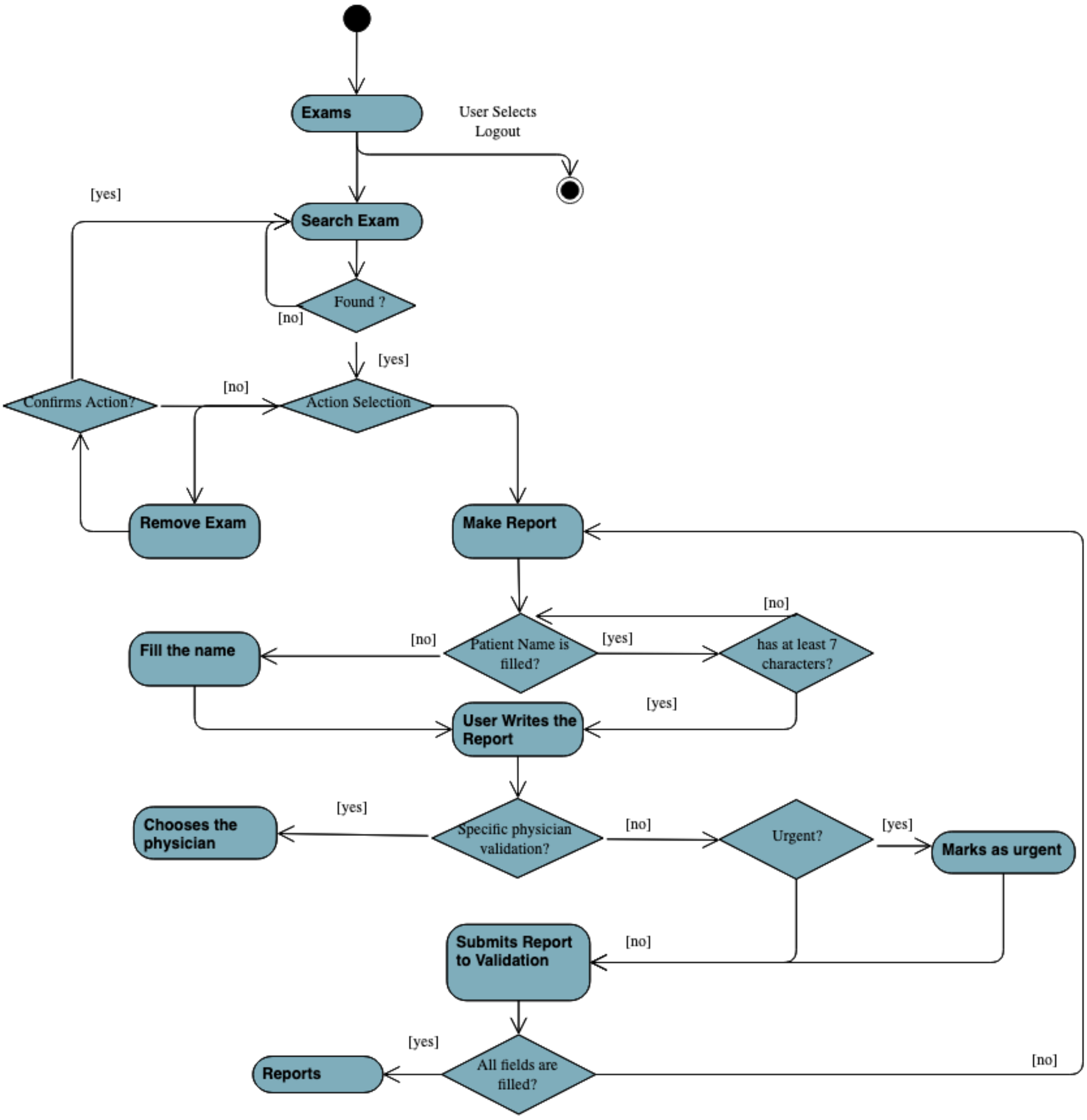


Figure 21: Workflow Diagram for the Exams Management Module.

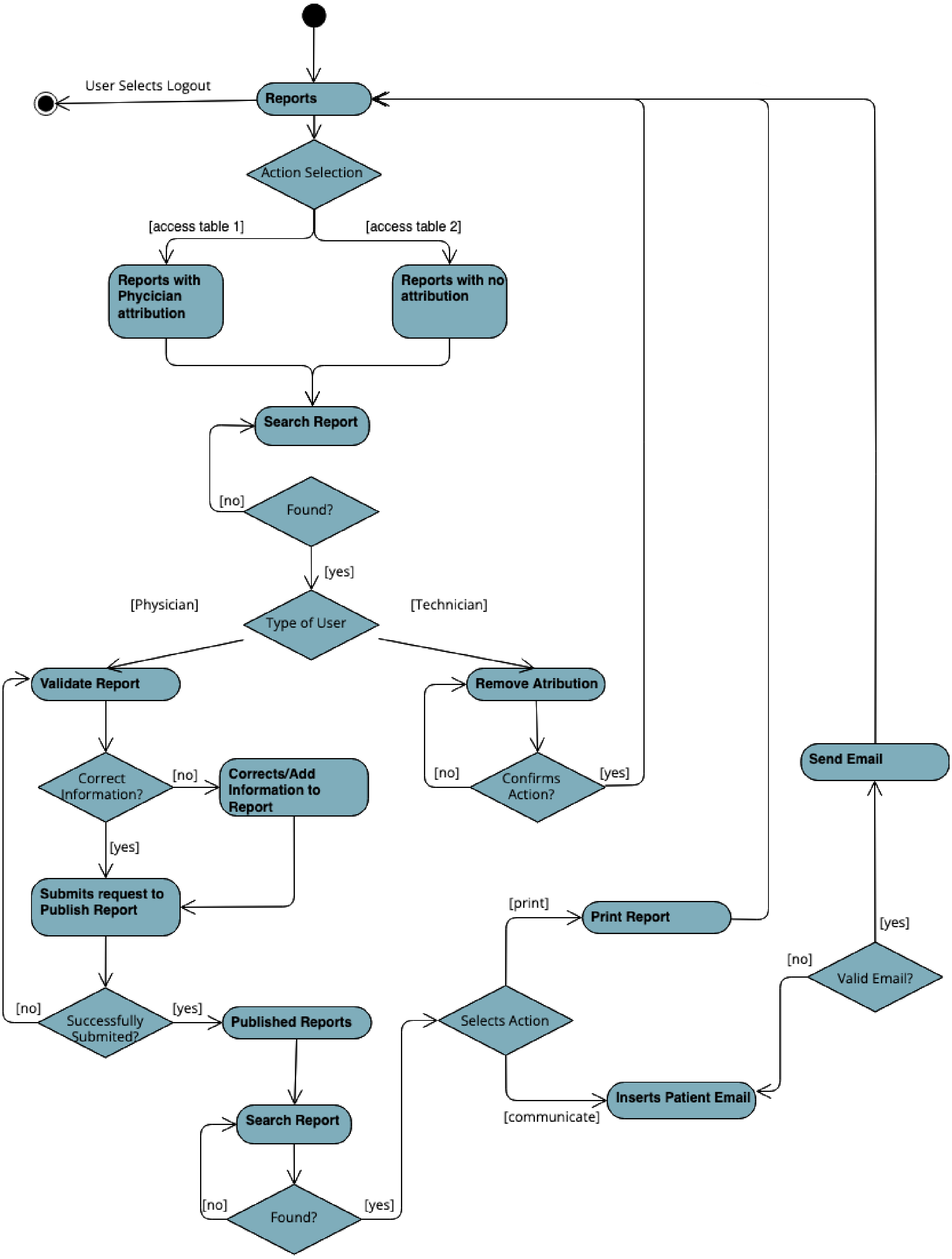


Figure 22: Workflow Diagram for the Reports Management Module.

In conclusion, the web application is comprised of various parts. Each user type (user role) has a unique set of permissions. So, only certain categories of users can access each module, and according to the user, each module gives a set of tasks that can be completed. Consequently, the rights provided to each user type per web module are presented in Table 8. It is vital to note the meaning of the following acronyms in the table:

- ”R”, denotes that the user has read access to the module, so he can inspect all of its contents.
- ”W”, denotes that the user has write permissions, which allows him to directly edit the data contained in the database module (insert or update) via the module.
- ”NA”, meaning ”Not Available”, denotes that the user has no access to the module or the functionality.

Table 8: Permissions applied for each type of user

	Manager	**Administrator**	**Physician**	**Technician**
Clients Module	RW	NA	NA	NA
Clinics Module Add Clinic	R	RW	NA	NA
Clinics Module Edit Informations	R	RW	NA	NA
Clinics Module Remove User	R	RW	NA	NA
Users Module Add User	RW	RW	NA	NA
Users Module Edit Informations	RW	RW	NA	NA
Users Module Remove User	RW	RW	NA	NA
Statistics Module	RW	R	NA	NA
Exams Module View Exam	NA	R	RW	RW
Exams Module Make Report	NA	NA	RW	RW
Exams Module Validate Report	NA	NA	RW	NA

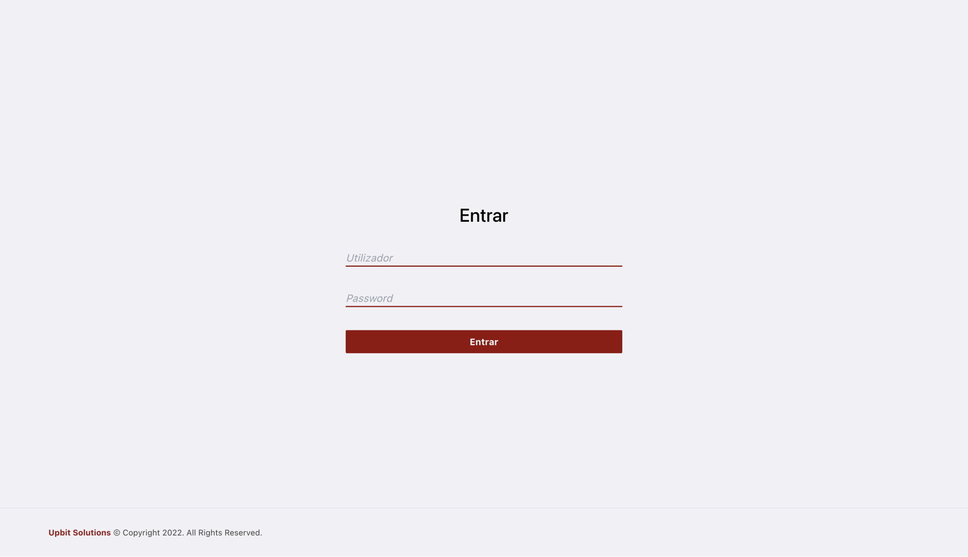
Table 8 continued from previous page

	Manager	**Administrator**	**Physician**	**Technician**
Exams Module Remove Atribution	NA	NA	RW	RW
Reports Module View Report	NA	NA	R	R
Reports Module Communicate	NA	NA	R	R
Reports Module Print	NA	NA	R	R

Throughout the development, security practices applied to Node.js such as modularization, clean code, code reuse, use of package manager npm to manage dependencies, use of ESLint to maintain code quality and consistency, use of process manager PM2 to manage and keep Node.js processes running in the background, use of Express framework to handle routing and middleware, use of environment variables to manage configuration settings, use of Winston library to log information about your application, database indexing, query optimization, normalization and encryption.

7.4 Results

As a result, the interface developed to comply with the workflows detailed in Section 7.3.2.4 is presented. The user's entry to the platform is pending registration and credential validation. Thus, a login page (Figure 23) was developed in order to allow access to the remaining pages.



Entrar

Utilizador

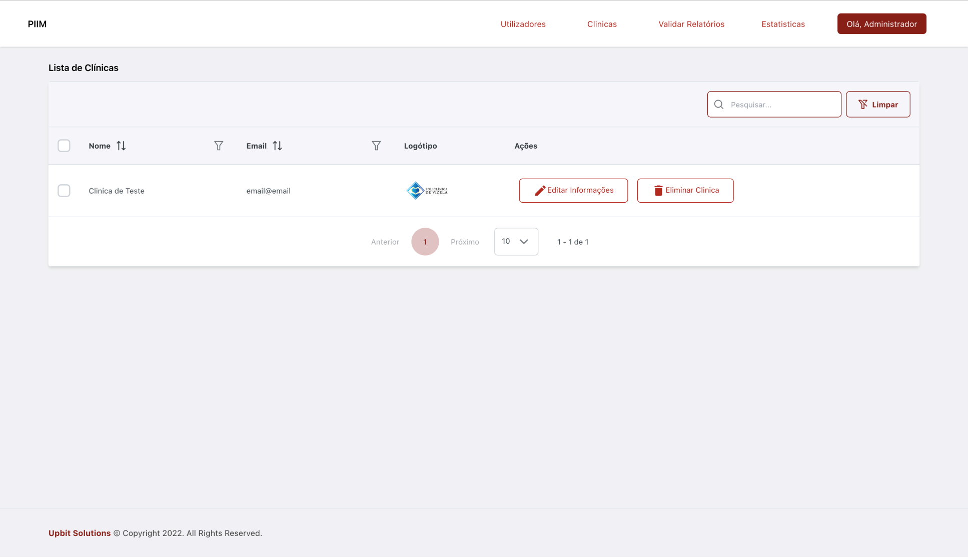
Password

Entrar

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 23: Login Page.

After validating the credentials and generating the token, the users will have access to the pages according to their type. Thus, if you are an Administrator, in the main navigation clinic management, illustrated in Figure 24 and user management (Figure 25) can be accessed.



Lista de Clínicas

Pesquisar...

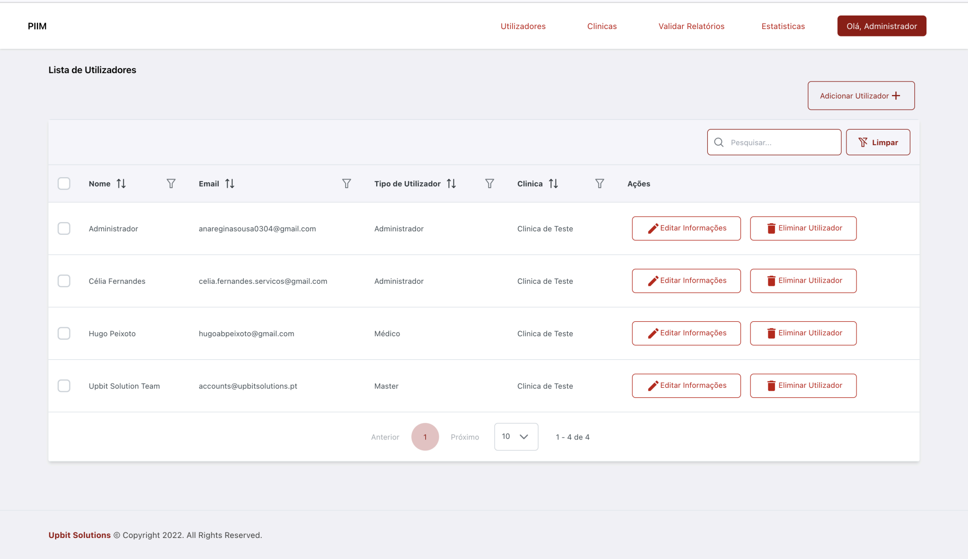
Limpar

□	Nome ⇅	Email ⇅	Logótipo	Ações
□	Clinica de Teste	email@email		Editar Informações Eliminar Clinica

Anterior 1 Próximo 10 1 - 1 de 1

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 24: Clinic Management Page.



Lista de Utilizadores

Nome	Email	Tipo de Utilizador	Clinica	Ações
Administrador	anareginasousa0304@gmail.com	Administrador	Clinica de Teste	Editar Informações Eliminar Utilizador
Célia Fernandes	celia.fernandes.servicos@gmail.com	Administrador	Clinica de Teste	Editar Informações Eliminar Utilizador
Hugo Peixoto	hugoabpeixoto@gmail.com	Médico	Clinica de Teste	Editar Informações Eliminar Utilizador
Upbit Solution Team	accounts@upbitsolutions.pt	Master	Clinica de Teste	Editar Informações Eliminar Utilizador

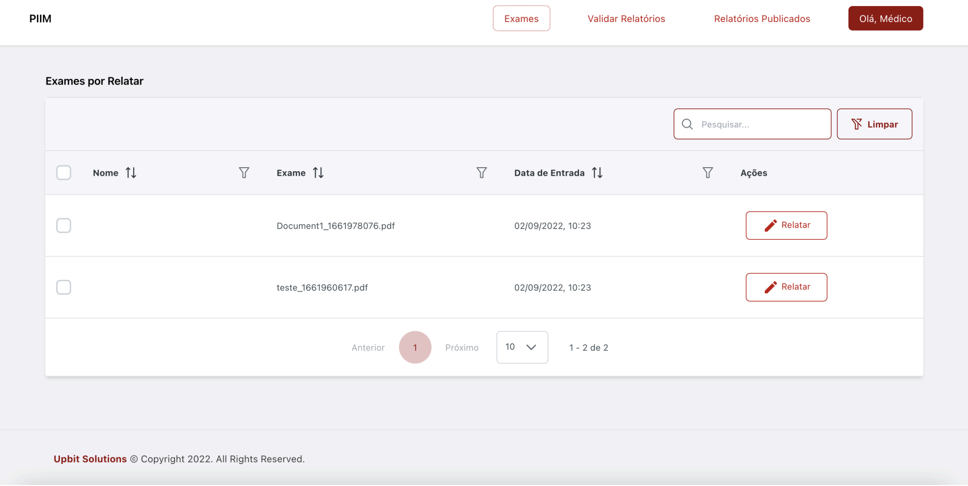
1 - 4 de 4

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 25: Users Management Page.

As illustrated in Figures 24 and 25, the Administrator can also see reports waiting for validation and a page with statistics for all the clinics.

If the user is a Technician or a Physician, the navigation bar will display exam management and report management (Figure 26). Both can report results using the button in the last column of the table on the main page (Figure 26). Then, a page will appear with the exam results and all the necessary fields to elaborate and send the report for validation, associating or not a specific physician to validate (Figure 27).



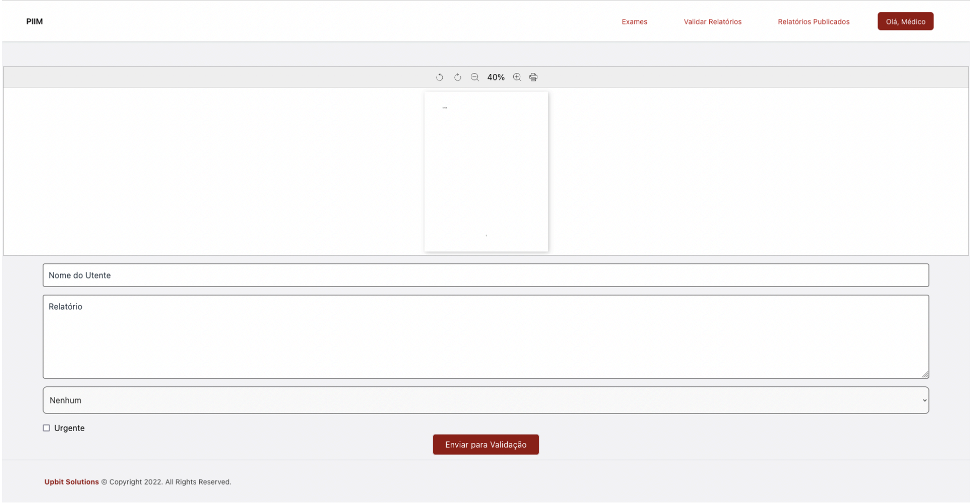
Exames por Relatar

Nome	Exame	Data de Entrada	Ações
	Document1_1661978076.pdf	02/09/2022, 10:23	Relatar
	teste_1661960617.pdf	02/09/2022, 10:23	Relatar

1 - 2 de 2

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 26: Exams Page (No report).



Nome do Utente

Relatório

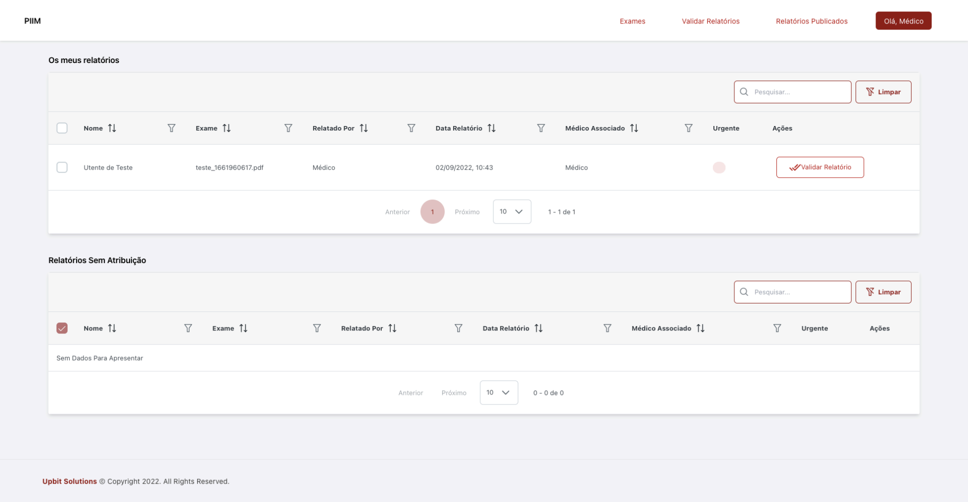
Nenhum

□ Urgente

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 27: Page to Make a Report.

After the report has been sent for validation, it appears for both the physician and the technician. For the physician, a page with two tables appears. In the first table, only the tests are associated with you, and therefore the validation is your responsibility. The reports without a specifically associated physician are in the following table, which is everyone's responsibility (Figure 28). The validation page shows all the information previously filled in by the technician, which can be corrected/modified by the physician before publishing (Figure 29).



Os meus relatórios

□	Nome	Exame	Relatado Por	Data Relatório	Médico Associado	Urgente	Ações
□	Utente de Teste	teste_1661960617.pdf	Médico	02/09/2022, 10:43	Médico	●	Validar Relatório

1 - 1 de 1

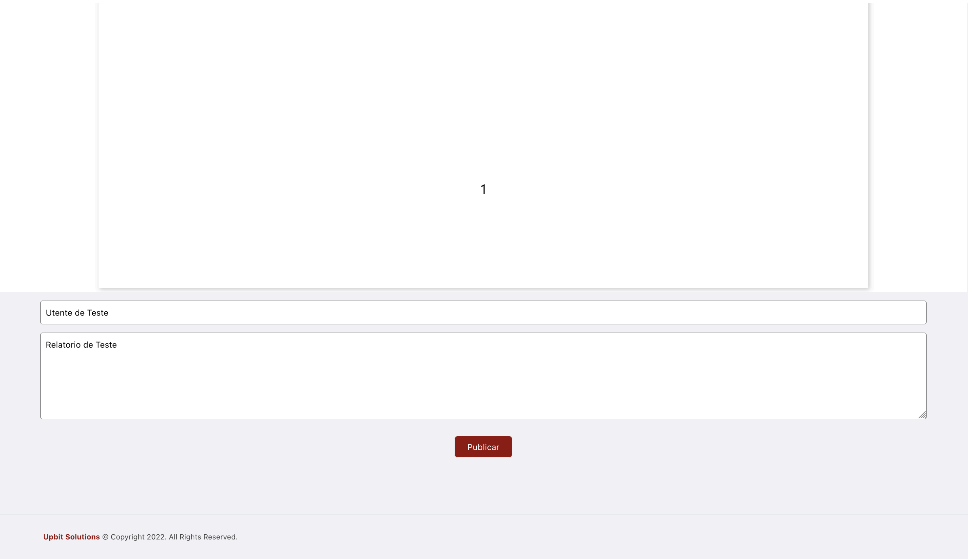
Relatórios Sem Atribuição

☑	Nome	Exame	Relatado Por	Data Relatório	Médico Associado	Urgente	Ações
Sem Dados Para Apresentar							

0 - 0 de 0

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 28: Pending Validation Page available for physicians.



1

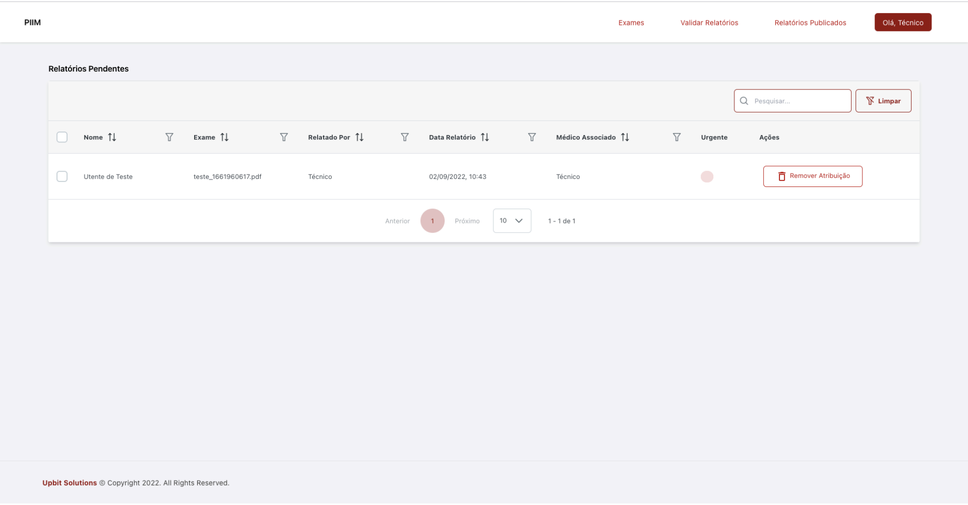
Utente de Teste

Relatorio de Teste

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 29: Verification Page.

In the case of the technician, he can only remove the assignment, thus allowing a faster flow of examinations or even brief corrections that may be needed (Figure 30).



Relatórios Pendentes

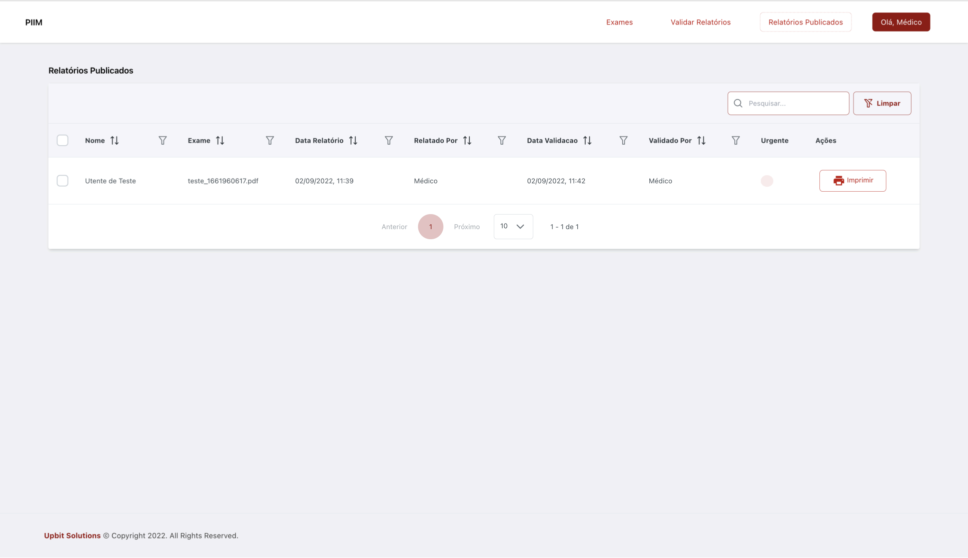
□	Nome ⇅	Exame ⇅	Relatado Por ⇅	Data Relatório ⇅	Médico Associado ⇅	Urgente	Ações
□	Utente de Teste	teste_1661960617.pdf	Técnico	02/09/2022, 10:43	Técnico	●	Remover Atribuição

1 - 1 de 1

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 30: Pending Validation Page available for technicians

Finally, the reports already published and the documents generated for them can be visualized and consulted, Figures 31 and 32.



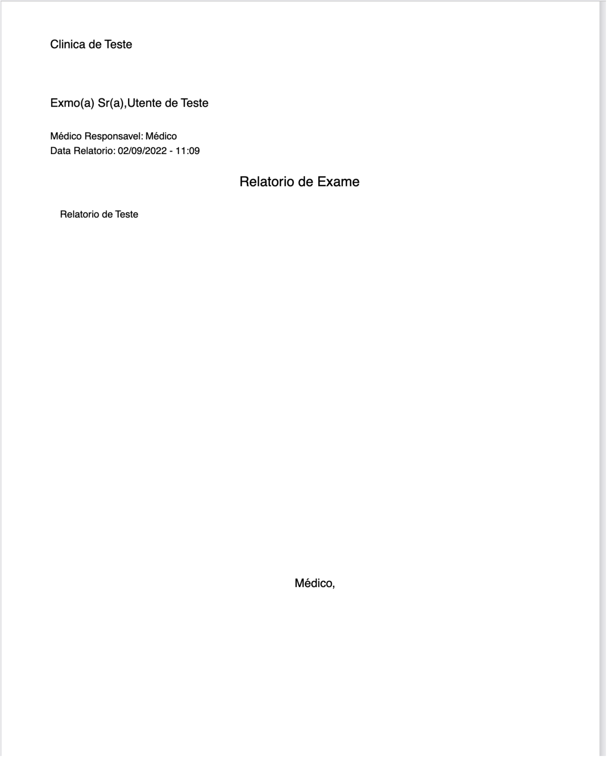
Relatórios Publicados

□	Nome	Exame	Data Relatório	Relatado Por	Data Validacao	Validado Por	Urgente	Ações
□	Utente de Teste	teste_1661960617.pdf	02/09/2022, 11:39	Médico	02/09/2022, 11:42	Médico	●	Imprimir

1 - 1 de 1

Upbit Solutions © Copyright 2022. All Rights Reserved.

Figure 31: Published Reports Page.



Clinica de Teste

Exmo(a) Sr(a),Utente de Teste

Médico Responsavel: Médico
Data Relatorio: 02/09/2022 - 11:09

Relatorio de Exame

Relatorio de Teste

Médico,

Figure 32: Example of the generated document.

7.5 Discussion and Conclusion

The developed solution intends to enhance the efficacy and accessibility of clinical test and analysis results reporting by providing a cost-effective and scalable choice for small and big institutions while avoiding the substantial software acquisition costs that are now necessary.

The solution employs the interoperability engine Mirth and the cloud computing platform Amazon to facilitate the integration of tests and results in real-time. The platform was built with React and Node.js, which are widely adopted open-source JavaScript technologies for creating online apps. Node.js facilitates backend development and server-side programming, while React focuses on constructing user interfaces. Using these technologies yields a robust platform capable of handling massive amounts of data.

The solution is highly scalable, allowing several clients to subscribe to plans and multiple users to access the program from any location. This is especially advantageous for healthcare businesses with multiple locations that need to share data. Cloud hosting increases the speed with which healthcare practitioners may obtain and implement results, hence enhancing patient outcomes. This method is highly adaptable, particularly in light of the new distant working paradigms.

Data security is a major problem in the healthcare industry, and this solution addresses this issue by incorporating multiple security mechanisms. Encryption guarantees that patient information is secure and accessible only to authorized individuals. Limited access and updates give an additional degree of protection by limiting data access and modification to authorized individuals. Also, the utilization of secure servers AWS adds another layer of protection. All sensitive data, including the Uniform Resource Locators (URL) of the API and the JSON web login token, is stored on the server. This ensures that data is safeguarded even if a security incident occurs. The PDF is secured using a code created at random.

Yet, there are also potential disadvantages and risks to consider. The system relies significantly on the AWS infrastructure and, by extension, the internet, which could be problematic in the event of a service interruption. Therefore, data privacy is a paramount concern. Even with security precautions in place, data breaches or illegal access to patient information is always possible. Integration of clinical outcome reporting with current systems can be difficult and may necessitate additional resources and training. Some organizations may be unable to deploy and sustain the solution since it may involve a substantial investment of time and resources.

8

Case Study III: The Covid-19 Influence on the Population Desire to Stay at Home

8.1 Introduction and Problem Identification

The Covid-19 pandemic has had a huge impact on people's daily life across the globe. On February 8, 2023, WHO had received reports of approximately 755 million confirmed cases of Covid-19, including approximately 7 million deaths. Considering the global population of 7500 million, only roughly 10% of the population reported having Covid-19 to the WHO [153]. However, the reality is that most of the population has already dealt with a serious case of Covid-19 and possibly even mortality. This is frightening and has implications for society as a whole.

One of the most visible changes has been how people interact with their surroundings, with many individuals present and, most significantly, opting to stay at home as much as possible to avoid getting the virus. This conduct is reasonable but has profoundly affected civilizations' economic, mental health, and social fabric.

This has resulted in a considerable decline in the number of individuals travelling to work, attending social events, and participating in recreational activities, which has greatly impacted the economy, mental health, and social cohesion. Remote work was implemented and eventually became a way of life. Increasingly, individuals have begun to place a higher value on their homes, preferring those with open and green spaces. Unfortunately, it is frequently impossible to execute work duties from home in various occupations, resulting in unemployment or the search for alternative employment [75, 23].

In light of this, by analyzing the causes and effects of the Covid-19 pandemic, this research aims to serve as a model for governments to better comprehend the many assessment criteria, particularly the adherence of individuals to stay at home and its effect on the economy. In order to lessen the severity of future pandemics, it is possible to develop more effective legislation for each country and culture by analyzing the population's preference to remain at home as well as other potential factors.

This chapter explores, using Big Data technologies, the impact of Covid-19 on the population's desire to stay at home, analyzing the reasons for this behaviour and the potential long-term consequences.

8.2 Objectives

The following objectives were outlined:

- Understand the effect of the Covid-19 pandemic on individuals' preferences for staying at home and the explanations behind these preferences.
- Identify the factors that contribute to people's decisions to stay at home and how these factors may vary between different demographic groups.
- Provide a baseline model for governments and policymakers to better understand the factors influencing people's adherence to staying at home and its economy impact.
- Provide insights and recommendations for governments and businesses to better support individuals during the Covid-19 pandemic.

8.3 Design and Development

The many aspects of the proposed new system's intended audience include government entities and health professionals, particularly those with administrative positions or authority. It is important to note that this system considers several functional requirements for development. Among all of them, the following stand out:

- Ability to collect data from various sources (static and/or dynamic).
- Identifying the selected data and standardizing it.
- Identification and treatment of null values and outliers.
- Dynamically data merging.
- Ability to analyze data that comes from static or dynamic sources.
- Building visually appellative and interactive dashboards.
- Ability to build and share reports.

To fulfil these requirements, the following tasks were outlined:

- Identify relevant and representative data sources.
- Data Collection.
- Standardization of codes and dates.
- Identifying null values and dealing with them.

- Merging data.
- Load the data to the visualization platform.
- Dashboard development.

8.3.1 General Architecture

The architecture used for the implementation of the case study, as depicted in Figure 33, consists of several steps generally featured in designs containing Big Data tools. As only public and free data repositories were used, the process begins with the capture and standardization of data through several changes before being shown on the visualization platform.

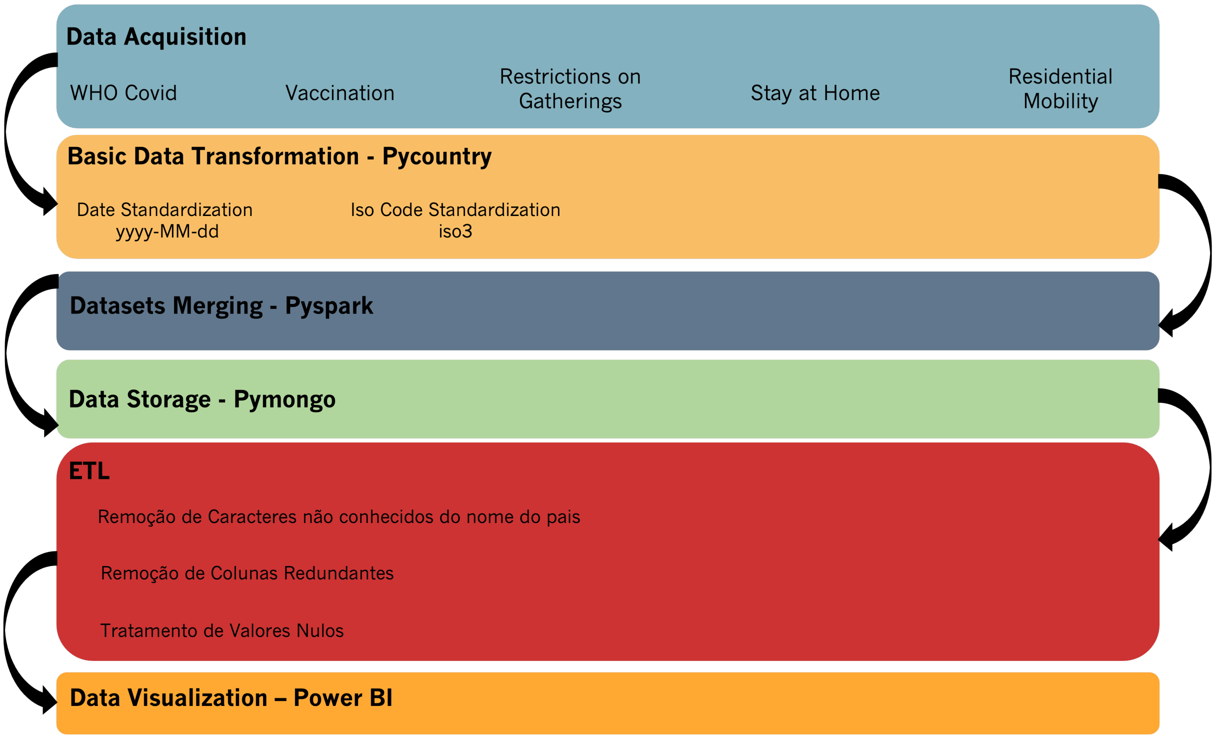

Data Acquisition
WHO Covid
Vaccination
Restrictions on Gatherings
Stay at Home
Residential Mobility
Basic Data Transformation - Pycountry
Date Standardization yyyy-MM-dd
Iso Code Standardization iso3
Datasets Merging - Pyspark
Data Storage - Pymongo
ETL
Remoção de Caracteres não conhecidos do nome do pais
Remoção de Colunas Redundantes
Tratamento de Valores Nulos
Data Visualization – Power BI

Figure 33: General architecture implemented in the study.

The steps shown will be detailed in depth in the sections that follow.

8.3.2 Development Stages

The suggested study's development phases were based on the normal data flow in a Big Data architecture. Thus, the five steps between data collecting and data visualization are outlined below.

8.3.2.1 Data Acquisition

Big Data solutions start with one or more data sources. Some examples include:

- Application data storage, such as relational databases.
- Static files produced by applications.
- Real-time data sources, such as IoT devices.

Five public datasets with recently collected data were used in this case study. Since the intent is to study the impact of the pandemic on society, the first dataset selected was from the World Health Organization and referred to cases of Covid-19. Then, a second dataset, also published by the World Health Organization, was chosen to complement the previous one, this one related to vaccination. Having said that, and since the case study focuses on people's preference to stay at home, three other datasets were selected, one related to contingency measures imposed by the government regarding gatherings, the second also with restriction measures but associated with going outdoors, and the last one related to people's willingness to stay at home. The five datasets are presented in detail below.

WHO Covid-19 The first data set includes global data associated with the Covid-19 pandemic in 236 countries. From January 3, 2020, to March 1, 2022, data were reviewed for the period under consideration. This dataset comprises as attributes the metadata related to the country: ISO country code, name, and World Health Organization (WHO) region, as well as the registration date. In addition, it includes attributes associated with the topic, such as the number of new Covid-19 cases each day, the number of cumulative Covid-19 cases, the number of new Covid-19 deaths per day, and the total number of Covid-19 deaths. 268286 records were evaluated in total.

WHO Vaccination The second data set illustrates the progress of the vaccination procedure in 235 countries between January 22, 2021, and March 27, 2022. Many metrics and features were employed to represent this immunization procedure. It should be emphasized, however, that not all observations have values for all of these characteristics, which is an additional risk. This dataset has 14 variables that represent the vaccination process. For the study of the evolution of vaccinations, the following characteristics were deemed important:

- country: name of the country.
- iso_code: International Standardization Organization (ISO) code associated with the country name.
- date: observation date.
- total_vaccinations: total number of doses administered. For vaccines that require multiple doses, each dose is counted.
- daily_vaccinations: new doses administered per day.
- people_vaccinated: total number of people who received at least one dose of vaccine.

- people_fully_vaccinated: total number of people who received all the doses prescribed by the initial vaccination protocol, i.e. 2 or 1 dose depending on the vaccine administered.
- total_boosters: total number of booster doses administered.
- daily_people_vaccinated: daily number of people who received the first dose of vaccine

A total of 94730 records were considered.

Restrictions on Public Gatherings The third dataset covers the public gathering limitations imposed on each day (from January 1, 2020, to March 21, 2022) in 186 countries. Hence, the dataset is separated into four attributes: country name, country code, day, and constraints applied. These constraints are denoted by values between 0 and 4, where:

- 0: no measures.
- 1: gatherings forbidden to more than 1000 people.
- 2: gatherings forbidden to more than 100 and less than 1000 people.
- 3: gatherings forbidden to more than 10 and less than 100 people.
- 4: gatherings forbidden to more than 10 people.

A total of 152791 records were considered.

Stay at home This data collection includes the daily recommendations to stay at home (from January 1, 2020, to March 21, 2022) in 186 nations. Hence, the dataset contains four columns: nation, country code, day, and restrictions applied. This indicator measures the increase in the percentage of individuals who opt to remain at home during the pandemic. This rise is comparable to the number of people who preferred to remain at home prior to the outbreak. These constraints are denoted by values between 0 and 3, where:

- 0: No measures.
- 1: Recommended not leaving the house.
- 2: Require not leaving the house with exceptions for daily exercise, grocery shopping, and 'essential' trips.
- 3: Require not leaving the house with minimal exceptions (only 1 member of the family can go out to the grocery store, for example)

152790 records were considered.

Residential Mobility This dataset links each day (from February 17, 2020, to January 31, 2022) and each of the 129 countries with a metric measuring people's willingness to stay home. This metric measures the increase in the percentage of people who prefer to stay at home during the Covid-19 pandemic, compared to typical behaviour before the pandemic.

For example, in the first dataset entry, it can be seen that on February 17, 2020, in Afghanistan, there was a 1.33% increase in people preferring to stay at home compared to the percentage of people before the pandemic. Apple mobility datasets and Google mobility reports from 2019 were used to calculate the percentage. The mobility reports were extracted before and after the covid, and the % difference in the movement was obtained.

The data was collated with government intervention data for granular breakdown. In total, 91933 records were considered.

8.3.2.2 Data Standardization and Merging

With the aforementioned dataset 1, the process began by reducing the country code (ISO code) from three to two characters. Due to the fact that the other datasets assign each country a 2-character code instead of a 3-character code, it was possible to unify the five datasets by this code. The Python library PyCountry was used to gather information about many countries. With this library, it was possible to retrieve the various names and related 2- and 3-character (alpha 2 and alpha 3) codes for each country. The 3-character code corresponding then replaced the column of dataset 1 with the code found in dataset 1.

In dataset 5, it was verified that the Day column had the dates in dd-MM-yyyy format, which is not concurrent with the other datasets. Thus, it was necessary to change the format of the dates in the fifth dataset so that all the datasets were compatible. The format adopted was yyyy-MM-ddd.

In addition, it was determined that some data records lacked an ISO code, requiring the completion of these instances so that the merge procedure would not disregard them. So, for each dataset, it was obtained a country list with no code. The countries identified in dataset 1 were Kosovo, Namibia, Bonaire Saba, St Eustatius, and Other. In the Pycountry library, only the ISO code for Namibia is recognized. The remaining records will therefore be processed during the Data Transformation stage. Laos and Côte d'Ivoire appeared without code in dataset 4. In this instance, the issue could only be resolved by aligning their names with the Pycountry library. Consequently, Lao People's Democratic Republic became Lao People's Democratic Republic, while Côte d'Ivoire became Côte d'Ivoire. There are no uncoded countries in datasets 2, 3, and 5.

Finally, iteratively, the merge of several datasets was performed. Previously it was necessary to rename the columns used in this joint to have the same name in the five datasets.

The implemented architecture uses Python's Pandas library to merge the five datasets. This library was chosen because it is appropriate for heterogeneous datasets consisting of tables containing several data types, allowing better use of these objects.

8.3.2.3 Data Storage

After standardizing and merging the selected datasets, the process advanced to the storage. Mongo was selected. MongoDB is a popular NoSQL database type. This type is frequently employed in Big Data applications. However, it is not a ”Big Data” database but a flexible and scalable database management system built to store and manage massive amounts of data. Nevertheless, it enables the storage and retrieval of unstructured data optimally for the types of data typically seen in Big Data use cases, such as log, time-series, and hierarchical data.

MongoDB is, a well-known tool for securely and versatilely storing JSON documents. This database offers a query language that enables you to easily extract information from the original document. MongoDB is frequently combined with other tools, such as Hadoop and Spark, to provide a comprehensive data processing, analysis, and storage solution in Big Data contexts.

As MongoDB is a document database, *to_dict* of the Pandas library was the method used to convert the dataframe containing the datasets to a dictionary. Next, the MongoClient method of the Pymongo library was used to establish a connection to the collection and database where the JSON file will be stored.

The *insert one* method of this library allowed the document to be inserted into the chosen collection. The MongoClient connection technique was used in conjunction with a query that returns the entire collection’s contents to obtain the stored data. This information will be saved to a Comma-Separated Values (CSV) file and processed.

8.3.2.4 Data Transformation

After storing the data in MongoDB, it was determined that the final dataset required treatment, and the treated data would be loaded into Power BI in the future. Notice that the stages and order of the provided architecture make this an Extract, Transform, Load (ETL) process, as the data previously saved in the MongoDB database are first extracted (Extract), then processed (Transform), and then loaded for the data visualization platform (Load).

PySpark was used to implement the transformation detailed in the following sections. This tool is best for iterative and sophisticated Big Data approaches. This library provides an Apache Spark API that applies the tool’s methods and functionalities in a Python context, retaining library consistency.

The transformation process begins with the creation of a Spark session. After that, all that’s left is to convert the dataset into a Pyspark dataframe object, on which the necessary transformations are performed. All the changes made will be described next:

1. **Country Renaming**: To facilitate subsequent data visualization, the names of the countries with unrecognized characters were changed. In the data of this case study, only the country Côte d’Ivoire (in Portuguese) presented this problem, and its name was changed from ”Côte d’Ivoire” to ”Cote de Ivoire”.

2. **Hiding Redundant Columns**: After a considered analysis, it was concluded that in the final dataset, there were some redundant columns or with little relevance for the case study in question, so they could be ignored. The hidden columns were: daily vaccinations raw, total boosters per hundred, daily vaccinations per million and people vaccinated per hundred.

3. **Treatment of Null Values**: It was necessary to analyze different cases and treat them in different ways, as explained in the sections below.

   a) **Records prior to the start of vaccination**: The initial null value treatment targeted records before vaccination for each country. Since the final dataset combines numerous datasets with different date ranges. Example: vaccine dataset. Only late 2020–early 2021 recordings exist (depending on the country in question). Hence, with all the data, entries before the previously indicated dates should have null values replaced by zero.

   b) **Countries with many missing values hidden**: Countries with insufficient data to generate synthetic values for missing fields have been identified. This approach ignores countries with columns with many null values since the data in these columns are not enough to generate realistic synthetic values that follow the same distribution. The dataset's null values were evaluated to find that 20% of fields in each country are empty. So, most countries could develop accurate synthetic values. Several countries with columns lack enough data to generate this data. For these circumstances, scrolling through the countries was performed, and the percentage of missing values for each column was calculated. It was decided to ignore the countries with more than 60% since this guarantees that the countries that remain in the dataset after this treatment have, for each column, a set of significant values above 50% that allow the generation of realistic synthetic values. Any records from a country with more than 60% missing values will be ignored.

   c) **Generation of synthetic values**: This transformational process was divided into two sections. In the initial phase, the fields were populated using the weekly mean values. In the second phase, missing values will be filled with the linear interpolation result using the previous and subsequent weeks' values. This transition began with the association of the year and week of the year in which each data was collected. Using this method, it was possible to group the records by country, year, and week of the year, and this grouping was utilized to generate the missing data in a realistic and reasonable way. The preceding procedure was only performed on columns with numeric values. When no values exist for any day of the week in a given nation, only linear interpolation using the values from the most recent preceding and/or succeeding weeks of that country was implemented.

8.3.2.5 Data Visualization

In this final phase of the architecture, the goal is to develop dashboards for analyzing the impact of the Covid-19 pandemic on the population's propensity to remain at home. PowerBi was used to design the dashboards, which enabled the creation of problem-specific diagrams in order to answer the use case. This tool was chosen since it is user-friendly, accessible, and free. When visualization data is not obtained in real-time, there is no need for more complicated tools.

After importing the information into PowerBi, the variables and attributes were connected to generate the problem-representing dashboards.

The analysis began with the development of a dashboard that enables the examination, over time and for each country, of the impact that the number of deaths and the number of individuals infected by Covid-19 had on the gathering and mobility restrictions. This dashboard now includes a graph for assessing the impact of the vaccination procedure and the values of new cases not only on the limits imposed but also on the preference of individuals to remain at home (residential stay). With this dashboard, one may assess how Covid-19 and the vaccination process influenced the limits imposed and people's willingness to remain at home.

8.4 Results

Analyzing the distribution of new cases recorded over the three years. In 2020, there are 46.1% of records, and in 2021, 46.2% of records. And in 2022, 7.7% of records. There is a positive correlation between the number of new cases and deaths, since countries with more cases also have more deaths.

The offered work analyzes three nations: the country where the study was conducted, Portugal (Human Development Index (HDI) - 0.866; 38th in the ranking), a developed country, USA (HDI - 0.926; 21st in the ranking), and a less developed nation, India (HDI - 0.633; 132nd in the ranking). As can be seen, the latter two also reflect the two nations with the most cases. Portugal had the highest number of instances per million residents, taking into account each country's approximate population at the time of the investigation (10 million for Portugal, 332 million in USA and 1408 million for India). Moreover, they are all among the 30 nations with the highest incidence of covid. In light of the diversity existing in these countries, it is believable that the study of these three countries will be representative of reality.

Two dashboards, depicted in figures 34 and 35, made it possible to conduct the analysis.

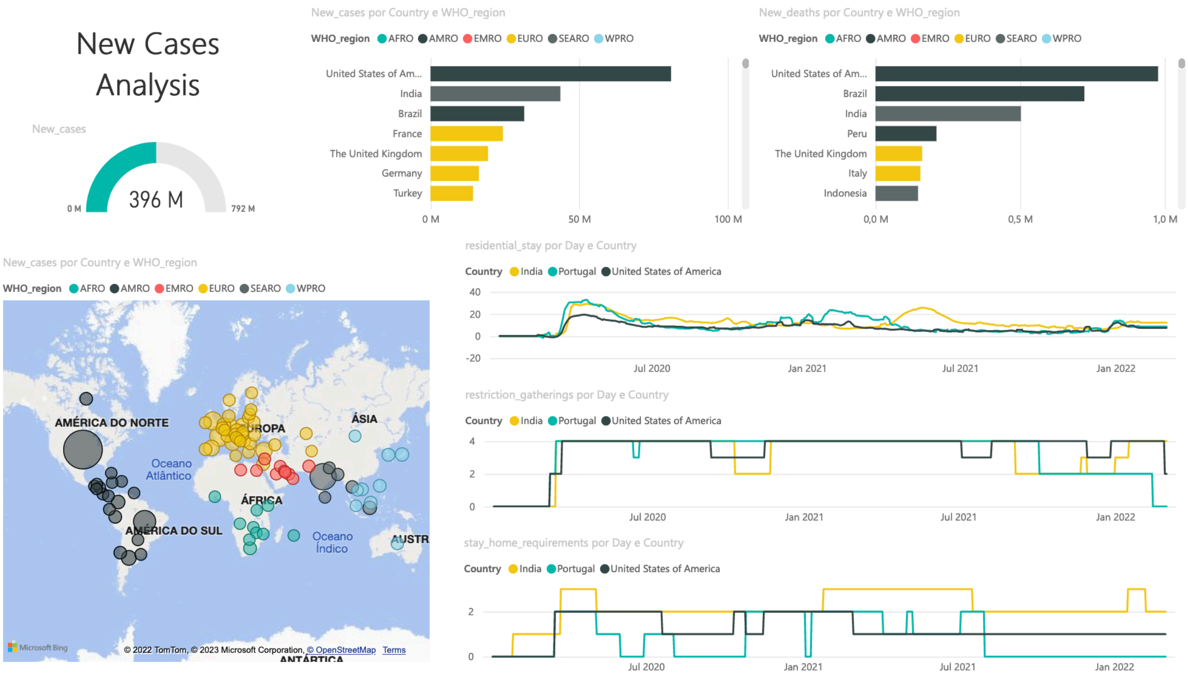

New Cases Analysis
New_cases
0 M
396 M
792 M
New_cases por Country e WHO_region
WHO_region ● AFRO ● AMRO ● EMRO ● EURO ● SEARO ● WPRO
United States of Am...
India
Brazil
France
The United Kingdom
Germany
Turkey
0 M
50 M
100 M
New_deaths por Country e WHO_region
WHO_region ● AFRO ● AMRO ● EMRO ● EURO ● SEARO ● WPRO
United States of Am...
Brazil
India
Peru
The United Kingdom
Italy
Indonesia
0,0 M
0,5 M
1,0 M
New_cases por Country e WHO_region
WHO_region ● AFRO ● AMRO ● EMRO ● EURO ● SEARO ● WPRO
AMÉRICA DO NORTE
Oceano Atlântico
EUROPA
ÁSIA
ÁFRICA
AMÉRICA DO SUL
Oceano Índico
AUSTR
ANTÁRTICA
Microsoft Bing
© 2022 TomTom, © 2023 Microsoft Corporation, © OpenStreetMap Terms
residential_stay por Day e Country
Country ● India ● Portugal ● United States of America
40
20
0
-20
Jul 2020
Jan 2021
Jul 2021
Jan 2022
restriction_gatherings por Day e Country
Country ● India ● Portugal ● United States of America
4
2
0
Jul 2020
Jan 2021
Jul 2021
Jan 2022
stay_home_requirements por Day e Country
Country ● India ● Portugal ● United States of America
2
0
Jul 2020
Jan 2021
Jul 2021
Jan 2022

Figure 34: General dashboard presenting the evolution of new cases by country against the metrics of interest.

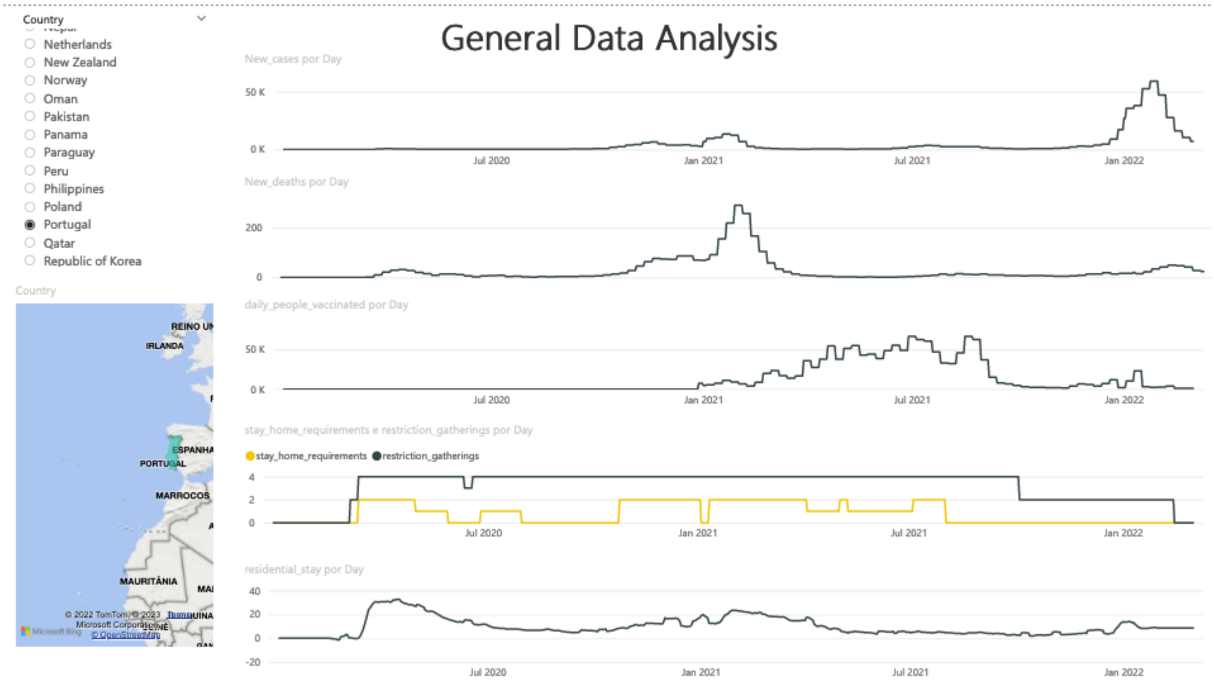


Figure 35: General dashboard presenting the evolution of the metrics of interest, worldwide.

Each of the selected countries will be analyzed in detail in the following sections.

8.4.1 Portugal

As indicated in the first graph of Figure 36, Portugal had two peaks in the number of new cases. On January 18, 2021, the first of about 13544 cases were documented. This equates to approximately 1,354 new cases per million people. The second peak occurs on January 25, 2022, and consists of roughly 59194 new cases. The ratio has already reached around 5,914 infected individuals per million, which is more than four times the previous peak.

On the other hand, a peak in the daily number of deaths was seen on January 27, 2021, as depicted by the second graph in Figure 36. Beginning in February 2022, there will be a second, not-so-significant hike. From the country's vaccination graph, it is evident that vaccination was gradual, beginning in December 2020 and ending in September 2021. By analyzing the first three graphs of Figure 36, it can be concluded that although the second peak of daily cases was much more pronounced than the first, this peak did not exacerbate the number of deaths because the majority of the population was already vaccinated and therefore more resistant to the virus's effects.

Figure 42 demonstrates that, of the three countries studied, Portugal has the greatest reduction in mortality, which may be explained by the fact that it is also the country with the highest vaccination rate. This is the expected behaviour; however, the study of actual data must confirm it. In this approach, proper hygiene and health behaviours can be promoted in the population in an informed manner during a future recurrence of this type by guaranteeing that government efforts and vaccination lead to improvements in the population.

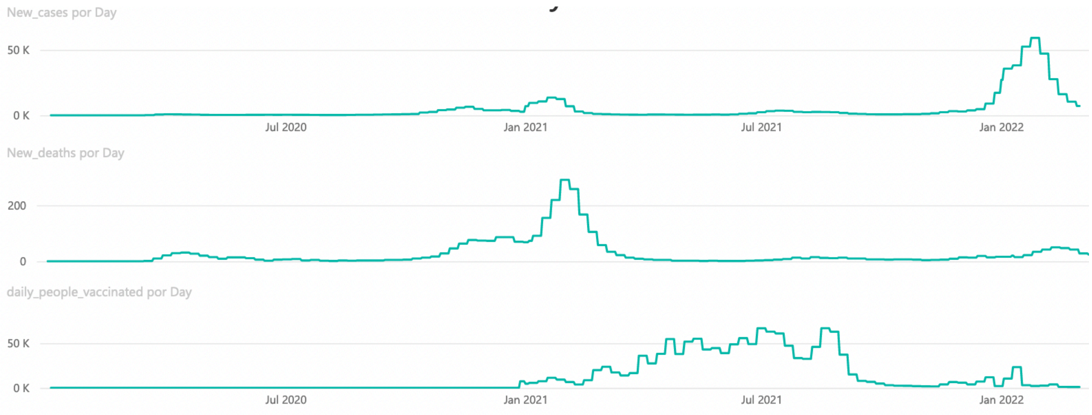


Figure 36: Variations in the number of new cases in Portugal throughout time.

The gatherings restrictions remained at their highest value, 4, until October 2021, when the vaccination process was already very advanced, and much of the population had at least one dose of vaccine. This analysis is supported by Figure 37. The confinement restrictions, on the other hand, varied over time, and the highest value reached was level 2. After April 2020, there was a reduction in the number of deaths, which translated into an easing of the containment restrictions, and the restrictions went from 2 to 1 and finally to 0. However, at the beginning of the vaccination process, there was also a peak in the number of deaths, which translated into a passage of the containment restrictions to levels 2 and 1. The increase in the number of vaccinated people translated into a decrease in the number of deaths. Consequently, the restrictions were relaxed to the value 0 - no measures. The conclusion is that, after rigorously analyzing the data, the governmental measures are not preventive but reactive. Probably if there were preventive measures, the number of deaths would have been lower since whenever the confinement restrictions are increased, the number of deaths decreases.

Having said this, and focusing on people's willingness to stay at home compared to the pre-pandemic period, it can be seen that the previously mentioned increase in the number of deaths and the resulting restrictions led to an increase in the population's willingness to stay at home and that this preference decreased when the number of deaths decreased and confinement was raised. In contrast, when the peaks of deaths occur in January 2021, the population's propensity to remain at home will increase. As the vaccination procedure advances and the percentage of the populationvaccinated population increases, people's willingness to stay at home declines to values near zero, resembling those observed before the pandemic. In addition to the huge increase in the number of instances in January 2022, there was also an increase in the propensity to stay at home, likely due to individuals feeling less safe outside.

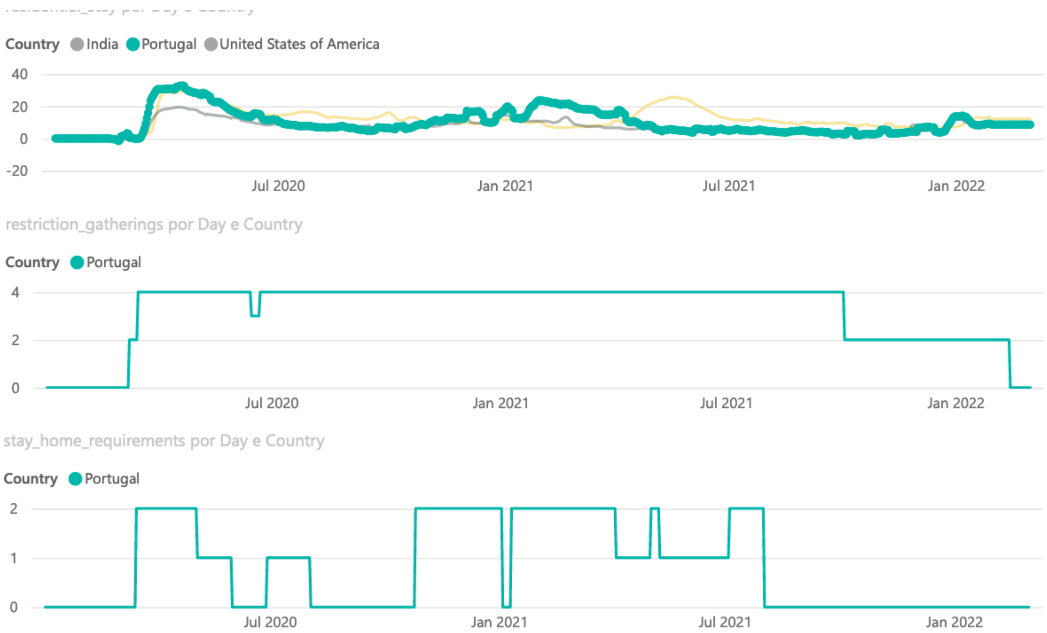


Figure 37: Variation of the preference index to remain at home in Portugal in relation to restrictions and contingency measures.

8.4.2 USA

There are three peaks on the chart illustrating the distribution of new cases of Covid-19 in the USA. The first was on January 4, 2021 (243226 new cases), the second on August 30, 2021 (176645 new cases), and the third on January 11, 2022 (far more severe) (850082 new cases). The ratios of new cases per million individuals for the three conditions are 732, 560, and 2560, respectively.

Similarly, a positive association exists between the previously investigated metric and the number of recorded deaths. The largest peaks in recorded deaths follow the largest peaks in new cases. Since a portion of the population was already partially immunized, the third peak of new cases was not proportional to the number of deaths, even though this is not the country with the highest vaccination rate (only about 76% of the population was immunized during the period of analysis). Peak vaccination rates occurred between January 10 and June 6, 2021.

Again, this is predicted, as vaccination has been demonstrated to reduce the severity of infection-related symptoms. In the future, these dashboards could be made accessible to the general public in order to improve literacy and awareness of the issue.

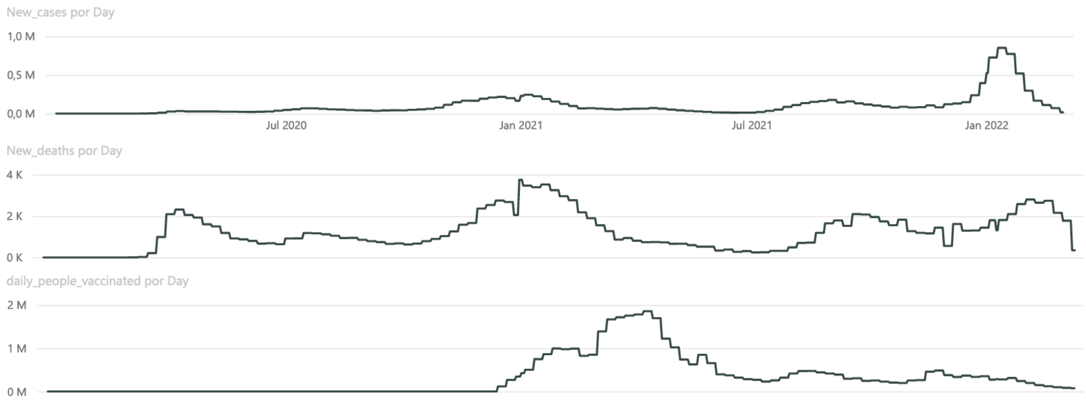


Figure 38: Variations in the number of new cases in USA, throughout time.

The restrictions on gatherings have largely remained between level 3 - no more than 10 and less than 100 people and level 4 - no more than ten people. This has an impact on the proliferation rate of the virus. Of course, it can be observed that there was also a tightening of the restriction measures at peaks of new cases, at least until vaccination started. It can also be seen that after the vaccination peak, the restriction to stay at home remained at level 1, which only represents a recommendation. Furthermore, there is a strong correlation between the restrictions implemented and the population's desire to stay at home. As can be seen in graph 39, the desire to stay at home peaked at the beginning of the pandemic, meaning that people preferred to stay at home during this period due to fear of the unknown. That said, it only increases significantly when the increase in new cases is also alarming.

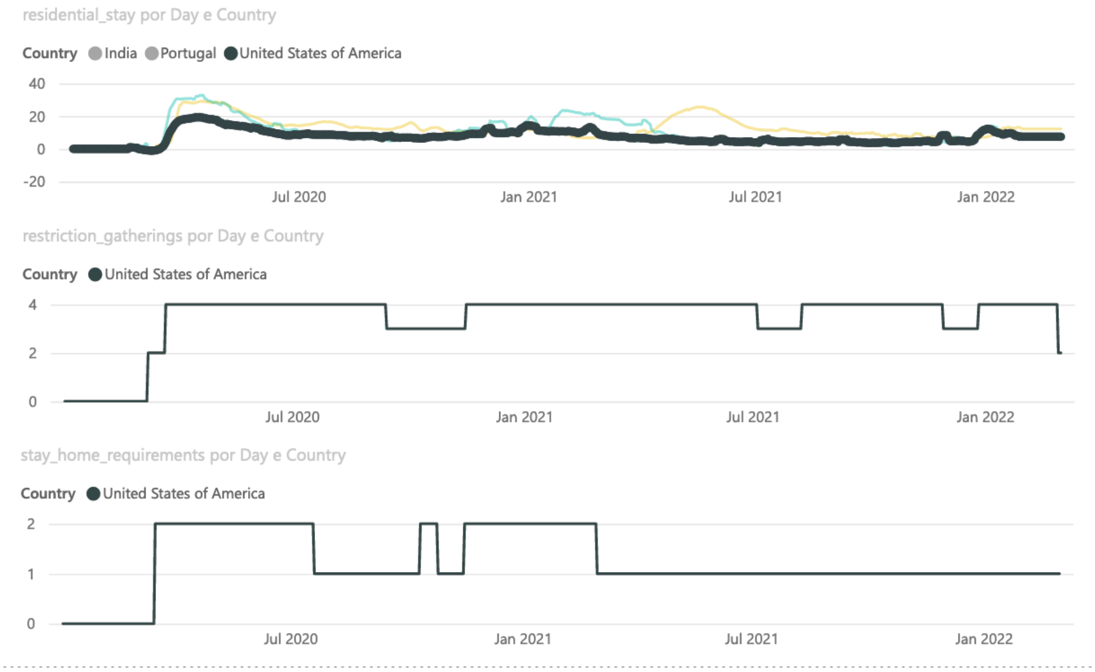


Figure 39: Variation of the preference index to remain at home in USA about restrictions and contingency measures.

8.4.3 India

Analyzing the behaviour of the country with the largest population. During the pandemic, India had three peaks of new registered cases. Although the numbers are frightening, the truth is that the ratios are lower than those observed in the two previous countries under study. The first peak was observed on September 7, 2020, with 94372 new cases. This translates to only approximately 67 cases per million. The second peak, on May 4, 2021, was the most drastic, with 401078 new cases, which translates into a ratio of 285 new cases per million inhabitants. The third case, on January 21, 2022, has a ratio of 255 new cases per million, which gives a total of 317532 new cases. The worrying factor observed in the analysis of the numbers from India regarding the number of deaths and the vaccinated population. The graph of deaths positively correlates with the graph of new cases. This may be since India, as noted in the figure, despite being the country with the most fully vaccinated people, is also the country with the lowest percentage of vaccinated people (69%).

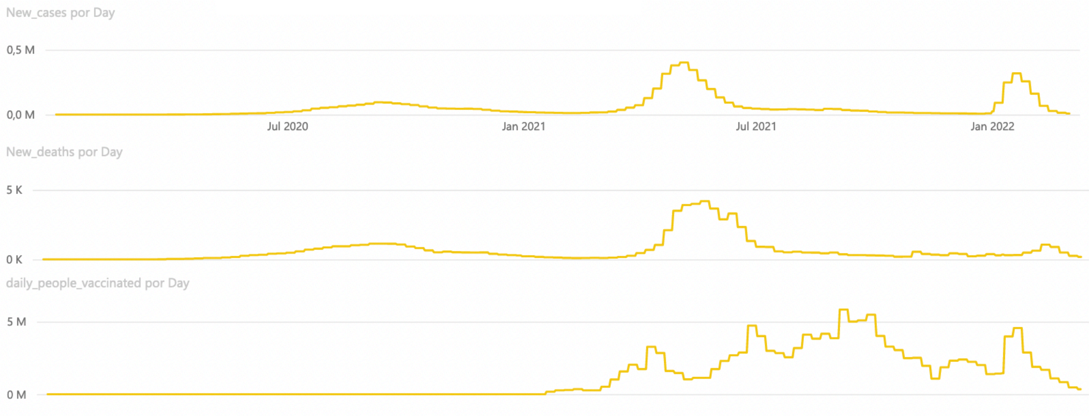


Figure 40: Variations in the number of new cases in India, throughout time.

Furthermore, it can be observed that from the very beginning of the pandemic, the gathering restrictions imposed were of the highest level, level 4 - no more than ten people allowed to gather together. However, the containment restrictions were not proportional, varying and increasing gradually. They only peaked in a short period in 2020, which may be a reflection of the high number of cases and deaths. The peaks in confinement restrictions coincide with the peaks in people's preference to stay at home. Concerning this metric, India has had the highest values for much of the time and has maintained them until the end, unlike the other countries.

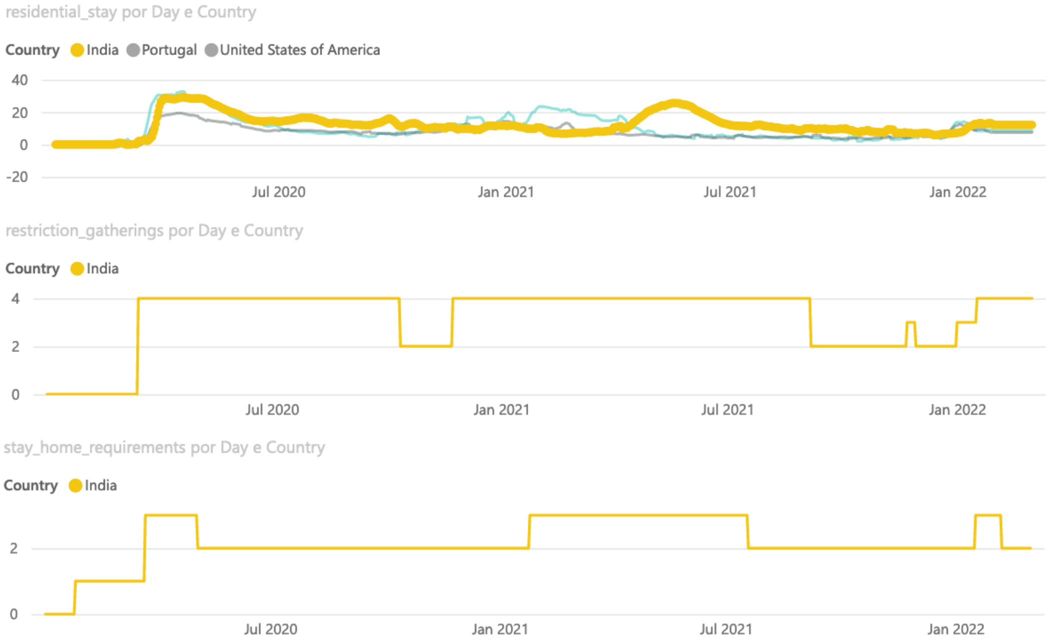


Figure 41: Variation of the preference index to remain at home in India in relation to restrictions and contingency measures.

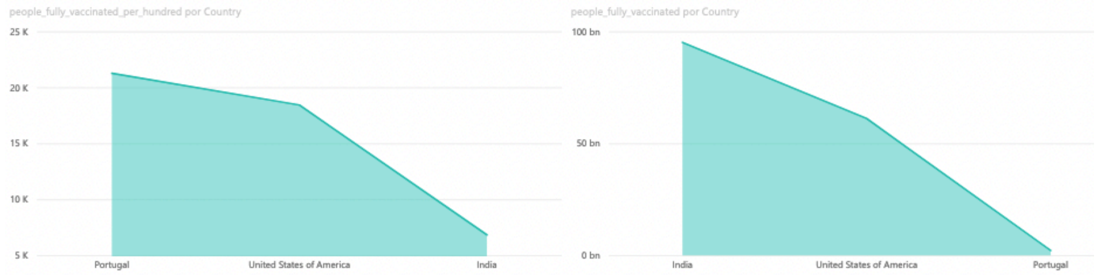


Figure 42: Population vaccinated per million and fully vaccinated, in the three studied countries.

Analyzing Figure 43, which illustrates in detail the most important metrics for the three countries studied, it can be inferred that in any of the countries, it is clear that the first impact was the one that scared the population the most, hence the big peak in preference to stay at home during the first month. Furthermore, in countries where the percentage of fully vaccinated people is higher, the preference to stay at home has a steep decline after the start of vaccination, showing only a slight increase in the period when the number of cases rose substantially (beginning of 2022). Overall, considering the restrictive measures imposed by governments, the population was complacent, so the two graphs have very similar distributions.

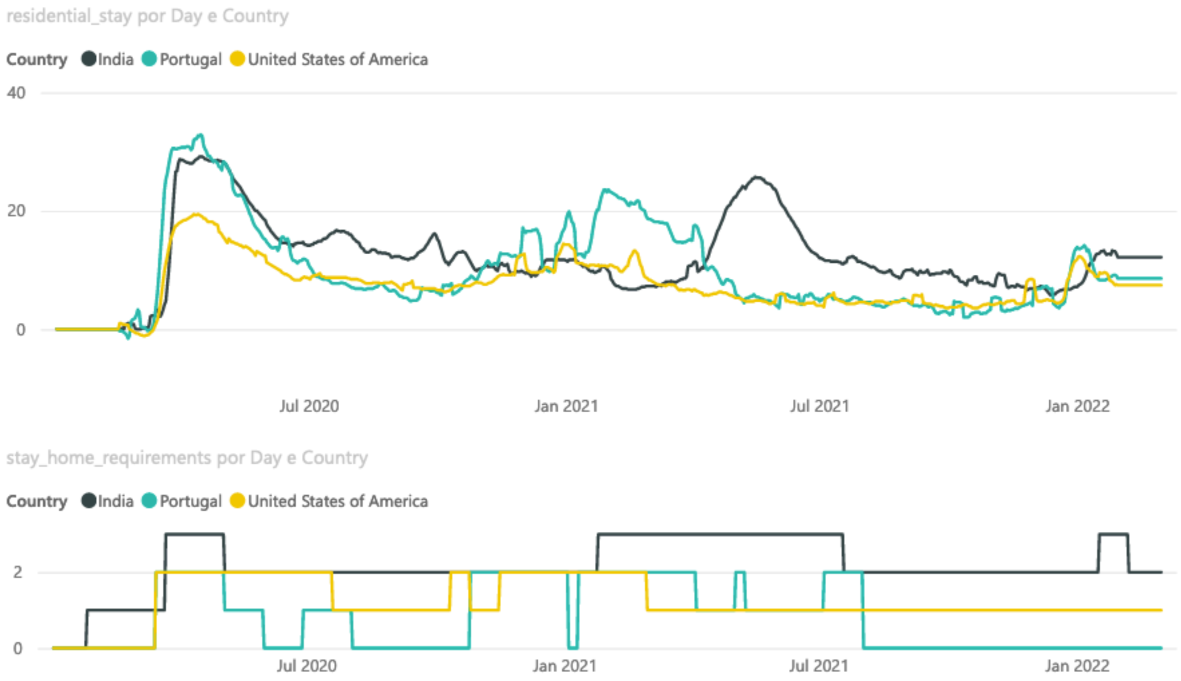


Figure 43: Summary of the variance in preference to stay at home as compared to the restriction measures in the three studied countries.

The detailed analysis presented above shows that, by using data mining and visualization tools, it is possible to build dashboards that support decision-making and that there can be very interesting cooperation between the data related to health institutions and their patients and the government measures that could be taken and shared in an informed way. Moreover, the centralization of data from various realities makes it possible to predict comparatively hypothetical scenarios and consequently prevent their occurrence, if necessary.

8.5 Discussion and Conclusion

The case study examined the effect of Covid-19 on people's preference to stay at home, specifically in Portugal, the United States, and India. Employing Big data tools, such as PySpark and Power BI, for the entire ETL process, the data were examined to identify how government policies, such as crowding and confinement limitations, and immunization affected the death rate and people's choice to stay at home.

The data analysis revealed that the government's actions are reactive rather than preventative. In addition, the preference of individuals to remain at home grew during periods of stricter restrictions but declined as vaccination progressed and the fatality rate decreased.

In conclusion, the findings of this study contribute to a greater understanding of the influence of government initiatives and vaccination on the mortality rate and people's inclination to remain at home during the Covid-19 pandemic. Architectures like the one built here can help future public policy and public health practices in public health literacy and decision support.

Part IV

Final Considerations

9

Discussion

9.1 Detailed Discussion

This doctoral thesis centered its organization on themes such as Big Data, Cloud Computing, Clinical Data Standards, and Real-Time Information Systems, which were explored to determine the optimal solution to the challenge mentioned above. In addition, taking into account the approach to the case studies, the study includes data standards for data organization and information exchange for each of the generated solutions to be sufficiently abstract for future applicability in other fields. Thus, the developed solutions are intended to provide health institutions with advancement in using available and accessible solutions to centralize the display and reporting of analysis and examination results without requiring substantial hardware and software investments.

Regarding the questions provided in the literature review (RQ1-RQ4), it is concluded that due to the volume and variety of data produced at any given moment in health institutions, Big Data is almost required to be considered. Moreover, analyzing these data in real-time determines the Development of personalized medicine and, consequently, the enhancement of the healthcare offered. According to Pramanik, Pal, and Mukhopadhyay, Big Data has the potential to impact the healthcare industry significantly. By analyzing huge amounts of data from various sources, healthcare organizations can structure the data, uncover patterns and hidden information for the user, and get decision-supporting information and knowledge. Hence, both the quality of the service and the cost of providing it will be enhanced [114]. Thus, the health sector's desire for research and innovation will improve.

The health sector is migrating towards new technological paradigms (IoT, Big Data, Cloud Computing, etc.), triggering a radical shift in mindset and approach to conventionally built applications. Historically, adopting new software has been financially and personally burdensome for institutions. Conventional systems involve costly investments in hardware and software licenses and a substantial amount of implementation effort from the institution's personnel. Given the financial and human resources limitations, cloud services offer a realistic option that permits scaling based on productivity and payment based on software usage [1].

Cloud-based systems are attractive because they provide on-demand access to unlimited computing and storage resources. This enables healthcare institutions to extend their resources as necessary to manage the growing volume of data without investing in costly infrastructure [118]. The most common scenario would involve purchasing infrastructure and licenses that optimally correspond to the institution's anticipated development pace. Regrettably, a lack of financial resources frequently limits the initial investment, limiting the solution's usefulness. Hence, it gradually becomes inadequate and unstable. The fact that cloud-hosted systems can be scaled anytime reduces their initial cost, making them more appealing to healthcare organizations.

Before adoption, this type of technology must be evaluated, and precautions must be taken regarding issues such as data security and privacy. In addition, the research suggests that these applications should be abstract and applicable to as many case studies as possible. Consequently, it is prudent to utilize data standards for information structuring and exchange [34].

In order to contribute to the community and as a result of applying the techniques, the Development of case studies began following the identification of the research need. DSR, which focuses on a problem-solving paradigm that tries to extend human knowledge through the production of innovative solutions, was chosen to answer research question RQ5 after taking into account the nature of the solution [12]. As such, it is consistent with this study's pragmatic and inductive methodology. Two more strategies were applied in addition to the DSR method. The initial case studies helped the design and development phase of the solution(s). PoC, on the other hand, is intended to accompany the demonstration phase of the developed solution's practicality. The selection of these approaches proved helpful during the research, as it improved the process of identifying and planning the tasks and the amount of time spent on each of them. In addition, adopting the DSR approach considerably enhanced the quality of the created solutions through iterative cycles between the development process and testing.

The three case studies developed were based on quite different architectures are:

- Case Study I: Building a Real-Time Data Repository based on Laboratory Test Results
- Case Study II: Novel Approach for a Software as a Service for Clinical Test Results Reporting
- Case Study III: The Covid-19 Influence on the Population's Desire to Stay at Home

The answer to research questions RQ6, RQ7 and RQ8 can be found below. The three architectures discussed next were studied and substantiated with the literature before implementation.

For the first case study, the architecture that proved to be the best suited for the problem is based on the gathering of data from several sources in N health institutions, which, after being integrated and preprocessed, are incorporated into the Data Warehouse. Although, as will be detailed later, there are other API-accessible data repositories, their data collecting and processing structures remain unknown. The processing schedule of the data determines the selection of the data warehouse. In this case study, the unprocessed data is transformed and stored, conserving a substantial amount of storage space. This data warehouse uses the traditional star schema, which proved the most effective in achieving the targeted goals [65]. The integration standard chosen for the execution of this architecture is HL7 because it is a recognized international standard for data format, and the integration engine Mirth Connect is user-friendly and capable of managing several channels with diverse data formats. In addition, it was evident that this engine was adaptive to the case studies. In addition, PySpark was used for data preprocessing, Node.js for API development, and SQL for data insertion into the repository. Swagger was employed to generate documentation about the API automatically. Unfortunately, a number of obstacles were encountered during the implementation of the intended architecture. Data quality assurance was one of the earliest difficulties to emerge. Since they come from several sources, it is vital to develop data standards and norms as well as validation methods to guarantee the quality of the final product. This approach supports the increasingly vital interoperability capacity of health information systems [133, 13]. Data security and privacy become a concern as soon as data processing begins. For this reason, data security measures have been designed to prevent unauthorized access, data breaches, and data loss and implemented data security methods, including encryption, anonymization, and access restriction, among others [105]. Further, it can be stated that the data warehouse was developed following the FAIR principles of becoming Findable, Accessible, Interoperable, and Reusable. These four characteristics are assured through the association of a unique key with each record, provision of generic user guidelines, use of internationally recognized terminologies and standards and Development of a API with multiple endpoints available and dynamic [64]. Additionally, this repository is capable of supporting a high number of users and handling massive volumes of data. The entire system was designed and built to be hosted in scalable distributed computing infrastructures in the cloud. In conclusion, the prior decisions enabled the construction of an interoperable solution in real-time, which is beneficial to the advancement of research in the field of health information systems. In addition, it is stated that the implementation of data standards enabled this solution to be globally extendable and even applicable to similar case studies.

The architecture of the second case study includes five key components. On the cloud side, there are the database, document storage, API, automatic interoperability engine, and web application. On the premise side, only clinic-hosted equipment and a semi-automatic integrator are present. AWS was chosen to host the cloud-based system. AWS is the oldest and, as a result, the most developed of the three platforms researched and compared; it also offers the most services. The characteristics that most reinforced the selection were dependability and adaptability. PostgreSQL was chosen as the data storage system since it is a common option for database administration in SaaS development. It enables concurrent connections, and the scaling procedure is relatively straightforward. Since AWS is being used to host the system, buckets from CSP have been implemented for the storage of documents, results, and reports. Node.js was chosen as the programming language for the API. This is built on JavaScript, one of the most widely used programming languages in the world, which enables the usage of the same language on the frontend (React) and backend, making the exchange of programming concepts easier. For the aforementioned reasons, the integration was then responsible for the HL7 and the Mirth communication engine for the automated portion. Python has been used to develop the semi-automatic intelligent agent. Microsoft and Pfizer are among the companies investing in this new programming paradigm. Although the developed architectures follow the same line of thought as the one shown here, the current research has abstracted the concept in a way that could be extremely beneficial. The Development of the second case study encountered many of the same obstacles as the first. The management of sensitive patient data necessitates a sophisticated security infrastructure that guarantees data protection, confidentiality, and privacy. This includes the Development of secure data storage and transfer mechanisms, two-factor authentication, access control protocols, and data encryption. The solution's greatest benefit can also be its most serious difficulty. Due to the absence of restrictions on the types of tests and analyses that can be integrated, integrating with a wide variety of clinical equipment and systems can be a complex procedure requiring a thorough understanding of numerous protocols and standards. The system must be able to receive, process, and transmit data in several formats, as well as convert them into a standard format for reporting and analysis. This obstacle was overcome by employing the open-source communication engine Mirth, which enables the creation of many channels with varying communication protocols. This integration engine proved to be the most complete and hence the most advantageous for this application. The user interface is another common concern. Nonetheless, it is claimed that professionally supported Development has made the system intuitive and user-friendly, with a clean and clear user interface that enables physicians to swiftly and easily locate the necessary information. The web app is optimized for a variety of platforms, such as desktop computers, laptops, tablets, and smartphones, enabling the mobility required by contemporary healthcare professionals and work models. This is a highly valuable asset in the present paradigm of hybrid work arrangements. Ongoing maintenance of this item is also a challenge. The system requires maintenance and support to guarantee that it stays dependable, up-to-date, and responsive to the changing needs of the user base. This includes monitoring performance, addressing bugs and security issues, implementing new integrations, and providing consumers with rapid technical help.

The architecture of the third case study starts with the collection of data from datasets that are publicly accessible. The next stage involves basic data transformation using the PyCountry package to match the ISO codes of the countries across all datasets. This standardization ensures data consistency across all datasets, hence easing data merging and analysis. During this process, the date format will also be standardized to provide uniformity across all datasets. The next stage entails combining the datasets by date and nation of pair. This will ensure that all relevant datasets data are integrated and can be used for future analysis. With PyMongo, the merged data will be stored in MongoDB. MongoDB is a good candidate for storing merged data because it is a document-oriented database that can store semi-structured data and scales easily. The fourth phase is ETL, which involves extracting data from MongoDB and transforming it with Python to remove special characters, handle redundant values, and handle null values. The converted data will be put into Power BI for dashboard creation and analysis. Using a systematic and structured method for gathering, processing, and evaluating data from public databases, the above-described architecture was determined to be appropriate. This case study needed caution in terms of data processing and purpose generalization. Thus, the first concern was the usage of public datasets. Although they facilitate the process, public datasets may have inconsistencies or be of poor quality, which might compromise the analyses' precision and dependability. To prevent this issue, we chose datasets published by scientifically reputable institutes, such as WHO. In addition, several of the datasets needed to be requalified due to missing, null, and inconsistent values. Lastly, it was required to define data formatting and normalization standards throughout the transformation in order to preserve consistency amongst datasets. Managing huge quantities of data is a very common challenge. Combining a scalable database and a distributed processing platform, such as Hadoop or Spark, assisted with data processing and analysis. The updating of the data is the weakness of this case study. Due to the limited data collecting periods of the public datasets, it was impractical to create a dynamic architecture to ensure that the most recent data was used in the analysis. In consideration of potential future implementations, preference was given to the usage of MongoDB as a document-oriented database due to its scalability.

Now answering RQ9, the three case studies stated can provide useful insights for healthcare organizations' real-time decision-making in various ways. By integrating laboratory test results and other pertinent data sources into a real-time data repository, healthcare institutions can get insights that enable faster and more accurate diagnosis and treatment, as well as improved patient outcomes. This can also help track and manage patient data, as well as establish more effective data analysis and reporting procedures. Lastly, it can provide information on how to integrate globally recognized data standards and other important data sources to develop a centralized platform for real-time data processing and analysis, not only in the context of clinical analysis outcomes but for any examination. Having said that, and considering the use of the established SaaS, the implementation of innovative ways for real-time clinical outcome reporting can assist healthcare institutions in conveying results more quickly and precisely, as well as allow improved collaboration among healthcare professionals. This approach can aid in analyzing several SaaS solutions and selecting the most suitable one for a certain institution. Studying the effect of Covid-19 on people's preference to stay at home can offer healthcare institutions information on how to enhance health and well-being, increase communication and interaction with patients, and create resilience and adaptation in response to future pandemics or emergencies. This insight can also help optimize processes for future events where the absence of proof reoccurs. Decision assistance based on real facts shown on intuitive dashboards and updated in real-time is the safest and most comfortable for everyone.

The tenth and final research question (RQ10) aims to discover the developed study's importance compared to similar solutions. Regarding the innovation of the first solution, which resulted from the first case study, most data repositories in the health field are restricted to the institutions themselves and are, therefore, not shared with the scientific community. There are, however, scarce examples such as National Health and Nutrition Examination Survey (NHANES), which provides data on the health, nutrition and socioeconomic status of the US population and offers a RESTful API that allows interested parties to search and retrieve data directly from the source. The catch is that new study editions are released only every two years. In situations where adversities change dramatically, such as a pandemic, the data are no longer representative [25]. In addition, there are a few healthcare service providers that offer an API for querying. ClinicalTrials.gov, for instance, has numerous government-sponsored and private clinical study datasets. Although new clinical trials are added daily to the general repository, the repository is updated each day. Each trial's data is only published once. WHO contains all global population health information. WHO provides an API for global health estimates data, giving access to both raw and prepared data. Usual update frequency is likewise daily [158]. In summary, although the identified repositories are extremely helpful from a scientific standpoint, they do not ensure the standardization of the offered data and do not provide real-time updates. As described in the study by Antonio Iyda Paganelli et al. in the article "*Real-time data analysis in health monitoring systems: A comprehensive systematic literature review*", the quality of diagnoses in healthcare facilities is highly dependent on data and the user data analysis techniques. Healthcare systems have extensively researched data analysis approaches. However, it is still being determined which strategies for real-time data processing are most appropriate in each circumstance. Thus, scientific research using data obtained in real-time, comparable to what occurs in daily life, is essential [109]. Hence, the developed data warehouse, API, and documentation have scientific value because they represent an innovation regarding real-time data repositories available to the scientific community. As one of the most important indications of disease, the breadth of clinical analysis results is of great importance. Consequently, the data in this repository can enable the real-time updating of AI models with treated and standardized real values, thereby enhancing their performance. The consequences of the developed models could be highly beneficial for the provision of medicine by, for instance, giving decision support tools that considerably enhance the provision of treatment in institutions. In addition, this architecture could be adapted to additional case studies, stimulating research, collaboration, and interoperability between several institutions.

Moving on to the results of the second case study, some solutions in the industry already apply the SaaS architecture to various purposes, revealing a widespread interest in this area. The solutions found and that have more similarities with the developed solution are the *LabCorp Patient Portal*, developed by LabCorp, it is a SaaS that allows the patient to view, download and print test results. Nevertheless, this solution has no impact on the daily life of the healthcare professional. The goal is to provide more accessible and convenient patient access. The *Ambra Suite* launched by Ambra Health, on the other hand, is a SaaS imaging test results management solution that allows healthcare professionals to access, view, share and collaborate on patient imaging exams. The goal is to improve collaboration among healthcare professionals in the interpretation and diagnosis of exam results. This software is the most similar to the one developed, also offering the ability to create exam reports. Still, it does not allow the integration of non-imaging exams. *Quest Diagnosis* has developed the *MyQuest Patient Portal*, which is a SaaS test results management that allows physicians to access patients' test results and review those results in real-time. Physicians can also submit new requests for exams or tests if needed. Compared to the solution developed in the second case study, the most significant limitation is that it does not allow for reporting and addenda. Finally and the most different, *Teladoc Medical Experts* is a telemedicine SaaS that allows physicians to view patients' test results and hold remote consultations to discuss those results and offer guidance and treatment to patients. This is an asset that could be implemented in future work. The solution developed is a solution of automatic integration or from a shared folder (intelligent agent), from various medical equipment, in real-time. Thus, this solution is generic and allows the integration of any analysis or clinical examination. Moreover, it provides a more efficient and cost-effective way to manage patient records, adapting to the new hybrid working model, which can significantly impact patient care. Being hosted in the cloud significantly decreases the investment required for a similar solution installed on-premise, which is very advantageous for healthcare institutions. It should also be added that this solution would be adaptable to any other industry since the intelligent agent would perform the integration function for any file.

Analyzing the results obtained from the last case study compared to what has already been investigated due to the impact of covid-19 on society, several studies have been published since 2020. The most studied topics, probably because they are the ones that have suffered most from the pandemic, are the economic impact, the impact on medical health, and the impact on education. As far as the economy is concerned, the Organization for Economic Cooperation and Development (OECD) published a report entitled "Coronavirus: The world economy at risk" in June 2020, in which it presented an analysis of the effects of the pandemic on the world economy. Some of the most important statements of the study were the overall reduction in Gross Domestic Product (GDP), the disproportionate impact on small and medium-sized businesses, and the reduction in global demand. This reduction was attributed to social distancing measures, business closures, and job losses. Since this first report, the organization has published regular updates on the economic impact. It can be seen that although the organization is releasing a report on possible warnings, the hard data to back up the claims is less explicit than it would like. It would be very advantageous if, for example, this organization had a system like the one proposed in this case study to adjust the dashboards according to the interest of the study. Moreover, this data must be updated in real-time. In this case, the visualization tool could even generate general reports automatically. Another great example, and very closely related to the study conducted in the present case study, is one published by the Kaiser Family Foundation (KFF) in September 2021. Entitled "The Implications of Covid-19 for Mental Health and Substance Use", it analyzed the impact of the pandemic on mental health and substance use in adults, specifically in USA. Some of the report's findings include increased symptoms of anxiety and depression (4 out of 10 adults reported symptoms), limited access to mental health services, and increased substance use (alcohol use increased by about 14% in 2020 compared to 2019). Once again, this study, while performing data analysis, is about past events and is unable to show analytical capabilities for decision support regarding preventative measures. In conclusion, the study of the impact of Covid-19 on people's preference to stay at home, besides being able to reveal new patterns or behaviours that were not previously known, leaves open a system that can be extrapolated and prepared to receive data from the future similar panoramas, serving as a decision support system and fostering the use of preventive practices.

9.2 Summary

By briefly and qualitatively examining the developed case studies, a comparison was made between the outlined objectives and the results obtained. Thus, if the objective is fully met, the evaluation is: ***; If there is still something to improve or future work that needs to be implemented, the evaluation is: **; In case the objective is fulfilled, but without fully meeting the need or expectation, the evaluation is: *.

9.2.1 Case Study I: Building a Real-Time Data Repository based on Laboratory Test Results

Table 9: Qualitative Analysis for Case Study I

Objective	**Score**	**Comment**
Collect and store large amounts of laboratory test results in an integrated location for easy access and analysis.	***	The second objective made feasible the collection of tremendous amounts of data. The proof of concept only included data from one healthcare organization since 2020, and given that the proof of concept only included data from one healthcare company. The data repository included 8.9 million records on January 31, 2023.
Implement interoperability to automatically receive and update the repository with new test results in real-time, eliminating the need for manual data entry.	**	Using the Mirth integration engine, creating several channels to receive HL7 messages and extracting and uploading the message data to the data repository was possible. Still, possible improvements could be implemented, for example, balancing the integration engine between two servers or hosting this task in the cloud.
Allow efficient data retrieval and analysis to improve patient care and support research.	***	By providing an API, AI algorithms can be trained and tested with real, accurate data, avoiding using synthetically generated data that does not always represent reality. The best results achieved in this model can improve the prediction or decision support system based on their information.

Table 9 continued from previous page

Objective	**Score**	**Comment**
Ensure data security and privacy through proper data management and compliance with relevant regulations.	**	Anonymization, Noisy Data, Access Restrictions To prevent unauthorized access and use of the information, security measures have been implemented, such as distributing the institutions through different channels (in the interoperability engine), anonymization, introducing noisy data, and restricted access. Even with security precautions in place, there is always the possibility of data breaches or illegal access and use of the information provided.
Improve communication and coordination between healthcare providers and organizations by providing a shared, real-time view of patient laboratory test results.	***	Implementing a globally recognized standard established in Portugal has made it possible to include several healthcare institutions. In addition, the selected standard (HL7) is also compatible with many of the medical equipment present in healthcare institutions.
Enable population health management and public health surveillance by analyzing large-scale laboratory test data.	***	By integrating a large part of the results of laboratory tests, in this case, Portuguese, studies can be returned to study the impact of certain factors on the results of the general population.
Support research by making large, diverse datasets of laboratory test results available for analysis.	***	The real-time interoperability makes the developed data repository much more up-to-date than the other public alternatives available for research. Although anonymized, the patient identifies the dataset even if the patient is noisy in his identification, which enables detailed and patient-centric analysis.
To extract insights from unstructured data, continuously improve the data repository by incorporating new data sources and functionalities, such as machine learning models and natural language processing.	**	Despite some tests, it was impossible to implement redundancy and server balancing, which would be an important improvement for it to be continuously available. In addition, it could be interesting to fill values that reach null with an AI algorithm.

9.2.2 Case Study II: Novel Approach for a Software as a Service for Clinical Test Results Reporting

Table 10: Qualitative Analysis for Case Study II

Objective	**Score**	**Comment**
Automate the process of generating and distributing clinical results, reducing the need for manual reporting and facilitating efficient reporting.	***	Software as a service, hosted in the cloud and customizable.
Reduce the need for initial investment, thus enabling small and medium-sized companies to computerize their healthcare systems.	***	Provision of a multi-tenant SaaS that can be used by several institutions without volume restrictions. SaaS offers pay-as-you-go service and regular scaling, hence making software for small businesses feasible.
Real-time access to clinical results to enable healthcare providers to quickly diagnose and treat patients, improving patient outcomes.	**	Although it is possible it depends on internet access.
Share clinical results with other healthcare providers, enabling better coordination of care and Enhancing communication and collaboration.	**	Only possible in the same institution.
Automate the reporting process to help ensure compliance with regulations and accreditation standards related to data management and reporting.	***	Implementation of the HL7 standard for information exchange. Use of security and privacy techniques.
Provide patients with access to their own clinical/exam results, helping to empower them to take a more active role in their healthcare.	*	The patient view is the least detailed. You will be sent an email with the result and a link to access the platform and see all your test reports.
Include built-in validation and quality checks to improve data accuracy, completeness, and consistency.	***	All the fields have a mandatory flow controlled. Moreover, the values filled in these fields are also subject to validation.

Table 10 continued from previous page

Objective	Score	Comment
Acquired data to be used to identify patterns and trends in patient populations, enabling healthcare providers to take a more proactive approach to population health management.	**	All the data obtained is encrypted in the database. Therefore, some further work would be required to decrypt and anonymize it for later availability.
Allow easy access for reports from any device, from any location, and at any time, improving the speed and quality of communication and decision-making.	**	Possible although conditional on internet access.
Development of a solution capable of fitting any exam or even analytic results.	***	Enabled through the two integration options (HL7 - automatic and intelligent agent-semi automatic). Files entering the software are in PDF format.
Ensure data security and privacy.	**	Access control, data encryption in the database, route protection, encryption of files stored in the cloud, use of secure servers, and regular updates.

9.2.3 Case Study III: The Covid-19 Influence on the Population Desire to Stay at Home

Table 11: Qualitative Analysis for Case Study III

Objective	Score	Comment
Understand the effect of the Covid-19 pandemic on individuals' preferences for staying at home and the explanations behind these preferences.	***	By processing the data and building dashboards, it was possible to draw conclusions about people's preference to stay at home during the pandemic period.
Identify the factors that contribute to people's decisions to stay at home and how these factors may vary between different demographic groups.	***	By having datasets that were geographically representative, it was possible to study the behaviour of different countries.

Table 11 continued from previous page

Objective	Score	Comment
Provide a baseline model for governments and policymakers to better understand the factors that influence people's adherence to staying at home and its impact on the economy.	**	The implemented architecture could be extrapolated to other case studies. However, additional work is needed.
Provide insights and recommendations for governments and businesses to better support and protect individuals during the Covid-19 pandemic.	***	The study carried out shows that through real-time data analysis, decision support systems can be implemented and help in the implementation of preventive rules.

In summary, the three case studies proved relevant and implementable in practice. They all bring innovation factors and hold potential for further research in their respective areas.

9.3 Proof of Concept

In the IT area, any research project must undergo testing and evaluation before it is made available to its end users by installing, testing and evaluating it in a non-production environment. It is, therefore, crucial to follow a series of guidelines to evaluate the proposed solution, particularly regarding whether or not the pre-defined requirements and objectives have been met. Thus, a PoC was performed to prove the proposed solution's viability and usefulness.

To outline the SWOT analysis, questionnaires were defined to allow the study of the acceptance of the proposed information in the developed solutions.

For the first case study, the questionnaire consisted in ten questions that can be found in Appendix A. In total, 71 responses were collected. The analysis of the responses reveals that despite working with data repositories, there is indeed difficulty in finding free data repositories, and the case that they are guaranteed to be standardized is rare (14.1%). Additionally, only four people know about real-time repositories, which are not public. 73.2% of the answers show that the update rate is often not even specified. Analyzing the nature of the data in the repositories, 50.7% are real data. Dimensionally, there is a dispersion, with the majority (52.1%) using medium size repositories (up to ten thousand records). For the interest in a work similar to the one developed here, only two people need to be made aware of the interest, 33.8% of the people consider it exciting, and 63.4% say it is advantageous.

Although the questionnaire was created for the second case study, no answers were collected because using the software as a service implies monetary resource spending. Therefore, it is intended to disseminate the questionnaire for future long-term users. The questions of the developed questionnaire as presented in Appendix B.

Case study III was evaluated through the questionnaire presented in Appendix C. 112 responses were collected. Analyzing the responses, it is concluded that the population's perception of decisions varies with the degree of satisfaction and the surrounding environment. Although a large part considers the measures to be preventive (55.1%), there is a large set of the sample, 48 people, who consider the measures to be reactive. Even though a significant part of the population, most of the time, has not felt anxiety, depression or loneliness (59.1%), nor has seen their alcohol (79.1%) and junk food (59.1%) consumption altered, they have also not felt confident with the political decisions. The majority, from 1 to 5, rated their degree of confidence as 3 (medium). Moreover, only 11.8% complied in full with government advice and guidelines. Finally, the last three questions show that the population can accept systems well if they are well-founded and transparent.

With the results described above, an internal and external analysis was performed to assess the strengths and weaknesses of the advances provided by the solutions developed and the external opportunities and threats.

Table 12: General SWOT Analysis

Internal Environment	**External Environment**
Strenghts	**Opportunities**
Interoperability Use of globally recognized and accredited healthcare standards Real-time data integration Ability to integrate large amounts of data Scalability Easy access to data Adaptability to the hybrid working model	Generalization and Adaptation (The developed solutions are general and adaptable to other scopes) Development of tools consumable by decision support systems that can improve patient outcomes Development of BI indicators Building a fully integrated, patient-centered SaaS (Ordering, Scheduling, Reporting, Monitoring and Tracking) Financial savings
Weaknesses	**Threats**

Table 12 continued from previous page

Internal Environment	External Environment
It depends on the internet connection Obsolete equipment, lacking integration capabilities require the use of semi-automatic methods New HL7 integrations depend on significant technical knowledge Managing patient data, even when security and privacy measures are taken, is risky Depends on the adaptation and acceptance of professionals and the institutions' willingness to innovate	New similar solutions Lack of specific law regulations Lack of technical knowledge Lack of confidence in systems based on data analysis

10

Conclusions

This manuscript ends with a detailed ”Conclusion and Future Work” chapter. This chapter presents the main conclusions of the research, providing an answer to the main research question (Section 10.1). Additionally, a dynamic analysis of the SWOT matrix will be performed in Section 10.2, highlighting future work. More collective projects and scientific publications that have been published, accepted, or submitted are presented in Section 10.3.

10.1 Conclusions

This research focused on discovering healthcare knowledge, especially when huge amounts of data with defined or indeterminate data structures are involved. Hence, specialized techniques and technologies must accommodate their variability under these conditions. Yet, a lack of financial, material or human resources can impede the ability of healthcare organizations to develop solutions capable of properly handling and interpreting this data. In addition, these technologies necessitate the participation of specialists with sophisticated expertise, who are uncommon in healthcare institutions or lack the time capacity to perform this type of maintenance. Cloud-based systems offer a solution to this issue by granting on-demand access to unlimited processing and storage resources. This simplifies the initial implementation by eliminating the need for complex installations and reduces the initial cost to the bare minimum.

Therefore, this doctoral thesis, in the Doctoral Program in Biomedical Engineering, was conducted in order to be able to answer the following main Research question:

”How can Big Data in healthcare institutions become relevant knowledge for real-time decision-making, with the ability to assist clinical management and generate clinical and performance indicators, improving processes and resources, especially in evidence-based medicine contexts?”

In order to answer this question, after the initial interviews, some current themes and paradigms were selected to be studied and used in the conception and development of the proposed archetypes, that is, systems that would incite the construction of knowledge, in real-time, in health institutions. The technologies studied that correspond to the initial hypotheses pondered are the following:

- Interoperability.
- Clinical Data Standards.
- Big Data.
- Cloud Computing.
- Real-Time Health Information Systems.
- Data Analytics.

According to the previously detailed methodology, three case studies were defined with distinct but complementary purposes. Thus, the case studies are:

- **Case Study I:** Building a Real-Time Data Repository based on Laboratory Test Results
- **Case Study II:** Novel Approach for a Software as a Service for Clinical Test Results Reporting
- **Case Study III:** The Covid-19 Influence on the Population Desire to Stay at Home

The first case study arises in the context of the need for data for research. The free data repositories available to the public have considerable limitations in terms of standardization, update and representativeness. Thus, Case Study one's main objective is to build a data repository, updated in real-time and able to receive clinical analysis results from several health institutions. Currently, in the data repository, there are 8.9 million anonymized and standardized records that can be accessed through an API, also developed within the scope of the case study. It is concluded that the objective was met even though it is not available since the project is still underway.

Case study 2 comes with the need to study the viability of implementing the cloud-based paradigm in healthcare institutions. Although they are quite promising, as discussed throughout the chapters, there are few solutions since success is not guaranteed. Therefore this case proposes a new approach that implements a SaaS that receives test results and analysis in real-time. In this way, it empowers healthcare professionals to access, view and report the results received. The result is very promising, presenting functionalities that allow the automation of the distribution process and visualization of clinical results with the ability to reduce the need for investment, thus allowing small and medium-sized healthcare institutions to computerize their systems.

The third case study serves as a study into the feasibility of implementing and accepting decision support systems developed with Big Data tools and how these can impact society. Therefore, a study was conducted to understand the effect of the pandemic, Covid-19, on individuals' preference for staying at home and the explanations behind these preferences. This study concludes that by using data visualization tools that can be updated in real-time, dashboards can be built that support decision-making by enabling cooperation between the data of institutions and their patients and governmental measures that could be taken and shared in an informed manner. Furthermore, centralizing data from various realities makes it possible to predict comparatively hypothetical scenarios and thus prevent their occurrence if necessary.

All case studies were developed generically and can be adapted to new case studies in healthcare or non-healthcare settings.

In summary, the development of the doctoral thesis allowed us to study how data can be treated to build knowledge in health institutions when there are material, monetary, or human resources limitations. It was found that, as discussed in the previous chapter, the most common solutions are not always those that best fit, even though some of the typical components play an essential role. Thus it is believed that development in the Cloud is the most advantageous for healthcare institutions, even though some processing must be done with tools such as Apache Spark. With the proposed solutions, it is possible that with some future work, a generic solution can be built that, based on Big Data principles such as speed, availability, redundancy, and volume, among many others, can build knowledge and provide very positive inputs to assist health professionals in the decision-making making process.

Concluding, it was concluded that knowledge construction, in real-time, in health institutions was made possible through Cloud solutions, scalable and with redundancy and availability implemented. Thus, implementing a SaaS enabled the reception and integration of clinical analysis results to be viewed and, if necessary, reported and validated by physicians. Still, considerable data protection and privacy concerns will be discussed further in future work. Until these innovations are deployed, storage for AWS has been designated in the European Union.

Thus, the thesis is relevant, significant, and original in healthcare information systems because it offers a unique perspective on these topics, provides a holistic view, and offers new knowledge with practical implications. Even so, there is still room for evolution and strategies that can and should be adopted towards leveraging these topics in health.

10.2 Future Work

To outline prospective future work, a dynamic analysis was conducted. The dynamic analysis implements the conclusions of the SWOT analysis through cross-referencing: Strengths-Opportunities(SO), Weaknesses-Opportunities(WO), Strengths-Threats(ST); Weaknesses-Threats(WT)

This also presupposes the continuity of the analysis when the solution evolves so that one can understand whether it is evolving positively. This new interpretation of the SWOT matrix allows the identification of the bets that should be made so that the strengths can respond to the opportunities identified (SO) of the constraints existing in the solution (WO) in order to identify evolutions so that the weaknesses can be overcome, of the warnings that need attention (ST) in order to overcome threats and enhance the strengths, and of the associated risks (WT) in order to try to circumvent the threats. This approach promises to enable the construction of more effective action plans to strengthen the competitive advantages of solutions.

So, the identification of more generic data standards that enable the application of solutions in other fields and industries would be highly tolerated in order to capitalize on the strengths and take more benefit of the recognized potential. Hence, new data repositories or data marts might be created for even more health-related domains. Also, the established SaaS may be stretched to new service areas where it was required, such as document management systems. Additional future tasks, such as developing monitoring or monitoring functionality, could also be highly useful for system adherence.

In order to overcome the threats and enhance the strengths, provide detailed documentation and workflows, manual of new integrations implementation. This way, the whole operation would be detailed and transparent to potential users, increasing confidence in using the software. The availability of the manual for new integrations would only require technical knowledge of a more basic level.

In order to overcome the weaknesses and make the most of the opportunities, the value should be added to the developed solutions, for example, through security, redundancy and availability of data access. This could be implemented, for example, with a time cache so that healthcare professionals can continue their work when there is no Internet connection. Furthermore, security should always be ensured by encrypting the stored data and restricting access to the data. Furthermore, a blockchain could be implemented so that the information is decentralized but protected. Finally, algorithms could be implemented to fill in missing values that would always have to be validated by professionals, for example, missing units of a certain result.

In order to respond to weaknesses and circumvent threats, there must be alternatives to automatic integration to overcome the problem of obsolete equipment. Furthermore, backups must be performed online, and it is fundamental that the SaaS does not depend on law regulations since the guidelines for development are not always clear. As for similar solutions, it is believed that with the continuity of support and updates to the systems, the work developed is a strong alternative in the market.

10.3 Scientific Contributions

In the framework of several individual and collective works carried out during the research, several publications were made in conferences, journals and book chapters. Moreover, a book entitled ”Big Data Analytics and Artificial Intelligence in the Healthcare Industry” was edited with Professor José Machado and researcher Hugo Peixoto. In addition, a patent application has been submitted in the framework of the automatic construction of ontologies according to large amounts of data, including files (for example, JSON). The collective projects that resulted in the publications described below were:

- Development and Research on Innovative Vocational Education Skills.

- Curriculum Development in Data Science and Artificial Intelligence.

- Factory of The Future - Smart Facturing.

- Integrated and Innovative Solutions for the well-being of people in complex urban centers.

The scientific contributions completed so far are:

- Journal
  - Drug–Drug interaction extraction-based system: An natural language processing approach. `https://doi.org/10.1111/exsy.13303`
  - Hierarchical Temporal Memory Theory Approach to Stock Market Time Series Forecasting. `https://doi.org/10.3390/electronics10141630`
  - Software Tools for Conducting Real-Time Information Processing and Visualization in Industry: An Up-to-Date Review. `https://doi.org/10.3390/app11114800`
  - Integrating a New Generation of Interoperability Agents into the AIDA Platform. `https://doi.org/10.33847/2686-8296.3.1_5`
  - Machine Learning Applied to Health Information Exchange. `https://doi.org/10.4018/ijrqeh.298634`
  - OpenEHR modelling applied to Complementary Diagnostics Requests. `https://doi.org/10.1016/j.procs.2022.10.148`
- Conference
  - Prediction models applied to lung cancer using Data Mining. `https://doi.org/10.1007/978-3-031-29104-3_22`
  - New Generation of Interoperable Artifacts in Medical Informatics.
  - Big Data in Healthcare Institutions: An Architecture Proposal. `https://doi.org/10.1007/978-3-031-33614-0_20`
  - Interoperability of Clinical Data through FHIR: A review. `https://doi.org/10.1016/j.procs.2023.03.115`
  - COVID-19 cases and their impact on global air traffic. `https://doi.org/10.1007/978-3-031-38204-8_2`
  - Steps Towards Intelligent Diabetic Foot Ulcer Follow-up based on Deep Learning.`https://doi.org/10.1007/978-3-031-38204-8_7`
  - The Impact of contingency measures on the COVID-19 reproduction rate. `https://doi.org/10.1007/978-3-031-38204-8_3`
  - Recommendation of Medical Exams to Support Clinical Diagnosis Based on Patient's Symptoms. `https://doi.org/10.1007/978-3-031-38204-8_8`

  - Implementing a Software-as-a-Service strategy in healthcare workflows. `https://doi.org/10.1007/978-3-031-38333-5_35`

- Book Chapters

  - Step Towards Monitoring Intelligent Agents in Healthcare Information Systems. `https://doi.org/10.1007/978-3-030-45697-9_50`
  - Contactless Human-Computer Interaction Using a Deep Neural Network Pipeline for Real-Time Video Interpretation and Classification. `https://doi.org/10.1007/978-3-030-90241-4_17`
  - Real-Time UCI Monitoring Using Apache Kafka. `https://doi.org/10.4018/978-1-7998-9172-7.ch001`
  - Improving the Effectiveness of Heart Disease Diagnosis with Machine Learning. `https://doi.org/10.1007/978-3-031-18697-4_18`
  - The Covid-19 Influence on the Desire to Stay at Home: A Big Data Architecture. `https://doi.org/10.1007/978-3-031-21753-1_20`
  - Sustainable and Social Energy on Smart Cities: Systematic Review. `https://doi.org/10.1007/978-3-031-20316-9_6`
  - Medical Recommendation System Based on Daily Clinical Reports: A Proposed NLP Approach for Emergency Departments. `https://doi.org/10.1007/978-3-031-21441-7_24`
  - A Comprehensive Study on Personal and Medical Information to Predict Diabetes. `https://doi.org/10.1007/978-3-031-20859-1_20`

It should be noted that other four other publications are accepted and awaiting publication:

- Collaborative Platform for Intelligent Monitoring of Diabetic Foot Patients - Colab4IMDF
- The interplay of Inflation, Healthcare Spending, and Suicide Rates: An Empirical Analysis
- Enhancing Data Science Interoperability: An Innovative System for Managing OpenEHR Structures

It can also be highlighted that some publications have been submitted to conferences and journals and are currently awaiting acceptance:

- Integrating SaaS and Multi-Agent Systems: A Case Study on Laboratory Results.
- Software as a Service for Real-Time Cloud-Based Clinical Test Results Reporting.
- Towards a Standardized Real-Time Data Repositories in Healthcare Institutions.

Bibliography

[1] G. Aceto, V. Persico, and A. Pescapé. “Industry 4.0 and Health: Internet of Things, Big Data, and Cloud Computing for Healthcare 4.0”. In: *Journal of Industrial Information Integration* 18 (2020), p. 100129. issn: 2452-414X. doi: `https://doi.org/10.1016/j.jii.2020.100129`. url: `https://www.sciencedirect.com/science/article/pii/S2452414X19300135` (cit. on pp. 20, 107).

[2] P. Agrawal. “Artificial intelligence in drug discovery and development”. In: *Journal of Pharmacovigilance* 6.2 (2018), 1000e173 (cit. on p. 49).

[3] S. C. Ahalt et al. “Clinical data: sources and types, regulatory constraints, applications”. In: *Clinical and translational science* 12.4 (2019), p. 329 (cit. on p. 32).

[4] M. R. Ahmed et al. “A literature review on NoSQL database for big data processing”. In: *Int. J. Eng. Technol* 7.2 (2018), pp. 902–906 (cit. on p. 27).

[5] S. Alotaibi and R. Mehmood. “Big data enabled healthcare supply chain management: opportunities and challenges”. In: *International Conference on Smart Cities, Infrastructure, Technologies and Applications*. Vol. 224. Springer. Springer, Cham, 2017, pp. 207–215. doi: `10.1007/978-3-319-94180-6_21` (cit. on p. 3).

[6] J. Andreu-Perez et al. “Big data for health”. In: *IEEE journal of biomedical and health informatics* 19.4 (2015), pp. 1193–1208 (cit. on p. 20).

[7] C. C. Austin et al. “Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements”. In: *IASSIST Quarterly* 39.4 (2016), pp. 24–24 (cit. on p. 24).

[8] K. Bakshi. “Considerations for big data: Architecture and approach”. In: *2012 IEEE aerospace conference*. IEEE. 2012, pp. 1–7 (cit. on p. 17).

[9] T. J. Barnes and M. W. Wilson. “Big data, social physics, and spatial analysis: The early years”. In: *Big Data & Society* 1.1 (2014), p. 2053951714535365 (cit. on p. 14).

[10] C. A. Baron et al. “NoSQL key-value DBs riak and redis”. In: *Database Systems Journal* 6.4 (2016), pp. 3–10 (cit. on p. 27).

[11] E. Bisong. “An overview of google cloud platform services”. In: *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (2019), pp. 7–10 (cit. on p. 47).

[12] J. vom Brocke, A. Hevner, and A. Maedche. “Introduction to design science research”. In: *Design science research. Cases*. Springer, 2020, pp. 1–13 (cit. on pp. 6, 108).

[13] T. Burns, J. Cosgrove, and F. Doyle. “A Review of Interoperability Standards for Industry 4.0.” In: *Procedia Manufacturing* 38 (2019), pp. 646–653 (cit. on pp. 33, 109).

[14] CareCloud. *Modern Healthcare Solutions*. 2022-10. url: `https://www.carecloud.com/` (cit. on p. 49).

[15] R. Čereš▯ák and M. Kvet. “Comparison of query performance in relational a non-relation databases”. In: *Transportation Research Procedia* 40 (2019), pp. 170–177 (cit. on p. 26).

[16] C. P. Chen and C.-Y. Zhang. “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data”. In: *Information sciences* 275 (2014), pp. 314–347 (cit. on p. 18).

[17] P.-T. Chen, C.-L. Lin, and W.-N. Wu. “Big data management in healthcare: Adoption challenges and implications”. In: *International Journal of Information Management* 53 (2020), p. 102078 (cit. on p. 30).

[18] P. Clayton et al. “Building a comprehensive clinical information system from components”. In: *Methods of information in medicine* 42.01 (2003), pp. 01–07 (cit. on p. 40).

[19] E. Codd. “An evaluation scheme for database management systems that are claimed to be relational”. In: *1986 IEEE Second International Conference on Data Engineering*. IEEE. 1986, pp. 720–729 (cit. on p. 24).

[20] M. Collier and R. Shahan. *Fundamentals of Azure*. Microsoft Press, 2016 (cit. on p. 46).

[21] M. Cornock. “General Data Protection Regulation (GDPR) and implications for research”. In: *Maturitas* 111 (2018), A1–A2 (cit. on p. 51).

[22] C. Costa and M. Y. Santos. “Big Data: State-of-the-art concepts, techniques, technologies, modeling approaches and research challenges”. In: (2017) (cit. on p. 16).

[23] C. Coughenour et al. “Changes in depression and physical activity among college students on a diverse campus after a COVID-19 stay-at-home order”. In: *Journal of community health* 46 (2021), pp. 758–766 (cit. on p. 89).

[24] N. Cozzoli et al. “How can big data analytics be used for healthcare organization management? Literary framework and future research from a systematic review”. In: *BMC health services research* 22.1 (2022), pp. 1–14 (cit. on p. 3).

[25] L. R. Curtin et al. "The National Health and Nutrition Examination Survey: Sample Design, 1999-2006." In: *Vital and health statistics. Series 2, Data evaluation and methods research* 155 (2012), pp. 1–39 (cit. on p. 111).

[26] M. Cusumano. "Cloud computing and SaaS as new computing platforms". In: *Communications of the ACM* 53.4 (2010), pp. 27–29 (cit. on p. 45).

[27] W. Dawoud, I. Takouna, and C. Meinel. "Infrastructure as a service security: Challenges and solutions". In: *2010 the 7th International Conference on Informatics and Systems (INFOS)*. IEEE. 2010, pp. 1–8 (cit. on p. 42).

[28] N. Deepa et al. "A survey on blockchain for big data: approaches, opportunities, and future directions". In: *Future Generation Computer Systems* (2022) (cit. on p. 3).

[29] C. Department. *A data-driven revolution in treatment*. url: `https://healthsciences.ku.dk/about/impact/translational-research/big-data-medicine/` (cit. on p. 22).

[30] Y. Duan, J. S. Edwards, and Y. K. Dwivedi. "Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda". In: *International journal of information management* 48 (2019), pp. 63–71 (cit. on p. 51).

[31] S. M. Erickson et al. "Patient safety: achieving a new standard for care". In: (2003) (cit. on p. 34).

[32] A. G. Fiks et al. "Electronic medical record use in pediatric primary care". In: *Journal of the American Medical Informatics Association* 18.1 (2011), pp. 38–44 (cit. on p. 40).

[33] B. Furht and F. Villanustre. "Big data technologies and applications". In: (2016) (cit. on p. 2).

[34] P. Galetsi, K. Katsaliaki, and S. Kumar. "Values, challenges and future directions of big data analytics in healthcare: A systematic review". In: *Social Science & Medicine* 241 (2019), p. 112533. issn: 0277-9536. doi: `https://doi.org/10.1016/j.socscimed.2019.112533`. url: `https://www.sciencedirect.com/science/article/pii/S0277953619305271` (cit. on p. 108).

[35] A. Gandomi and M. Haider. "Beyond the hype: Big data concepts, methods, and analytics". In: *International journal of information management* 35.2 (2015), pp. 137–144 (cit. on p. 15).

[36] S. Garfinkel. *The cloud imperative*. 2011-10. url: `https://www.technologyreview.com/2011/10/03/190237/the-cloud-imperative/` (cit. on p. 42).

[37] T. Garrido et al. "Making the business case for hospital information systems–a Kaiser Permanente investment decision". In: *Journal of Health Care Finance* 31.2 (2004), pp. 16–25 (cit. on p. 41).

[38] C. Gaudet-Blavignac et al. "Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for processing free text in health care: systematic scoping review". In: *Journal of medical Internet research* 23.1 (2021), e24594 (cit. on p. 38).

[39] M. Golfarelli and S. Rizzi. “From star schemas to big data: 20+ years of data warehouse research”. In: *A comprehensive guide through the Italian database research over the last 25 years* (2018), pp. 93–107 (cit. on p. 29).

[40] V. Gonçalves and P. Ballon. “Adding value to the network: Mobile operators’ experiments with Software-as-a-Service and Platform-as-a-Service models”. In: *Telematics and Informatics* 28.1 (2011), pp. 12–21 (cit. on p. 45).

[41] A. González-Ferrer and M. Peleg. “Understanding requirements of clinical data standards for developing interoperable knowledge-based DSS: A case study”. In: *Computer Standards & Interfaces* 42 (2015), pp. 125–136 (cit. on p. 34).

[42] A. Gorelik. *The enterprise big data lake: Delivering the promise of big data and data science*. O’Reilly Media, 2019 (cit. on p. 29).

[43] L. O. Gostin, L. A. Levit, S. J. Nass, et al. “Beyond the HIPAA privacy rule: enhancing privacy, improving health through research”. In: (2009) (cit. on p. 39).

[44] S. Goyal. “Public vs private vs hybrid vs community-cloud computing: a critical review”. In: *International Journal of Computer Network and Information Security* 6.3 (2014), pp. 20–29 (cit. on p. 43).

[45] E. A. Gray and J. H. Thorpe. “Comparative effectiveness research and big data: balancing potential with legal and ethical considerations”. In: *Journal of comparative effectiveness research* 4.1 (2015), pp. 61–74 (cit. on p. 3).

[46] B. Gupta, P. Mittal, and T. Mufti. “A review on Amazon web service (AWS), Microsoft azure & Google cloud platform (GCP) services”. In: *Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, 27-28 February 2020, Jamia Hamdard, New Delhi, India*. 2021 (cit. on pp. 46, 47).

[47] E. Gürel and M. Tat. “SWOT analysis: A theoretical review.” In: *Journal of International Social Research* 10.51 (2017) (cit. on pp. 9, 10).

[48] C. Gyorödi, R. Gyorödi, and R. Sotoc. “A comparative study of relational and non-relational database models in a Web-based application”. In: *International Journal of Advanced Computer Science and Applications* 6.11 (2015), pp. 78–83 (cit. on p. 26).

[49] D. Haak et al. “DICOM for clinical research: PACS-integrated electronic data capture in multi-center trials”. In: *Journal of digital imaging* 28 (2015), pp. 558–566 (cit. on pp. 36, 37).

[50] R. Han, L. K. John, and J. Zhan. “Benchmarking big data systems: A review”. In: *IEEE Transactions on Services Computing* 11.3 (2017), pp. 580–597 (cit. on p. 2).

[51] J. L. Harrington. *Relational database design and implementation*. Morgan Kaufmann, 2016 (cit. on pp. 24, 26).

[52] I. A. T. Hashem et al. "The rise of "big data" on cloud computing: Review and open research issues". In: *Information systems* 47 (2015), pp. 98–115 (cit. on p. 18).

[53] T. Hill and R. Westbrook. "SWOT analysis: it's time for a product recall". In: *Long range planning* 30.1 (1997), pp. 46–52 (cit. on p. 9).

[54] HIMSS. *Interoperability and Health Information Exchange*. 2020-08. url: `https://www.himss.org/interoperability-and-health-information-exchange` (cit. on p. 32).

[55] HIMSS. *Interoperability in Healthcare*. 2021-08. url: `https://www.himss.org/resources/interoperability-healthcare` (cit. on pp. 33, 34).

[56] S. M. Huff et al. "Development of the logical observation identifier names and codes (LOINC) vocabulary". In: *Journal of the American Medical Informatics Association* 5.3 (1998), pp. 276–292 (cit. on p. 38).

[57] I. Ibnouhsein et al. "The big data revolution for breast cancer patients". In: *European journal of breast health* 14.2 (2018), p. 61 (cit. on p. 22).

[58] IEEE. "IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries". In: *IEEE Std 610* (1991), pp. 1–217. doi: `10.1109/IEEESTD.1991.106963` (cit. on p. 32).

[59] J. Iivari and J. R. Venable. "Action research and design science research-Seemingly similar but decisively dissimilar". In: (2009) (cit. on p. 6).

[60] W. H. Inmon. "What is a data warehouse". In: *Prism Tech Topic* 1.1 (1995), pp. 1–5 (cit. on p. 29).

[61] S. International. *5 Step Briefing*. url: `https://www.snomed.org/snomed-ct/five-step-briefing` (cit. on p. 38).

[62] ISO. 2021-02. url: `https://www.iso.org/standards.html` (cit. on p. 34).

[63] Y. Al-Issa, M. A. Ottom, and A. Tamrawi. "eHealth cloud security challenges: a survey". In: *Journal of healthcare engineering* 2019 (2019) (cit. on p. 3).

[64] A. Jacobsen et al. *FAIR principles: interpretations and implementation considerations*. 2020 (cit. on pp. 39, 109).

[65] K. Jameel, A. Adil, and M. Bahjat. "Analyses the Performance of Data Warehouse Architecture Types". In: *Journal of Soft Computing and Data Mining* 3.1 (2022), pp. 45–57 (cit. on p. 108).

[66] C. Ji et al. "Big data processing: Big challenges and opportunities". In: *Journal of Interconnection Networks* 13.03n04 (2012), p. 1250009 (cit. on p. 29).

[67] J. S. Johnson, S. B. Friend, and H. S. Lee. "Big data facilitation, utilization, and monetization: Exploring the 3Vs in a new product development process". In: *Journal of Product Innovation Management* 34.5 (2017), pp. 640–658 (cit. on p. 15).

[68] A. Joint and E. Baker. “Knowing the past to understand the present–issues in the contracting for cloud based services”. In: *Computer Law & Security Review* 27.4 (2011), pp. 407–415 (cit. on p. 45).

[69] M. A. Kamal et al. “Highlight the features of AWS, GCP and Microsoft Azure that have an impact when choosing a cloud service provider”. In: *Int. J. Recent Technol. Eng* 8.5 (2020), pp. 4124–4232 (cit. on p. 47).

[70] M. Karatas et al. “Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives”. In: *Expert Systems with Applications* (2022), p. 116912 (cit. on p. 22).

[71] P. Kaushik et al. “Cloud computing and comparison based on service and performance between amazon aws, microsoft azure, and google cloud”. In: *2021 International Conference on Technological Advancements and Innovations (ICTAI)*. IEEE. IEEE, 2021, pp. 268–273. doi: 10.1109/ICTAI53825.2021.9673425 (cit. on p. 47).

[72] M. J. Kavis. *Architecting the cloud: design decisions for cloud computing service models (SaaS, PaaS, and IaaS)*. John Wiley & Sons, 2014 (cit. on pp. 44, 45).

[73] K. Kawamoto et al. “Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success”. In: *Bmj* 330.7494 (2005), p. 765 (cit. on p. 2).

[74] N. Khan et al. “The 10 Vs, issues and challenges of big data”. In: *Proceedings of the 2018 international conference on big data and education*. 2018, pp. 52–56. doi: 10.1145/3206157.3206166 (cit. on p. 15).

[75] R. Lakhan, A. Agrawal, and M. Sharma. “Prevalence of depression, anxiety, and stress during COVID-19 pandemic”. In: *Journal of neurosciences in rural practice* 11.04 (2020), pp. 519–525 (cit. on p. 89).

[76] C. Lam. *Hadoop in action*. Manning Publications Co., 2010 (cit. on p. 18).

[77] H. Landi. *Mayo Clinic launches 2 new companies to use patient data and AI to advance early disease detection*. 2021-04. url: https://www.fiercehealthcare.com/tech/mayo-clinic-launches-two-new-companies-to-advance-early-disease-detection-using-ai (cit. on p. 22).

[78] D. Laney. “3D data management: Controlling data volume, velocity and variety”. In: *META group research note* 6.70 (2001), p. 1 (cit. on p. 15).

[79] M. Langarizadeh et al. “Effectiveness of Anonymization Methods in Preserving Patients’ Privacy: A Systematic Literature Review.” In: *eHealth* 248 (2018), pp. 80–87 (cit. on p. 51).

[80] E. J. Larson. “The Myth of Artificial Intelligence”. In: *The Myth of Artificial Intelligence*. Harvard University Press, 2021 (cit. on p. 14).

[81] N. R. Lawrence and S. H. Bradley. “Big data and the NHS–we have the technology, but we need patient and professional engagement”. In: *Future Healthcare Journal* 5.3 (2018), p. 229 (cit. on p. 22).

[82] M. Lehne et al. “The Use of FHIR in Digital Health-A Review of the Scientific Literature.” In: *GMDS* September (2019), pp. 52–58 (cit. on p. 39).

[83] S. S. Manvi and G. K. Shyam. “Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey”. In: *Journal of network and computer applications* 41 (2014), pp. 424–440 (cit. on p. 44).

[84] S. T. March and V. C. Storey. “Design science in the information systems discipline: an introduction to the special issue on design science research”. In: *MIS quarterly* (2008), pp. 725–730 (cit. on p. 6).

[85] B. Marr. *Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance*. John Wiley & Sons, 2015 (cit. on pp. 14, 15).

[86] A. Martins et al. “An evaluation of how big-data and data warehouses improve business intelligence decision making”. In: *World Conference on Information Systems and Technologies*. Springer. 2020, pp. 609–619 (cit. on p. 29).

[87] P. S. Mathew and A. S. Pillai. “Big Data solutions in Healthcare: Problems and perspectives”. In: *2015 International conference on innovations in information, embedded and communication systems (ICIIECS)*. IEEE. 2015, pp. 1–6. doi: `10.1109/ICIIECS.2015.7193211` (cit. on p. 17).

[88] V. Mayer-Schönberger and K. Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013 (cit. on p. 3).

[89] C. J. McDonald et al. “LOINC, a universal standard for identifying laboratory observations: a 5-year update”. In: *Clinical chemistry* 49.4 (2003), pp. 624–633 (cit. on p. 37).

[90] A. Meier and M. Kaufmann. “Nosql databases”. In: *SQL & NoSQL databases*. Springer, 2019, pp. 201–218 (cit. on p. 27).

[91] M. Al-Mekhlal and A. A. Khwaja. “A synthesis of big data definition and characteristics”. In: *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*. IEEE. 2019, pp. 314–322. doi: `10.1109/CSE/EUC.2019.00067` (cit. on p. 2).

[92] P. Mell, T. Grance, et al. “The NIST definition of cloud computing”. In: (2011) (cit. on p. 43).

[93] S. B. Merriam. *Case study research in education: A qualitative approach.* Jossey-Bass, 1988 (cit. on p. 8).

[94] N. Miloslavskaya and A. Tolstoy. “Big data, fast data and data lake concepts”. In: *Procedia Computer Science* 88 (2016), pp. 300–305 (cit. on p. 29).

[95] F. Miranda et al. “Machine Learning Applied to Health Information Exchange”. In: *International Journal of Reliable and Quality E-Healthcare (IJRQEH)* 11.1 (2022), pp. 1–17 (cit. on pp. 33, 38).

[96] M. A. Mizani. “Cloud-based computing”. In: *Key advances in clinical informatics*. Elsevier, 2017, pp. 239–255 (cit. on p. 49).

[97] C. M. Mohammed, S. R. Zeebaree, et al. “Sufficient comparison among cloud computing services: IaaS, PaaS, and SaaS: A review”. In: *International Journal of Science and Business* 5.2 (2021), pp. 17–30 (cit. on p. 45).

[98] A. Moniruzzaman and S. A. Hossain. “Nosql database: New era of databases for big data analytics-classification, characteristics and comparison”. In: *arXiv preprint arXiv:1307.0191* (2013) (cit. on p. 26).

[99] D. L. Moody and M. A. Kortink. “From enterprise models to dimensional models: a methodology for data warehouse and data mart design.” In: *DMDW*. 2000, p. 5 (cit. on p. 29).

[100] F. Nargesian et al. “Data lake management: challenges and opportunities”. In: *Proceedings of the VLDB Endowment* 12.12 (2019), pp. 1986–1989 (cit. on p. 29).

[101] *National Patient Register*. url: `https://www.socialstyrelsen.se/en/statistics-and-data/registers/national-patient-register/` (cit. on p. 31).

[102] K. Y. Ngiam and W. Khor. “Big data and machine learning algorithms for health-care delivery”. In: *The Lancet Oncology* 20.5 (2019), e262–e273 (cit. on p. 3).

[103] NHS. *Spine*. 2022-10. url: `https://www.digital.nhs.uk/services/spine` (cit. on p. 31).

[104] C. Ochs et al. “Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies”. In: *Journal of the American Medical Informatics Association* 22.3 (2015), pp. 507–518 (cit. on p. 38).

[105] O. I. Odabi and S. C. Oluwasegun. “Data security in health information systems by applying software techniques”. In: *Journal of Emerging Trends in Engineering and Applied Sciences* 2.5 (2011), pp. 775–781 (cit. on p. 109).

[106] C. Oliveira et al. “Improving the Effectiveness of Heart Disease Diagnosis with Machine Learning”. In: *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Complex Systems Simulation. The PAAMS Collection: International Workshops of PAAMS 2022, L’Aquila, Italy, July 13–15, 2022, Proceedings*. Springer. 2022, pp. 222–231 (cit. on p. 51).

[107] D. Oliveira et al. “OpenEHR modeling: improving clinical records during the COVID-19 pandemic”. In: *Health and Technology* 11.5 (2021), pp. 1109–1118 (cit. on p. 52).

[108] W. H. Organization. *Ageing and health*. 2022-10. url: `https://www.who.int/news-room/fact-sheets/detail/ageing-and-health` (cit. on p. 61).

[109] A. I. Paganelli et al. "Real-time data analysis in health monitoring systems: A comprehensive systematic literature review". In: *Journal of Biomedical Informatics* (2022), p. 104009 (cit. on p. 112).

[110] K. Peffers et al. "A design science research methodology for information systems research". In: *Journal of management information systems* 24.3 (2007), pp. 45–77 (cit. on p. 7).

[111] B. Phadermrod, R. M. Crowder, and G. B. Wills. "Importance-performance analysis based SWOT analysis". In: *International Journal of Information Management* 44 (2019), pp. 194–203 (cit. on p. 9).

[112] A. G. Picciano. "The evolution of big data and learning analytics in American higher education." In: *Journal of asynchronous learning networks* 16.3 (2012), pp. 9–20 (cit. on p. 14).

[113] N. Pimenta et al. "A Comprehensive Study on Personal and Medical Information to Predict Diabetes". In: *Distributed Computing and Artificial Intelligence, 19th International Conference*. Springer. 2022, pp. 197–207 (cit. on p. 51).

[114] P. K. D. Pramanik, S. Pal, and M. Mukhopadhyay. "Healthcare big data: A comprehensive overview". In: *Research Anthology on Big Data Analytics, Architectures, and Applications* (2022), pp. 119–147 (cit. on pp. 15, 30, 107).

[115] K. H. Pries and R. Dunnigan. *Big Data Analytics: A practical guide for managers*. Auerbach Publications, 2015 (cit. on p. 15).

[116] *Projections of resident population in Portugal*. 2020-03. url: https://www.ine.pt/xportal/xmain?xpid=INE%5C&amp;xpgid=ine_destaques%5C&amp;DESTAQUESdest_boui=406534255%5C&amp;DESTAQUESmodo=2%5C&amp;xlang=en (cit. on p. 61).

[117] W. Puangsaijai and S. Puntheeranurak. "A comparative study of relational database and key-value database for big data applications". In: *2017 International Electrical Engineering Congress (iEECON)*. IEEE. 2017, pp. 1–4 (cit. on p. 27).

[118] L. Rajabion et al. "Healthcare big data processing mechanisms: The role of cloud computing". In: *International Journal of Information Management* 49 (2019), pp. 271–289 (cit. on pp. 3, 108).

[119] M. Ramesh and A. S. Bali. "42The Remarkable Healthcare Performance in Singapore". In: *Great Policy Successes*. Oxford University Press, 2019-09. isbn: 9780198843719. doi: 10.1093/oso/9780198843719.003.0003. url: https://doi.org/10.1093/oso/9780198843719.003.0003 (cit. on p. 22).

[120] R. Reubens. "To craft, by design, for sustainability: Towards holistic sustainability design for developing-country enterprises". In: (2016) (cit. on p. 6).

[121] RevCycleIntelligence. *Big Data Tool saves CMS* $1.5 B by preventing Medicare fraud$. 2016-06. url: https://revcycleintelligence.com/news/big-data-tool-saves-cms-1.5b-by-preventing-medicare-fraud (cit. on p. 22).

[122] I. Robinson, J. Webber, and E. Eifrem. *Graph databases: new opportunities for connected data.* " O'Reilly Media, Inc.", 2015 (cit. on p. 27).

[123] F. S. Saleh et al. "Effective use of synthetic data for urban scene semantic segmentation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer Cham, 2018, pp. 84–100. doi: 0.1007/978-3-030-01246-5 (cit. on p. 51).

[124] R. Santos et al. "Real-Time UCI Monitoring Using Apache Kafka". In: *Big Data Analytics and Artificial Intelligence in the Healthcare Industry*. IGI Global, 2022, pp. 1–37 (cit. on p. 27).

[125] P. Saranya and P. Asha. "Survey on Big Data Analytics in health care". In: *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE. 2019, pp. 46–51. doi: 10.1109/ICSSIT46314.2019.8987882 (cit. on p. 3).

[126] M. Saunders, P. Lewis, and A. Thornhill. *Research methods for business students*. Pearson education, 2009 (cit. on p. 6).

[127] S. Schulz, R. Stegwee, and C. Chronaki. "Standards in healthcare data". In: *Fundamentals of Clinical Data Science* (2019), pp. 19–36 (cit. on pp. 35, 36).

[128] S. Senthilkumar et al. "Big data in healthcare management: a review of literature". In: *American Journal of Theoretical and Applied Business* 4.2 (2018), pp. 57–69 (cit. on p. 2).

[129] A. B. Sergey, D. B. Alexandr, and A. T. Sergey. "Proof of concept center—a promising tool for innovative development at entrepreneurial universities". In: *Procedia-Social and Behavioral Sciences* 166 (2015), pp. 240–245 (cit. on p. 9).

[130] S. Shaikh and D. Vora. "YARN versus MapReduce—A comparative study". In: *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE. IEEE, 2016, pp. 1294–1297. doi: 10.23919/INDIACom54597.2022 (cit. on p. 18).

[131] D. W. Simborg. "The case for the HL7 standard." In: *Computers in Healthcare* 9.1 (1988), pp. 39–40 (cit. on p. 38).

[132] H. A. Simon. *The Sciences of the Artificial*. 3rd ed. MIT Press, 1996 (cit. on p. 6).

[133] N. Skyttberg et al. "How to improve vital sign data quality for use in clinical decision support systems? A qualitative study in nine Swedish emergency departments". In: *BMC medical informatics and decision making* 16 (2016), pp. 1–12 (cit. on p. 109).

[134] R. Sousa et al. "Contactless Human-Computer Interaction Using a Deep Neural Network Pipeline for Real-Time Video Interpretation and Classification". In: *Advanced Research in Technologies, Information, Innovation and Sustainability: First International Conference, ARTIIS 2021, La Libertad, Ecuador, November 25–27, 2021, Proceedings*. Springer. 2021, pp. 209–220 (cit. on p. 30).

[135] R. Sousa et al. "Hierarchical temporal memory theory approach to stock market time series forecasting". In: *Electronics* 10.14 (2021), p. 1630 (cit. on p. 51).

[136] R. Sousa et al. “Medical Recommendation System Based on Daily Clinical Reports: A Proposed NLP Approach for Emergency Departments”. In: *Artificial Intelligence XXXIX: 42nd SGAI International Conference on Artificial Intelligence, AI 2022, Cambridge, UK, December 13–15, 2022, Proceedings*. Springer. 2022, pp. 315–320 (cit. on p. 51).

[137] R. Sousa et al. “Software tools for conducting real-time information processing and visualization in industry: An up-to-date review”. In: *Applied Sciences* 11.11 (2021), p. 4800. doi: `10.3390/app11114800` (cit. on pp. 18, 52).

[138] R. Sousa et al. “Step towards monitoring intelligent agents in healthcare information systems”. In: *Trends and Innovations in Information Systems and Technologies: Volume 3 8*. Springer. 2020, pp. 510–519 (cit. on p. 32).

[139] R. Sousa et al. “Sustainable and Social Energy on Smart Cities: Systematic Review”. In: *Advanced Research in Technologies, Information, Innovation and Sustainability: Second International Conference, ARTIIS 2022, Santiago de Compostela, Spain, September 12–15, 2022, Revised Selected Papers, Part II*. Springer. 2022, pp. 72–84 (cit. on p. 2).

[140] R. Sousa et al. “The Covid-19 Influence on the Desire to Stay at Home: A Big Data Architecture”. In: *Intelligent Data Engineering and Automated Learning–IDEAL 2022: 23rd International Conference, IDEAL 2022, Manchester, UK, November 24–26, 2022, Proceedings*. Springer. 2022, pp. 199–210 (cit. on pp. 15, 17).

[141] *Spine (NHS IT)*. 2022-04. url: `https://psnc.org.uk/digital-and-technology/systems-apps/spine-nhs-it/` (cit. on p. 31).

[142] G. Strawn and C. Strawn. “Relational databases: Codd, stonebraker, and Ellison”. In: *IT Professional* 18.2 (2016), pp. 63–65 (cit. on p. 24).

[143] J. Surbiryala and C. Rong. “Cloud computing: History and overview”. In: *2019 IEEE Cloud Summit*. IEEE. 2019, pp. 1–7 (cit. on pp. 42, 43).

[144] C. Tankard. “Big data security”. In: *Network security* 2012.7 (2012), pp. 5–8 (cit. on p. 30).

[145] D. Thara et al. “Impact of big data in healthcare: A survey”. In: *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 2016, pp. 729–735. doi: `10.1109/IC3I.2016.7918057` (cit. on p. 15).

[146] B. Thuraisingham. “Big data security and privacy”. In: *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*. 2015, pp. 279–280 (cit. on p. 30).

[147] Today. *The big read: Big data making a great difference in healthcare*. 2015-02. url: `https://www.todayonline.com/singapore/big-data-making-great-difference-healthcare` (cit. on p. 21).

[148] S. A. Tovino. “The HIPAA privacy rule and the EU GDPR: illustrative comparisons”. In: *Seton Hall L. Rev.* 47 (2016), p. 973 (cit. on p. 39).

[149] N. E. Transformation Directorate. *NHS AI Lab To Speed Up Cancer And Heart Care*. 2020-09. url: https://transform.england.nhs.uk/news/nhs-ai-lab-speed-cancer-and-heart-care/ (cit. on p. 22).

[150] D. Ucuz et al. “Comparison of the IoT platform vendors, microsoft Azure, Amazon web services, and Google cloud, from users’ perspectives”. In: *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE. 2020, pp. 1–4 (cit. on pp. 46, 47).

[151] R. Uzwyshyn. “Research data repositories: the what, when, why and how”. In: *Computers in Libraries* 36.3 (2016), pp. 8–21 (cit. on p. 23).

[152] WHO. “International classification of diseases”. In: *WHO [Internet]* (1992) (cit. on p. 37).

[153] WHO. *Who coronavirus (COVID-19) dashboard*. url: https://covid19.who.int/ (cit. on p. 89).

[154] M. D. Wilkinson et al. “A design framework and exemplar metrics for FAIRness”. In: *Scientific data* 5.1 (2018), pp. 1–4 (cit. on p. 39).

[155] M. D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1 (2016), pp. 1–9 (cit. on p. 39).

[156] M. Wittig and A. Wittig. *Amazon web services in action*. Simon and Schuster, 2018 (cit. on p. 46).

[157] L. Youseff, M. Butrico, and D. Da Silva. “Toward a unified ontology of cloud computing”. In: *2008 Grid Computing Environments Workshop*. IEEE. 2008, pp. 1–10 (cit. on p. 42).

[158] D. A. Zarin et al. “The ClinicalTrials. gov results database—update and key issues”. In: *New England Journal of Medicine* 364.9 (2011), pp. 852–860 (cit. on p. 112).

[159] N. Zeinali, A. Asosheh, and S. Setareh. “The conceptual model to solve the problem of interoperability in health information systems”. In: *2016 8th International Symposium on Telecommunications (IST)* (2016), pp. 684–689 (cit. on p. 33).

[160] M. L. Zeng. “Interoperability”. In: *KO Knowledge Organization* 46.2 (2019), pp. 122–146 (cit. on p. 33).

[161] P. Zikopoulos and C. Eaton. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011 (cit. on pp. 15, 16).

[162] N. Zulkarnain and M. Anshari. “Big data: Concept, applications, & challenges”. In: *2016 International Conference on Information Management and Technology (ICIMTech)*. IEEE. IEEE, 2016, pp. 307–310. doi: 10.1109/ICIMTech.2016.7930350 (cit. on p. 17).

A

Case Study I - Importance and Value of Data, its quality, veracity and availability in Research

A.1 Questionnaire and Acquired Responses

This section presents the questions that compose the questionnaire for the third case study as well as the collected responses.A total of 71 responses were collected, submitted by researchers, university professors, and workers in enterprises with contact in research.

1. What category of institution are you affiliated with?

    - University
    - Research Center
    - Enterprise
    - Enterprise with Research Department
    - Healthcare Institution
    - Other

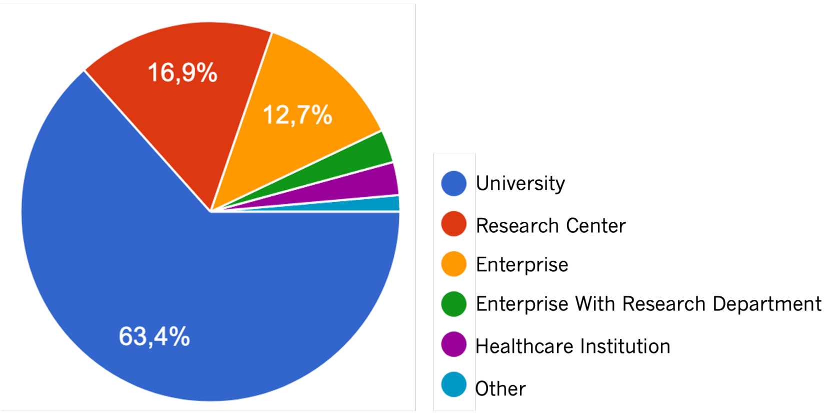


Figure 44: Collected responses for the instititution category.

2. How many research projects, dependent on the existence of data, have you met?

   - 1-10
   - 10-100
   - 100-1000
   - None

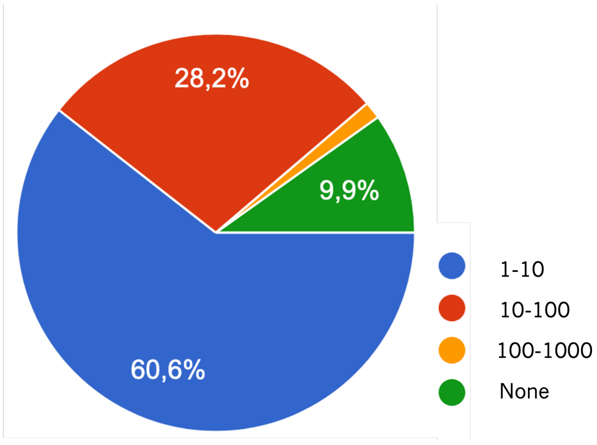


Figure 45: Collected responses for the known number of data dependent projects.

3. Are repositories typically accessible to the scientific community at no cost?

   - Yes
   - No
   - Sometimes

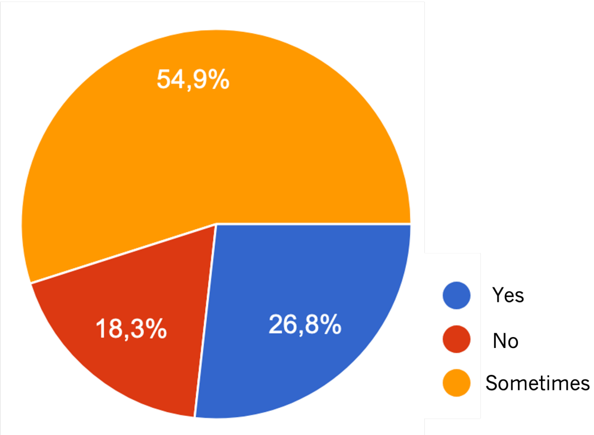


Figure 46: Collected responses for the repositories accessibility.

4. Are the repositories you know standardized?

- Yes
- No
- Not specified

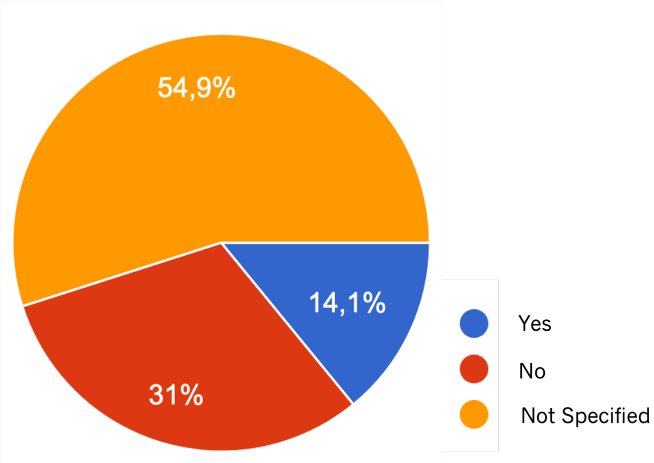


Figure 47: Collected responses for the standardization question.

5. How often are the repositories updated?

   - Never
   - 1 x per year
   - 1 x per week
   - 1 x per day
   - Real Time
   - Not specified

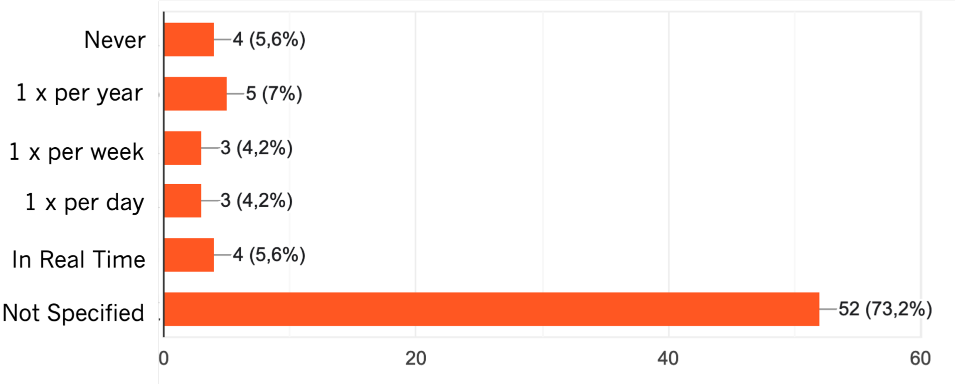


Figure 48: Collected responses for the most common update rate.

6. If you answered, "In real time", in the previous question, please specify which.

- No answers were provided. We ask people and it was due to the confidential contracts. So, the data repositories were not public.

7. Are the data present in the repositories real or synthetically generated? Answer considering most of them.

- Real Data
- Synthetic Data
- Not specified

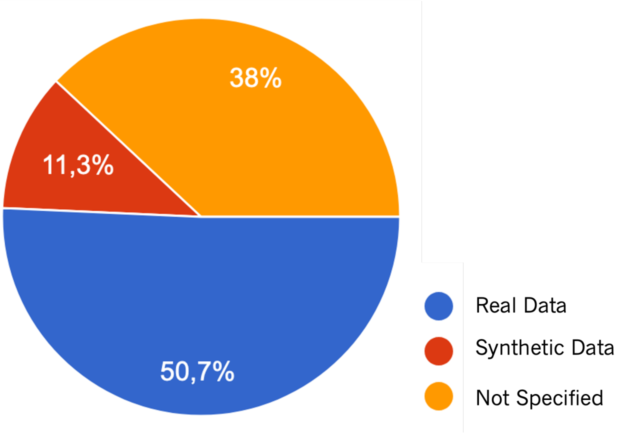


Figure 49: Collected responses for the typical data nature in repositories.

8. How large are the repositories? Very Small - up to one hundred records; Small - up to one thousand records; Medium - up to ten thousand records; Large - more than one hundred thousand records; Very Large - more than 1 million records.

- Very Small
- Small
- Medium
- Large
- Very Large

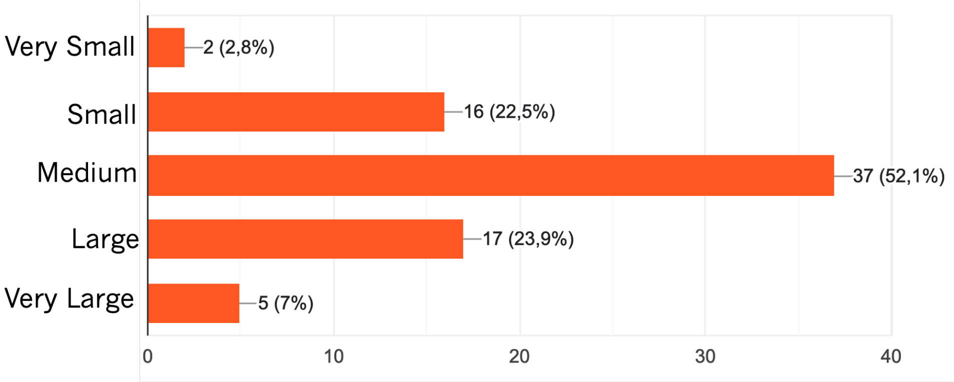


Figure 50: Collected responses for the repositories dimension.

9. How would you classify a repository, updated in real time, with anonymized and standardized data, made available through a free API?

   - Irrelevant
   - I'm not aware of the interest
   - Interesting
   - Extremely useful

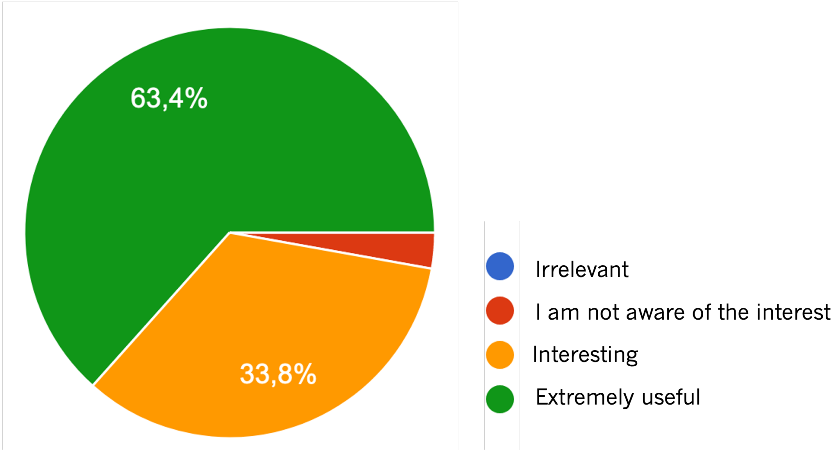


Figure 51: Collected responses for the possible Project Novelty.

10. Please indicate what improvements could be implemented.

   - The response to this question was not mandatory. Only 10 answers were submited where suggestions like metrics implementation, documentation, low response time, possibility to view some samples, possibility to configure the requests to have only relevant information.

Thank you for your cooperation.

B

Case Study II - Evaluation of the acceptance and usability of a SaaS or other cloud based systems

Table 13: Structure of the Questionnaire about SaaS

<table>
<tr><td colspan="11"><b>Please rate your level of satisfaction with the following aspects of our software:</b></td></tr>
<tr><td></td><td colspan="2">Very Satisfied</td><td colspan="2">Somewhat Satisfied</td><td colspan="2">Neutral</td><td colspan="2">Somewhat Unsatisfied</td><td colspan="2">Very Unsatisfied</td></tr>
<tr><td>Ease of access</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Hardware Compatibility</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Security</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Reliability</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Ability to integrate with other systems</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Ease of use</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Look and Feel</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Collaborate with team</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Documentation</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Value for money</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Overall Performance</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Support Quality</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr>
</table>

Table 13 continued from previous page

Please rate your level of satisfaction with the following aspects of our software:										
How likely is It that you would recommend this software to a friend or relative?										
	Comment									
Do you have any thoughts on how to improve this software?										

Thank you for your cooperation.

C

Case Study III - Evaluation of the acceptance and usability of management and decision support systems in critical situations

C.1 Questionnaire

This section presents the questions that compose the questionnaire for the third case study as well as the collected responses. A total of 112 responses were collected, submitted by the general population.

1. Age Group

   - < 18 years
   - 18 a 40 years
   - 40 a 65 years
   - > de 65 years

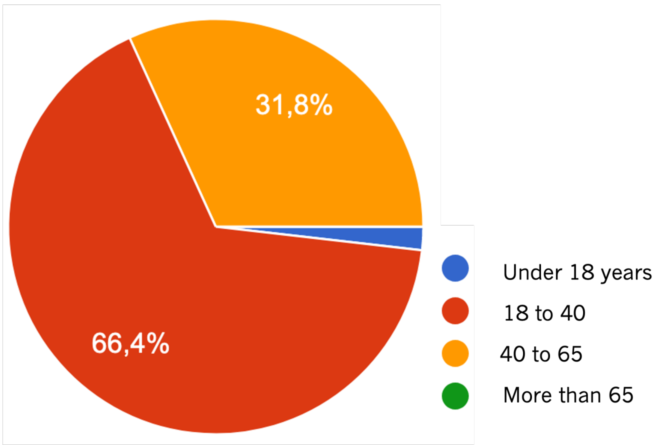


Figure 52: Distribution of the Age Group.

2. Residence Country

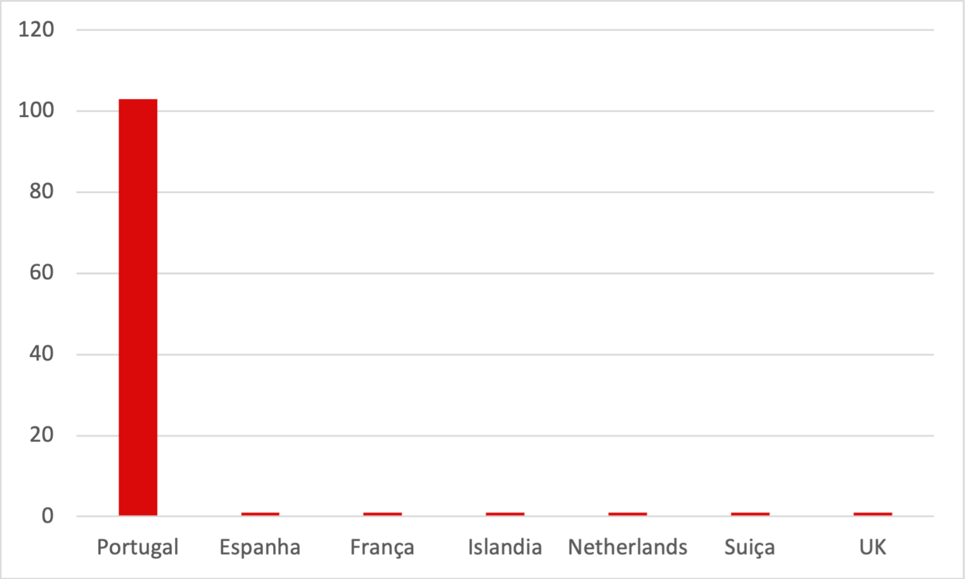


Figure 53: Residencial Country.

3. On a scale of 0 to 10, how satisfied were you with your life during Covid-19?

- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

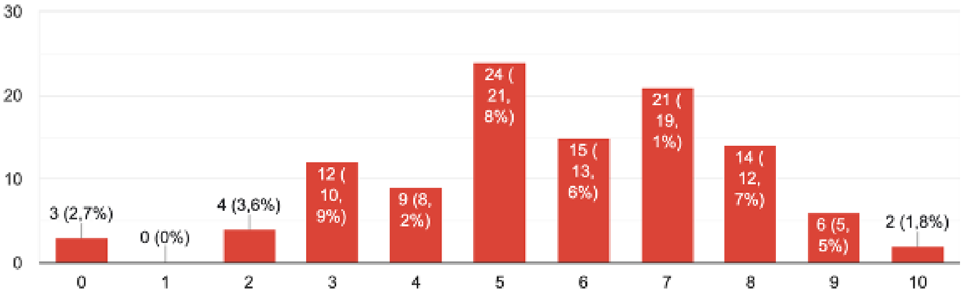


Figure 54: Degree of Satisfaction with life, during Covid-19.

4. During the confinement period how often did feelings like: Anxiety, Depression and/or Loneliness arise?

- Everyday
- Every Week
- Very Occasionally
- Never

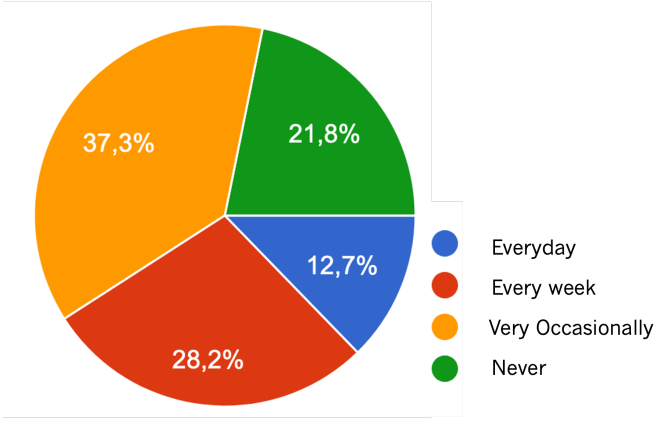


Figure 55: Periocity of feelings such as anxiety, depression, and loneliness.

5. Thinking about the situation since the introduction of government restrictions: Has the frequency of alcohol consumption changed?

   - Increased
   - Decreased
   - Did not change

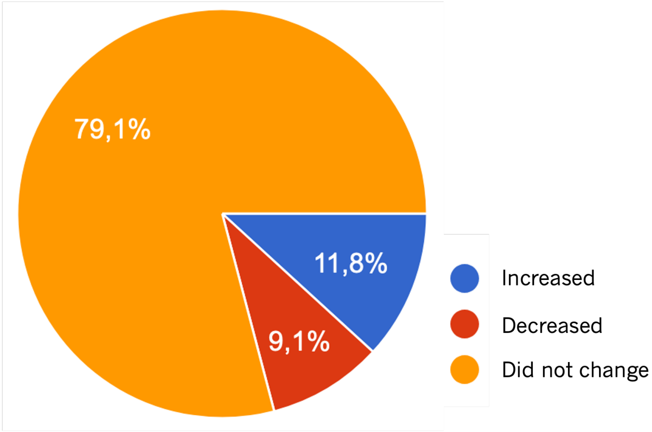


Figure 56: Change in the frequency of alcohol consumption.

6. Thinking about the situation since the introduction of government restrictions: Has the consumption of junk food and sweets changed?

- Increased
- Decreased
- Did not change

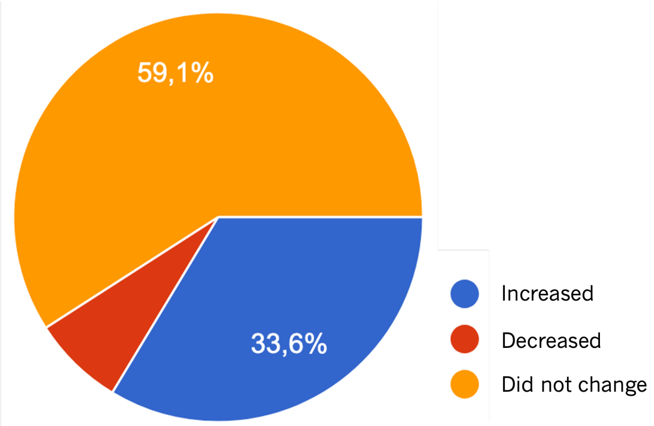


Figure 57: Change in the frequency of junk food consumption.

7. Before the pandemic, did you work from home?

- Yes
- No
- Sometimes

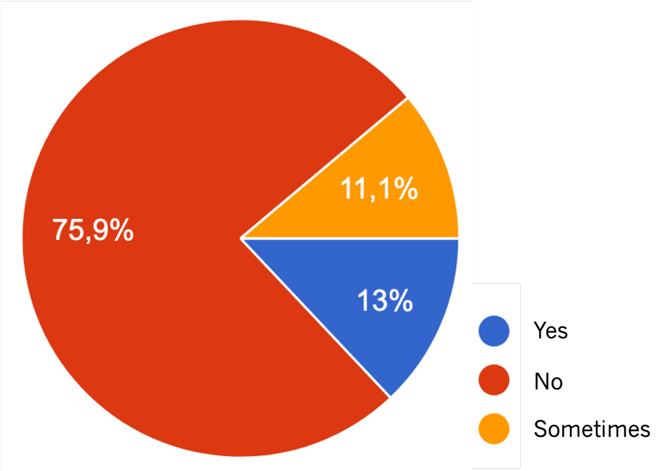


Figure 58: Preference in Pre-pandemic work arrangements.

8. When the Covid-19 restrictions were lifted, which would be more advantageous:

- Work from home
- Return to the workplace

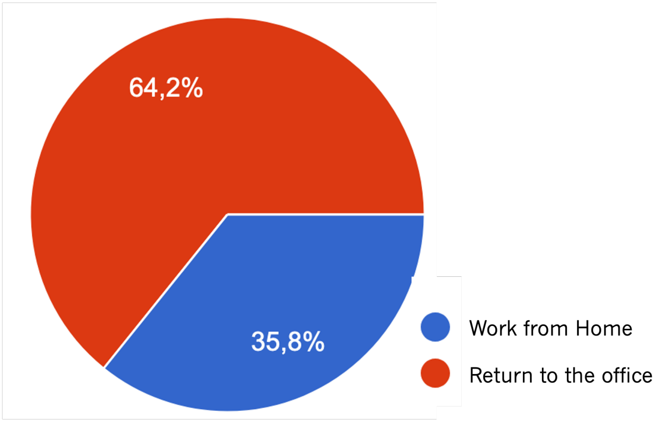


Figure 59: Preference Post-pandemic work arrangement preference.

9. On a scale of 0 to 10, where 0 is nothing and 10 is entirely, how well have the government's advice and guidance regarding Covid-19 been respected?

- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

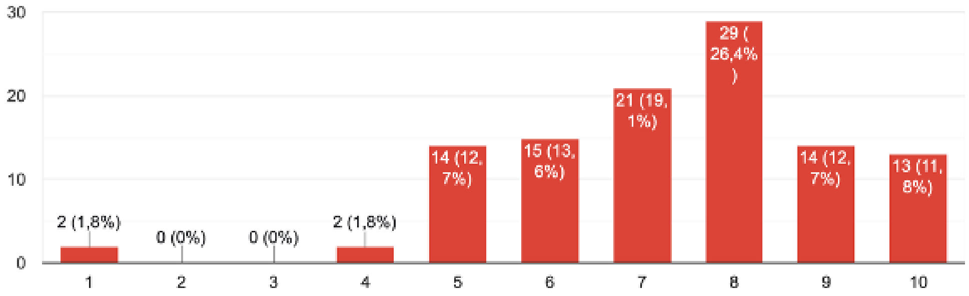


Figure 60: Compliance with imposed restrictive measures.

10. Were the restriction measures preventive or reactive?

- Preventive
- Reactive

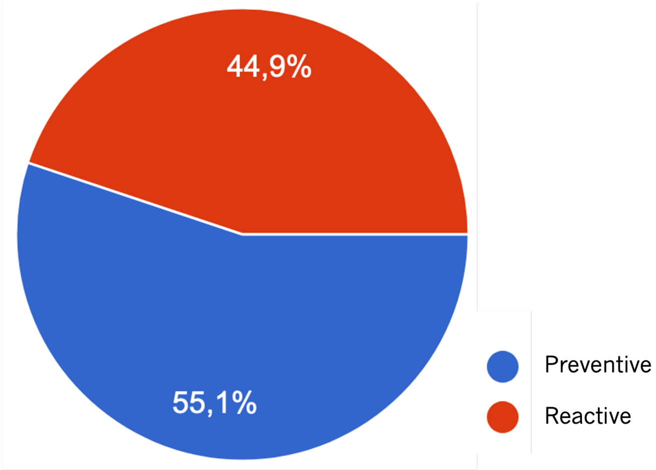


Figure 61: Perception of the type of measures adopted.

11. What information sources were used for the daily update on Covid-19?

    - Radio
    - Television
    - Social Media
    - All
    - Others

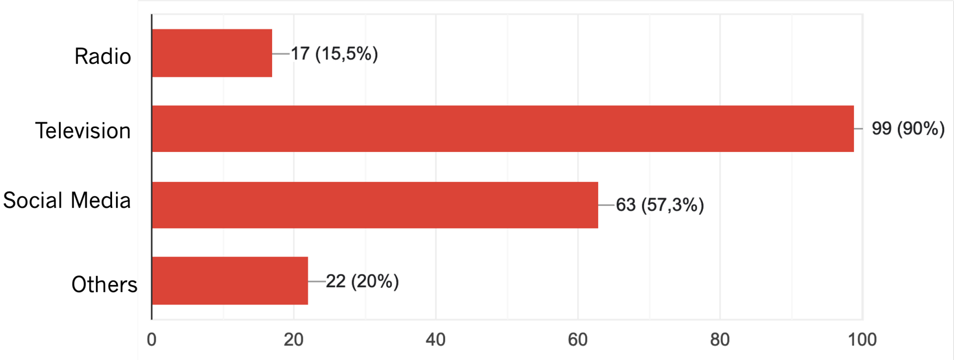


Figure 62: Information sources daily used.

12. On a scale of 1 to 5, where 1 is none and 5 is total, how confident are you in political decisions during Covid-19?

    - 0
    - 1
    - 2
    - 3

- 4
- 5

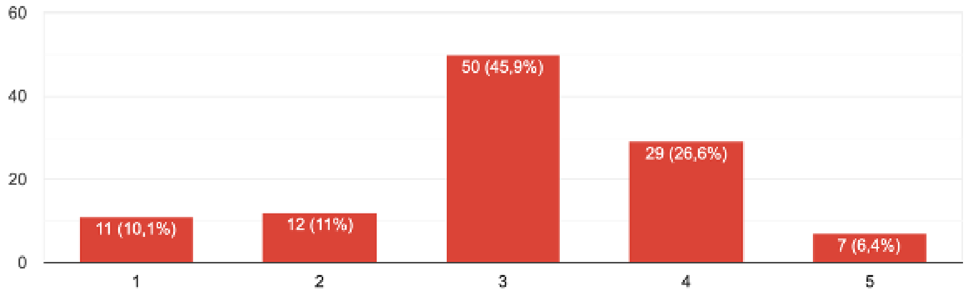


Figure 63: Trust degree in political decisions, during Covid-19.

13. Which guidelines were most likely to break?

    The response to this question was not mandatory. Only 66 responses were collected. The responses were classified into: None, All, Stay at home, Inside Travel, Vaccination, Face Mask, and others described in Figure 64.

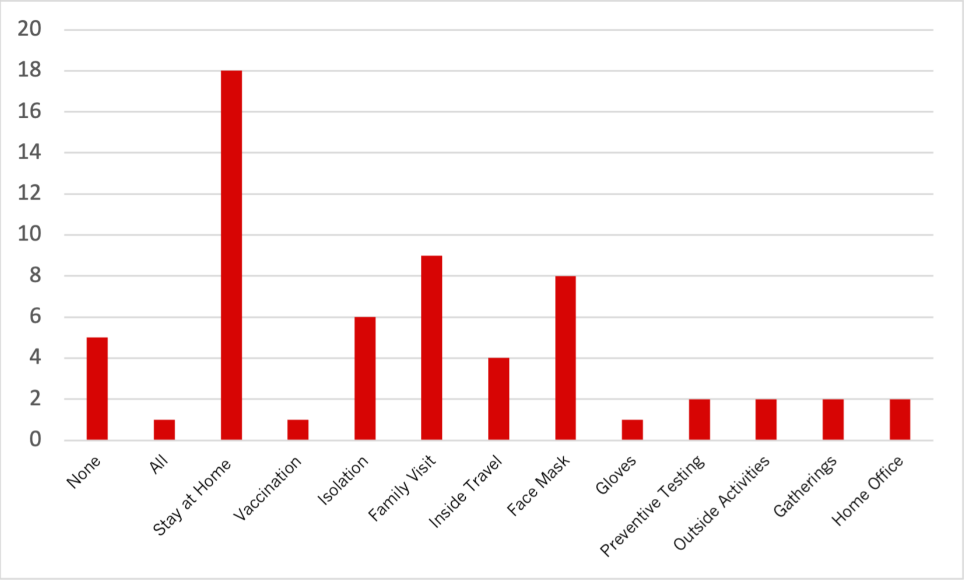


Figure 64: Government guidelines most likely to break.

14. If government restrictions were based on a real-time data analysis system, would the degree of trust increase?

    - No
    - Slightly
    - Yes

- Totally

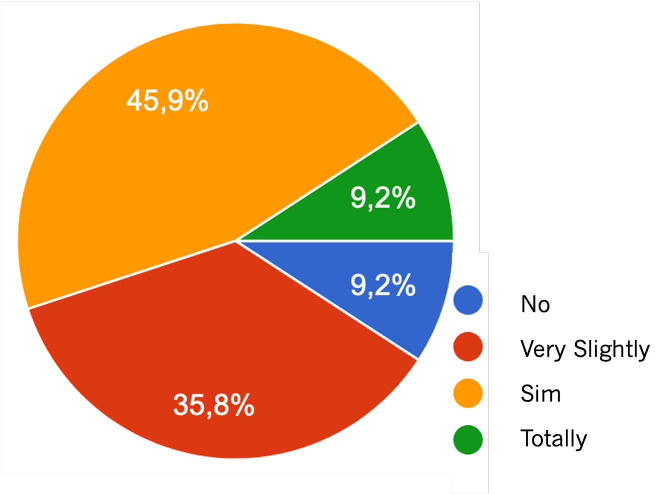


Figure 65: Change in the degree of confidence when using decision support systems.

15. If there was an automatic data analysis system, based on artificial intelligence, which according to, for example, the number of cases and deaths, adjusted the restriction measures, would you trust it?

    - Yes
    - No
    - Maybe

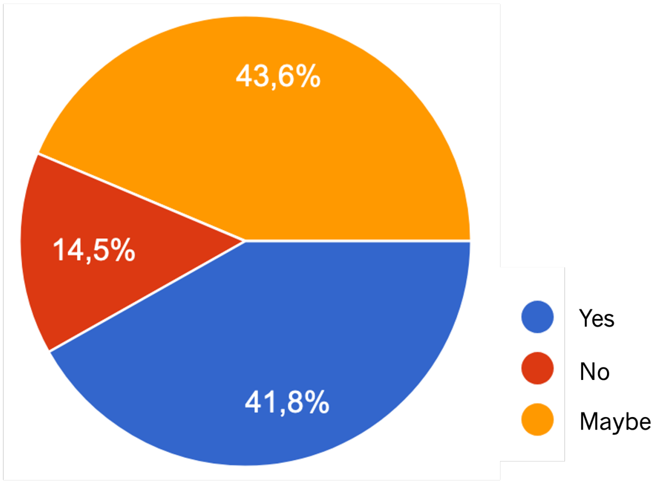


Figure 66: Change in the degree of confidence when using substantiated artificial intelligence methods.

16. If the decision-making process were more transparent, allowing the population to view the data, would trust increase?

- Yes
- No
- Maybe

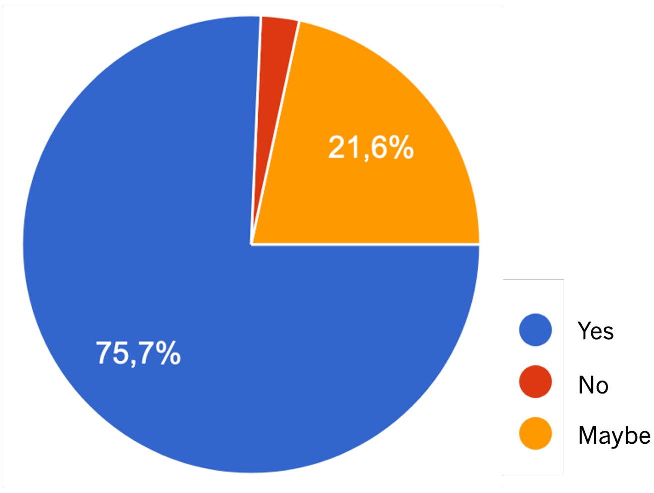


Figure 67: Change in the degree of confidence when using decision support systems with transparency for the population.

Thank you for your cooperation.