UMinho | 2024

Rui Miranda **Real-Time Information Proessing**


**Universidade do Minho**
Escola de Engenharia

Rui Pedro Mesquita Miranda


# Real-Time Information Proessing


June, 2024

**Universidade do Minho**
Escola de Engenharia

Rui Pedro Mesquita Miranda

# Real-Time Information Proessing

Doctorate Thesis
Doctorate in Informatics

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

June, 2024

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.27) (Lourenço, 2021).

# Acknowledgements

I would like to express my heartfelt gratitude to everyone who has supported me in my doctoral journey. Firstly, I extend my sincere thanks to my esteemed supervisors, José Machado and António Abelha, for their guidance, mentorship and invaluable insights throughout the course of my research. Their unwavering assistance and dedication have been instrumental in helping me achieve this milestone in my academic career.

I would like to extend my sincere appreciation to my colleagues at the Knowledge Engineering Group, whose unwavering encouragement, motivation, and camaraderie have been of immense help to me. I wholeheartedly thank Bosch Car Multimedia, the Municipal Council of Guimarães, and the *Hospital de Santo António* for their invaluable aid and funding that have enabled me to carry out this research.

I would also like to thank Raquel Rainha and Johann Behrens, my dear friends whose unwavering encouragement has been invaluable to me. Finally, I extend my gratitude to my friends from Ad Astra and the Velvet Room, two Discord communities who have supported me by providing a platform for socialization and relaxation during my doctoral journey.

Lastly, I would like to thank my family for their love and encouragement, which have been the driving force behind my academic pursuits. Without their consistent support and guidance, this achievement would not have been possible. Once again, I express my sincere gratitude to all those who have played a role in making this dissertation a reality.

## 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)

__________________________________________

(Rui Pedro Mesquita Miranda)

*"You cannot teach a man anything; you can only help him discover it in himself."*

**— Galileo**, Somewhere in a book or speach
(Astronomer, physicist and engineer)

# Resumo

## Processamento de Informação em Tempo Real

Todos os dias, grandes quantidades de dados são criadas a partir de dispositivos e sensores Internet of Things (IoT). Estes dados podem ser transformados em conhecimento de grande valor que pode ser utilizado para melhorar a capacidade de decisão e reduzir custos. Além disso, técnicas de Machine Learning podem ser utilizadas para identificar tendências a partir da informação recolhida. Por conseguinte, vários sectores podem beneficiar destas tecnologias, tais como a área da indústria, cuidados de saúde e cidades inteligentes.

Na área da indústria, existe um desejo pela sinergia entre humanos e máquinas autónomas, aproveitando a capacidade intelectual para aumentar a eficiência operacional. Em cuidados de saúde, são diariamente armazenadas vastas informações de pacientes, provenientes de uma grande variedade de fontes. Os dados recolhidos permitem analisar tendências e criar modelos de previsão que consigam prevenir potenciais problemas de saúde. Por último, cidades inteligentes podem detetar e responder a várias desafios, tais como o congestionamento de tráfego, segurança pública, consumo de energia e poluição ambiental, com a ajuda de dados de sensores, câmaras e outros dispositivos.

Consequentemente, as fontes de dados podem ser analisadas e estudadas para descobrir potenciais perspetivas e previsões. Estas abordagens incluem a utilização de sensores e sistemas autónomos, e introduzindo *edge computing* de uma forma distribuída e inteligente.

No âmbito deste projeto de dissertação, o objetivo é conceber, desenvolver, e explorar novas plataformas de análise e apoio à decisão em tempo real. Estas soluções foram implementadas com sucesso na indústria automóvel, em cidades inteligentes e no sector da saúde, demonstrando resultados positivos na recolha e análise de dados, na otimização de processos e na capacidade de tomada de decisões. As metodologias e abordagens desenvolvidas nesta investigação foram adaptadas para garantir a precisão dos dados, a sua segurança e a possibilidade de se chegar a resultados viáveis.

**Palavras Chave**: Big Data, Machine Learning, Sistemas de Suporte à Decisão, Data Mining, Business Intelligence (BI), Data Warehouse (DW)

# Abstract


## Real-Time Information Proessing

Large quantities of information from Internet of Things (IoT) devices and sensors are created every day. These data can be transformed into valuable knowledge that can be used to improve decision-making and reduce costs. Furthermore, Machine Learning techniques can be employed to identify hidden patterns from data. Thus, many domains can benefit from these technologies, such as manufacturing, healthcare, and smart cities.

In the manufacturing industry, there is a wish for an increased synergy between humans and autonomous machines, harnessing human brainpower to increase process efficiency. In healthcare, vast amounts of patient information are recorded on a daily basis from a wide variety of sources, for example, medical equipment and electronic health records. The data collected can be used to analyze trends and create predictive models that are able to prevent potential health problems. Finally, smart cities are able to identify and respond to various issues, such as traffic congestion, public safety, energy consumption, and environmental pollution, by using data from sensors, cameras, and other devices.

Thus, data sources can be explored and exploited to discover potential insights and predictions. These solutions include the usage of sensors and autonomous systems, and introduce edge computing in a distributed and intelligent manner.

In the context of this doctorate's thesis, the aim is to design, develop, and explore new platforms for real-time analytics and decision support. These solutions have been successfully implemented in the manufacturing industry, smart cities, and healthcare settings, and have shown positive outcomes in data collection and analysis, process optimization, and decision-making capabilities. The results obtained from the case studies presented in this dissertation show the effectiveness of real-time information processing solutions in increasing efficiency, providing accurate overviews of operations, and reducing communication gaps. The methodologies and approaches developed in this research have been tailored to ensure data accuracy, security, and actionable findings.



**Keyworks**: Big Data, Machine Learning, Decision Support Systems, Data Mining, Business Intelligence (BI), Data Warehouse (DW), Industry 5.0, Smart-Cities, Healthcare

# Contents

# List of Figures

# List of Tables

# Acronyms

**ADL** Archetype Definition Language *(p. 51)*
**AI** Artificial Intelligence *(p. 52)*
**AIDA** Agência para Integração, Difusão e Arquivo de Informação Médica *(pp. 78, 83, 94)*
**API** Application Programming Interface *(pp. xi, 12, 48, 71)*

**BI** Business Intelligence *(pp. vi, vii, 5, 15, 16, 21)*

**CEP** Complex Event Processing *(pp. 3, 5, 19, 25, 30, 56–59, 61, 63, 88)*
**CHUSA** Centro Hospitalar Universitário de Santo António *(pp. 1, 3, 4, 12, 13, 78, 84, 86, 91, 93, 96, 97)*
**CT** Clinical Terms *(p. 50)*

**DBMS** Database Management System *(pp. 17, 57)*
**DICOM** Digital Imaging and Communications in Medicine *(p. 50)*
**DSR** Design Science Research *(pp. 1, 5, 6, 13)*
**DW** Data Warehouse *(pp. vi, vii)*

**EHR** Electronic Health Record *(pp. 50, 51, 79, 92, 94, 97)*
**ETL** Extract, Transform, and Load *(pp. xi, 3–5, 11, 14–18, 48, 56, 57, 62, 66, 67, 79, 91, 96)*

**FHIR** Fast Healthcare Interoperability Resources *(pp. 50, 94)*

**GDPR** General Data Protection Regulation *(pp. 19, 33, 83)*

**HDFS** Hadoop Distributed Filesystem *(p. 20)*
**HL7** Health Level 7 *(pp. 13, 50, 79, 94)*

**HTTP** HyperText Transfer Protocol *(p. 69)*

**IIoT** Industrial Internet of Things *(pp. 52, 55)*
**IoT** Internet of Things *(pp. vi, vii, 1, 3, 5, 30, 34, 37, 41, 44, 52, 53, 55, 68, 77, 78, 93, 95, 96)*
**ISO** International Organization for Standardization *(p. 50)*
**IT** Information Technology *(pp. 3–5, 8, 49, 78, 91, 93, 95)*

**JDBC** Java Database Connectivity *(pp. 23, 28)*

**LAMP** Linux, Apache, MySQL, PHP/Perl/Python *(p. 59)*

**MCS** Mobile Crowdsencing *(pp. 36–40)*

**OLAP** Online Analytical Processing *(p. 17)*
**OLTP** Online Transaction Processing *(p. 17)*
**OSI** Open Systems Interconnection *(p. 50)*

**PoC** Proof of Concept *(pp. 8, 13)*
**PRISMA** Preferred Reporting Items for Systematic Reviews and Meta-Analyses *(pp. 32, 40–42)*

**RDBMS** Relational Database Management System *(p. 16)*
**REST** REpresentational State Transfer *(p. 69)*

**SQL** Structured Query Language *(p. 17)*
**SWOT** Strengths Weaknesses Opportunities Threats *(pp. xi, 8, 9, 14, 66, 76, 82)*

**TAM** Technology Acceptance Model *(pp. 8–10)*

**XDK** Cross-Domain Development Kit *(pp. xi, 55, 60, 61, 87)*

# 1

# Introduction & Research Methodologies

The present document describes the research and analysis for the development and exploration of next-generation platforms for decision support and unified Big Data analysis, as well as the processing of data streams in real-time. These projects arose within the scope of the Doctoral Program in Informatics at the University of Minho. This research was developed at the Centro ALGORITMI, a research unit of the University's School of Engineering, along with Bosch Car Multimedia, the Municipal Council of Guimarães, and the Centro Hospitalar Universitário de Santo António (CHUSA) in Porto.

This introductory chapter is divided in multiple sections. A brief contextualization of this project is presented, including references to key topics of discussion (Section 1.1), as well as the main motivation that led to its creation (Section 1.2). In addition, this chapter will give an overview of the research methodologies used in this project, particularly focusing on the DSR approach. It will also discuss various research techniques, highlighting the need for improved knowledge engineering in real-time. This chapter also includes the main objectives proposed for the development of this doctorate's thesis, thereby offering a description of the work performed (Section 1.6). Lastly, this chapter ends with the presentation of the document's structure to simplify its reading (Section 1.7).

## 1.1 Context and Scope

The development of new data producers, such as Internet of Things (IoT) devices, is revolutionizing the way data are collected, processed, and explored. Data are now gathered at an unprecedented level of precision and scale, enabling data-driven decision-making and insights across a range of industries and domains. Furthermore, by leveraging the connectivity of IoT devices, businesses can explore new opportunities to gain insight into their operations and customers, and develop innovative solutions to address their needs. As the technologies of these devices continue to evolve, the opportunities for businesses to benefit from the data they generate grow exponentially.

The manufacturing industry is under a paradigm shift, from a digital manufacturing model, where products control their assembly process using modular and efficient manufacturing systems, to a personalized manufacturing model, where products will be designed for an individual based on their personal requirements (Doyle-Kent & Kopacek, 2020; Lasi et al., 2014). Furthermore, consumers are increasingly more aware of their environmental impact. Thus, future solutions must employ technologies to increase environmental sustainability, helping to save the environment (Nahavandi, 2019, p. 3). The term 'Industry 5.0' was adopted to represent this transformation, representing the 'fifth industrial revolution', as well as continuing the earlier generation naming scheme, alluring to how software is versioned (Oztemel & Gursev, 2020, p. 1).

New technological advancements, from sensor data interoperability, more mechanization and automation, to the convergence of the physical and digital worlds, and virtual training will support this transition, delivering extensive changes in manufacturing systems (Nahavandi, 2019, pp. 5–6). In addition, ever-changing requirements demand for flexible tools and modules that are able to adapt to people, machines, and processes (Robert Bosch GmbH, 2020). These solutions will enable "human-robot coworking", that is, humans will focus on tasks requiring creativity, while robots will do the rest of the work (Demir et al., 2019, p. 3).

As urban populations grow, the need to modernize cities and give high-quality essential services to communities has become increasingly important. This has led to the rise of Smart Cities, where technology and infrastructure are combined to enhance the quality of life in a sustainable and transparent way (Kirimtat et al., 2020; Nižetić et al., 2020). The processing of information in real time is a crucial component of smart cities, as it enables them to improve public safety, manage traffic flow, reduce energy consumption, and optimize the delivery of public services. With the right infrastructure and technology in place, real-time processing of data can also help cities identify opportunities for economic development and create new job opportunities. Thus, real-time processing of data can be used to develop and implement predictive analytics that can help cities better prepare for future scenarios.

Real-time information processing is becoming an integral part of healthcare, enabling healthcare professionals to respond to changes in patient health and offer timely, effective care. Many use cases have been applied to healthcare, including monitoring vital signs such as heart rate, blood pressure, and respiration rate, checking patient adherence to treatment plans, and detecting potential health risks before they become critical (Afreen Banu & Rajamani, 2023; Alharbi et al., 2021). In addition, machine learning algorithms have been used to analyze large amounts of clinical data and identify patterns that can help healthcare professionals make better decisions (F. Hassan et al., 2020). However, the use of these technologies in healthcare is still in its infancy, and there are many challenges that need to be addressed before they can be widely adopted, such as data security and privacy, data quality, and data interoperability.

Thus, data processing in real time is becoming increasingly important to many businesses, as it allows them to analyze large amounts of data and make decisions. This is especially true about Big Data, where the vast amounts of data that are generated were previously thought to be useless or computationally expensive to process can now be of great value. Technologies such as Apache Storm, Apache Flink, Apache Spark, and Apache Kafka are being used to process data in real time and give insights into business operations (Sousa et al., 2021).

However, these technologies have their own set of challenges, such as ensuring data semantics and data quality. Data semantics and data quality are essential elements of successful data processing, while data semantics refer to the meaning and context of data, as well as the relationships between different data points (Techopedia, 2018). Data quality involves ensuring that data are accurate, reliable, and up to date. With high data quality, in particular with time granularity, "... real-time process control of the company can be implemented on the basis of the automatically recorded data, which are directly available for management via fast network connections, regardless of where the decision-makers would like to retrieve them" (Q. Hassan, 2018, p. 33). In contrast, poor data quality can result in inaccurate results and faulty decision-making (Q. Hassan, 2018, p. 35). Therefore, understanding and optimizing data semantics and quality are essential for successful data processing and leveraging its value.

This thesis aims to explore the importance of real-time data processing, by developing platforms that had data semantics and interopeabilty as their main focus. This research was developed in collaboration with several institutions, with each having their own use-case. The first use-case was built along with Bosch Car Multimedia in Braga, with the goal of developing an integrated environment for analysis, visualization and reporting of business indicators, namely by analytical dashboards for interactive exploration and analysis of production lines, with the usage of Complex Event Processing (CEP) technologies and ontologies. The second use-case was built along with the Municipal Council of Guimarães, where a platform for creating, managing, and visualizing geofences is presented and discussed. Finally, the third use-case was developed with the Centro Hospitalar Universitário de Santo António (CHUSA), aiming to build a unified patient monitoring analytics platform, implementing streaming data processing pipelines, along with Publisher/Subscriber patterns to support all present needs, and adaptable to future ones.

## 1.2 Motivation

Accordingly, the main motivation for this project focuses on developing and exploring a new generation of tools in the field of Information Technologies (ITs) in order to gather and store large quantities of data, as well as the ability to process that data in real-time, enabling new analytical solutions and artificial intelligence approaches.

These data, retrieved from sensors, Internet of Things (IoT) devices, and data producers will be used to tackle multiple challenges, from intelligent failure prediction and avoidance to data-driven process optimisation. Therefore, multiple solutions, including web applications and enterprise communication protocols will be developed, using several methodologies and technologies now available and viable for the conception of the defined IT solutions, also called IT artefacts.

As a first case study (**Case Study nº 1**), a new platform for sensorial Extract, Transform, and Load (ETL), a framework that is able to store and organise all data from sensors and Internet of Things (IoT) devices, a web application for real-time analytics and alerts, and new communication protocols based on semantic interoperability will be developed. The platform and framework will be capable of aggregating and storing large quantities of data provided from sensors and equipment, and preparing that information to be used timely. The web application will support the production lines and will be developed by adopting modern technologies and frameworks. Lastly, the new communication protocols will be integrated with existing and ones, ensuring all current and future systems that are part of the organisation's landscape can interoperate.

For the second case study (**Case Study nº 2**), the platform for creation, management, and visualization of *geofences* is presented and discussed. This platform is composed of a web application that allows the creation and management of geofences for the Municipality of Guimarães, as well as the visualization of metrics; a mobile application, available for *Android* and *iOS* devices, which allows the notification of users when they enter a geofence; and web services that enable interactions from external applications in a fast and secure way.

The aim of the third case study (**Case Study nº 3**), in collaboration between the Centro Hospitalar Universitário de Santo António (CHUSA) and the University of Minho, focuses on the development and exploration of an architecture for the processing and storage of data in real- time for decision support at the healthcare institution, with a case study about patient monitors telemetry. Its main goals are to discover CHUSA's main necessities and obstacles, identify methods to process data from patient monitors in real-time, develop new approaches for aggregation, extraction, transformation, and loading of information in real-time, create a real-time analytics dashboard that incorporates an overview of the platform and its impact on decision-making, develop an architecture to support the persistence of data for the sensorial ETL system through stateful data streaming, create methodologies to quantify the efficiency and effectiveness of real-time processing protocols, and establish automatic alerting to medical professionals when patient parameters are outliers. In addition, the case study will implement communication protocols based on semantic interoperability, and enable external applications to access and interact with the developed solution.

## 1.3 Research Methodologies

Research methodologies are an essential part of a doctoral degree. They offer the structure for researhers to collect and analyze data, as well as to develop theoretical frameworks for their research By comparing, evaluating, and synthesizing existing information and theories, new knowledge and understanding can be created. Furthermore, research methodologies help to ensure that the results of doctoral research are valid, reliable, and replicable. Research methodologies can also offer the opportunity to examine their research questions from a variety of perspectives, enabling them to develop an in-depth understanding of their research topics. Thus, research methodologies are vital to the success of a doctoral degree, as they offer the foundation for students to develop their own unique and valuable research insights.

A research project requires the adoption of methodologies to define its phases, processes, and methods. As such, every IT study must carefully research and analyze multiple methodologies and technologies that are both available and viable for the conception of a solution (Brandão et al., 2016). A systematic review of the literature was conducted to identify the most relevant technologies and sofware tools for CEP, their advantages and limitations, and how they can be applied to the established use cases. The final decision should be based upon their main advantages and limitations, as well as their compliance and security requirements.

A literature review on multiple topics, including Business Intelligence (BI), Big Data, Data Warehousing, machine learning algorithms and their application in BI was conducted. As such, searches using keywords on abstract and citation databases were consulted, in particular Scopus and Web of Science. Primary sources, namely reports, articles, and dissertations were consulted, as well as secondary sources, for example, books and magazines. In addition, internet searches were performed to confirm and clarify concepts. By using a deductive approach, it is expected the development of theory thought a well-defined research strategy that will be tested and confirmed with data (Martinez, 2016, p. 145).

Martinez noted that *"A well-thought-out and consistent set of assumptions will constitute a credible research philosophy, which will underpin your methodological choice, research strategy and data collection techniques and analysis procedures"*, enabling researchers to design a coherent research project. An ontological philosophy was adopted, referring to assumptions about the nature of reality (Martinez, 2016, p. 143). Hence this project will base upon the Pragmatism research paradigm, as it avoids unnecessary debates between the different philosophies, instead focusing on a problem and aiming to contribute with practical solutions that inform future practice (Martinez, 2016, p. 143).

To answer all proposed Research Questions, a Research Design applied to engineering and sciences was formulated, leading to the adoption of the Design Science Research methodology. This methodology is used for the development of solutions and approaches in ITs, creating knowledge than can be shared and discussed.

For this thesis project, the DSR methodology is used to ensure the solution to the problem accomplishes all needs and requirements of professionals, and benefits society.

Data will be collected from multiple sources from production lines, including from sensors, Internet of Things (IoT) devices , and data producers. This knowledge will be used to develop new methodologies to Extract, Transform, and Load (ETL) information from sensors in real-time (sensorial ETL), as well as architectures to support the persistence of information for sensorial ETL systems. Hence a mixed investigation methodology will be adopted, using quantitative and qualitative methods for data collection and analysis. This project is additionally characterised by its cross-sectional time horizon, as it involves *"...the study of a particular phenomenon (or phenomena) at a particular time"* (Martinez, 2016, p. 200).

This way, the institutions are provided with an appropriate and well-founded solution, based on methodologies and technologies already explored and proven suitable for the solution to their problem in question, while elucidating new knowledge both to the institution and to the scientific community.

### 1.3.1 Design Science Research

Design Science Research (DSR) can be described as a *"research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artefacts, thereby contributing new knowledge to the body of scientific evidence"* (Hevner & Chatterjee, 2010, p. 5). Two primary activities improve and elaborate on the behaviour of aspects in Information Systems, namely the *" creation of new knowledge through design of novel or innovative artefacts (things or processes) and the analysis of the artefact's use and/or performance with reflection and abstraction."* (Kuechler & Petter, 2017, p. 1).

This paradigm is relevant for guiding the design and development of Information Systems, as they give a framework for understanding, identifying and resolving complex problems (March & Storey, 2008, pp. 2, 3). Moreover, the iterative design process allows for the continuous improvement of existing artefacts, leading to the creation of even more effective Information Systems.

This methodology uses an iterative approach, allowing it to have the freedom to adapt the structure and evaluation of the developed artefact until an ideal solution is found (Kuechler & Petter, 2017, pp. 8–10).

The Design Science Research (DSR) methodology was represented by (Peffers et al., 2007) in Figure 1.

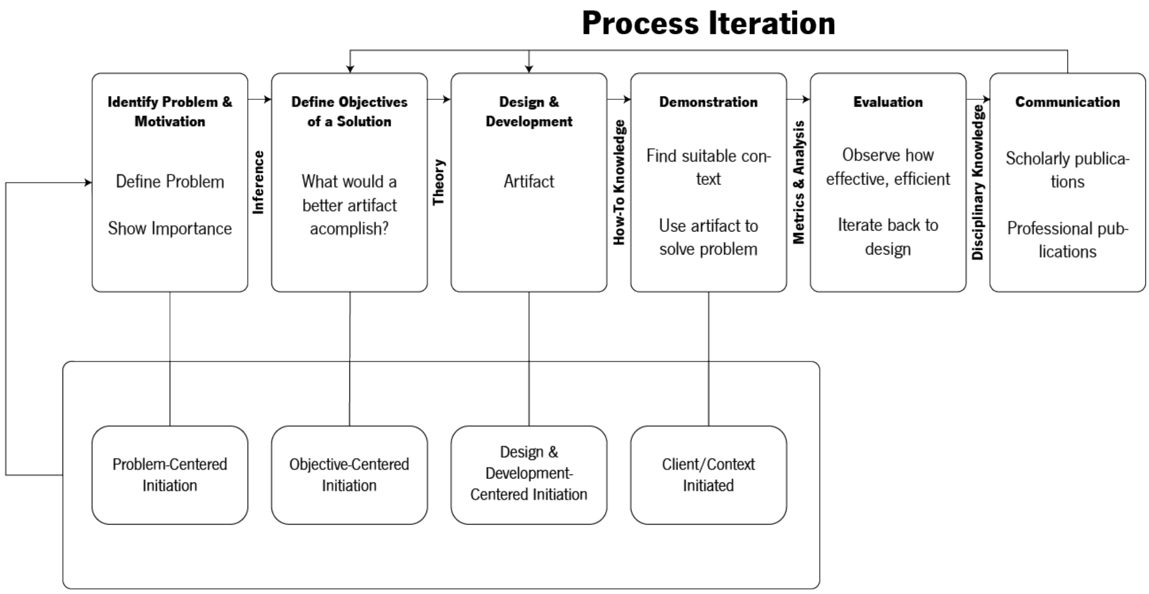


Figure 1: Design Science Research Methodology. Adapted from (Peffers et al., 2007, p. 15)

This research employed various strategies to adhere to the DSR methodology. Its exploratory nature seeks to bring new perspectives and approaches to address issues noticed in a certain setting. This involved examining existing information exchange systems and storage in several organizations and domains, and how real-time information processing approaches can benefit users and professionals. Through this, gaps were identified, which led to the generation of new platforms for information processing, storage , and visualization.

Both quantitative and qualitative methodologies were used, including content analysis and critical studies. The project was also characterized by its long-term temporal horizon.

## 1.4 Case Study Approach

A case study approach is a research method that involves a comprehensive and in-depth examination of a particular topic or phenomenon. This method is particularly useful for examining complex real-world situations and provides a rich and detailed account of the subject under investigation. To conduct a case study, it's essential to select a case that's relevant, interesting, and representative of the research question. A theoretical framework should be constructed to guide a data collection and analysis, either by utilizing existing theories or developing new ones based on literature review. Data should be collected from a variety of sources, including interviews, documents, and observations, to ensure the validity and reliability of the findings. Following data collection, the case should be described and analyzed using the theoretical framework, highlighting key characteristics, meanings, and implications for the research question (Benbasat et al., 1987; Gillham, 2010; Yin, 2018).

Thus, this project employed a case study research approach, ethnography, and a cross-sectional methodology to conduct an in-depth and extensive research of several subjects of study (the cases), as well as their associated contextual conditions, at different moments.

The goal was to examine many institutions to get a broad set of information, with the aim of producing and constructing a universal artefact that's applicable to the greatest number of similar environments as possible. To best answer the research topic, information was collected and analyzed using both quantitative and qualitative methodologies. Primary data were collected through participant observation, focus groups, and semi-structured interviews. Secondary data were gathered through archival research, primarily from primary (dissertations and conference proceedings) and secondary (journals and books) sources. Using grounded theory, theories and models were generated based on the collected and analyzed data. The approach was both inductive and deductive. It was inductive as the construction of theories and models could only occur after a data collection and analysis, which was then compared to literature, and deductive as it was difficult to perform a solid and credible study that relied on case studies alone through an inductive strategy.

A literature review of the state of the art is essential for the preparation of this doctoral thesis, leading to the formation of hypotheses based on existing theories and the development of a research method to evaluate those hypotheses using a deductive approach.

## 1.5 Proof Of Concept

A Proof of Concept (PoC) is a demonstration to confirm whether a concept or theory is feasible for real-world applications (Singaram & Jain, 2018; Techopedia, 2019). A methodology is made up of practical models that can prove or validate a concept through analysis and development. Proof of concept research *"presents a discovery about our knowledge of the world and the structure of existence"* (Kendig, 2016).

A Proof of Concept is considered a key procedure in the process of design, development, implementation, and proposal of software solutions in the field of ITs. They validate whether a project fulfils its main features and defined objectives, as well as identifies potential flaws or mistakes in the developed solution. This includes iterative usability testing with different user types, as well as a cross-sectional study to assess the system's feasibility and effectiveness, to measure the overall impact. In addition, the proof of concept in research aims to answer a question whose answer has wide applicability in areas beyond those tested for (Kendig, 2016).

In this dissertation project, the viability and functionality discussion of the developed tools for decision support and insights was corroborated by the application of a Proof of Concept methodology. In order to evaluate the technical and functional quality of the system, the Technology Acceptance Model (TAM) method was used to assess the Perceived Usefulness, Perceived Ease of Use, Behavioral Intention, and Use Behaviour (Chooprayoon & Che, 2010). The acceptability, feasibility, and utility of the research project presented in this doctoral thesis were justified by employing a Strengths Weaknesses Opportunities Threats (SWOT) analysis and a TAM evaluation. These processes are outlined below:

### 1.5.1 SWOT Analysis

The Strengths Weaknesses Opportunities Threats (SWOT) analysis planning technique is used to develop strong strategic plans by analysing the strengths and weaknesses (internal factors), in addition to opportunities and threats (external factors), of a business or project plan (Mind Tools Content Team, 2019). A SWOT analysis promotes a fact-based and data-driven analysis of the internal and external factors that are advantageous or detrimental, helping reveal prospective opportunities to exploit and understand potential weaknesses (contributors, 2019; Mitchell, 2019).

SWOT analysis groups key data into two main categories: internal and external. Internal factors are the present strengths and weaknesses of an organisation while external factors are the opportunities and threats offered by the environment, out of control of the organisation. SWOTs can be used as inputs for the creation of strong, well-calculated strategies (Hay & Castilla, 2006).

In the context of applying a Strengths Weaknesses Opportunities Threats (SWOT) analysis to the developed platform and the food recommender module, each study can thus be based on four main components:

- Strengths: internal attributes of the platform that are helpful towards its successful development and integration;

- Weaknesses: internal attributes of the platform that are harmful towards its successful development and integration;
- Opportunities: external attributes of the platform that are helpful towards its successful development and integration;
- Threats: external attributes of the platform that are harmful towards its successful development and integration.

A carefully thought out SWOT analysis can produce valuable information that helps in decision-making. When the potential implications of this information on the organisation are fully assessed, meaningful analysis emerges (Hay & Castilla, 2006).

The SWOT matrix was represented by (Mitchell, 2019) in Figure 2, where each characteristic is made up in one quadrant, providing a quick overview of a project's position.

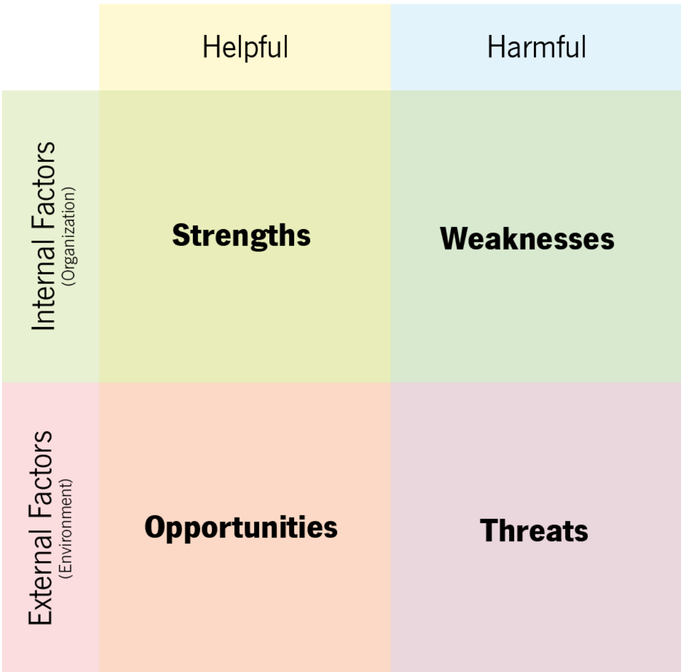


Figure 2: SWOT Matrix (Mitchell, 2019)

### 1.5.2 Technology Acceptance Model

The Technology Acceptance Model (TAM) is a widely used theoretical framework for understanding and predicting user acceptance and adoption of new technologies. It was first proposed by Davis in 1989, and he hypothesizes that the perceived usefulness and ease of use of a new technology are the two main factors that influence a user's intention of using it (Davis, 1989). According to TAM, *perceived usefulness* is the degree to which a user believes that technology will enhance their job performance, while *perceived ease of use* is the degree to which a user believes that a technology is easy to use and learn. These factors can be influenced by a range of external variables, such as social influence and technology-related anxiety (Chooprayoon & Che, 2010).

The usefulness of TAM for research lies in its ability to give a framework for understanding and predicting user acceptance of new technologies. By utilizing TAM, a researcher can identify the key factors that decide whether professionals or other stakeholders will adopt or resist new technologies. This, in turn, can help policymakers to design more effective strategies for implementing them and improving outcomes. Furthermore, TAM can be customized to suit specific populations, technologies or contexts, and provides a set of variables that have been well tested and validated in many other fields. This makes it a powerful tool for empirical research across different industries, and its utility extends beyond academia as it can be useful to practitioners in the digital innovation and technology industries (Venkatesh, 2000b).

There are several versions of the Technology Acceptance Model (TAM) that have been developed over the years. Some main versions of TAM and their key aspects are:

- **TAM1**: This is the original version of TAM developed by Davis in 1989. It proposed that **Perceived Usefulness** (PU) and **Perceived Ease of Use** (PEOU) were the two key factors that influenced users' acceptance of technology. PU refers to the perceived benefits of using technology, while PEOU refers to the perceived ease of using technology (Davis, 1989).

- **TAM2**: This version of TAM was developed by Venkatesh in 2000, and it expanded on the original model by including social factors like subjective norm and cognitive instrumental processes like perceived behavioral control. Subjective norm refers to the influence of social norms and expectations on an individual's behavior, while perceived behavioral control refers to an individual's perception of their ability to perform a particular behavior (Venkatesh, 2000a).

- **TAM3**: This version of TAM was developed by Venkatesh in 2008, and it introduced the idea of trust as a key factor that affects users' acceptance of technology. Trust is referred to users' confidence in the technology and the organisation providing it (Venkatesh, 2000b; Venkatesh & Bala, 2008).

Technology Acceptance Model (TAM) 3 outlines various determinants that influence an individual's perceived usefulness of a technology. These determinants include how easy they think it will be to use; what important people in their life think about it; how using it will affect their social status; if they think it's relevant to their job, will do a good job at their tasks, and if they can see and show others the results of using it (Venkatesh & Bala, 2008). As for the determinants of perceived ease of use, they include how confident someone is in their ability to use a computer for a particular task; whether someone believes the organization and technology they have access to will support their ability to use a particular system; how scared or worried someone is about using a computer; how creative or playful someone is in their interactions with a computer; how much someone enjoys using a specific computer system just for the sake of using it, not just for the outcome; and how easy it is to complete specific tasks using a computer system, based on actual effort instead of just perceptions (Venkatesh & Bala, 2008).

## 1.6 Objectives

The present iteration of the dissertation has been the development and exploration of a new generation of software solutions for decision support and insights in several domains as its main objective. With this project, it's expected to obtain methodologies to quantify the efficiency and effectiveness of real-time processing protocols, new approaches for the aggregation, extraction, transformation, and loading (sensorial ETL) of information in real-time, a viable architecture to support the persistence of data for sensorial ETL systems, and new communication protocols based on semantic interoperability sustained in the development of ontologies.

Therefore, in the context of this dissertation project, the primary Research Question that pervades the work has arisen:

**How can real-time information processing solutions be used to increase efficiency of a wide range of processes, allow for a more representative and accurate overview of current operations, and decrease communication gaps?**

- **Research Question 1:** What are the manufacturing industry's main necessities and obstacles, and how can the extraction and analysis of real-time sensorial data help optimise its workflow and reduce costs?

- **Research Question 2:** How to extract the untreated data from multiple sensors, process said data, and load it to relevant repositories having into account their context?

- **Research Question 3:** How can real-time sensor data processing deliver new opportunities to optimise production and reduce costs?

To answer these questions, the following Case Study and its objectives are highlighted:

- Case Study 1: Development of a real-time Extract, Transform, and Load (ETL) system that's able to gather and store large quantities of data provided by sensors and equipment, as well as to prepare information to be used in a timely manner:
  - Revision of literature regarding machine learning and real-time data processing
  - Elicitation of existing systems at Bosch Car Multimedia in Braga related with this project
  - Requirements elicitation for data processing in real-time
  - Analysis of tools and methodologies for data warehousing and ETL processes
  - Development of a real-time Extract, Transform, and Load (ETL) system that's able to extract, prepare, and load all information to data warehouses

  - Development of a framework for organisation and persistence of large quantities of data from sensors
  - Development of a web application for a visualisation dashboards, allowing an intuitive overview of the system in operation
  - Development of a module to support machine learning approaches, learning and adapting as data arrives
  - Ensuring the proposed platform is compatible with the current systems in place at Bosch

- **Research Question 4:** What types of data can be effectively processed and stored in real-time on a platform for decision support?
- **Research Question 5:** How can real-time sensor data processing deliver new opportunities to Smart Cities and contribute to the development of a more sustainable and efficient urban environment?

In order to answer these questions, the following Case Study and its objectives are proposed:

- Case Study 2: Development of a platform for the processing and storage of data in real-time for automatic geofencing and decision support at the Municipaility of Guimarães
  - Develeoplent of a backoffice application for the planning and creation of geofences for a given location, the visualization of metrics coming from the mobile applications, and the administration of users of the platform.
  - Development of a mobile application for the notification of users when they enter or leave a geofence.
  - Development of Application Programming Interfaces (APIs) for the integration of the platform with other systems.

- **Research Question 6:** What are the use cases for a real-time analytics dashboard, increasing user productivity, as well as optimising CHUSA's resources?
- **Research Question 7:** How can real-time sensor data processing deliver new opportunities to optimize patient's quality of care?
- **Research Question 8:** Can interoperability and ontologies help develop new communication protocols for distributed real-time communication protocols?

In order to answer these questions, the following Case Study and its objectives are proposed:

- Case Study 3: Development and exploration of an architecture for the processing and storage of data in real- time for medical support at the Centro Hospitalar Universitário de Santo António (CHUSA)

- Analysis of new approaches for the aggregation, extraction, transformation, and loading of information in real-time
- Development of a real-time analytics dashboard that incorporates an overview of the platform and its impact in decision making
- Development of communication protocols based on interoperability sustained in the development of integrations for Health Level 7 (HL7)
- Development of automatic alerting systems to medical professionals when patient parameters are outliers
- Publish an anonymized repository to the scientific community
- Development of methodologies to quantify the efficiency and effectiveness of real-time processing protocols
- Deployment and testing at CHUSA

Henceforth, the objectives that determine the project are seen as Research Questions, serving as a guide for solving the studies presented, and thus, which ones will be searched for solutions.

## 1.7 Document Structure

The following document is structured in 4 chapters: Introduction & Research Methodologies, Background, Case Studies, and Final Considerations.

- **Chapter I - Introduction & Research Methodologies:** This chapter aims to help readers to understand the focus and organization of the work, to give a detailed description of the research methodologies used in this dissertation, and to follow its progression as it's presented. An overview of the research area and relevant background information will offer context for this dissertation. Afterward, the main motivation for research is discussed, including the specific problems or gaps in the literature that the work aims to discuss. A description of the research methods and approaches that were used in this study will be offered. This will include a discussion of the Design Science Research (DSR) methodology, as well as the data collection and analysis techniques that were used. Next, the Case Study approach will be described in detail, including the steps that were followed and the tools and techniques that were used to conduct the research. Finally, the Proof of Concept (PoC) evaluation methodology will be discussed, and how it will be used to validate the work for real-world applications. After that, the main objectives of the study will be stated and described in detail. Finally, the structure of the document will be outlined, including a description of the various chapters, and the relationship between them.

- **Chapter II - Background:** This chapter provides an overview of real-time information processing technologies and their importance in enabling institutions to make decisions efficiently and accurately. It covers several topics such as Business Intelligence, Big Data, data warehousing systems, and ETL processes. It examines the potential of these concepts in the context of this research, implications, and challenges of their use. The chapter also discusses several Business Intelligence projects that have implemented machine learning systems and the importance of data quality, interoperability, and semantics. Additionally, it includes a section on crowdsensing and its potential for solving real-world problems, as well as a section on healthcare and the importance of interoperability and semantics in medical data. This chapter serves as a foundation for later chapters of this thesis, which will explore their use in more detail.

- **Chapter III - Case Studies:** This chapter presents the case studies that were conducted as part of this research. Each case study begins with an introduction to the problem that was addressed, and the motivation for the study. The methodology that was used to conduct the work is described in detail, including the solution's architecture, and the tools and techniques that were used. Finally, the results of the study are presented in detail, including a Strengths Weaknesses Opportunities Threats (SWOT) analysis of the final product.

- **Chapter IV - Final Considerations:** The last chapter of this document discusses and presents the main conclusions and contributions obtained through the development of this platform. The main results of the use-cases are presented and discussed in detail, answering all research questions. Future work to improve the platforms and its applications are proposed, including the identification of possible new applications, the development of new methods, and the evaluation of existing methods. In addition, the chapter discusses the proposed platform's potential implications for other research areas and its relevance to the broader field of information technology.

- **Appendix A - Publications:** All scientific contributions, namely, all publications and thesis elaborated on throughout this project are described.

# 2

# Background

This chapter will offer an overview of the current state-of-the-art in real-time information processing technologies. As data have become increasingly prevalent, real-time information processing has become an essential tool for businesses, institutions, and organizations. By allowing for fast and efficient data analysis, real-time information processing enables organizations to make decisions eficientyly and accurately. However, the success of real-time information processing depends on the quality of the data, as well as the ability to effectively interpret and use the data.

After an introduction of Business Intelligence is presented, a showcase regarding Big Data is performed, describing its main methodologies and commonly-used technologies. Furthermore, data warehousing systems are also presented, along with the Extract, Transform, and Load process. Several Business Intelligence projects that have implemented machine learning systems in literature are also discussed.

To ensure that data can be processed accurately and efficiently, interoprability, data quality, and semantics must be taken into consideration.

## 2.1 Big Data and its Applications

Businesses are now aware of the importance of data to discover hidden insights more than ever. A better understanding of their customers, the market, supply, and resources can give a competitive edge over their competitors. Business Intelligence (BI) technologies help provide "..., current, and predictive views of business operations" (Han et al., 2012, p. 27).

BI technologies include "..., online analytical processing, business performance management, competitive intelligence, benchmarking, and predictive analytics" (Han et al., 2012, p. 27). By using data mining approaches, businesses are able to perform effective market analysis, discover potential strengths and weaknesses of their competitors, and other smart business decisions. This way, they are able to transform data into valuable knowledge for strategic decision taking, based on their past experiences.

Big Data is a growing industry, as Big Data and business analytics solutions were forecasted to reach $189.1 billion in 2019 (Shirer & Goepfert, 2019). Core components of the Business Intelligence (BI) ecosystem include operational source systems, Extract, Transform, and Load (ETL) systems, data presentation frameworks, and business intelligence applications (Kimball & Ross, 2013, p. 18). Operational source systems are “*the operational systems of record that capture the business's transactions*”(Kimball & Ross, 2013, p. 18), prioritizing processing performance and availability. These may include multiple databases within the organisation, as well as external vendors (Chaudhuri et al., 2011) As data quality may vary, data cleansing and standardisation processes are a necessity in BI. As such, ETL tools are used to obtain data, perform a series of cleansing techniques, from correcting misspellings, to binning data from multiple sources and data de-duplication, and structuring and load it into the presentation area's target dimensional models (Kimball & Ross, 2013, pp. 19, 20). Data warehouses are used in BI architecture to store and process large amounts of structured information. Finally, BI applications query the data in the data warehouse, and leverage their presentation area for analytic decision-making. Thus, two main concepts related to BI technology, namely the ETL process and data warehousing are discussed in the following sections.

As more data is generated, there is an increasing desire to design “*low-cost data platforms that can support much larger data volume than that traditionally handled by Relational Database Management Systems (RDBMSs)*” (Chaudhuri et al., 2011, p. 90). Big Data engines aim to tackle this necessity, introducing sophisticated approaches to distributed processing of large data-sets. As such, Big Data systems will be further discussed in the following section.

The term Big Data can be applied to information that cannot be processed or analysed using traditional processes or tools (Zikopoulos et al., 2011, p. 3). Currently organisations have access to a large wealth of data, but cannot extract value from them due to its format and quantity. Thus, approaches and technologies were developed to help identify patterns and extract valuable knowledge from these data sources.

Zikopoulos et al. have identified three main characteristics of Big Data: volume, variety, and velocity. The first one, volume, can be attributed to the amount of data that can be stored. Companies try to track and store every interaction as possible, but need the proper tools to manage and analyse these data. The second characteristic, variety, represents a shift from traditional analytics, to now include raw, semi-structured, and unstructured data in account. Finally, the velocity as these data are generated has increased dramatically. Big Data tools now have to perform analytics to data in motion, enabling insights based on emerging information (Zikopoulos et al., 2011, pp. 5–9).

Big Data approaches are now used on multiple use cases, from search ranking, ad tracking, and location tracking to financial account fraud detection, building sensors, and loan risk analysis (Kimball & Ross, 2013, pp. 527–528). Each of these use cases have their specific needs and requirements but have the goal of using previously useless data to extract insights. Big Data complements Data Warehousing, as data might initially be seen of low quality, but may be key to future questions (Zikopoulos et al., 2011, p. 11). That data might be on Big Data repositories until its value is discovered, then it might be migrated into data warehouses when its value is proven and sustainable (Zikopoulos et al., 2011, p. 12).

### 2.1.1 Data Warehousing

A data warehouse can be defined as a "a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decision" (Inmon, 2002, p. 31). Hence, data warehousing solutions use decision support technologies to "enable knowledge workers to make better and faster decisions" (Chaudhuri & Dayal, 1997). The data warehouse is maintained separately from operational databases, as their functional and performance requirements are different. While operational databases, supporting Online Transaction Processing (OLTP) applications, are designed for consistency and recoverability, data warehouses, employing Online Analytical Processing (OLAP) applications, are designed for decision support, prioritising query throughput and response times. Thus, historical, summarised, and consolidated data are more important on data warehouses than individual records, with orders of magnitudes of data larger than operational databases (Chaudhuri & Dayal, 1997, p. 1).

Depending on the adopted DBMS, multiple-dimensional models can be implemented. On relational DBMS, the most common dimensional models are the star schema, due to their resemblance, while dimensional models implemented in multidimensional database environments are referred to as OLAP cubes (Kimball & Ross, 2013, p. 8).

OLAP cubes store and index data using formats and techniques specifically designed for dimensional data. The OLAP cubes are able to create and manage performance aggregations and summary tables. Thus, they are capable of delivering better query performance due to their pre-calculations, indexing strategies, and optimisations (Kimball & Ross, 2013, pp. 8–9). Furthermore, they may also provide advanced analytical functions, exceeding those available with SQL.

However, as data may arrive from multiple data sources and external sources, additional processes to clean and transform it before loading it to the data warehouse. That process is performed using ETL tools, and is discussed in the following subsection.

### 2.1.2 Extract, Transform and Load

Extract, Transform, and Load (ETL) is a process that is used to integrate data from various sources into a central location, such as a data warehouse. The term refers to the three phases of the process: **extraction**, **transformation**, and **loading**. This essential process helps organisations to get a complete view of their data, which in turn helps them to make informed decisions (Chaudhuri et al., 2011, p. 96).

The first step is accessing data from various sources and bringing it into the ETL system. The extraction process can either be full or incremental (Kimball & Ross, 2013, pp. 19–20). Full extraction involves extracting all the data from the source system and loading it into the data warehouse. Incremental extraction involves extracting only the new or modified data from the source system since the last extraction (Kimball & Ross, 2013, p. 512). The extraction process can be automated or manual, depending on the complexity of the system. Some sources of data may require complex queries to extract the desired data, while others may be simple and straightforward.

After the data has been extracted, it needs to be transformed to fit the data model of the data warehouse. This is the transformation phase of the ETL process (Kimball & Ross, 2013, pp. 19–20). There are several steps involved in the transformation phase, including data cleansing, data validation, data aggregation, and data integration (Han et al., 2012, pp. 84–85). Data cleansing involves removing duplicates, correcting errors, and filling in missing data (Han et al., 2012, p. 88). Data validation ensures that data are accurate, complete, and consistent (Han et al., 2012, p. 94). Data aggregation combines data from multiple sources into a single dataset (Han et al., 2012, p. 99). Data integration involves merging data from different sources and removing any inconsistencies (Han et al., 2012, p. 111). Transformations can be simple, such as changing the format of a date field, or complex, such as joining data from multiple sources to create a single dataset.

After the data has been extracted and transformed, it is then loaded into the target system, typically a data warehouse. This is the loading phase of the ETL process (Kimball & Ross, 2013, pp. 19–20). The data warehouse consists of several layers, including the staging area, the integration layer, and the presentation layer (Oracle, 2023). The staging area is where the data are first loaded. It is a temporary storage area where data are stored before it's transformed and loaded into the target system. The integration layer is where the data from multiple sources are integrated and transformed. The presentation layer is where the data are presented to the end-users in the desired format. The loading process can be automated or manual. Automation is preferred as it reduces the chances of errors and saves time (Inmon, 2002, p. 19). Automated loading processes can be scheduled to run at regular intervals so that the data warehouse is always up-to-date.

### 2.1.3 Challenges in Big Data Implementations

Big Data implementations are characterized by various challenges that organisations must address to achieve their desired objectives effectively. As the volume, variety, velocity, and veracity of data continue to grow exponentially, developers face enormous challenges in capturing, storing, processing, analyzing, and managing massive amounts of information (Abualkishik, 2019). Although Big Data presents opportunities to make data-driven decisions that enhance performance and create value, the challenges it imposes are equally critical to address (Lalmi & Adala, 2021). Several challenges include ensuring processing speed and scalability, data quality, data security and privacy, and costs (Gkoulalas-Divanis & Bettini, 2018; Gorelik, 2019; Rai, 2022, p. 7). This, choosing the right Big Data platform is a critical decision for stakeholders and developers (Gorelik, 2019, pp. 7–12). This section presents a brief overview of these challenges, and discusses how they can be addressed.

The large volume of data that are generated by organisations and individuals is the most significant challenge for Big Data systems. Institutions should ensure that they have enough storage and processing power to manage the growing data volumes (Gavin, 2019, p. 3). Furthermore, the velocity of data generation is also a challenge for Big Data systems. Sources like sensors and social networks can generate data at a high rate and must be processed and analysed in real -time (Luraschi et al., 2918). However, the faster the data are generated, the shorter the time window available for analysis. This presents a challenge for enterprises that require real-time processing of data to make fast decisions. The storage systems used should also be scalable, able to tackle growing volumes of data as needed (Gorelik, 2019, p. 87). Traditional relational databases are thus not suitable for handling large volumes of unstructured data generated by data producers. Instead, institutions should adopt advanced data storage and processing technologies, such as NoSQL databases and data streaming tools that allow for faster processing of data in near real-time (Maas & Garillot, 2019). A review on existing software tools for real-time information processing is presented in Section 2.2.

Data privacy and security are also concerns in Big Data implementations. Big Data may include sensitive information that requires protection from unauthorized access or disclosure (Kale et al., 2020; Mavriki & Karyda, 2020; Raja et al., 2020). Moreover, data privacy regulations require that organisations adhere to data protection laws such as GDPR (Kunigk et al., 2018, p. 511). Addressing data privacy and security challenges requires investments in data encryption, access controls, and firewalls to ensure that data are secured from unauthorized access or disclosure (Gorelik, 2019, p. 58).

Big data implementations require robust technical infrastructures that can handle large, complex data sets. Moreover, technical infrastructures require maintenance, upgrades, and scalability to accommodate growing data needs (Kimball & Ross, 2013). Addressing infrastructure challenges requires investments in cloud technologies, distributed computing, and software-defined infrastructure that can support big data processing.

Finally, Big Data implementations require large investments in technology, infrastructure, and human resources. Moreover, costs may increase due to data storage, hardware, software, and maintenance. Addressing cost challenges requires prioritizing expenditure on critical aspects of Big Data implementations, such as data quality, security, and analysis, while limiting unnecessary expenses (Verma & Bhatt, 2021).

## 2.2 Software Tools for Real-Time Information Processing

The challenge of processing complex events in real-time in order to improve decision-making processes can be addressed by utilizing applications designed for Complex Event Processing (CEP) (Sousa et al., 2021). These applications have the ability to analyze massive data sets and extract useful information, making it easier to develop decisions. This section will thus analyze different technologies available for CEP in real-time and their different characteristics.

To identify the most appropriate tools for this review, selection criteria have been established in the form of inclusion and exclusion criteria (Haddaway & Macura, 2018; Wilson, 1994). Inclusion criteria include features like the ability to work in real-time, the capacity to handle complex event processing, and integrations with popular data sources (e.g. Microsoft Excel, Microsoft SQL Server, Oracle and MySQL). Exclusion criteria include availability, support, cost, licensing and scalability.

### 2.2.1 Apache Haddop

Multiple technological solutions are available, with different approaches to processing and storing extremely large quantities of data. The most impactful solutions are offered by the Apache Software Foundation, an organisation that provides open-source software, by supporting software project communities and limiting potential legal exposure for their project committers.

The most popular framework for Big Data processing and storage is Apache Hadoop. Hadoop is an open-source project by the Apache Software Foundation, with the aim of creating a computing environment built on top of a distributed file system, allowing for large-scale data operations (Zikopoulos et al., 2011, p. 54). It is comprised of two main components, the Hadoop Distributed Filesystem (HDFS) and MapReduce.

The HDFS is a data storage and data processing filesystem designed to store and provide access to large amounts of data quickly. HDFS is able to spread files across a cluster of nodes, allowing for redundancy in case of a server failure (Vohra, 2016b, p. 3).

MapReduce is a programming paradigm originally developed by Google for performing web page searches across thousands of physically separated machines (Kimball & Ross, 2013, p. 530). This paradigm is comprised of two main tasks, namely the map phase and the reduce phase. The map phase uses one or more mappers to process the input data, while the reduce phase uses zero or more reducers to process the data output during the map phase (Vohra, 2016b, p. 5). Hadoop uses a Java implementation of the MapReduce paradigm, enabling scalability across servers in a cluster.

The Apache Hadoop ecosystem is comprised of several projects providing additional functionalities from querying and indexing to streaming and messaging, among more (Vohra, 2016b, p. 3). These include:

- **Apache Ambari**, a suite of tools for provisioning, managing, and monitoring Apache Hadoop clusters
- **Apache Avro**, a data serialization system
- **Apache Cassandra** and **Apache HBase** as databases
- **Apache Chukwa**, a data collection system for monitoring large distributed systems
- **Apache Hive**, providing SQL-like queries for data aggregation and analysis
- **Apache Oozie**, a workflow scheduler system to manage Apache Hadoop jobs
- **Apache Flume**, a data warehouse for managing large datasets stored in HDFS
- **Apache Mahout** for creating scalable machine learning applications
- **Apache Pig**, a high-level Hadoop programming language

- **Apache ZooKeeper**, providing coordination services for distributed applications

The architecture of the Apache Hadoop ecosystem was represented by (Abualkishik, 2019) in Figure 3.

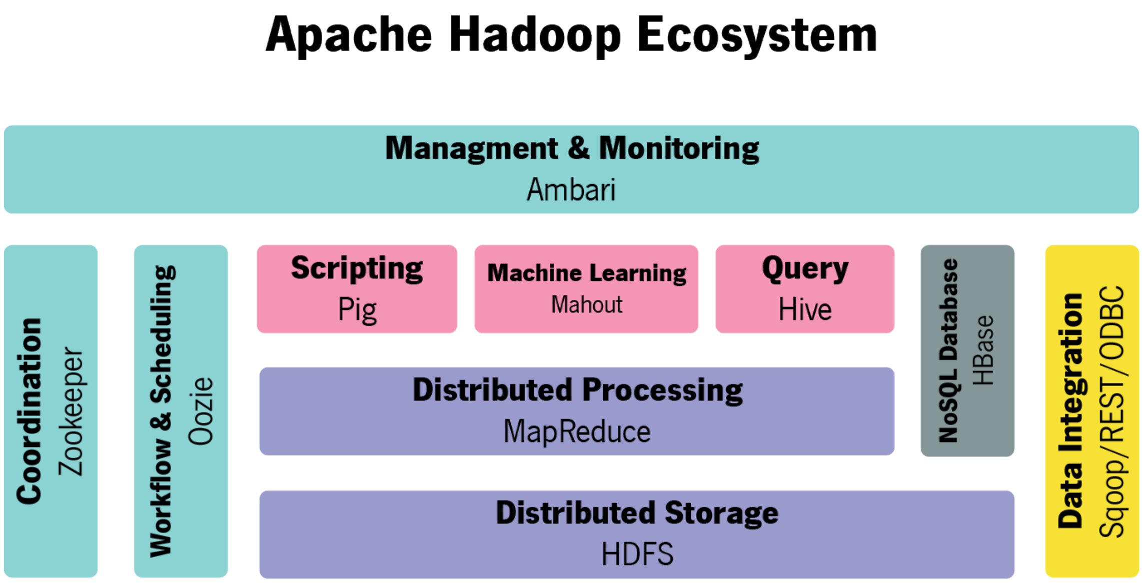


Figure 3: Apache Hadoop ecosystem (Abualkishik, 2019)

While Hadoop's role on Big Data is critical, more sophisticated projects have been arisen, with different approaches and goals. Furthermore, there's an increasing need to support BI tasks in real-time, in other words, "*make business decisions based on the operational data itself*" (Chaudhuri et al., 2011, p. 90). In the following sections, we will discuss some of the most relevant projects in the Apache Software Foundation ecosystem, including Apache Spark, Apache Kafka, and Apache Flume.

### 2.2.2 Apache Flume

#### 2.2.2.1 Concept

Apache Flume is a framework designed to stream data and collect, aggregate, and move large amounts of data in a distributed environment (Foundation, 2020b). It was initially developed by Cloudera, then acquired by the Apache Software Foundation in 2012 (Hoffman, 2013). Its primary purpose is to ingest event data into HDFS (Hadoop Distributed File System) in a simple and automated manner (Vohra, 2016a). Flume also supports various data reading mechanisms, such as Avro, Thrift, and Syslog (Foundation, 2020a).

#### 2.2.2.2 Components

Apache Flume features three base components (Source, Channel, and Sink) and three additional components (Interceptors, Channel Selectors, and Sink Processors). Source is responsible for data entry, accepting only unstructured events. Channel acts as an intermediary and stores data temporarily before it is sent to the destination (e.g. HDFS). Sink sends the data to its destination. Interceptors modify or inspect events as they are transferred between Source and Channel. Channel Selectors determine which Channel will be used when multiple are available. Lastly, Sink Processors invoke a particular sink from a group of sinks, creating fail-over paths, and providing load balancing.

Figure 4 illustrates the architecture when all components are in use.

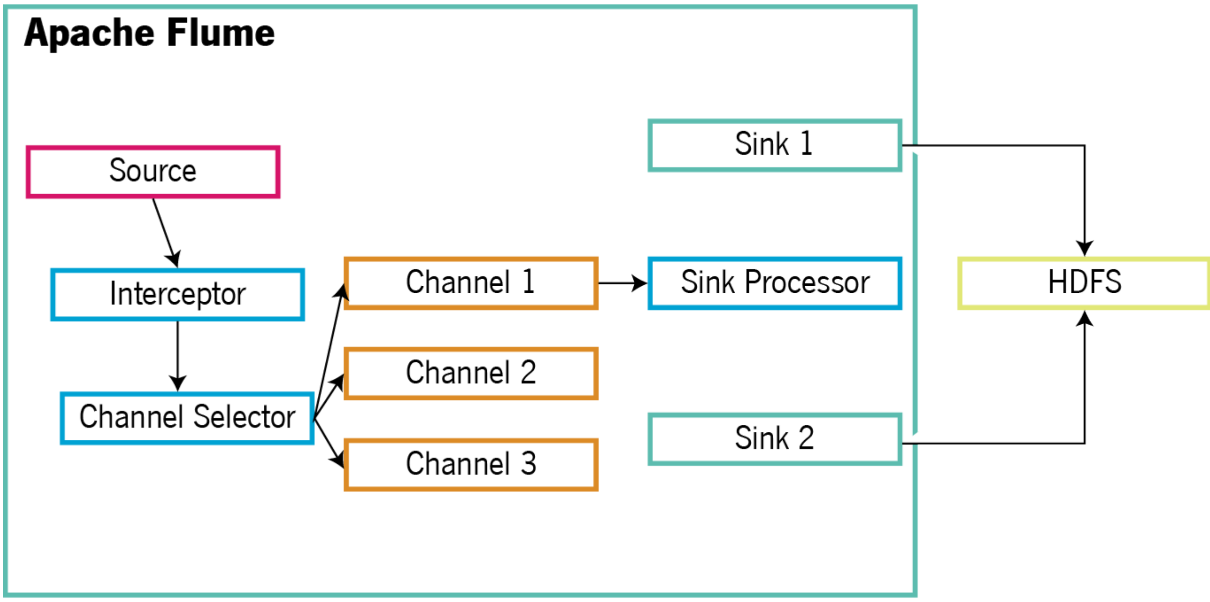


Figure 4: Flume architecture. Adapted from (Foundation, 2020a)

#### 2.2.2.3 Advantages and Disadvantages

The advantages of Apache Flume are numerous (John & Misram, 2017). Some main ones include: it is open-source and based on streaming data flows; it can ingest unstructured data; it is highly extensible, available, and horizontally scalable; it offers reliable message delivery; it is economical to install and operate; it has a declarative configuration; and it provides complete documentation with many examples.

Disadvantages include not guaranteeing end-to-end reliability or 100% unique messages; scalability is limited; the throughput depends on the backing store; and topology reconfiguration is challenging (EDUCBA, 2020).

## 2.2.3 Apache Sqoop

### 2.2.3.1 Concept

Sqoop is a tool for transferring data between relational databases and Hadoop. It is developed by the Apache Software Foundation (The Apache Software Foundation, 2019). It is designed to import and export data between RDBMSs and the Hadoop ecosystem (Arvind, 2012).

It has two major components: connectors and drivers. Connectors allow Sqoop to bridge the differences between SQL dialects and optimize data transfer. Drivers refer to the Java Database Connectivity (JDBC) Driver, the standard Java API for accessing RDBMSs and some data warehouses (Vohra, 2016c).

Figure 5 shows the basic architecture for Apache Sqoop.

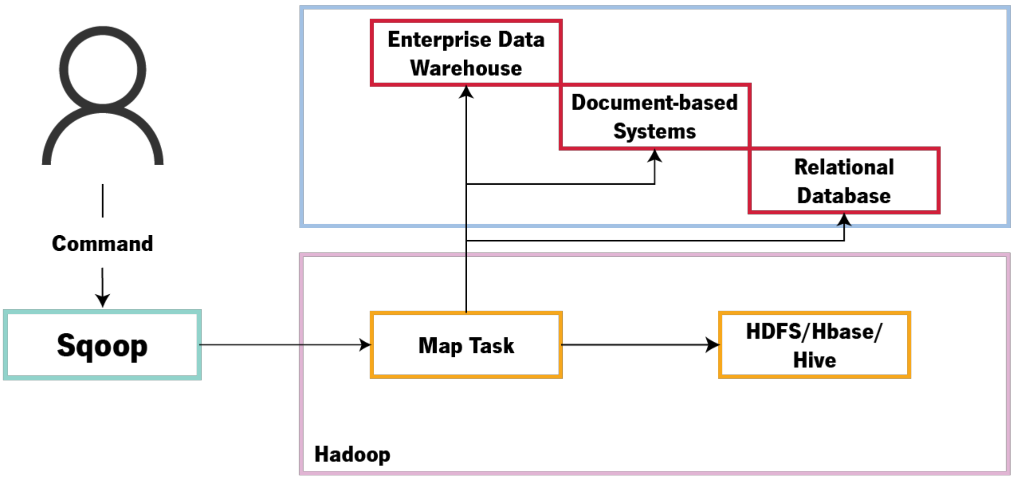


Figure 5: Sqoop architecture

Apache Sqoop only makes sense when used in the Hadoop ecosystem. Its advantages include full or incremental load, parallel import/export, compatibility with SQL queries, data compression, Kerberos security integration, loading data directly into HIVE/HBase, automated processes, robustness, and regular updates.

Disadvantages include its reliance on Java Database Connectivity (JDBC) connections, lack of pause/resume capability, export performance depending on hardware configuration, slow speed due to MapReduce, lack of GUI, and lack of support for certain databases.

## 2.2.4 Solace PubSub+

Solace Systems, now known simply as Solace, is a Canadian middleware company that manufactures and sells message-oriented devices and software for routing information. It was founded in 2001 by Craig Betts with the goal of incorporating messaging systems into hardware (Corporation, 2020).

PubSub+ is the company's primary tool, which is a unified message broker that supports publishing/subscribing, queuing, requesting/responding, and streaming using open APIs and protocols through hybrid cloud and IoT environments (Solace, 2021b).

PubSub+ is composed of five layers. The first layer is the PubSub+ Event Broker, which is responsible for loading an event mesh that allows for the sharing of events across environments and components of the distribution company (Solace, 2021a). This layer is broken down into three sub-layers: PubSub+ Event Broker-Software, PubSub+ Event Broker-Appliance, and PubSub+ Event Broker-Cloud. The software version allows for the use of various communication protocols, such as AMQP, JMS, MQTT, REST, and WebSocket; the appliance version is designed with high-speed FPGAs and network processors that enable low and predictable latency; and the cloud version allows for the quick implementation of software event brokers.

The second layer is PubSub+ Event Mesh, which is responsible for dynamically routing events from one application to another. The third layer is PubSub+ Streaming APIs and Integrations, which provides on and off-ramps and proprietary APIs for messaging. The fourth layer is PubSub+ Event Portal, which is an event managing tool that allows for the discovery, construction, visualization, sharing, and management of aspects of the Event-Driven Architecture. Finally, the fifth layer is PubSub+ Platform Security, which enables message architectures that obtain consistent multi-protocol authentication of the client, as well as security clearance management.

These layers are organized as illustrated in Figure 6.

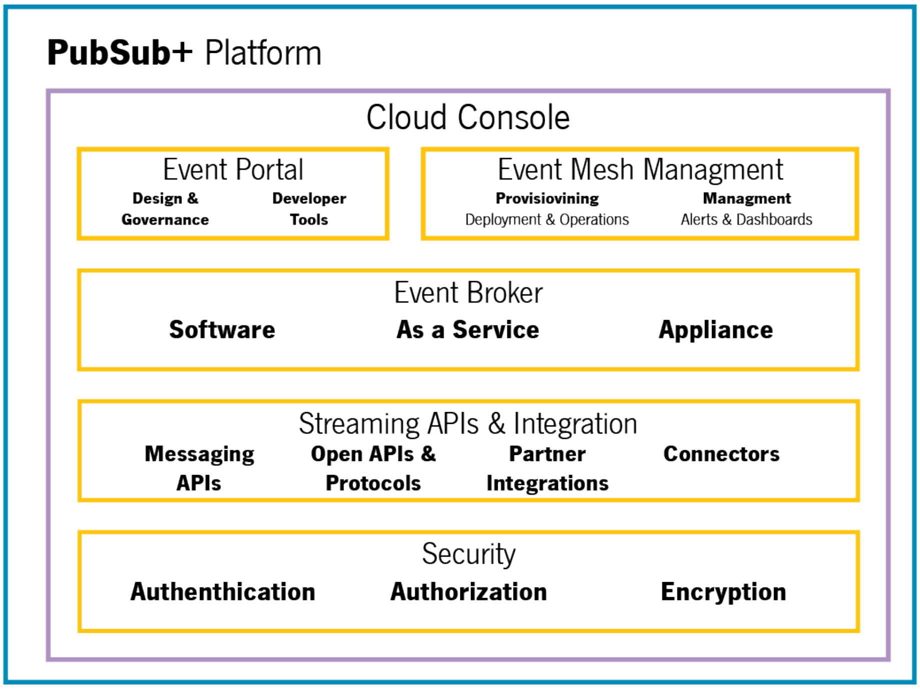


Figure 6: Layout of the internal layers of PubSub+. Adapted from (Solace, 2021b)

In a technical context, a message is information that can be shared between computer systems without the need for direct links or knowledge of each other's location. The components required for sending messages include a publisher (also known as a producer), the message itself, a subscriber (also known as a consumer), a topic (where the message is intended to be consumed by more than one subscriber), and a queue (where the message is intended to be consumed by at most one subscriber). These components are organized as seen in Figure 6.

Solace PubSub+ has many advantages: easy JavaScript development with existing Solace JavaScript API; all microservices and web clients use one IP address; no specific development is needed for messaging, web connectivity, authentication, entitlement or message recording; development efforts focus on main tasks; short and easy projects with low risk, low cost, and high transparency; no specific development of web services needed; high fanout; guaranteed messaging; low latency; WAN optimization; high-speed caching; content routing and transformation; legacy integration and migration; one API; and one management environment. Disadvantages include poor technical support and support for only a few message VPNs (5 VPNs).

### 2.2.5 Apache Kafka

Apache Kafka is an open-source platform developed in Java and Scala for distributed streaming processing. It was initially created by LinkedIn to handle large amounts of data with low latency and process data in real-time. Afterward, it was donated to the Apache Software Foundation, its current owner.

Kafka is a scalable, fault-tolerant, and low-latency event streaming platform that can process millions of events per second, and can be used for distributed streaming, event-driven applications, big data ingestion pipelines, and pub/sub messaging. It's especially suitable for complex event processing needs since messages are published directly to the distributed event log, with no need for centralized control. Moreover, different CEP services can subscribe to the same data streams without blocking each other.

Kafka runs as a cluster and can contain many data centers, allowing for efficient communication between data producers and consumers using message-based topics. Each topic consists of a key, a value, and a date/time stamp. The platform has four core APIs: Producer API, Consumer API, Streams API, and Connector API. With the Producer API, an application can publish a stream of records to one or more Kafka topics. The Consumer API enables an application to subscribe to one or more Kafka topics and process streams of records produced to those topics. The Streams API allows an application to process data in Kafka using a stream processing paradigm, while the Connector API connects applications or data systems to Kafka topics.

Figure 7 shows the basic architecture of Kafka.

Apache Kafka has 8 components: topics, producers, consumers, brokers, partitions, replicas, leaders, and followers. Topics are categories that records are published to and producers are responsible for deciding which record goes to which partition. Consumers join consumer groups and each record published to a topic is delivered to one consumer instance in each consumer group. Brokers are servers that

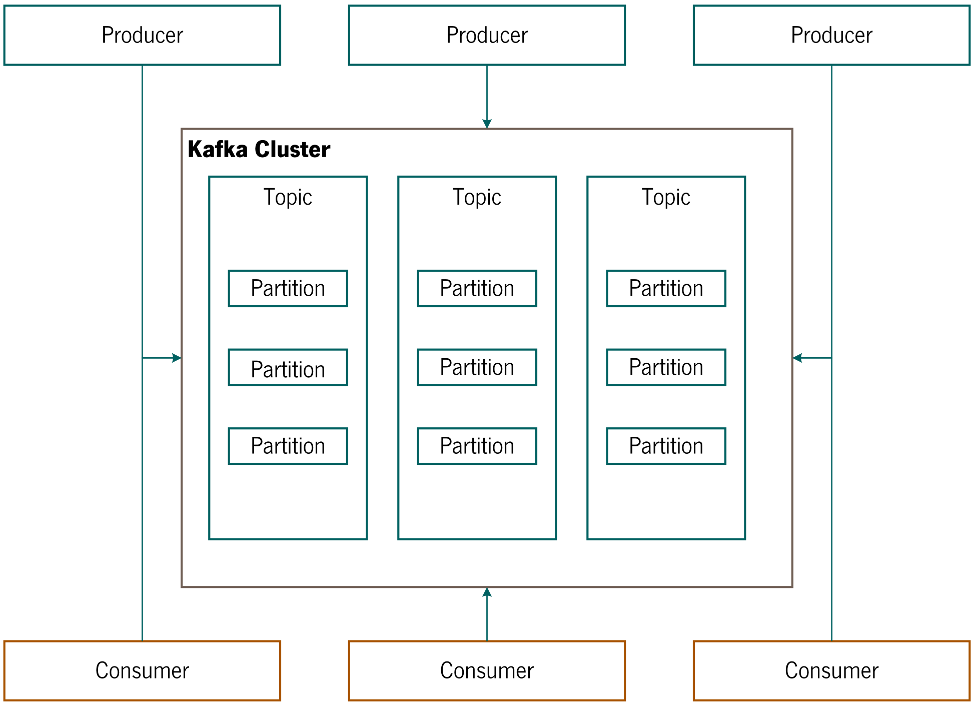


Figure 7: Apache Kafka basic architecture

manage Kafka clusters and each topic is divided into partitions which are replicated across brokers. Replication is important for designing resilient and highly available deployments and leaders are responsible for ensuring that followers have the same records.

Figure 8 provides a simplified view of these components and their interrelations.

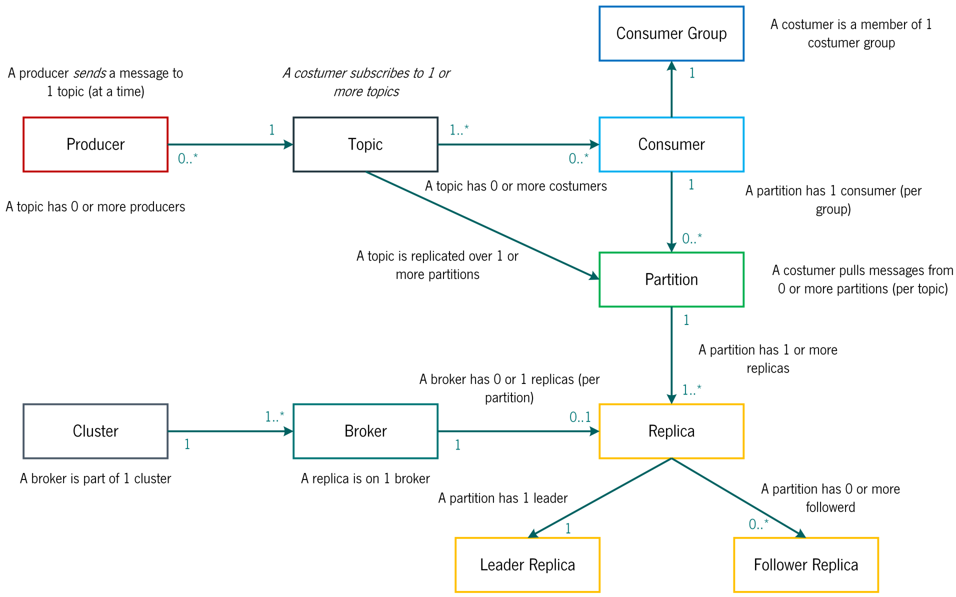


Figure 8: Relation between Kafka components

Apache Kafka offers many advantages, such as allowing senders and receivers to operate independently, supporting programming languages such as Java and Python, employing data encryption, and allowing messages and data to be integrated into applications using APIs. It also has features such as replication and partitioning, making it scalable and resistant to node failure.

However, Kafka lacks a full set of management and monitoring tools, and it does not support wildcard topic selection, making it unsuitable for certain use cases. Additionally, when the number of queues increases, performance may suffer, and certain messaging paradigms, such as request/reply, are not supported.

## 2.2.6 Apache Spark

### 2.2.6.1 Concept

Apache Spark is a powerful Big Data tool for processing large data sets in a parallel and distributed manner. It extends the MapReduce programming model made popular by Apache Hadoop, allowing for the development of large data processing applications. Compared to Hadoop, Spark is much faster, achieving performance increases of up to 100x in some cases (The Apache Software Foundation, 2021).

A great advantage of Spark is that all its components are integrated into the tool itself, such as Spark Streaming, Spark SQL, and GraphX, whereas Hadoop requires additional tools like Apache Hive to run properly. In addition, it supports programming in Java, Scala, and Python (Frampton, 2015).

Spark Core is the main component, providing basic processing functions like map, reduce, filter, and collect. Other components include Spark SQL, Spark Streaming, MLlib, GraphX, and SparkR (see Figure 9):

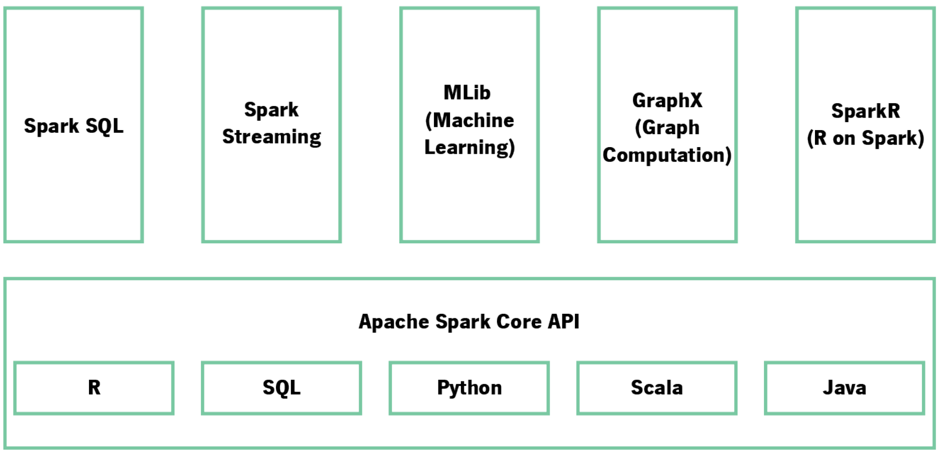


Figure 9: Apache Kafka main components

Apache Spark is a unified analytics engine that has several components. At its core, Spark is embedded with a special collection called RDD (resilient distributed dataset). RDD is an abstraction of Spark which handles partitioning data across nodes in a cluster and holds them in the memory pool as a single unit. Two operations are carried out on RDDs: Transformation and Actions.

Spark SQL provides the ability to access Spark data sets through a Java Database Connectivity (JDBC) API, allowing users to execute SQL queries and use traditional BI and visualization tools. Spark Streaming is used to process streaming data in real-time, based on micro-batch computing.

MLlib is Apache Spark's scalable machine learning library, offering algorithms and pipelines. GraphX is a component of Spark that is specifically used for graphs and graph-parallel computation. It reuses the Spark RDD concept and simplifies graph analysis with properties attached to each vertex and edge. Finally, SparkR is an R package that provides a light-weight frontend to use Apache Spark from R, and also supports distributed machine learning using MLlib.

Apache Spark offers several advantages, including in-memory computing, compatibility with Hadoop and its ecosystem, real-time stream processing, cost efficiency, speed, ease of use, advanced analytics, dynamic nature, multilingual capabilities, increased access to big data, and an open-source community (KnowledgeHut, 2019).

Some drawbacks of Apache Spark are its lack of automatic optimization process, issues with file management systems, fewer algorithms, small files problem, window criteria, and lack of suitability for multi-user environments (Frampton, 2015).

### 2.2.7 Microsoft Power BI

Microsoft Power BI was launched on July 24, 2015. It's a set of software services, applications, and connectors that together transform raw data into visually appealing, interactive, and meaningful information. This data can come from various sources with different formats, either hosted locally or in the cloud. Table 1 lists the highlighted data sources supported by Power BI.

Table 1: Power BI data sources

| | |
|---|---|
| Excel | Text/CSV |
| XML | JSON |
| Oracle Database | IBM DB2 Database |
| MySQL Database | PostgresSQL Database |
| Sybase Database | Teradata Database |
| SAP HANA Database | SAP Business Warehouse server |
| Azure SQL Database | Google Analytics |
| Facebook | Github |

The biggest advantage of Power BI is the ability to provide a personalized interface for each user, with different parameters according to their position in the company or the ability to share data between employees. The created dashboards and reports link to datasets that collect all relevant data in one place.

Figure 10 shows the basic architecture of this tool's functions.

Microsoft Power BI is composed of several parts, each of which is essential to the platform. These components can be used individually or combined. The main components of Power BI are Power Query (a self-service ETL tool), Power Pivot (an in-memory data modeling component for fast aggregation and

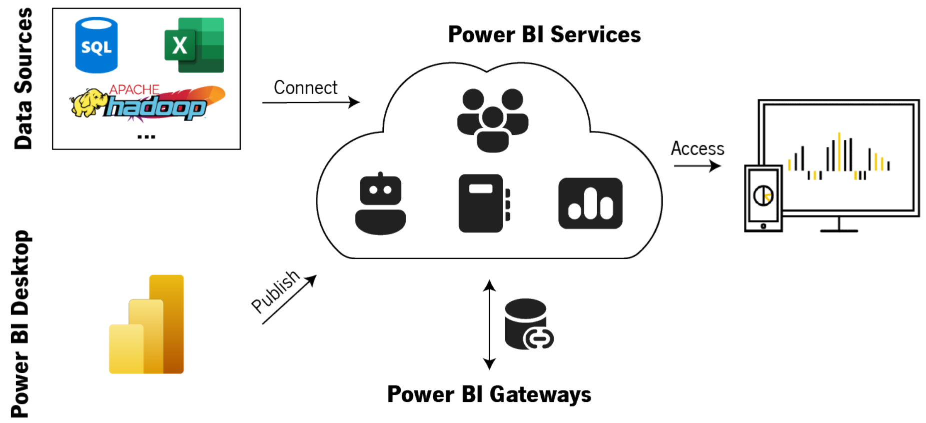


Figure 10: Power BI architecture. Adapted from (Technologies, 2018)

calculation), Power View (an interactive visualization tool in Excel workbooks), Power Map (a 3D data visualization tool in Excel), Power Q&A (which recognizes words to find answers), Power BI Desktop (a report-creating tool with drag-and-drop features), and Power BI Website (which allows users to share reports and dashboards). Additionally, Power BI has mobile apps for Android and iOS to view reports and dashboards. Combining these components yields great benefits.

Microsoft Power BI has a range of attractive visualizations, allows for data retrieval from multiple sources, provides customizable dashboards and tiles, as well as informative reports and DAX data analysis. Additionally, it provides helpful tools such as ad-hoc reporting and feedback, and trend indicators. Power BI Desktop is free, custom visualizations are available on the Microsoft Marketplace, integration with Excel is enabled, and users have access to rich personalized dashboards with regular updates. There are no memory and speed constraints, and it can extract intelligence from data.

Power BI is not ideal for handling complex relationships between tables, custom visuals are not configurable, the user interface can be crowded, formulas may be rigid, and it has a limited capacity for large data volumes. Futhermore, it does not provide data cleansing solutions and can be difficult to understand and master.

### 2.2.8 Tableau

Tableau is a software dedicated to providing information and creating indicators to help with decision-making. It was created in 2003 by Professor Pat Hanrahan and his PhD student Chris Stolte in their computer department. It is a business intelligence tool that has won multiple awards, including "Best Overall in Data Visualization" by DM Review, "Best of 2005 for Data Analysis" by PC Magazine, and "2008 Best Business Intelligence Solution (CODiEaward)" by the Software & Information Industry Association.

Tableau has several components, such as Tableau Desktop, Tableau Server, Tableau Online, Tableau

Public, and Tableau Reader. Tableau Desktop is a business intelligence application that offers analysis with a simple drag workflow. Tableau Server is an online and mobile business intelligence solution used to publish dashboards with Tableau Desktop and share them throughout an organization. Tableau Online is a hosted version of Tableau Server with storage of 100 Gigabytes for the commercial license. Tableau Public is for people who want to tell interactive data stories on the web. Tableau Reader is a free desktop application to open and view visualizations built in Tableau Desktop.

Tableau has many advantages, such as remarkable visualization capabilities, ease of use, high performance, multiple data source connections, thriving community and forum, easy online access, no need for technical or programming knowledge, and quick response for making a dashboard. Its disadvantages include high cost, poor after-sales support, poor BI capabilities, and time and resource-intensive staff training.

### 2.2.9 Discussion and Conclusions

Complex Event Processing (CEP), also known as event or stream processing, is the use of technology to query data before or without ever storing it in a database (Databricks, 2020). CEP systems address the "velocity problem" of big data, where data is a stream of predefined events. The sliding window approach of CEP systems keeps only a portion of actual data in the main memory, while discarding or archiving old events (JBoss Community Documentation, 2020). This allows for efficient analysis of the most recent events, without having to fit all data into system memory (Liu et al., 2015).

CEP is used in various applications such as business activity monitoring, network sensors that monitor industrial facilities, and market data analysis (AIMultiple, 2020). It is mainly used for low latency requirements, high input event volumes, and complex event patterns such as those based on temporal or spatial relationships (Databricks, 2020).

Numerous tools for CEP have been introduced and showcased in this section, including Apache Spark and Apache Kafka for streaming processing, and Microsoft Power BI and Tableau Suite for data transformation. Other solutions include Amazon Kinesis Analytics, Apache Flink, Apache Samza, Apache Storm, Hadoop, Microsoft Azure Stream Analytics, and Oracle Stream Analytics (Databricks, 2020).

## 2.3 Crowdsensing on Smart Cities

The emergence of the internet and the Internet of Things (IoT) has led to the development of Smart Cities, which leverage connected devices and data to enhance the well-being of their residents. This paradigm shift has also brought about the idea of crowdsensing, allowing individuals to actively participate in various initiatives, while geofencing enables location-based notifications to be sent to citizens when they're in proximity to areas of interest. These technological advancements haven't only evolved to focus on technical aspects but also on their potential to enhance citizens' quality of life. Consequently, this section aims to conduct a comprehensive review of existing literature, seeking to identify past studies that have explored the implementation of crowdsensing and geofencing technologies in the context of Smart Cities.

The primary objectives of this review are twofold: first, to understand the underlying purpose of incorporating geofences and location-based notifications in smart cities, and second, to examine the strategies and tools utilized to integrate these technologies. To achieve these objectives, a rigorous search process was undertaken, employing the Scopus citation and abstract database. Through this meticulous approach, seventeen papers were selected as pertinent to the discussion and subsequently analyzed.

By conducting this review, the intention is to uncover valuable insights into crowdsensing and geofencing technologies in the context of smart cities. The findings from this study will contribute to a deeper understanding of the potential benefits and challenges associated with these technologies, ultimately helping in the development of more efficient and effective smart city initiatives.

### 2.3.1 Introduction

With the continuous growth in urban population and the increasing need for essential community services, the modernization of cities and their infrastructure has become crucial (Kirimtat et al., 2020). Consequently, the concept of Smart Cities has emerged, integrating technological advancements and infrastructure to enhance the quality of life in a sustainable and transparent manner. Smart Cities place particular importance on the domains of Internet of Things and Internet of People due to the constant evolution of related technologies and the growing interest in these areas worldwide, not only among the academic community but also among industry and civil society (Nižetić et al., 2020).

The study and development of Smart Cities has garnered significant interest for several years, although initial research primarily focused on technological concepts. More recently, attention has shifted towards a holistic perspective on the associated information systems, emphasizing the quality of life for citizens and the impact of smart technologies on their lives, as well as the social, economic, and environmental sustainability of cities (Ismagilova et al., 2019).

One area of notable development within Smart Cities is crowdsensing. Crowdsensing leverages personal mobile devices and their sensors to collect and process information in a distributed and collaborative manner, without imposing more burdens on management or administrative entities.

The field of crowdsensing is diverse and has undergone significant evolution, with approaches ranging from collecting sensory data through users' devices, where citizens contribute to the collective, to providing tailored intelligent information to individual citizens, allowing them to benefit from the collective.

In the context of citizens benefiting from the collective, recent works and developments have focused on the utilization of geofences, which are virtual perimeters defined for specific geographic regions. Geofences enable devices and applications to provide citizens with relevant information when they are in proximity to areas of interest (Google, 2022). Incorporating geofences to deliver personalized information sensitive to each citizen's context, either through mobile applications or notifications, serves as a goal for Smart Cities' development and implementation, as well as for expanding the user base of crowdsensing.

Therefore, the aim of this section is to explore recent scientific publications that present a comprehensive overview of the current state of development regarding the application of geofencing and smart notification in Smart Cities. The research follows the principles of a systematic review and addresses a main research question: **What is the role and purpose of crowdsensing and crowdsourcing, specifically geofences and location-based notifications, in smart cities?**

### 2.3.2 Methodology

This review adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement and checklist, chosen for its wide acceptance within the scientific community in the computer science and engineering field. Following the guidelines provided by PRISMA, the research approach encompasses the following steps (Page et al., 2021a):

1. Identification of research questions and relevant keywords.
2. Creation of a research query.
3. Definition of eligibility criteria to filter and narrow down the articles sample.
4. Analysis of the resulting set of studies and papers.
5. Presentation and discussion of the obtained results.

The initial phase of the research was conducted on April 25, 2022, utilizing Scopus as the data source. Scopus was chosen due to its comprehensive coverage of publications, quality assurance, and large repository of scholarly materials.

To conduct the bibliographic research, a range of keywords was identified as a starting point. These keywords were applied in the fields of title, abstract, and keywords. To facilitate search organization, the identified keywords were divided into two groups, each combined by a conjunction method. Within each group, the keywords were combined by disjunction. This approach ensures that all documents containing at least one keyword from each group are included in the results.

The first keyword group pertains to areas and technical subjects directly aligned with the research topic, such as Smart Cities, Crowdsensing, Smart Notifications, and Location-based Notifications. The second keyword group aims to filter the broader scope of technological aspects, including Information Communication Technology, Information Systems, and Mobile Computing. This grouping strategy helps focus the results within the context of information systems and agents.

By implementing the aforementioned search strategy, the following research query was formulated:

( TITLE-ABS-KEY ( smart AND cities ) OR TITLE-ABS-KEY ( crowdsensing ) OR TITLE-ABS-KEY ( smart AND "notifications" ) OR TITLE-ABS-KEY ( location AND based AND "notifications" ) AND ( TITLE-ABS-KEY ( information AND communication AND technology ) OR TITLE-ABS-KEY ( information AND systems ) OR TITLE-ABS-KEY ( mobile AND computing ) ))

This research employed a set of exclusion criteria to effectively screen the collected articles and studies for eligibility. Any document that met any of these criteria was excluded from the analysis, ensuring the reliability and relevance of the final dataset. The exclusion criteria were as follows:

1. Publications that were not accessible in Open Access.
2. Publications that were not produced within the last five years (from 2018) or have not yet been fully published.
3. Publications that were not related to the fields of Computer Science or Engineering.
4. Publications that were not categorized as Articles or Reviews/Surveys and were not written in English.
5. Publications that did not primarily focus on the variables studied or were considered irrelevant to the research context.
6. Publications that were not in compliance with the European Union's General Data Protection Regulation (GDPR).

These exclusion criteria were carefully selected to ensure that only high-quality, recent, relevant, and legally compliant publications were included for analysis. By adhering to these criteria, the study aimed to maintain a rigorous and comprehensive approach to its literature review.

### 2.3.3 Results

The initial literature review encompassed a comprehensive search across various bibliographic databases, yielding a total of 14,840 relevant studies. Subsequently, these studies were subjected to a meticulous evaluation against predetermined eligibility criteria. The initial five criteria were applied utilizing the advanced filtering capabilities of Scopus. Thus, a subset of articles and reviews about the domains of engineering and computer science, which satisfied the following extra criteria: namely open accessibility, English language publication, and citation or authorship in the past five years, was derived.

Out of the resulting subset comprising 575 studies, a subset of 526 was deemed unsuitable for inclusion due to failure to follow the data extraction, manipulation, privacy, and data protection standards mandated by the European Union.

The remaining 49 studies were subjected to a thorough and comprehensive analysis, entailing a detailed reading of each publication. Only 17 studies were deemed pertinent to the topic of crowdsensing and crowdsourcing applications in smart cities. Specifically, these studies focused on the utilization of geofences and location-based notifications, excluding those developed within the confines of enterprise environments. As a consequence, the remaining 32 studies were deemed irrelevant and therefore excluded from the review.

For improved clarity on the overview of the research selection process, Figure 11 presents a PRISMA flowchart delineating the various stages of study screening and selection.

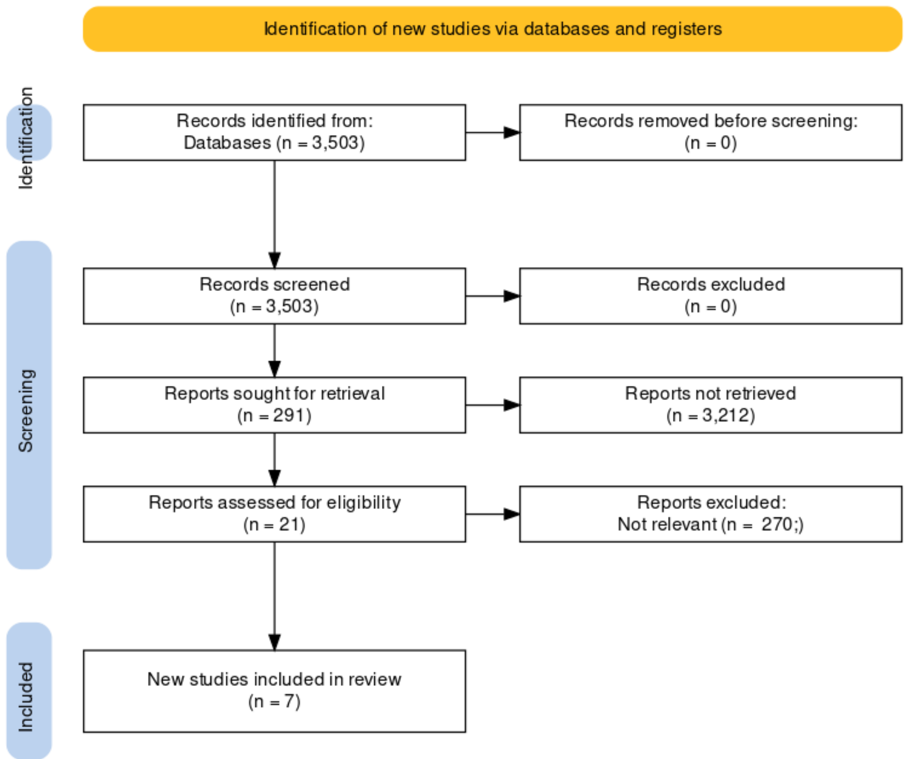


Figure 11: PRISMA flowchart - adaptation applied in this study (Page et al., 2021a)

Next, a summary of the results found in the review to answer the research questions will be presented.

In their comprehensive study, Ande et al. give a thorough overview and reference guide focusing on IoT systems, with particular emphasis on security concerns. The authors demonstrate that the Internet of Things has emerged as a result of advancements in computing, communication technologies, and the Internet, all intertwined with the innate human aspiration to enhance our quality of life (Ande et al., 2020).

Amaxilatis et al. contribute to the field by introducing a solution for deploying and managing crowdsensing campaigns and experiments across federated cities in the OrganiCity facility. Through their investigation, the authors conclude that the implemented system affords developers the ability to rectify their campaign design if they identify underperformance, thereby enabling strategic decision-making (Amaxilatis et al., 2018).

Cheng et al. propose a collaborative geofence site selection (CGSS) method aimed at assisting city management entities in designating suitable geofence sites for dock-less shared bikes within urban areas. In this context, geofences are employed to determine optimal locations for renting bicycles and scooters, ensuring maximum user satisfaction with service availability and comprehensive coverage for future expansions of the city (Cheng et al., 2019).

In their work, Pánek provides a comprehensive account of the developmental and implementation process of an emotional mapping approach, which serves as a valuable methodology for engaging individuals in map-based dialogues pertaining to the present and future conditions of public spaces, by solely relying on their personal experiences. Pánek concludes that emotional maps, when employed as a crowdsourcing tool, offer a user-friendly environment for social participation, while simultaneously fostering a sense of identity and belonging within specific social groups or communities (Pánek, 2019).

Building upon this notion, Fernandes et al. have developed a mobile application-based platform that enables users to obtain pertinent information concerning the existing and forthcoming state of urban areas, while also delivering intelligent notifications equipped with contextual relevance for every individual user. By leveraging the capabilities of a mobile application, the authors have designed a comprehensive solution that caters to the diverse needs and preferences of users seeking real-time updates and personalized information regarding their city's resources and services (Fernandes et al., 2020).

In a similar vein, Foschini et al. introduce ParticipAct, a pioneering mobile crowdsensing technology that harnesses the power of edge nodes to compute potentially hazardous scenarios within crowded environments. Central to this approach is the installation of a sensing client application on users' smartphones, which enables the collection and transmission of relevant data to a centralized cloud server. At the core of ParticipAct's functionality lies the notion of targeted sensing campaigns, which are formulated and implemented by researchers and platform administrators. By leveraging edge-enabled mobile crowdsensing techniques, ParticipAct presents a novel solution for identifying and addressing potential crowd-related risks in a timely manner (Foschini et al., 2021).

Kousiouris et al. conducted a study that used social network data to identify large crowd concentration (LCC) events. They integrated user data collected from a Smart Transportation platform with social network activity peaks to identify events that could potentially impact the user journey. The results of their research demonstrated that this approach could be particularly beneficial for caregivers of disabled users, enabling them to modify planned journeys to avoid congestion, confusion, and limited mobility circumstances (Kousiouris et al., 2018).

Roman et al. proposed a scalable and manageable mobile on-street parking spot identification system as an alternative to stationary sensor methods. The authors developed a supervised learning system that analyzed the structure of sonar traces to distinguish between parked automobiles and road clutter. The system transmitted this information to a central server, which subsequently generated a map of parking occupation. Notably, the authors emphasized the potential for this mobile sensor system to evolve into a crowdsourcing approach, where members of the public could voluntarily install sensors on their vehicles and contribute data. In return, participants could receive reduced parking costs or free access to parking apps as a form of compensation (Roman et al., 2018).

In a different domain, Picaut et al. proposed a novel crowdsourcing-based method for assessing noise environments. They developed a smartphone application and a data infrastructure specifically designed for gathering both physical and perceptual information about the sound environment. The data contributed by each user was used to populate a community database, enabling a more comprehensive representation of the sound environment in terms of spatial and temporal aspects compared to traditional numerical modeling methodologies. This crowdsourcing approach holds promise for enhancing our understanding of noise pollution and developing effective strategies for its mitigation (Picaut et al., 2019).

In their study, Hu et al. devised a three-stage Stackelberg game to address the issue of fluctuating participant numbers in the MCS context. To conceptualize the MCS scenario as a sensory data market, the players in the game were categorized into two groups: monthly-pay participants and instant-pay participants. This categorization ensures the continuous contribution of sensory data from participants who pay a monthly fee. Moreover, to preserve fairness within the sensory data market, the game incorporates a secure reward distribution mechanism supported by blockchain technology (Hu et al., 2020).

In their comprehensive analysis, Kirimtat et al. critically examined and summarized various smart city projects, with a specific focus on key concepts and strategies for data management. The authors noted a noticeable gap in the literature, specifically the lack of scientific reports on smart floating cities despite their potential as viable candidates for future smart cities. Based on their evaluation of current scientific studies, they concluded that further research is needed to shed more light on this emerging area (Kirimtat et al., 2020).

Pilloni provided a detailed overview of the fundamental aspects pertaining to Industry 4.0. In their exploration, they delved into the impact, advantages, and challenges that various technologies such as the Internet of Things (IoT), crowdsourcing and crowdsensing, and big data will have on industrial processes. By examining these technologies and their potential implications, the authors aimed to provide insights into the transformational effects that Industry 4.0 will bring to industrial contexts (Pilloni, 2018).

In their study, Yang et al. offer a perspective that presents crowdsensing as a technique wherein many individuals use mobile devices equipped with sensors to share sensory data for measuring, analyzing, or inferring issues of common interest. This approach highlights the potential of crowdsensing to gather and use data from a vast network of individuals (Yang et al., 2019).

Shahrour and Xie delve into the role of the Internet of Things (IoT) and crowdsourcing in the development of smart cities. They emphasize the significance of data in smart city solutions and explore how the IoT and mobile crowdsourcing contribute to various applications in smart cities. The authors argue that further collaboration is required between research focused on smart city technologies centered on the IoT and research centered on citizen participation through crowdsourcing in order to fully realize the potential of smart city development (Shahrour & Xie, 2021).

L. Wang and Alexander narrow their focus to the recruitment of citizens through social networks for crowdsensing projects. They recognize that citizen engagement is pivotal to the success of such projects and investigate strategies for effectively involving citizens in crowdsensing initiatives. This area of study highlights the importance of citizen participation and its impact on the outcomes and effectiveness of crowdsensing projects (L. Wang & Alexander, 2019).

Ismagilova et al. adopt an Information Systems perspective to discuss various aspects of smart cities, including smart mobility, smart living, smart environment, smart citizens, smart government, and smart architecture. They analyze existing research findings on these topics and shed light on the potential shortcomings and future possibilities within the field of smart cities. This comprehensive analysis contributes to a deeper understanding of the different dimensions and applications of smart cities (Ismagilova et al., 2019).

Nižetić et al. review and discuss the Internet of Things (IoT) and its potential, challenges, and concerns in the context of building smart and sustainable futures. The authors aim to enhance the understanding of both the technological advancements in IoT application domains and the environmental implications associated with the increasing use of IoT products. This review provides valuable insights into the current state and potential future directions of IoT technologies about sustainability and environmental impact (Nižetić et al., 2020).

### 2.3.4 Discussion

This subsection aims to analyze and expound upon the discoveries attained through the review of the literature.

The initial subsection elucidates the pivotal role and purpose of crowdsourcing, specifically regarding geofences and location-based notifications, within the context of smart cities. By outlining the fundamental principles and significance behind these mechanisms, a comprehensive understanding of their practical applications in urban environments is presented.

Moving forward, the following subsection elucidates the strategies and tools that have already been devised and utilized for the seamless integration and implementation of crowdsourcing mechanisms in the functioning of smart cities. By examining the methodologies adopted and the technological resources employed, an in-depth examination of the successful realization of crowdsourcing endeavors within urban contexts is expounded upon.

Lastly, the subsequent subsection focuses on the ways in which crowdsensing implementations can significantly enhance the lives of citizens residing in smart cities. This entails exploring the potential improvements in various aspects of citizens' lives, while simultaneously highlighting the necessary measures to actively engage citizens as contributors in the crowdsourcing process. Through this exploration, the potential benefits and citizen empowerment associated with crowdsensing initiatives within a smart city setting are brought to the forefront.

#### 2.3.4.1 Role and Purpose of Crowdsensing and Crowdsourcing

According to Pilloni, Mobile Crowdsencing (MCS) encompasses both mobile crowdsensing and mobile crowdsourcing (Pilloni, 2018). Mobile crowdsensing entails the use of mobile sensors to gather data from diverse sources, while mobile crowdsourcing requires users to fulfill specific tasks, often involving the provision of feedback. The integration of MCS has found extensive application in various domains such as smart cities, personal health care, and environmental monitoring (Hu et al., 2020).

Regarding its significance and purpose within smart cities, relevant literature suggests that crowdsensing plays a vital role in facilitating the collection and dissemination of a large volume of data. This data enables the tracking of citizen habits and movements within an urban environment (Amaxilatis et al., 2018; Foschini et al., 2021). Furthermore, crowdsourcing initiatives have been recognized for encouraging both active and passive participation from citizens in the collection of data pertaining to the functioning of the city, its amenities, quality of life, and the environment (Shahrour & Xie, 2021).

The development of mobile crowdsensing and crowdsourcing technologies offers a promising avenue for constructing economical and high-fidelity monitoring systems tailored for urban infrastructures, services, and the environment. This integration holds immense potential to hurry the implementation of smart city initiatives, while also facilitating the collection of valuable insights from citizens. By garnering the ideas and sentiments of individuals, these innovative approaches enable a comprehensive examination of the true impact of smart city projects on the quality of life experienced by the inhabitants (Foschini et al., 2021; Shahrour & Xie, 2021; Yang et al., 2019).

#### 2.3.4.2 Strategies and Tools

The development and integration of mechanisms for crowdsensing and crowdsourcing in smart cities are of the utmost importance to effectively harness the power of collective intelligence. In this regard, various strategies and tools have been identified in the analyzed papers, which are considered to be highly relevant (as shown in Table 2).

One notable solution for deploying and managing crowdsensing campaigns, as proposed by Amaxilatis et al., offers a valuable feature that allows developers to adapt and improve their campaign design if they observe any underperformance (Amaxilatis et al., 2018). This capability enables continuous refinement and optimization of the crowdsensing process, ensuring the collection of high-quality data.

In contrast, the approach put forth by (Pánek, 2019) not only facilitates the collection of data but also empowers members of the community by encouraging them to think spatially about their surroundings (Pánek, 2019). By enabling individuals to map their opinions, perceptions, and feelings, this approach fosters a sense of belonging to a community and promotes active citizen engagement in shaping the development of smart cities.

Overall, these strategies and tools contribute to advancing the implementation and integration of crowdsensing and crowdsourcing mechanisms in smart cities, leading to more efficient and citizen-centric urban management.

#### 2.3.4.3 Leveraging Crowdsensing and Activating Citizen Engagement

Mobile crowdsensing (MCS) has emerged as a promising approach to enhancing citizens' daily lives and providing urban civilizations with new perspectives (Capponi et al., 2019; Nižetić et al., 2020). In the

Table 2: Strategies and tools used to implement and integrate mechanisms of crowdsensing and crowdsourcing in smart cities(R. Miranda et al., 2022).

| **Authors(s)/Article** | **Strategies/Tools** |
|---|---|
| Amaxilatis et al. (Amaxilatis et al., 2018) | A tool to assist individuals organizing crowdsensing campaigns in precisely defining their desired input from participants in such endeavors. |
| Pánek (Pánek, 2019) | The emotional mapping methodology enables both individuals and municipalities to engage in a map-based conversation about the present and future conditions of public spaces, drawing from their personal experiences in those areas. |

context of smart cities, MCS enables citizens not only to interact and engage with various services but also to actively contribute data through crowdsensing (Ismagilova et al., 2019). The availability of more information leads to more comprehensive conclusions and the creation of high-quality data sets (Foschini et al., 2021). Thus, understanding how to effectively recruit citizens and incentivize their data contributions has become a crucial research topic for the success of MCS programs (Capponi et al., 2019; L. Wang & Alexander, 2019).

Capponi et al. identified several application domains where MCS can significantly improve the quality of life in smart cities. These domains include emergency management and prevention, environmental monitoring, health care and wellbeing, e-commerce, indoor localization, intelligent transportation systems, mobile social networks (MSNs), public safety, unmanned vehicles, urban planning, waste management, and Wi-Fi characterization (Capponi et al., 2019).

To engage many citizens as active contributors, various strategies have been proposed. Table 3 presents the most relevant strategies identified in the literature. These strategies focus on leveraging social incentives, such as reputation systems and gamification, financial incentives, cognitive and emotional incentives, and altruistic motives (Capponi et al., 2019).

In the studies examined, a consensus among the authors was reached regarding the positive impact of crowdsensing initiatives on enhancing the quality of life in smart cities. Moreover, the literature outlined various application domains that hold significant potential for mobile crowdsensing to exert its influence.

Furthermore, the literature delved into the strategies employed to actively engage citizens as contributors in such initiatives. Notably, methods were identified that leverage gamification techniques, wherein sensing tasks are transformed into interactive games, thereby incentivizing participation. Additionally, offering services or monetary rewards in exchange for data contributions emerged as viable tactics in motivating individuals to participate. It is widely acknowledged among the authors that the success of any crowdsensing implementation hinges upon securing the involvement of many individuals.

Overall, the literature attests to the manifold benefits that can be derived from the integration of crowdsensing initiatives within smart cities. The strategies put forth in the literature offer insightful guidance on fostering citizen engagement, thereby bolstering the efficacy of crowdsensing efforts.

Table 3: Strategies focused on encouraging citizens to actively contribute.

| **Authors(s)/Article** | **Strategies** |
|---|---|
| Capponi et al. (Capponi et al., 2019) | There are three categories of strategies for stimulating participation and encouraging citizens to be active contributors: entertainment, service, and monetary. The entertainment category involves using methods that turn sensing tasks into games to stimulate participation. The service category involves providing users with services in exchange for their data. The monetary category involves rewarding users with money for their contributions. |
| Foschini et al. (Foschini et al., 2021) | Reward programs encourage and stimulate users to complete tasks, thereby augmenting the quantity and quality of gathered information, making it a common strategy for increasing and reinforcing participation through gamification techniques. |
| Hu et al. (Hu et al., 2020) | In order to maintain a consistent number of participants in the MCS situation, a three-stage Stackelberg game was developed, treating it as a market for sensory data and guaranteeing the ongoing contribution of sensory data from monthly paying participants. |

### 2.3.5 Conclusion

The concept of smart cities plays a crucial role in promoting the sustainability and development of densely populated areas. To ensure the successful transition to smarter cities, involving residents in the process, becomes essential. One of the key mechanisms to achieve this is crowdsensing and crowdsourcing strategies.

To comprehensively analyze the existing literature on smart cities and their application through crowdsensing and crowdsourcing, a systematic review was conducted. After the PRISMA model, the review drew upon the Scopus database as the primary source. Out of the initial pool of 20,240 studies, the rigorous application of the PRISMA methodology resulted in a final selection of 17 relevant studies that were aligned with the predefined research questions.

Based on these research questions, the state of the art was evaluated about the role and purpose of crowdsensing and crowdsourcing in smart cities. The compiled studies point towards the potential of crowdsensing to gather and share large volumes of data. Such data can be utilized to track urban dwellers' habits and movements, among other variables. On the other hand, crowdsourcing initiatives enable citizens to actively and passively participate in data collection about the city's services and quality of life, thereby enhancing their performance.

This research has been conducted as part of a larger project aimed at investigating the practical implementation of crowdsensing, crowdsourcing, and geofences within the infrastructure and framework of smart cities. Building upon the valuable insights gained through the systematic review, future work will involve the development of a platform that integrates these concepts into a viable and effective information system. This platform is expected to bring significant value to the smart city paradigm and improve the lives of its citizens when implemented.

natbib

## 2.4 Data Platforms for Real-time Insights in Healthcare

The usage and popularity of Internet of Things (IoT) devices in healthcare has grown exponentially, thanks to advancements in Big Data technologies and machine learning algorithms. This has created new opportunities for data engineers in the field of healthcare and continuous care (Hak et al., 2022). However, there is a pressing need to bridge the gap between the creation of information and the ability to offer actionable insights and interventions timely. Implementing a large-scale data processing architecture presents its own set of challenges, especially when it comes to choosing the most appropriate tools and determining how to apply them effectively. Fortunately, streaming systems have reached a level of maturity that allows hospitals worldwide to leverage their vast datasets and work together with data producers to predict and influence future events.

This section's goal is to explore the current state-of-the-art in data platforms for healthcare, specifically those that enable the creation of real-time metrics and actions. To ensure transparency and consistency, the reporting of this review follows the PRISMA guideline. By adhering to this guideline, technological advancements made in this critical sector can be better understood and evaluated. To gather relevant literature, a comprehensive search was conducted on the Scopus abstract and citation database on May 13, 2022. Specific keywords were used to identify potential documents published from January 2020 onwards. These documents were required to be written in English and accessible through the B-ON consortium, which grants Portuguese students legal access to most publishers' content.

From the retrieved articles and papers, seven studies were selected for deeper discussion based on their relevance and impact for this review. These selected studies offer insights into their main objectives, data sources, tools used, as well as their approaches to interoperability and support for machine learning algorithms in decision support. The collected articles highlight the current use of Big Data in healthcare institutions of all sizes. However, the ability to process large volumes of data from sensors and events and notify stakeholders still remains in its early stages. There is a clear need for further advancements to ensure the swift processing of data and the timely dissemination of critical information to improve patient outcomes and enhance the efficiency of healthcare delivery.

### 2.4.1 Methods

This section outlines and justifies the methodology employed in conducting this systematic review. By examining earlier papers and articles, valuable insights can be gained regarding the approaches taken, the obstacles encountered, and the strategies employed to overcome them. Such an analysis enables a comprehensive understanding of the existing body of knowledge in the field.

To ensure the accuracy and transparency of this review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist was adhered to. This checklist serves as a guideline to effectively communicate the main objectives, methodology, and results of the review (Page et al., 2021b).

The data sources utilized for this review primarily consist of online databases, which offer a wide range of scholarly information from various disciplines. Additionally, specific inclusion and exclusion criteria were carefully defined to determine the eligibility of documents for analysis. This systematic approach aims to ensure the selection of relevant and high-quality sources.

Moreover, a meticulous process was employed to identify and extract pertinent information from the selected documents. The documentation of this process guarantees transparency and allows for reproducibility and verification of the findings.

By adopting explicit, rigorous, and accountable methods, this study aims to contribute to the interpretation of research findings and the advancement of research knowledge in a manner that can ultimately benefit society as a whole (Gough et al., 2012). Through the careful application of these robust methodologies, the potential for meaningful insights and impactful recommendations can be maximized.

#### 2.4.1.1 Information Sources

The Scopus[1] abstract and citation database was selected as the primary data source for this study due to its extensive size, comprehensive coverage across various publication areas, and rigorous quality assurance measures. The database was last accessed on September 1, 2022, ensuring the inclusion of the most up-to-date information available. Moreover, the impact of journals on the highlighted publications was assessed by consulting the SCImago Journal & Country Rank[2] on September 15, 2022. This additional source serves as a valuable reference for evaluating the influence and prestige of the journals included in the analysis.

#### 2.4.1.2 Search Strategies

To begin the literature search, a preliminary set of keywords was identified as a starting point. These keywords will be applied to the title, abstract, and keywords fields (TITLE-ABS-KEY) to capture relevant literature. In an effort to further refine the search process, the keywords were categorized into two groups. These groups are interconnected by the **AND** operator, meaning that both groups must be satisfied for a document to be retrieved.

The use of the **OR** operator within each group allows for inclusivity, as it ensures that all documents containing at least one keyword from each group will be considered. The rationale behind this decision is to maximize the number of potential references while maintaining a level of specificity.

[1] https://www.scopus.com
[2] https://www.scimagojr.com

The first group pertains to technical areas and subjects directly relevant to the research topic, while the second group aims to encompass broader areas within the field of technology, specifically in the context of healthcare and continuous care. This strategy enables the results to be contextualized within the healthcare domain while retaining a focus on relevant technological aspects.

It should be noted that this review is restricted to the subject areas of Computer Science, Engineering, and Medicine. Moreover, only articles published between 2019 and 2022 will be included in the analysis. These restrictions align with the desired scope of the research and ensure the most current and relevant literature is examined.

#### 2.4.1.3 Inclusion and Exclusion Criteria

After retrieving the initial set of documents, a set of selection criteria were defined to refine the results based on specific inclusion or exclusion criteria. The selection criteria applied in this research, presented in the order of their application, are as follows:

Firstly, the accessibility criterion requires that the documents must be available at B-ON, a consortium that provides Portuguese universities with access to most publishers. This ensures that only documents that can be accessed through this platform are included in the research.

Secondly, the timeliness criterion mandates that the documents must have been published from 2019 onwards. This criterion confirms that only the most recent and up-to-date information is considered for the analysis.

Next, the research area criterion states that the papers should fall within the scope of Computer Science and Engineering in a healthcare environment. This criterion restricts the selection of papers that are relevant to the specific context of this review.

Lastly, the document type criterion specifies that the selected documents must have already been published, be in either an Article or Paper format, and be written in English. This criterion ensures that only finalized research publications are included, thereby maintaining the academic rigor and quality of the selected documents.

Upon identifying articles that initially appeared relevant for further evaluation, the full text of those articles was downloaded and thoroughly read. It is worth noting that although the B-ON consortium provides access to a wide range of paid articles, some articles had to be excluded due to their prohibitive costs.

As a result of the aforementioned search strategy, the initial set of documents was substantially filtered, resulting in a total of 3503 documents for further analysis. A visual representation of the research process can be seen in Figure 12. The figure illustrates the systematic and iterative nature of the search process, highlighting the application of the aforementioned selection criteria.

#### 2.4.1.4 Synthesis Methods

To conduct an in-depth analysis of the various technologies and methodologies for real-time information processing in healthcare environments, a comprehensive categorization system was devised to assess the selected documents. These categories were specifically formulated to address the research inquiries proposed in this study:

1. **Main Objectives:** This category aims to identify and compile the primary objectives and goals outlined in each document.
2. **Data Sources:** In this category, the sources of data used in the studies are examined, which may include data collected from Internet of Things (IoT) devices, manually generated data by medical professionals, or synthetic data for experimentation.
3. **Tools Used:** To understand the technological aspect of the proposed solutions, this category scrutinizes the frameworks and programming languages adopted or suggested for the development of the proposed systems.
4. **Semantic Interoperability:** This category encompasses all measures taken to ensure semantic interoperability, including the implementation of relevant standards and protocols to enable seamless exchange and interpretation of data.
5. **Machine Learning Support:** This category evaluates whether the proposed platforms ease the development and integration of machine learning algorithms, which can augment decision-making processes in healthcare.
6. **Limitations:** Here, the vulnerabilities and limitations identified in the documents are presented. These shortcomings may have been acknowledged by the authors themselves or recognized by the reviewers. Additionally, potential future steps to address these limitations are also discussed.

By employing this categorization system, a comprehensive understanding of the available technologies and approaches in real-time information processing in healthcare can be achieved. Additionally, this system allows for the identification of research gaps and potential areas for improvement in this domain.

### 2.4.2 Results and Discussion

A total of 3503 documents were retrieved from Scopus using the previously mentioned research query. From this initial set, the most relevant documents were selected, resulting in a subset of 291 documents. These 291 documents were then subjected to a more detailed screening process, wherein the abstract, results, and conclusion were thoroughly analyzed. The outcome of this screening process yielded a final set of 7 documents that align with the intended research themes and context. These seven documents will serve as the basis for the literature review and state of the art analysis.

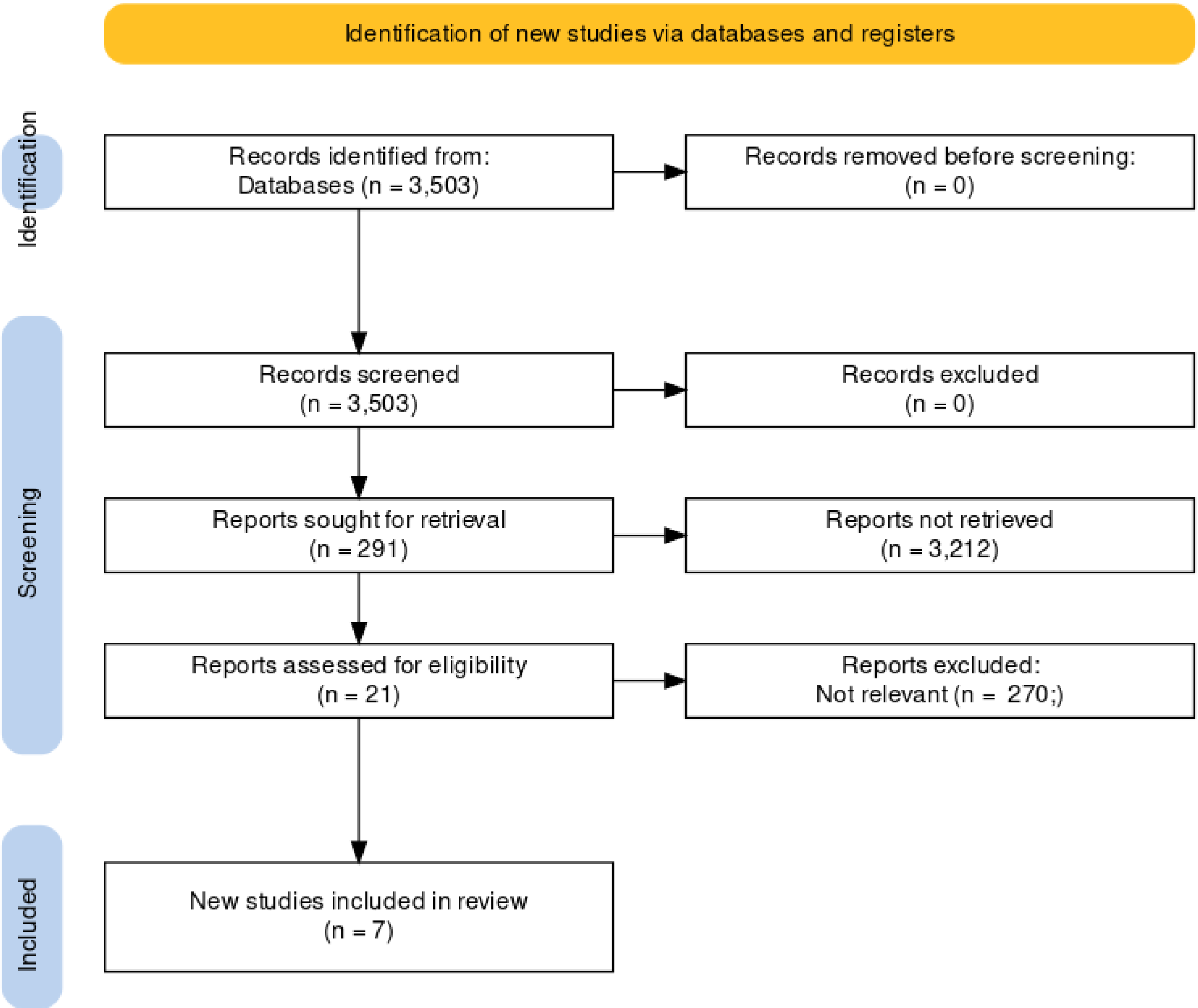


Figure 12: Research progress flow diagram (created using Haddaway et al., 2022)

The primary aim of this section is to identify the current state-of-the-art in data platforms for healthcare that enable the real-time creation of metrics and actions, thereby facilitating prompt and effective patient care. Additionally, the analysis will consider key aspects such as the utilization of machine learning approaches for detecting patterns in medical data to extract further knowledge. Furthermore, attention will be given to topics concerning data semantics and standardized formats, as these elements play a vital role in facilitating interoperability and validating the viability of the platform.

The third use-case seeks to explore the main benefits associated with the development and implementation of data platforms that provide real-time insights for the healthcare community and patients themselves.

After scrutinizing a total of twenty-two articles, it was found that the authors cover a diverse range of topics.

In their study, El-Ganainy et al.) put forward a novel system that offers real-time clinical decision support without the requirement of an offline training system with extensive datasets, unlike traditional machine learning systems. The system proposed by the authors encompasses two stages, with the first stage utilizing Hierarchical Temporal Memory (HTM) and the second stage enabling real-time flow processing and providing unsupervised predictions. The results of the study demonstrate that the proposed system surpasses logistic regression in terms of classification accuracy, recall, precision, and area under the receiver operating curve (El-Ganainy et al., 2020).

Shehab et al. tackle the improvement of health monitoring systems through the adoption of IoT-enabled devices in diverse environments, such as smart homes and smart hospitals. These systems necessitate a dynamic analysis of critical patient flows, such as electrocardiogram (ECG) flow. The researchers' findings highlight that a per-user scheduler effectively resolves resource limitations of Fog Computing nodes and maximizes the utilization of the Fog Computing infrastructure, thereby facilitating the safe expansion of the system up to twice its original capacity (Shehab et al., 2021).

Meanwhile, Alharbi et al. propose an innovative real-time heart rate prediction system designed to aid doctors and patients in anticipating heart rate risks in real-time. The system consists of two phases: an offline phase aimed at developing the model using various prediction techniques to identify the approach yielding the smallest root mean square error, and an online phase employing Apache Kafka and Apache Spark for advanced prediction of heart rate based on the optimized model (Alharbi et al., 2021).

In a study conducted by Saleh et al., a system for predicting systolic blood pressure in real time was explored. The system utilized both offline and online modes, employing deep learning models and Apache Kafka and Apache Spark, respectively. The offline mode involved training several deep learning models to minimize the root square error, while the online mode utilized the best deep learning model and streamed time series data from the systolic blood pressure system. The researchers found that the BI-LSTM model with three hidden layers demonstrated the best performance (Saleh et al., 2021).

Another study by Thanh et al. focused on the development of an Internet of Health Things (IoHT) platform specifically designed for the healthcare environment. The platform utilized a brokerless, microservices architecture and placed significant emphasis on data collection, users and device management, and remote device control. To address various challenges, such as limited processing capacity, power savings, data collection speed and accuracy, security mechanisms, and scalability, real-time alerts for the medical team were incorporated into the IoHT platform. The authors concluded that the proposed platform performed effectively, with no errors and unaffected by geographical distance (Thanh et al., 2021).

Afreen Banu and Rajamani designed an IoT-based multisensory online vital monitor, named VITALS, for bedside monitoring of physiological parameters including heart rate, body temperature, blood pressure, and peripheral oxygen saturation. The system extracted vital signs at regular intervals and transmitted them to a Big Data analytics system over Wi-Fi for further analysis. Various technologies, such as Apache Kafka, Apache Spark, Hadoop Distributed File System (HDFS), Spark SQL, Hive, and Matplotlib, were utilized to collect live data streams, categorize vital signs, archive data for long-term storage, and enable caregivers to access and visualize relevant information. The authors also developed a mobile application. According to the authors, the proposed system offers enhanced care solutions, particularly for individuals with limited access to care services (Afreen Banu & Rajamani, 2023).

F. Hassan et al. introduced a system that leveraged machine learning streaming models, specifically streaming linear regression with stochastic gradient descent (SGD), to analyze streaming integrity data events received via Kafka threads. The system was tested on both historical medical datasets (diabetes, heart disease, and breast cancer) and generated datasets simulating wearable medical sensors. The experimental results demonstrated the system's ability to learn online and update the model as new data arrived, taking into account the window size (F. Hassan et al., 2020).

### 2.4.3 Conclusions and Future Work

Healthcare environments are widely acknowledged for their complexity and high levels of stress, particularly due to the critical nature of the situations involved. Thus, it is imperative to leverage technological advancements to identify potential events and scenarios and to notify all relevant stakeholders in a timely manner. This holds true not only for hospitals but also for other health institutions.

One such technology that has gained significant attention is Big Data, which enables the use of extensive data collected from past patients to gain knowledge and insights. However, the power of Big Data is magnified when combined with real-time information processing approaches. By integrating these two technologies, medical professionals can receive immediate notifications when novel events occur, enabling them to take swift action.

To gain a comprehensive understanding of the current state-of-the-art in real-time information processing technologies in the healthcare sector, a systematic review was conducted. This systematic review strictly adhered to the guidelines provided by the PRISMA 2020 statement, which facilitates the accurate reporting of systematic reviews. The primary source of documents used for this review was the Scopus database, renowned for its comprehensive collection of academic and professional literature.

Through this rigorous review process, the aim was to not only explore the existing technological landscape in real-time information processing in healthcare but also to identify gaps and potential areas for future research and development. By shedding light on the current advancements and challenges in this domain, our findings will contribute to the ongoing efforts in improving healthcare delivery and saving more lives.

## 2.5 Interoperability and Semantics

Interoperability can be summarised as the ability of various systems, devices, and software applications to communicate and exchange information with each other seamlessly, while semantics refers to the meaning of language constructs, as opposed to their form (syntax) (Oliveira, Miranda, Hak, et al., 2021). Semantics describes processes a computer follows when executing a program in a specific language by describing relationships between input and output or explaining how a program will be executed on a certain platform (Techopedia, 2018).

Interoperability and semantics are crucial for real-time information because they ensure that data are captured, stored, and transmitted accurately, consistently, and efficiently. With interoperability, real-time information can be seamlessly exchanged between different systems and platforms without the need for complex and time-consuming manual processes, enabling organizations to make timely and informed decisions based on up-to-date data. Semantics, on the other hand, ensures that data are interpreted correctly and in the right context. This is important because real-time information is often highly time-sensitive and can have significant implications for business operations and decision-making.

To ensure data quality in real-time information, organizations must establish robust frameworks that include data quality controls, data standards, and data validation processes. They must also invest in technologies that ensure interoperability, such as APIs, service-oriented architectures, and data integration platforms.

In this section, the importance of interoperability and semantics in real-time information systems, as well as quality in real-time information systems are discussed. To further illustrate the importance of these concepts, the importance of interoperability and semantics in the healthcare and manufacturing sectors are also discussed. Furthermore, openEHR, a widely used open-source standard for interoperability in healthcare is presented.

### 2.5.1 Data Quality in Real-Time Information Systems

Data quality was refered about ”data that are fit for use by data consumers” (R. Y. Wang & Strong, 1996). It can be characterized by its accuracy, completeness, consistency, and reliability (Analide et al., 2011). In real-time information systems, data quality is particularly critical, as decision-makers rely on timely and accurate data to make informed decisions.

Data quality is important for real-time information systems because it affects the decisions made by the users of the system. However, if the data are inaccurate or inconsistent, the decisions made may not be the best, possibly impacting the organization. This can be exemplified in the healthcare sector, where the quality of data can lead to wrong diagnoses, and have a significant impact on the patient’s health (Duarte et al., 2010).

Thus, data quality is important for real-time information systems as it ensures accurate decision-making, trust in the system, efficient operations, improved data analysis, and reduced costs. Organizations must prioritize data quality to maximize the benefits of real-time information systems, by implementing data validation, data cleansing, and data standardization solutions, and commit to a culture where data quality is important at all levels of the organization (Kimball & Ross, 2013).

An ETL system should have a data cleansing system to ensure that the data are accurate and consistent. A data clensing system is a process that identifies, captures, and corrects inaccurate, incomplete, or inconsistent data in a data warehouse. This subsystem should be able to identify and correct data quality issues, such as missing values, duplicate values, and inconsistent values, include requirments for source systems and integrations to supply better data, and give descriptions of quality issues expected to be encontered (Kimball & Ross, 2013).

Another important aspect of data quality is data standardization. Data standardization refers to the process of defining and imposing a consistent format, structure, and quality requirements on the data within an organization (He et al., 2019). This standardization process ensures that data are reliable, accurate, and consistent across all systems and departments within an organization. Furthermore, data standardization enhances data sharing and integration across different systems and platforms, enabling organizations to merge and leverage their data, to make data-informed decisions, and improve their operational efficiency and effectiveness. Thus, data standardization is an important aspect of interopability in real-time information systems, as a set of standards would be required to ensure that data are interpreted correctly and in the correct context in all applications (He et al., 2019). Examples of data standardization include formatting dates, times, and temperatures in a consistent manner, using standardized codes for data elements, and using common naming definitions (Kimball & Ross, 2013).

More steps to ensure data quality in real-time information systems include the establishment of data governance policies that define who handles data quality, what policies to follow, and what checks to perform; the development and maintainance of data dictionaries that define the structure of the data, its meaning, and its context.

### 2.5.2 Interoperability in Healthcare

Information technology systems are changing the healthcare sector, from the discovery of cures for diseases and the development of new treatment techniques to the improvement of patients' diagnoses and the enhancement of effective health care delivery (Brito et al., 2019). From health centers to large-scale hospitals, the increased use of IT approaches in clinical procedures is improving the patients' quality of care, as well as optimizing the health institution's resources (Cardoso et al., 2014; Duarte et al., 2011; Neves et al., 2018). Furthermore, healthcare institutions are now understanding the value and impact data can have in the patient's quality of care. This extremely important asset is mostly used for operational record and clinical decision making. Consecutively, the next generation of healthcare services will employ IoT devices, equipped with sensors and wireless connections, gathering large amounts of data (Oliveira, Miranda, Hak, et al., 2021).

However, as healthcare units keep growing, so does the complexity of their medical and information technologies, increasing the possibility of medical errors (M. Miranda et al., 2009). Thus, interoperability is an imperative when developing and integrating software solutions, allowing the exchange of important data, and initiate actions between different products without any additional effort on the user (Cardoso et al., 2014).

As expected, data standards are vital for interoperability in healthcare, as they define the format and structure of clinical data, and ensure that they're interpreted correctly and in the correct context in all applications (He et al., 2019). International organizations are developing standards and guidelines for electronic health records to support the exchange, integration, and retrieval of electronic health information. However, achieving semantic interoperability (Oliveira, 2018), which involves agreements on standards, information models and terminologies, is the most difficult level. This level should initially be implemented in priority clinical areas of high relevance to patient safety. Some of the most widely used data standards in healthcare include HL7, DICOM, and SNOMED CT (Cardoso et al., 2014).

Health Level 7 (HL7) is a set of international standards for the exchange, integration, sharing, and retrieval of electronic health information. It is a messaging standard that ensures standardized communication between different healthcare systems, Electronic Health Record (EHR) systems, and other healthcare-related systems (HL7 International, 2023). Its name is derived from the highest level of the ISO communications model for Open Systems Interconnection (OSI) model, the application layer. The main objectives of Health Level 7 (HL7) are to develop an open and extensible framework for healthcare data exchange that can integrate various healthcare applications and systems, to improve healthcare outcomes by promoting interoperability and data sharing among disparate systems, and to standardize the exchange of electronic health information to improve patient care, reduce waste and increase efficiency (M. Miranda et al., 2009). Furthermore, its Fast Healthcare Interoperability Resources (FHIR) specification combines the best features of HL7 with the latest web technologies to create a modern, flexible, and scalable framework for healthcare data exchange (HL7 International, 2023). HL7 defines standards for the exchange of various types of patient-related data, including laboratory results, clinical observations, demographics, diagnoses, and treatment plans, among other things. HL7's focus on standardization also plays an essential role in ensuring seamless and efficient health data exchange among healthcare providers, patients, or researchers.

The Digital Imaging and Communications in Medicine (DICOM) standard is a set of international standards that helps to regulate storage, communication, retrieval and display of medical images and associated data. DICOM aims to ensure that medical images and information can be communicated and shared across various medical devices and systems. It is a standard that helps to regulate storage, communication, retrieval and display of medical images and associated data. It is an essential tool used in medical imaging, particularly in radiology, cardiology, and oncology (Ferreira, 2018).

SNOMED Clinical Terms (CT) is a comprehensive clinical terminology system used to describe and identify various medical concepts, terms, procedures, and diagnoses in clinical practice. The SNOMED terminology is developed by SNOMED International, a non-profit organization, with the aim of standardizing and improving communication between healthcare professionals, researchers, and patient care systems globally (Alahmar & Benlamri, 2020). It does this by providing a standardized language that ensures accurate and consistent patient data recording, sharing, and analysis across healthcare settings, resulting in safer, efficient, and cost-effective healthcare delivery. SNOMED is a structured system designed for Electronic Health Record (EHR) and is used extensively throughout the clinical domain. It is a hierarchical taxonomy, with a tree structure and concepts organized by related features, based on an ontology of healthcare concepts that is both detailed and comprehensive. SNOMED has many advantages over other medical coding systems, including its ability to handle ambiguous and evolving terminology, to support multilingual descriptions of medical concepts, and to allow for the ease of multi-axial classification of concepts (Shafqat et al., 2020). In addition, SNOMED is a standardized terminology used across countries, which allows for sharing and comparison of clinical data on an international level.

openEHR is an open-source standard for Electronic Health Records (EHRs), created with the aim of providing a flexible and standardized way of capturing, storing, and exchanging health data (openEHR Foundation, 2021). The standard is based on the principles of semantic interoperability, modularity, and extensibility, which allows for the creation of clinically relevant health information models that can be used across different healthcare settings (Li & Tsui, 2020). The architecture of openEHR is designed around a set of archetypes and templates, which define the structure and content of the health data (Hak, Oliveira, et al., 2020). The archetypes describe clinical entities, such as a diagnosis or medication, providing a standard definition of their attributes, relationships, and constraints. The templates define how the archetypes are combined to create a specific clinical document or data structure. Both archetypes and templates are created using a domain-specific language, known as the Archetype Definition Language (ADL) (Hak, Guimarães, et al., 2020).

The architecture of openEHR is divided into three layers: the reference model, the archetype model, and the service model. The reference model defines the core data structures and operations for representing clinical data, including the composition, which is the container for clinical information (Alves et al., 2019; Oliveira, Miranda, Leuschner, et al., 2021). The archetype model defines the clinical knowledge and semantics of the data that make it possible to represent a wide range of clinical scenarios in a standard manner. The runtime environment implements the reference and archetype models, providing an interface for capturing, storing, and exchanging health data(Brandão et al., 2016; The openEHR Foundation, 2015). Figure 13 presents a diagram illustrating the openEHR specification.

The openEHR standard provides guidelines and open specifications at clinical and software implementation levels, and can be used to build information and interoperability solutions for healthcare (Meredith, 2021). It has been deployed on several national healthcare systems and hospitals, including the Ministry of Health at the Republic of Slovenia, with more than 85% of Slovenia's national heath data saved on the openEHR format (Better, 2021). In Portugal, the openEHR has been used during the COVID-19 pandemic, interoperating with the institution's legacy information systems, with the aim of responding to the pandemic's evolution (Oliveira, Miranda, Hak, et al., 2021; Oliveira et al., 2020). The integration of openEHR in a pandemic had demonstrated its versatility on existing systems.

### 2.5.3 Interoperability in Manufacturing

Manufacturing is a complex process that involves the integration of different systems and technologies. Whren these components work together, they can allow for innovative and efficient manufacturing processes, and achieve a high level of efficiency and productivity. One of the primary benefits of interoperability in Manufacturing 5.0 is that it facilitates real-time data collection and analysis (Demir et al., 2019). By integrating various sources, manufacturers can gain a better understanding of their operations and

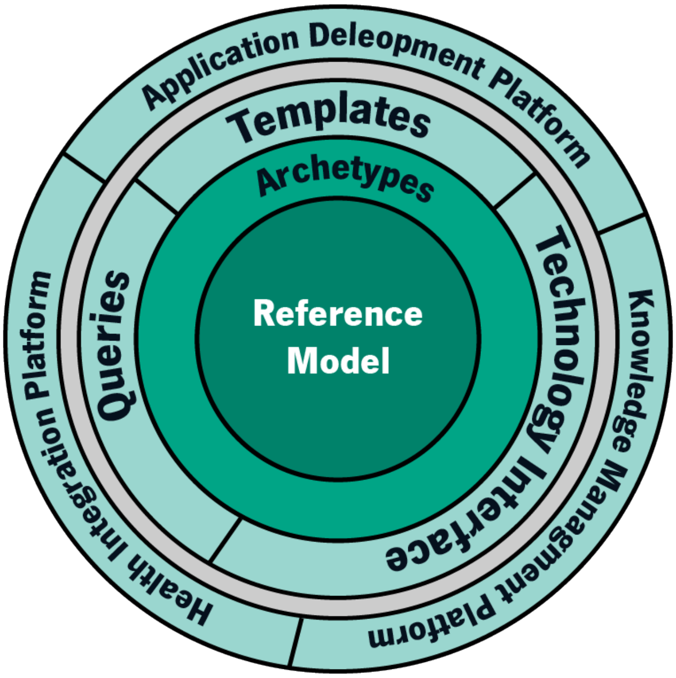


Figure 13: openEHR specification. Adapted from (The openEHR Foundation, 2015)

identify areas where they can improve their efficiency and profitability (Lu, 2017). For instance, data collected from sensors embedded in various machines and systems can be analyzed to detect patterns and identify anomalies (Mujahid Ghouri et al., 2021). This can help manufacturers optimize their production processes, reduce waste, and improve quality control.

Another important aspect of interoperability in Manufacturing 5.0 is the ability to integrate various systems and technologies into a single, cohesive network. By integrating these systems, manufacturers can reduce their reliance on manual processes and improve the accuracy and speed of their operations. This is particularly important in complex manufacturing environments where the use of multiple systems and technologies is common. By leveraging interoperability, manufacturers can streamline their operations and reduce the risk of errors and delays.

Several technologies and solutions are used to achieve interoperability in manufacturing, including Industrial Internet of Things (IIoT), cloud computing, Artificial Intelligence (AI), and edge computing. The Industrial Internet of Things (IIoT) is a network of interconnected devices, sensors, and systems that communicate with each other to share data and insights (Doyle-Kent & Kopacek, 2020). By leveraging IoT technologies, manufacturers can gain real-time insights into their operations, automate various processes, and optimize their production processes (Thanh et al., 2021). IoT technologies enable the seamless integration of different systems and technologies, making it easier for manufacturers to innovate and improve their operations.

Other important technologies that enable interoperability in Manufacturing 5.0 include cloud computing and edge computing (Salesforce, 2020). Cloud computing involves the delivery of computing resources, including software, storage, and processing power, on -demand over the internet (Runsewe & Samaan, 2021). These resources are provided by a third-party vendor who hosts and manages them, enabling organizations to access them from anywhere in the world with an internet connection. Cloud services can be used to store and analyze data from various sources, making it easier for manufacturers to access the insights they need to make more informed decisions (Thames & Schaefer, 2016). In contrast, edge computing, also known as fog computing or edge analytics, is a distributed computing framework that uses resources closer to data sources for efficient data processing and analysis (Elhadad et al., 2022). Edge computing can play a vital role in achieving interoperability by enabling faster and more efficient data processing, analysis, and decision-making, thereby facilitating the exchange of information between different devices and systems (Firouzi et al., 2022).

Thus, interoperability is a critical aspect of manufacturing 5.0, allowing manufacturers to integrate different systems, devices, and machines to improve their operations and gain a competitive advantage. However, achieving interoperability is often challenging because of the complexity and heterogeneity of manufacturing systems, which use different protocols, interfaces, and data formats. Technologies that enable interoperability, such as IoT, cloud computing, and edge computing are essential to realizing the full potential of Manufacturing 5.0 and achieving its benefits.

# 3

# Case Studies

This chapter provides an overview of the use cases for real time information processing in a distributed computing environment. This chatper will discuss the challenges of designing, deploying and managing efficient and reliable real-time systems, and how developers can use such systems to support various types of applications from business intelligence to analytics, stream processing and data warehousing, and from IoT to edge computing. A discussion of the different approaches for designing and deploying a real-time system will be presented, along with the associated best practices. Finally, it will explore the potential of using real time systems for different types of applications, as well as the potential risks and challenges associated with their implementation.

## 3.1 Bosch and the Factory of the Future

### 3.1.1 Introduction

Bosch Car Multimedia is a division of Robert Bosch that specializes in developing and manufacturing electronic components for vehicles aimed at enhancing the driving experience and providing better connectivity. Bosch Car Multimedia has factories in various parts of the world, including Thailand and Portugal. The products and services offered by Bosch Car Multimedia include infotainment systems, navigation systems, instrument clusters, as well as telematics solutions to enable smart vehicle communications and connectivity to the internet.

Industry 5.0 is the latest wave of technological advancement in the manufacturing industry. It is characterized by the integration of physical manufacturing systems with intelligent networks and real-time information processing. Adopting Industry 5.0 technologies offers many advantages for the manufacturers, such as increased efficiency, higher productivity, quicker decision-making, better quality control, and reduced waste. Manufacturers like Bosch can benefit from Industry 5.0 as it would allow them to streamline their manufacturing processes, improve product quality, and increase their competitive edge in the market. By integrating data analytics, machine learning, and other intelligent technologies into its operations, Bosch could leverage real-time information processing to optimize its production line, reduce downtime, and respond to changing market demands.

In this context, the *Factory of Future* project has emerged, resulting from the collaboration between Bosch Car Multimedia in Braga, Portugal, and the University of Minho. This project was composed of five subprojects and their lines of research, namely Quality Control, Logistics, Production Process, Industrial Operations Management, and Business Intelligence. Thus, the *Factory of Future* project has generated scientific and technological advances that have translated into a suite of processes, systems, and tools, offering a high level of flexibility and quality, as well as competitive costs to the main operations related to different phases of Industry 5.0 initiatives. This case study corresponds to the P55 subproject, titled 'Real Time Information Processing'.

The architecture and scope of the Factory of Future project is showcased in Figure 15.

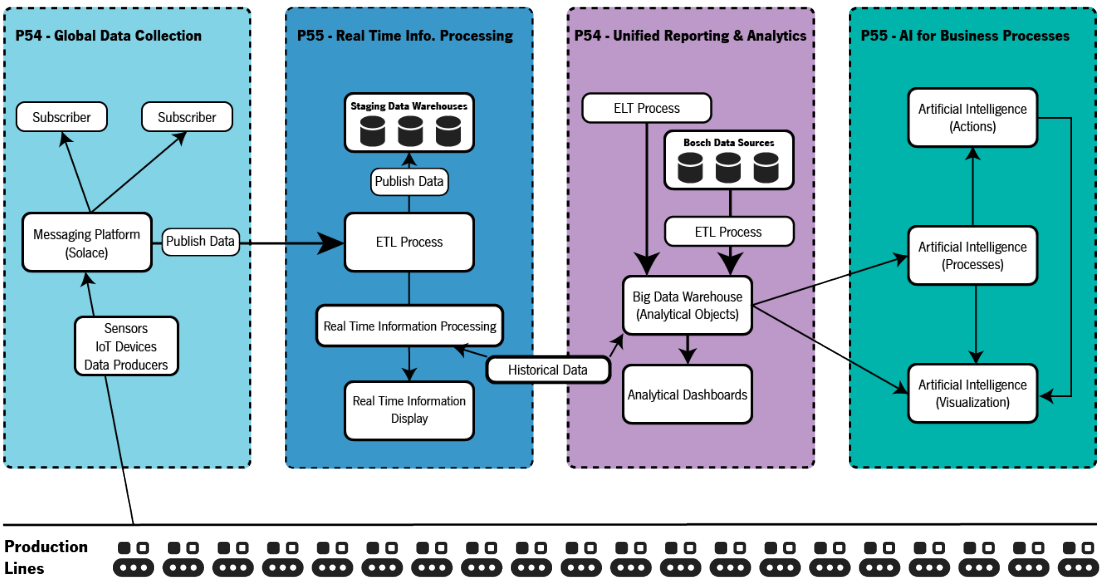


Figure 14: Factory of Future project scope & architecture

### 3.1.2 Methods

The production lines of Bosch Car Multimedia in Braga were deployed with Bosch XDKs and telemetry to gather data from various devices across their organization. Bosch Cross-Domain Development Kits (XDKs) are a development kit that offers a comprehensive range of sensors and functions for Internet of Things (IoT) applications. It is a versatile platform that enables developers to create applications for a variety of different use cases, from Industrial Internet of Things (IIoT) to smart city projects. The devices come equipped with a wide range of sensors, including temperature, humidity, pressure, acceleration, magnetometer, and gyroscope, as well as a light sensor. Additionally, the kits have built-in Wi-Fi and Bluetooth connectivity, as well as a MicroSD card slot for data storage. To manage this influx of data, they are utilizing Solace PubSub for an efficient and scalable data distribution. With Solace PubSub, the company can reliably and securely send data to the appropriate receivers in real-time, ensuring that their operations are always in sync. Additionally, Solace is providing the company with valuable insights into their data usage, allowing them to make data-driven decisions and optimize their operations. The combination of Bosch XDKs, telemetry, and Solace PubSub is enabling Bosch Car Multimedia to efficiently and effectively manage their devices and data, giving them a competitive edge in their industry.

To achieve the defined goals, a platform was developed that is composed of various technological components, including Complex Event Processing (CEP) for querying and processing data before storing it within a database, a web application for appropriate data representation, and the use of ontologies for effective communication. Additionally, a Staging Data Warehouse will be required for data loading, which will also act as a data repository. Finally, a data loader and standardization part will be developed to load data into the staging data warehouse.

The web application developed as part of this project aims to provide a user-friendly interface for interacting with the system. The application is designed to display information about the factory processes in a graphical format, allowing users to quickly and easily identify patterns and trends. The web application is also integrated with CEP, providing a real-time ETL system that can effectively provide sensorial interaction with the user. The web application is also designed to be scalable and adaptable, allowing for the addition of new functionalities and features in the future. This adaptability is critical as the needs of the factory change and evolve over time.

The staging data warehouse serves as a repository for data that has been processed and transformed by the CEP system. This data warehouse allows for historical analysis and querying of past events, as well as providing a secure location for storing data that may not be suitable for immediate processing by the CEP system. The integrated use of these components results in a powerful and efficient ETL system capable of providing sensorial interactions with the user through a web application interface. This interface allows data analysts to interact with the system in real-time, providing valuable insights and information about the factory processes.

In order to achieve these goals, this project adopted the employment of ontologies, which are formalized sets of concepts used to represent and categorize knowledge in a specific domain. Ontologies give a common vocabulary and set of categorizations for a particular domain, allowing for standardized and consistent data representation and communication between different systems and users. The use of ontologies, in combination with a CEP system, allows for real-time data processing and analysis by providing a standardized framework for organizing and interpreting complex events. By defining a set of concepts and relationships between them, Ontologies enable the CEP system to identify and react to situations that may not have been anticipated beforehand.

Figure 15 shows the architecture of the P55 Business Intelligence subproject.

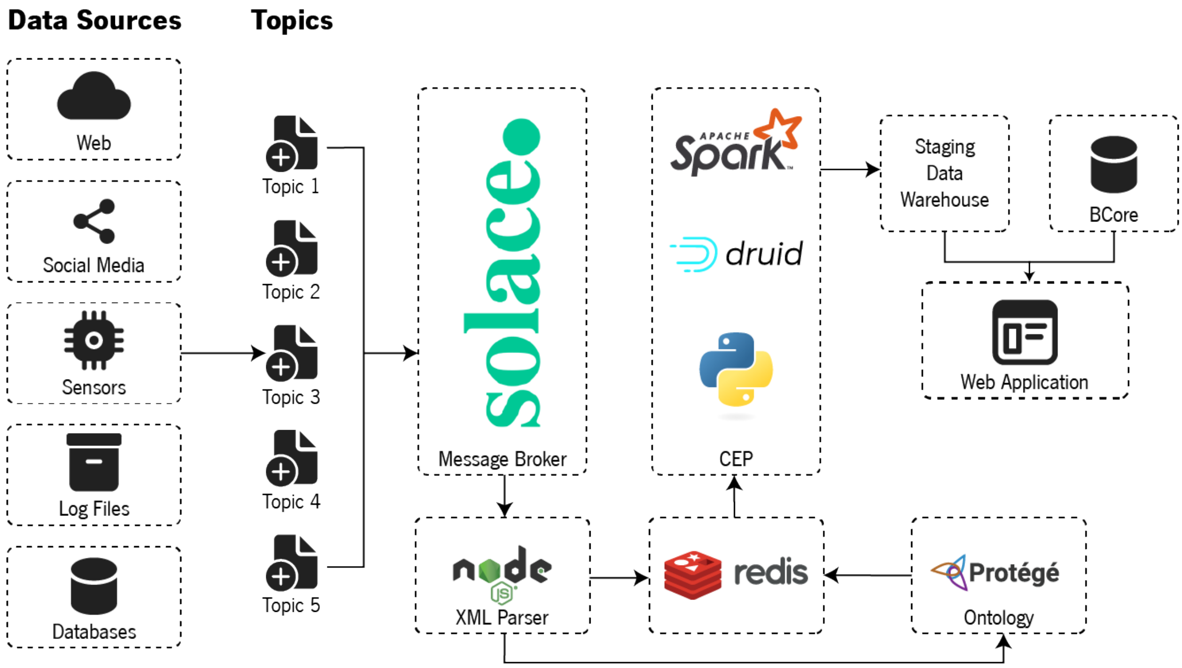


Figure 15: Architecture of the case study

### 3.1.3 Results

With the specification and architecture of the sensorial ETL system in place, the development of the platform began. The first step was to develop the staging data warehouse, as well as a data loader and standardizer. Afterward, a web application was developed to offer a user-friendly interface for interacting with the system. Finally, the CEP system was developed to provide real-time data processing and analysis. This system was also integrated with the web application, allowing for real-time sensorial notifications. The use of ontologies was also implemented to give a standardized framework for organizing and interpreting complex events. In this section, an overview of all these components will be provided.

#### 3.1.3.1 Staging Data Warehouse

The staging data warehouse is a data repository that stores all data that has been retrieved from machines and sensors, as well as configuration and metadata for the web application. This data warehouse is used to store short and medium-term data, as long-term data is stored in an approppriate data lake. The staging data warehouse was built using the MySQL, of the most used relational Database Management System (DBMS) in the world due to its compatibility with multiple operating systems, its versatility, and a wide variety of third-party integrations and utilities.

Furthermore, to load data into the staging data warehouse, a data standardizer was developed. For the development of this tool, the Node.js runtime environment was used, enabling JavaScript code to run on the server uncoupled from a browser (Brown, 2014). This standardizer is responsible for connecting to a Solace PubSub broker and subscribing to the appropriate topics. When it receives a message from a topic, it parses the message, performs any necessary transformations, and stores the message in the staging data warehouse. The developed tool re-publishes the processed message to a new topic, where it can be consumed by the CEP system, as well as a Redis server for notification purposes on the web application.

To store all data regarding production lines, the staging data warehouse was designed using a star schema, as shown in Figure 16. This schema is composed of a fact table, which stores all messages from machines, sensors, and other devices, as well as a set of dimension tables, which store metadata and configuration information.

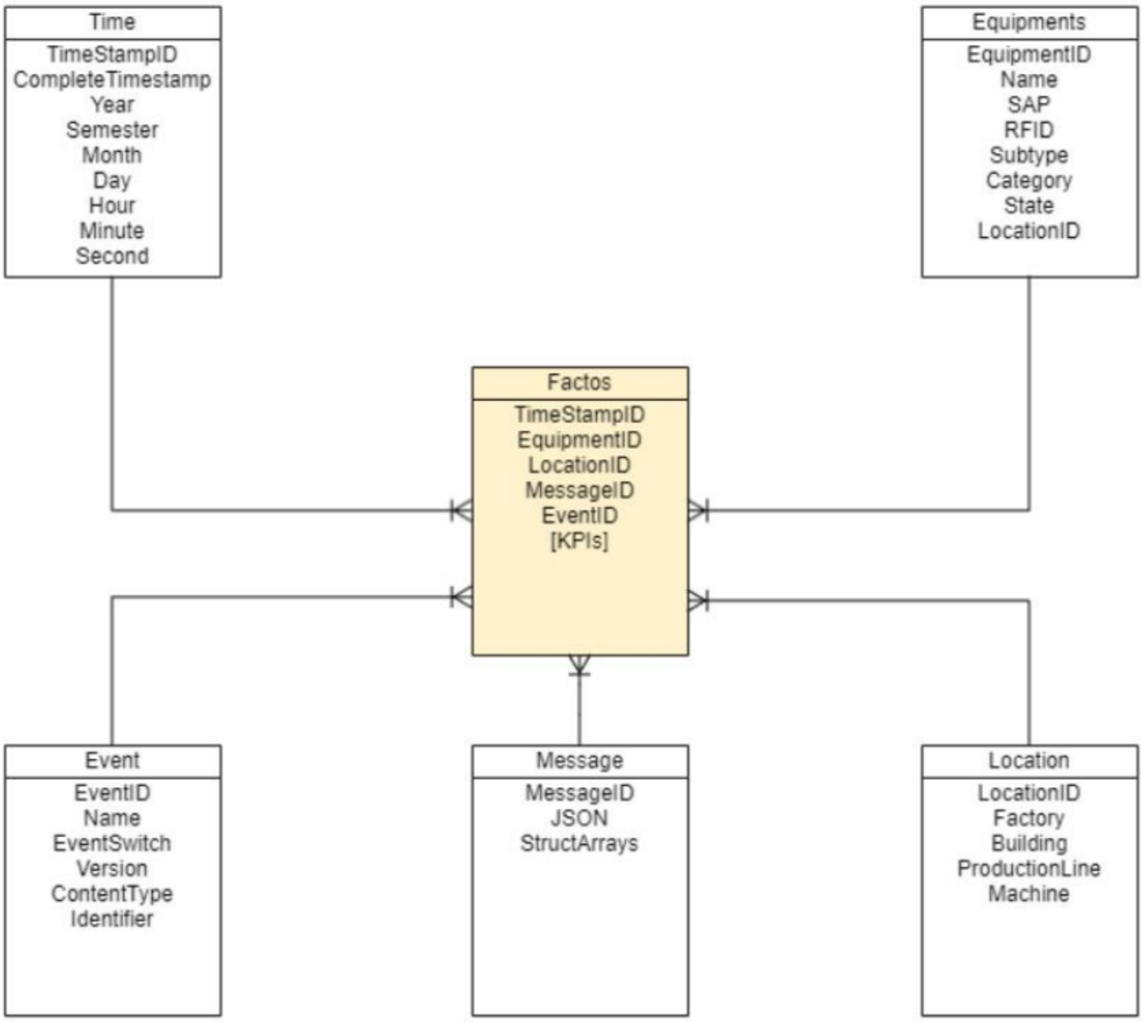


Figure 16: Data structure for staging data warehouse

#### 3.1.3.2 Web Application

The developed web application offers a suite of tools for the professionals at Bosch Car Multimedia to interact with production lines and equipment, as well as the data the machines create. This includes visualization dashboards for analytics and monitoring, with an intuitive interface. Along with integrations with the CEP system, the web application is able to notify users in real-time regarding any abnormality in the production lines. This application was developed using the LAMP stack, a combination of Linux, Apache, MySQL, and PHP. In addition, the Laravel framework, a PHP framework used for web application development, was extensively used across the project. Laravel's range of features and advantages, along with its documentation and community support, made it an ideal choice for this project.

The web application is composed of a backoffice and a frontoffice, each with its own set of functionalities and requirements. The backoffice is designed for users who need to manage Bosch Car Multimedia's factories and its production lines. Several modules were developed, namely the management of plants, buildings, floors, and production lines that are displayed to the users in the frontoffice. It was designed in a way that makes it easy to create interactive floor plans through the different layers of the factory, allowing end-users to navigate through the factory.

The frontoffice was designed for employees to monitor in real-time the parts that pass through production lines and equipment. Users can drill down into the factory plant, buildings, production lines, down to a specific machine. A set of graphics with relevant information about the readings of the processed parts in real-time were developed. Users are able to customize the graphics they want to see to the data the web application is receiving from the production lines. In addition, the web application is able to notify users in real-time regarding any abnormality in the production lines. This was acomplished using Lavarel Echo, a JavaScript library that allows the web application to receive real-time broadcasts from the CEP system.

Figures 17 and 18 show screenshots of the web application and the process of drilling down into the factory plant. A user can navigate through the different layers of the factory, from the factory to a specific machine. Users who have access to the backoffice can create and edit these interactive floor plans in an intuitive way, thus allowing the application to be easily adapted to the factory's needs.


Figure 17: Interactive factory plant map

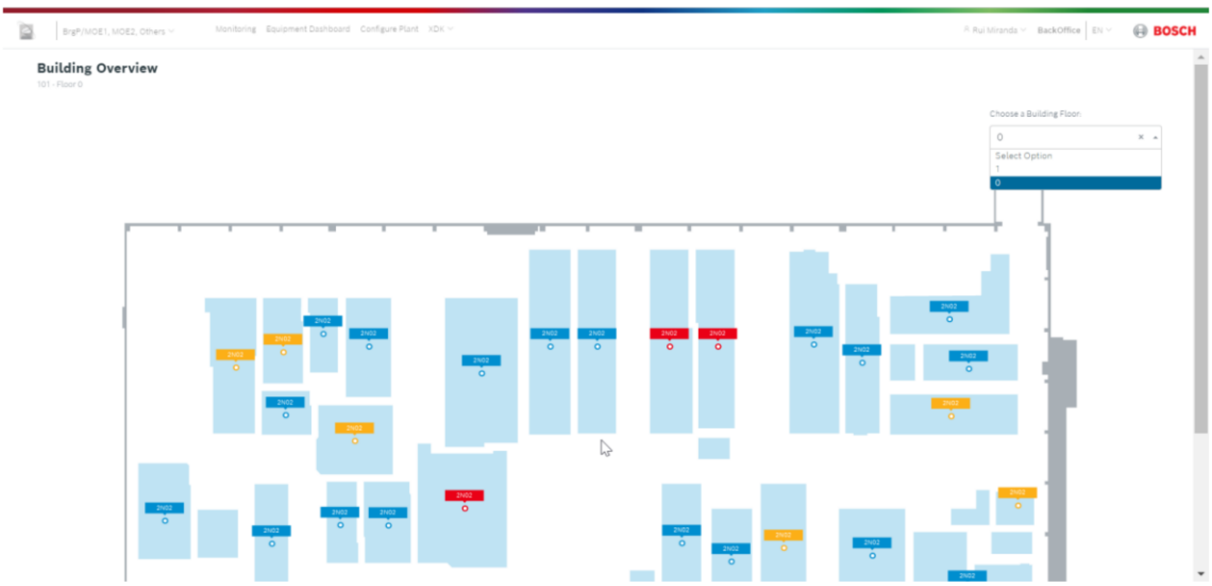


Figure 18: Interactive map of building production lines

Figure 19 and shows a screenshot of the web application, displaying the real-time dashboard from a Bosch XDK.

Two different types of dashboards were developed, one for XDK devices and another for generic equipment. The XDK dashboard shown on Figure 19 can display the real-time data from the built-in sensors, while Figure 20 shows how graphs can be configured to display data from a machine.

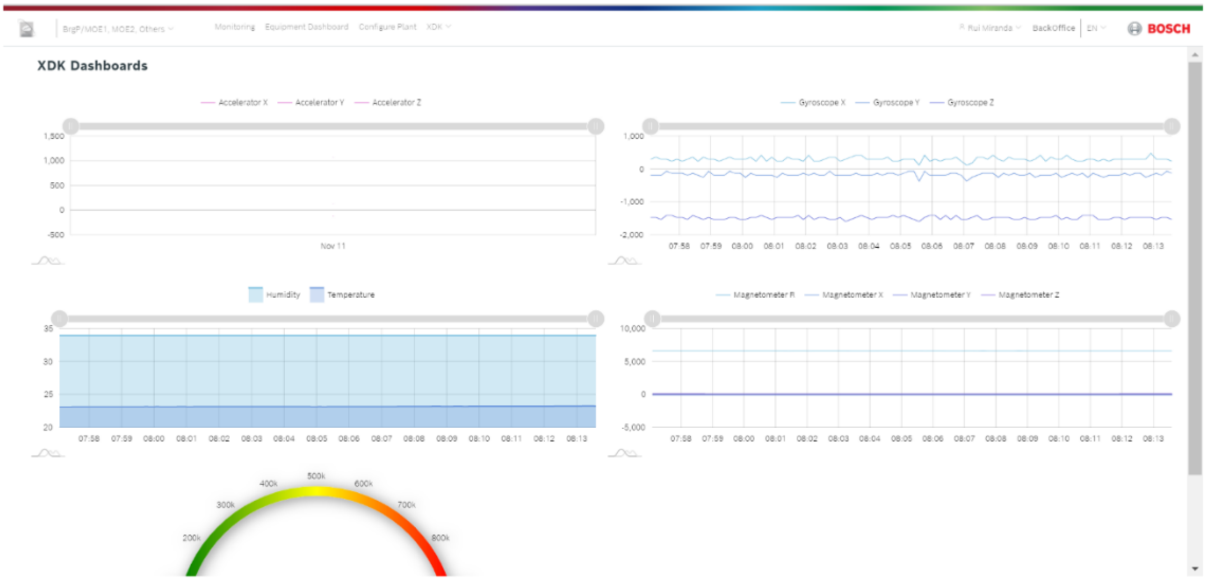


Figure 19: Real-time representation of XDK data flow

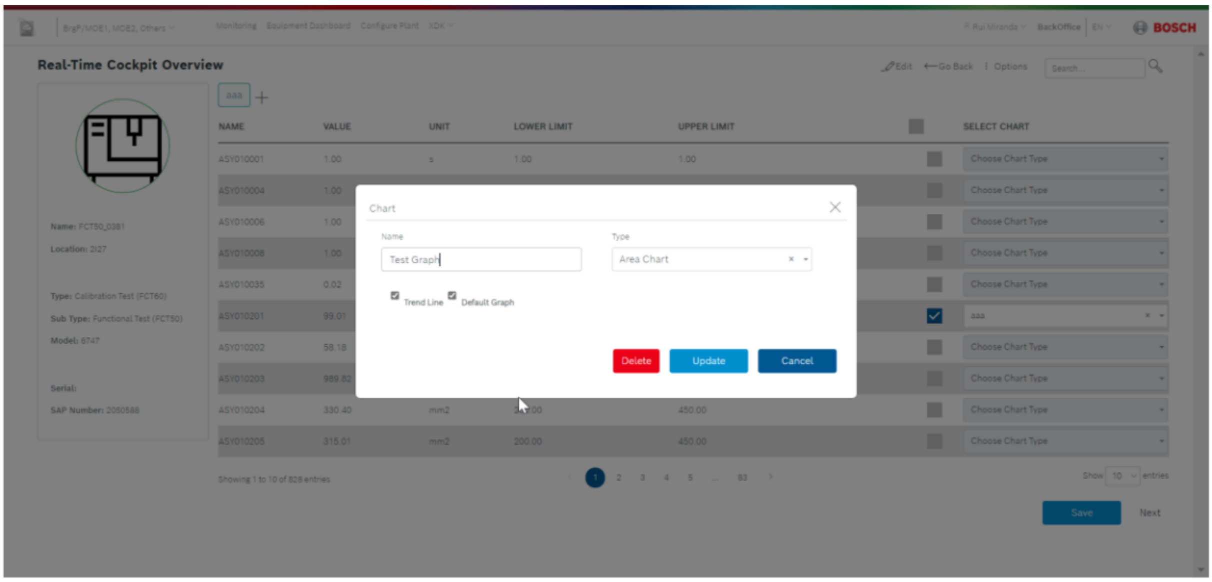


Figure 20: Configuration of a graph on a cockpit

#### 3.1.3.3 Complex Event Processing

Complex Event Processing plays an essential role in this project and has been designed to execute a sequence of procedures based on real-time rules. The developed CEP system leverages Redis, Apache Spark, Apache Druid, Apache Kafka, and Jupyter as its main tools.

Redis, an open-source in-memory data structure store, functions as a database, cache, and message broker. Redis is used to stream data from the processed parts of the production line in real-time from

Solace PubSub to Spark and Druid. Data from Redis, received through the developed data standardizer, is processed and stored in ingested into Druid or dispatched to Laravel Echo to alert users.

Furthermore, Spark processes all the data from MySQL (historical data) and Redis (real-time data). Spark is used to perform data analysis and machine learning. Jupyter is deployed to execute Python scripts that process the data on Spark. To automate the streaming of data from Spark to Druid, a Kafka instance was deployed.

Bosch has developed multiple use cases for Complex Event Processing and established rules for each use case, which trigger a set of actions when verified. The actions performed are sent as alerts to managers with information about the production line and equipment, along with a description of the alert generated. All proposed use cases were successfully implemented, requiring no significant data transformations as the data was already in a clean, usable format. UC1 focuses on identifying high energy consumption in equipment that is active but not in production or maintenance. UC2 deals with identifying shared display mismatches that can be reused, reducing waste in the process, while UC3 aims to detect and quantify gap filler waste and estimate its occurrence. These use cases can generate alerts that are sent to the web application, notifying users in real-time. The following figure shows the transformation that was done to the data gathered in real-time.

```
def processRealTimeDataFrame(messageType, dataframe):
  ## NEW additon here to process received telegram message ##
  df3 = dataframe.rdd.map(lambda x: \
    (x.event, x.header["timeStamp"], x.header["eventId"], x.header["eventSwitch"], x.header["eventName"], x.header["contentType"],x.header["version"],\
     x.items["fuNo"],x.items["statIdx"],x.items["productClass"],x.items["workPos"],x.items["dataType"],x.items["processNo"],x.items["processName"],x.items["processDescription"],\
     x.items["toolPos"],x.items["statNo"], x.items["name"], x.items["value"], x.items["groupFlag"],\
     x.location["fuNo"],x.location["statIdx"], x.location["lineNo"], x.location["application"],\
     x.location["workPos"],x.location["processName"], x.location["toolPos"], x.location["statNo"], x.location["processNo"],\
     x.structs["result"],x.structs["machineID"],x.structs["workingCode"],x.structs["typeNo"],x.structs["batch"],x.structs["typeVar"],x.structs["nioBits"]))\
     .toDF(["event","timeStamp","eventId", "eventSwitch", "eventName", "contentType", "version", \
           "fuNo", "statIdx", "productClass", "workPos", "dataType", "processNo", "processName", "processDescription",\
           "toolPos", "statNo", "name", "value", "groupFlag", "fuNoLocation", "statIdxLocation", "lineNoLocation", "applicationLocation", "workPosLocation", "processNameLocation",\
           "toolPosLocation", "statNoLocation", "processNoLocation", "result", "machineID", "workingCode", "typeNo", "batch", "typeVar", "nioBits"])

  df2 = dataframe.select(col("processData"), explode(col("processData")), col("header"))

  df4 = df2.rdd.map(lambda x: \
    (x.header["eventId"], x.col["timeStamp"], x.col["checkType"], x.col["result"], x.col["upLim"], x.col["unit"], x.col["dataType"], \
     x.col["setValue"], x.col["loLim"], x.col["name"], x.col["details"], x.col["value"])
    ).toDF(["eventId","processDataTimeStamp","processDataCheckType", "processDataResult", "processDataUpLim", "processDataUnit", "processDataDataType", "processDataSetValue", \
            "processDataLoLim", "processDataName", "processDataDetails", "processDataValue"])
  df_merged = df3.join(df4, "eventId")
  final_df = df_merged.withColumn("isHistory", lit('false'))
  #df_merged.show()


  ## Store processed telegram message to be ingested in druid ##
  final_df.coalesce(1).write.mode('append').format('json').save('/home/vea1brg/CEP/productionLine')
  print("Real time data stored successfully!")
```

Figure 21: Real-time ETL process

#### 3.1.3.4 Ontologies

In this project, ontologies were used in two ways, one about knowledge about factory data and the other regarding an invention report.

Two specific types of ontological representations are of particular interest to the outcomes achieved, namely relationship charts and instance graphs. Relationship charts are graphical tools that allow users to view the relationships between two or more classes, with shading increasing, according to the amount connections between classes. The number of connections for each class is also displayed. Figure 22 shows a relationship chart for the ontology developed for this project.

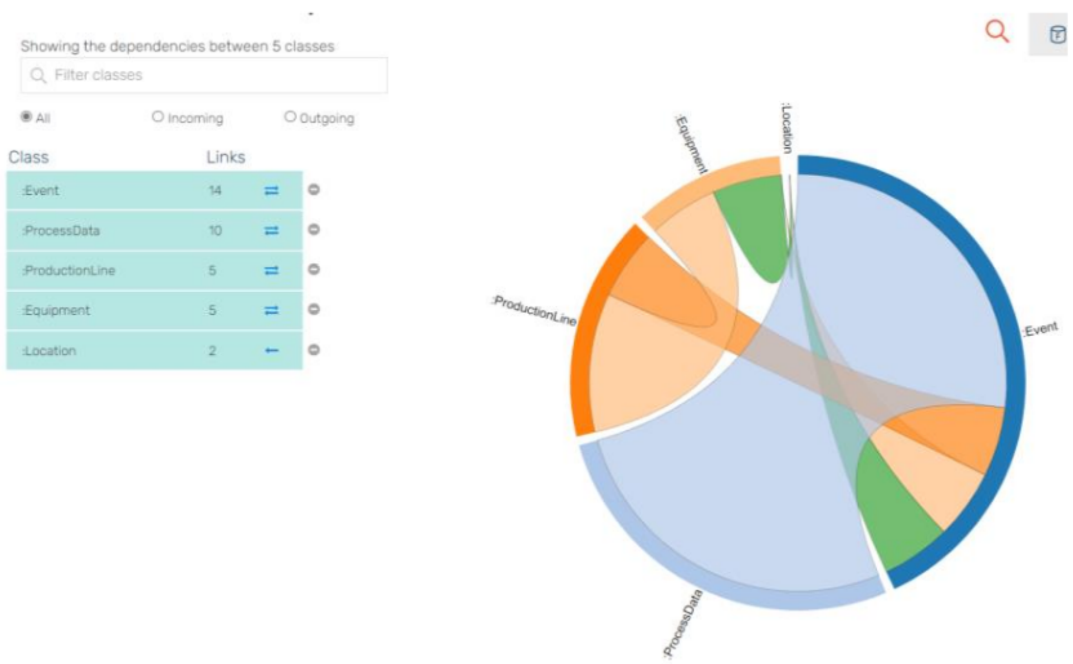


Figure 22: Relationship chart

On the other hand, an instance graph is a graphical tool that allows users to view the values of each class for each instance. To improve the state of ontologies, all attributes have data types, and two new classes have been built. Measurements and maintenance orders have been added to enhance the connection with the CEP system. An instance graph for the ontology created for this project is shown in Figure 23.


Figure 23: Instance graph

As a result, the graph of relationships was obtained, which blends the features of a relationship chart and an instance graph, resulting in the depicted figure.

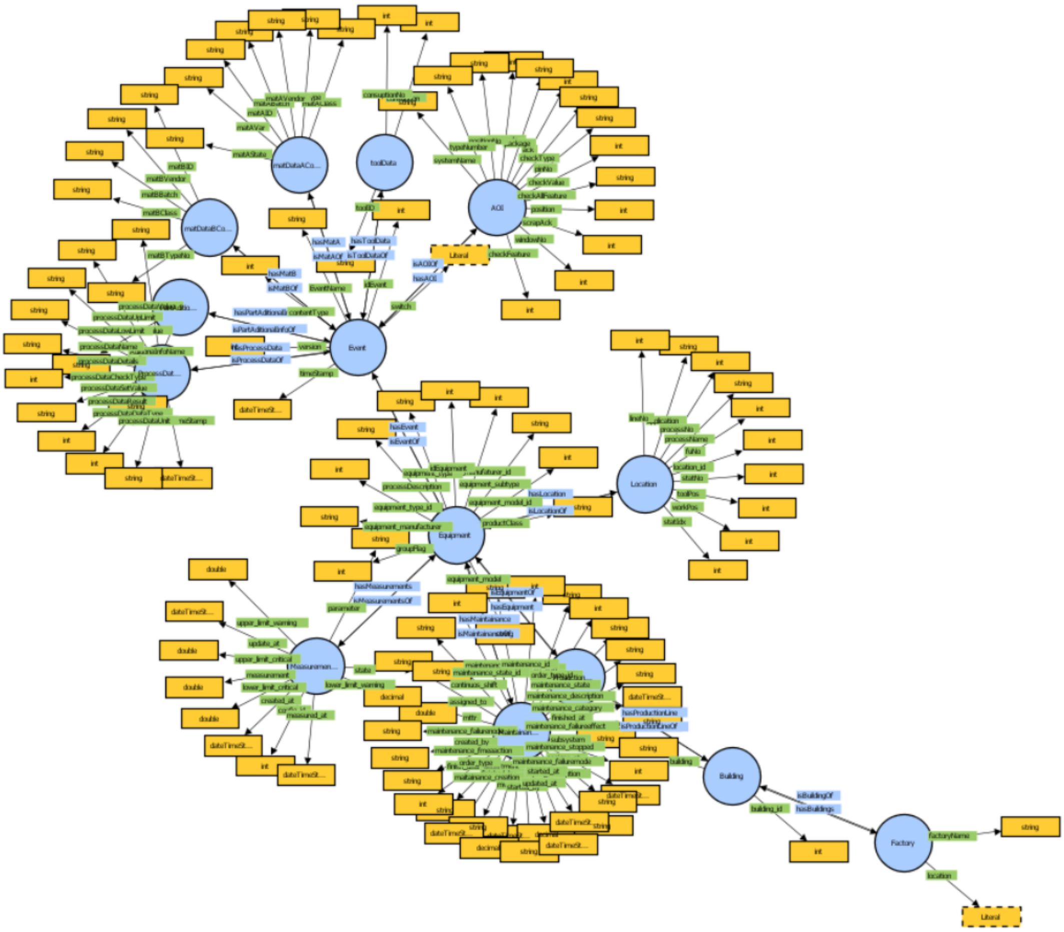


Figure 24: Entity relationship graph

Mid-July of 2021 marked the start of the development of the Ontology Invention Report, a collaborative effort between the University of Minho and the Bosch Car Multimedia teams, with each organization sharing equal copyright of the proposed invention. The aim behind the creation of the report was to showcase a novel technology for data processing within an industrial context. Over six months were spent ideating and drafting the report, which was finally approved for submission by both teams in November.

Upon receiving the document, Bosch relayed in December that their legal team would be reviewing it for a potential patent submission. The report underwent several phases, which included outlining the specifics of the product's intended purpose, researching inventions and technologies similar to the proposed invention, carrying out patent searches using relevant keywords, making incremental improvements via feedback and suggestions from both teams, executing proposed improvements to enhance the innovation's agility and robustness, reviewing and analyzing the final proposal, and developing a functional prototype.

### 3.1.4 SWOT Analysis

An intenal and external SWOT analysis was performed to assess the developed system's internal strengths and weaknesses, as well as the external opportunities and threats.

- **Strengths**
  - The sensorial ETL system can capture and analyze large amounts of data in real-time, allowing for rapid decision-making and increased efficiency.
  - The system can detect even the slightest changes in environmental and operational factors, leading to early detection of potential issues or opportunities for improvement.
  - Industry 5.0 environments prioritize the integration of advanced technologies like sensorial ETL systems, meaning there is significant market demand and potential for growth.
  - The system can be customized and optimized to meet the specific needs of different industries and organizations.
- **Weaknesses**
  - Implementing a sensorial ETL system requires significant investment in hardware, software, and training, making it difficult for some organizations to adopt.
  - The system is only as effective as the sensors and data it relies on, meaning there is a risk of inaccurate or incomplete data leading to faulty analysis.
  - Depending on the level of automation and data analysis, there may be concerns around data privacy and security.
  - The implementation of a sensorial ETL system may be seen as disruptive or overcomplicated by some employees or stakeholders, leading to resistance or reluctance to adopt.
- **Opportunities**
  - The increasing demand for Industry 5.0 solutions and advanced technologies presents significant growth opportunities for companies offering sensorial ETL systems.
  - As more organizations prioritize sustainability and energy efficiency, there is potential for sensorial ETL systems to help identify areas of improvement and reduce waste.
  - The system can be integrated with other technologies like artificial intelligence and machine learning to create even more advanced analysis and predictive capabilities.
  - There may be opportunities to partner with other organizations or industries to develop customized applications or solutions for specific use cases.
- **Threats**

- The rapid pace of technological change means that newer or more advanced systems may emerge and replace sensorial ETL systems.
- There is potential for data breaches or other security issues, which could lead to damaged reputation and loss of trust among customers.
- Depending on the industry or application, there may be regulatory barriers or standards that must be met in order to implement a sensorial ETL system.
- Some organizations may choose to rely on more traditional data collection and analysis methods, making it difficult to convince them of the value of a sensorial ETL system.

### 3.1.5 Conclusions

The fourth industrial revolution (Industry 4.0) was the beginning of a massive digital transformation for companies across the globe. Now, with the emergence of Industry 5.0, manufacturing companies are set to experience an even greater wave of technological advancements. The focus of Industry 5.0 is on the interaction between humans and machines. While still utilizing the latest technologies like artificial intelligence and machine learning, Industry 5.0 aims to keep the human touch on the manufacturing process. This approach emphasizes the importance of real-time data and its role in decision-making, supply chain management, and quality control. This paradign aligs with Bosch Car Multimedia's vision of creating intelligent solutions that enable drivers and passengers to interact with vehicles in the most intuitive way.

To stay ahead of the competition, Bosch Car Multimedia proposed the development of a real-time information processing approach that enables continuous analysis of data. A sensorial Extract, Transform, and Load (ETL) system was developed to collect and process data from various sources, including sensors, and transform it into a unified format that provides valuable insights into the manufacturing process. By accessing real-time data, Bosch Car Multimedia can detect and address issues as they arise, rather than waiting for them to become bigger problems. This allows the company to improve its quality control and reduce the cost of production significantly. It also enables businesses to better understand customer preferences and adjust production to meet demand.

In regard to the main research questions posed at the beginning of the project, the developed platform can increase the manufacturing process's efficiency, productivity, and quality by providing real-time alerts that can be used to detect and address issues as they arise. It can also improve safety and sustainability by reducing the amount of waste produced and the amount of energy consumed during the manufacturing process. In addition, the developed web application can be used to visualize production lines in real-time, providing a comprehensive overview of the manufacturing process.

For future work, the developed platform can be extended to include a wider variety of sensors, and video analytics could also be used to detect and classify defects in the manufacturing process. In addition, more sophisticated machine learning models can be developed to predict future trends and improve the quality of the products being manufactured. Finally, the platform can be integrated with other systems to provide a comprehensive view of the factory.

## 3.2 Crowdsensing on Smart Cities using Geofenced Notifications

### 3.2.1 Introduction

Smart cities are one of the latest trends in urban development. Using technology and innovations, smart cities can give higher efficiency and better living standards for its citizens. With features that include public transportation, waste management, and energy consumption, smart cities aim to offer a more sustainable and reliable living environment to those living in urban areas.

One aspect of smart cities that is gaining more attention is crowdsensing. Crowdsensing involves collecting data through various sensors and devices, including smartphones, to detect and interpret information about the environment. This information can include traffic congestion, air quality, and noise levels. By gathering this data, smart cities can develop strategies to improve the quality of life for their citizens, optimize public transportation, manage traffic congestion, and reduce pollution levels.

Another powerful feature that can be added to smart cities is geofencing. Geofencing technologies use positioning systems to create geographic boundaries, which can trigger notifications or alerts when a device enters or exits a specific area. This technology can be applied to smart cities to develop applications that send citizens notifications about relevant city updates or emergencies. For example, citizens can receive real-time updates about road closures, track the location of public transportation, and receive alerts about natural disasters or other emergencies.

Guimarães is a picturesque city located in the north of Portugal, known for its well-preserved medieval architecture and rich cultural heritage. It was designated as the European Capital of Culture in 2012 and has been recognized as a UNESCO World Heritage Site. With the aim of becoming a smart city, the city has become a hub for experimentation and innovation in the field of Internet of Things (IoT) in collaboration with the University of Minho and other partners. Some of the areas where sensors have been deployed in Guimarães include public transportation, shared bicycles and scooters, street lighting, energy production and consumption, air quality, and traffic flow.

In this context, a crowdsensing platform for smart cities that uses geofencing technology was proposed. This project aims to analyse and classify the data collected by sensors in Guimarães, automatically create geofences based on past insights, and send relevant notifications to citizens. These notifications can have a significant positive impact on the lives of citizens by making it easier to adjust their daily routines based on changing city conditions. For instance, drivers can receive alerts about traffic congestion or available parking spaces, and citizens can be warned about natural disasters or other emergencies.

This case study aims to develop the components for the management and storage of geofences, as well as the development of a mobile application that allows citizens to interact with the platform. That way, other research groups were responsible for the development of the data collection and analysis components, and the developed platform can be used to test and validate their research.

### 3.2.2 Methods

The proposed methodology involves several technologies to create a crowdsensing platform for smart cities. Firstly, data is collected from various sensors strategically placed in Guimarães. This data is then analysed and classified to extract useful insights and identify patterns. The insights obtained are then used to automatically create geofences in specific locations that are relevant to citizens, such as areas with high traffic congestion or areas prone to natural disaster.

Thus, the platform is composed of three main components, namely web services, a mobile application, and a backoffice. The web services are an interface that allows external applications to create, edit, and delete geofences. The mobile app allows citizens to interact with the platform, by exploring the city map and receiving notifications when they enter a geofence. Finally, the backoffice is a web application that allows the platform administrators to manage the geofences created.

For the development of these components, several tools and services from Firebase were used. Firebase is a mobile and web development platform by Google that provides a set of easy-to-use and scalable backend tools and services for app developers. The tools and services used in this project include Cloud Functions, Cloud Firestore, Firebase Hosting, and Firebase Authentication. Firestore is a cloud-hosted NoSQL database that allows developers to store and synchronize data in real-time across multiple clients and platforms. It was used to store the geofences created by administrators and external applications. Cloud Functions allow developers to run code in response to events triggered by Firebase features and HTTP requests. They were used to develop the RESTful web services that allow external applications to interact with the platform. Firebase Hosting is a web hosting service that allows developers to deploy web apps and static content to a global content-delivery network. It was used to host the backoffice web application. Finally, Firebase Authentication was used on the backoffice, proviving an easy-to-use and secure authentication system. Figure 25 shows the architecture of the platform, and technologies used in each component.

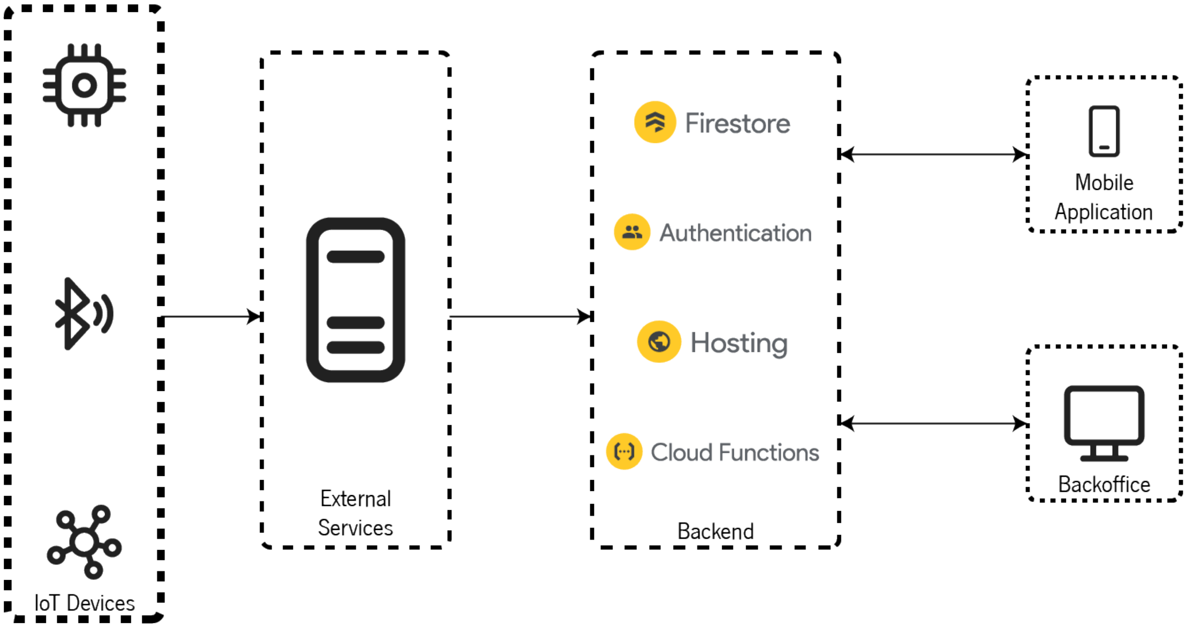


Figure 25: Architecture of the platform

### 3.2.3 Results

#### 3.2.3.1 Web Services

To develop the platform's web services, Firebase's Cloud Functions were used. Cloud Functions allow for the development of backends without having to worry about servers. These are single-purpose JavaScript functions that are executed in a secure environment on Node.js. They're only executed when a specific event to watch for is triggered. Deploying the code on Google's servers requires a single command. After that, Cloud Functions automatically scale computing resources to match the application's usage patterns.

Web services allow external applications to interact with the platform in a simple and secure way. The OpenAPI specification, formerly known as the Swagger specification, was used to document the endpoints developed. The OpenAPI specification is defined using YAML, allowing the definition of methods, parameters and models of the platform. Furthermore, by adopting the Swagger framework, an interactive documentation is be generated automatically, helping to keep documentation, clients, and code in sync. The following figure shows the Swagger web interface of the developed platform, allowing a live document of the system's web services.

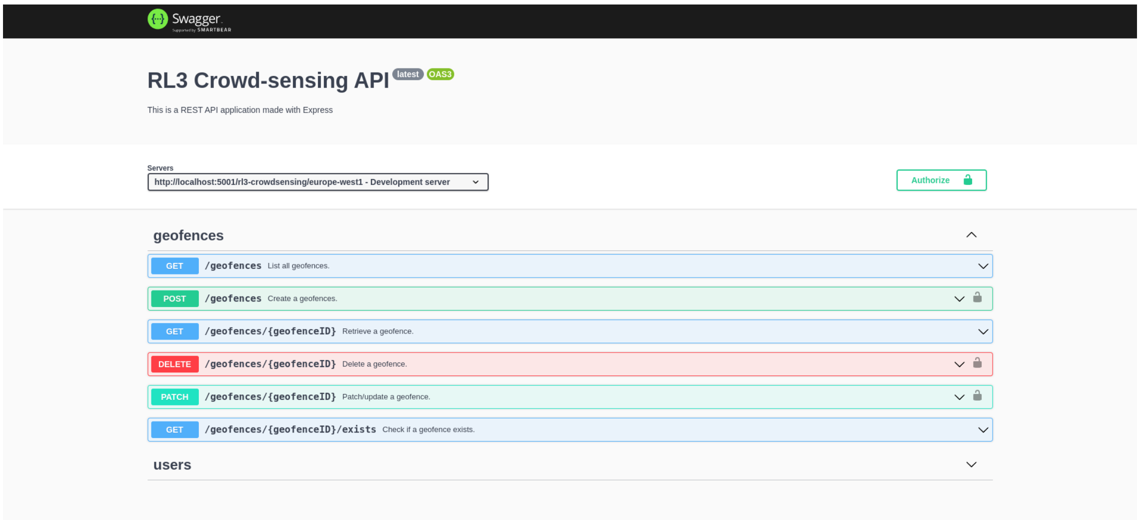


Figure 26: Web API documentation leveraging Swagger

The platform data, in particular the created geofences, are stored in Cloud Firestore. Cloud Firestore is a document database that allows data to be stored, synchronized and queried for mobile and web applications. The data is structured with collections and documents, allowing the creation of hierarchies to store and retrieve data using expressive queries. Cloud Firestore comes with SDKs for mobile and web devices, and a comprehensive set of security rules to make it possible to access the database without having to maintain servers.

#### 3.2.3.2 Mobile Application

The designed mobile application allows users to interact with the created geofences. The user can explore the city map and receive notifications when they enter a geofence on their phone. Since it will be developed using React Native, this application supports the two most commonly used mobile operating systems - Android and iOS - using the same code, saving time and resources. In addition, the Expo platform, which includes a set of tools and services for React Native, was used to develop the application.

To develop the prototypes of the mobile app, the InVision Studio screen design tool was adopted. InVision Studio allows teams to design, prototype and animate in the same application, while using the InVision platform to share and work together with each other.

The mobile application is divided into multiple components, namely onboarding, main screen, events, and settings. The onboarding component is the first screen the user sees when they open the application, showing a brief introduction to the platform, and its main features. The main screen is the main component of the application, showing the map of the city, and active geofences. The events component presents the user the events that are happening in the city, and the settings component allows users to change the application's settings. The following figures show screens of the developed mobile application.

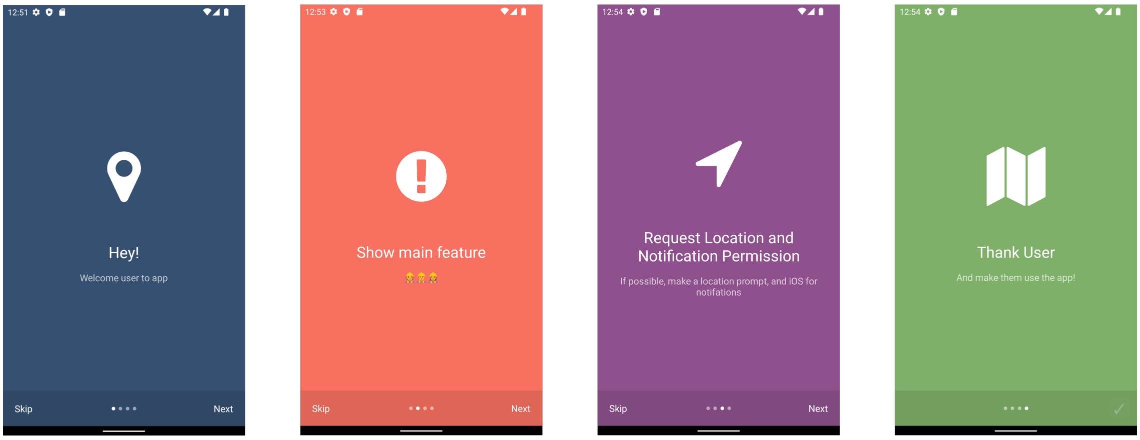


Figure 27: Onboarding interface running on an Android simulator

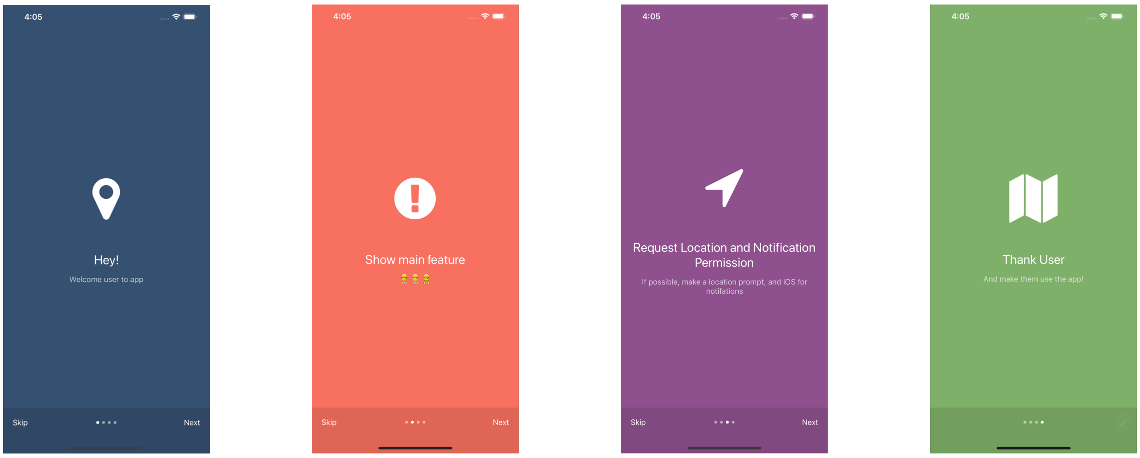


Figure 28: Onboarding interface running on an iOS simulator

The following images show the same app running on both iOS and Android operating systems:

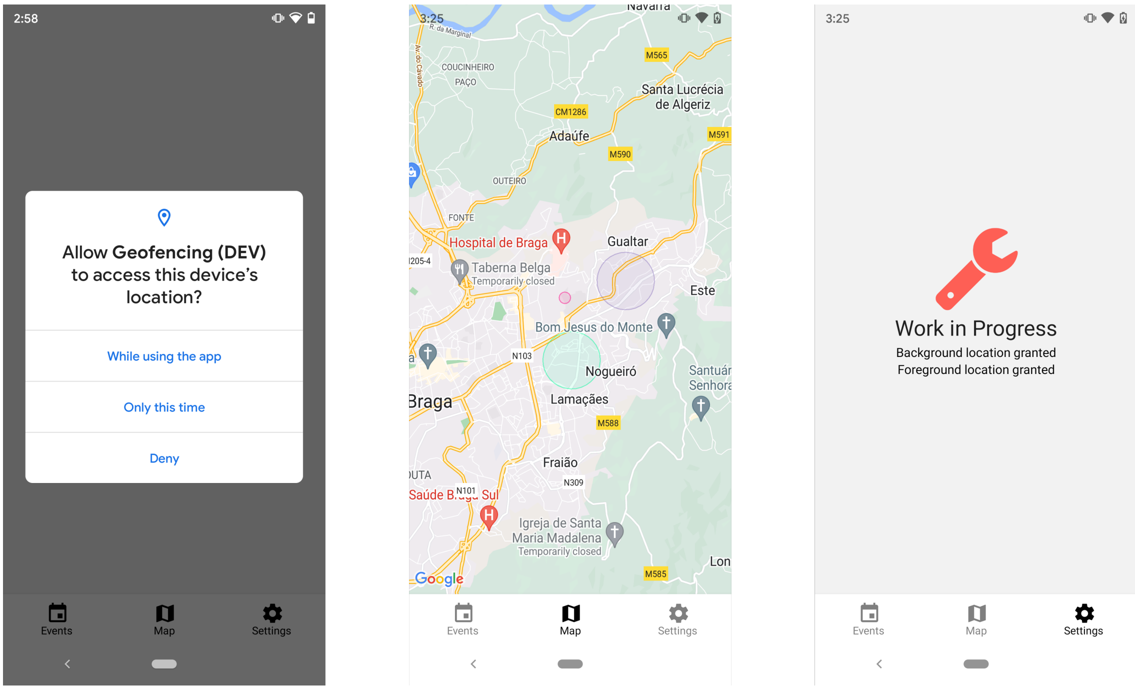


Figure 29: Application running on an Android device

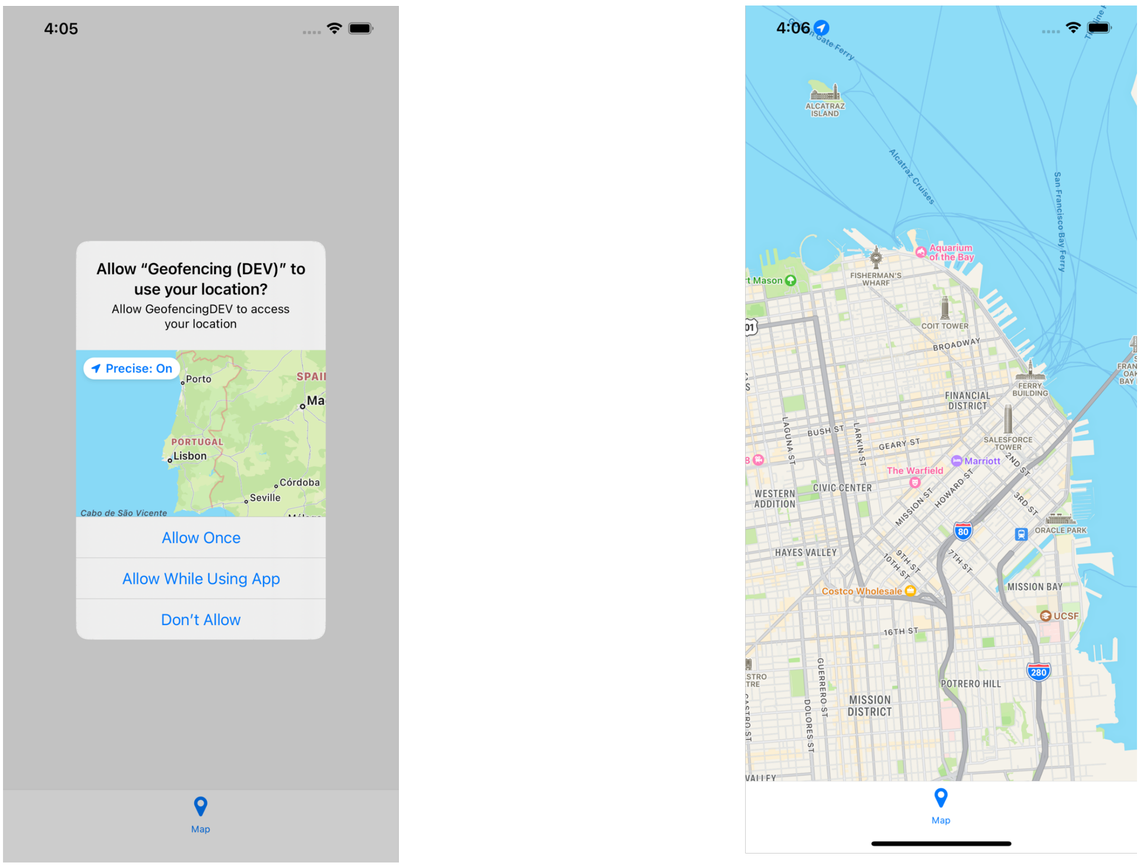


Figure 30: Application running on an iOS device

When a user enters a geofence, they receive a notification. The following image shows a screenshot of a phone receiving a notification when it enters or exits a created geofence:

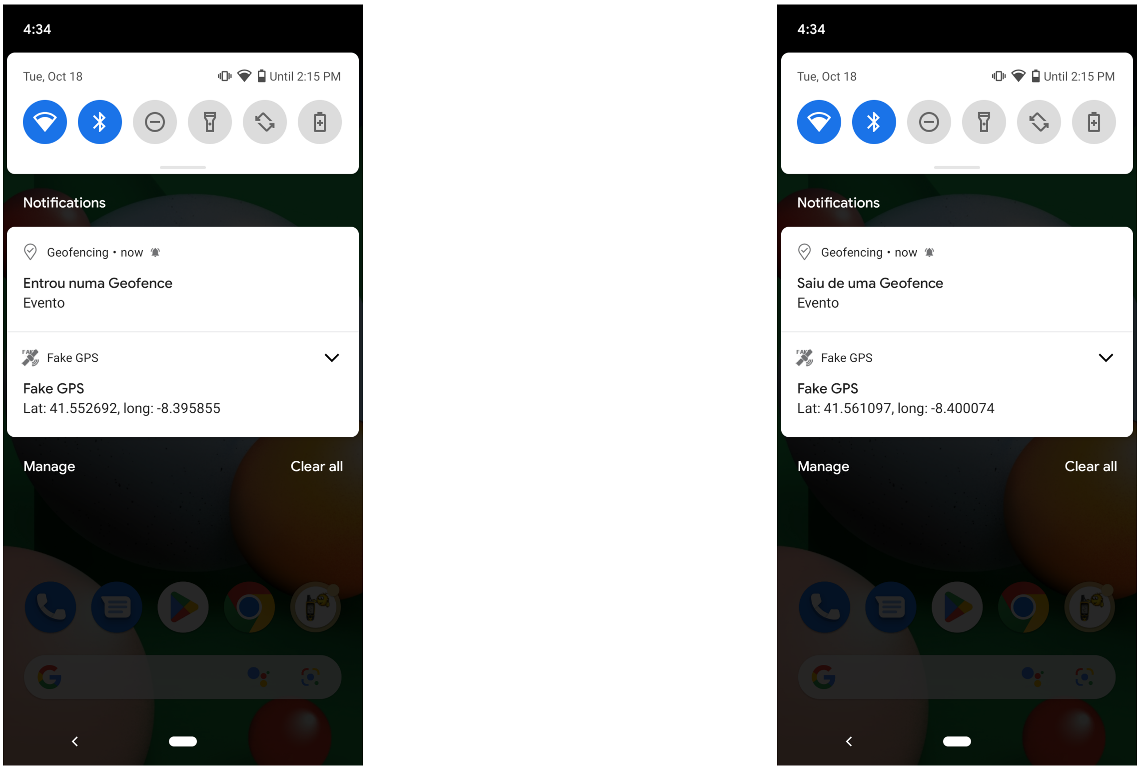


Figure 31: Notifications when an Android device enters or leaves a geofence

#### 3.2.3.3 Web Application

The developed web application will support administrators by offering a set of tools and modules for geofence planning and management.

By using the React library for JavaScript, a web application has been created that operates on a single HTML page and dynamically updates as user engagement changes (Wasson, 2013). By leveraging modern technologies and frameworks, additional features and modules can be seamlessly integrated. Common libraries that were often employed include Axios for facilitating HTTP requests, Formik for streamlining form building tasks, and React Router for declarative routing. The component structure of Tailwind UI was employed to create an aesthetically pleasing and responsive application. Moreover, the create-react-app tool was adopted, allowing for a modern development setup with no configuration.

The authentication system for the Backoffice application is fully functional, using Firebase Authenthication features, and hosted on Firebase Hosting. Figure 32 shows the login interface for the Backoffice application.


Figure 32: Login interface for backoffice

Using Tailwind UI and Leaflet components, an application interface was developed. After logging in, a page is displayed where it is possible to visualize the geofences stored in Firebase on a map. When the user clicks on each geofence, it is possible to view certain information such as the category and event type, as well as the respective coordinates of the geofence epicenter. Figures 33 and 34 show the developed interface.

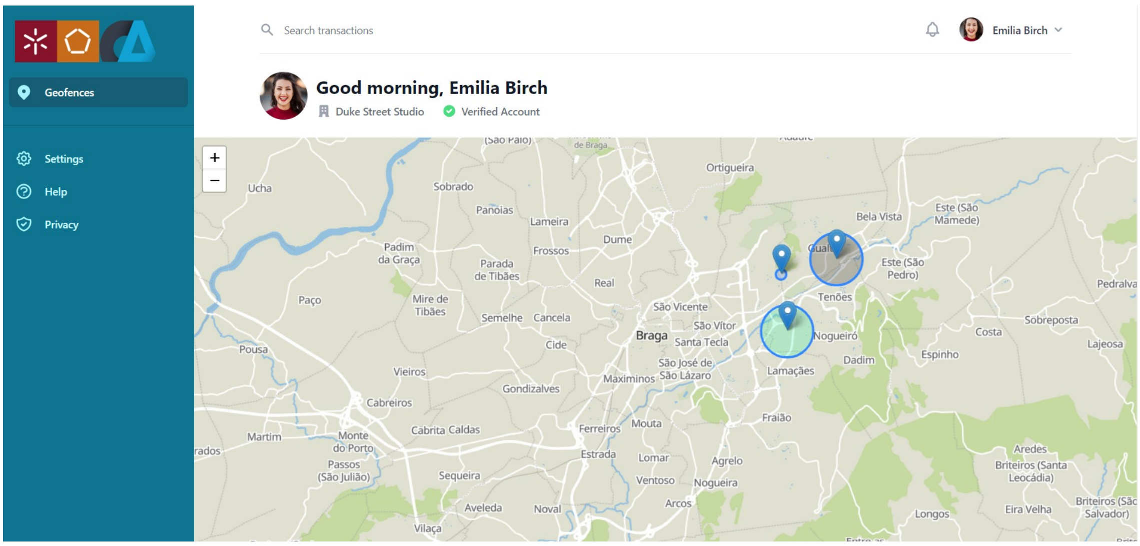


Figure 33: Backoffice

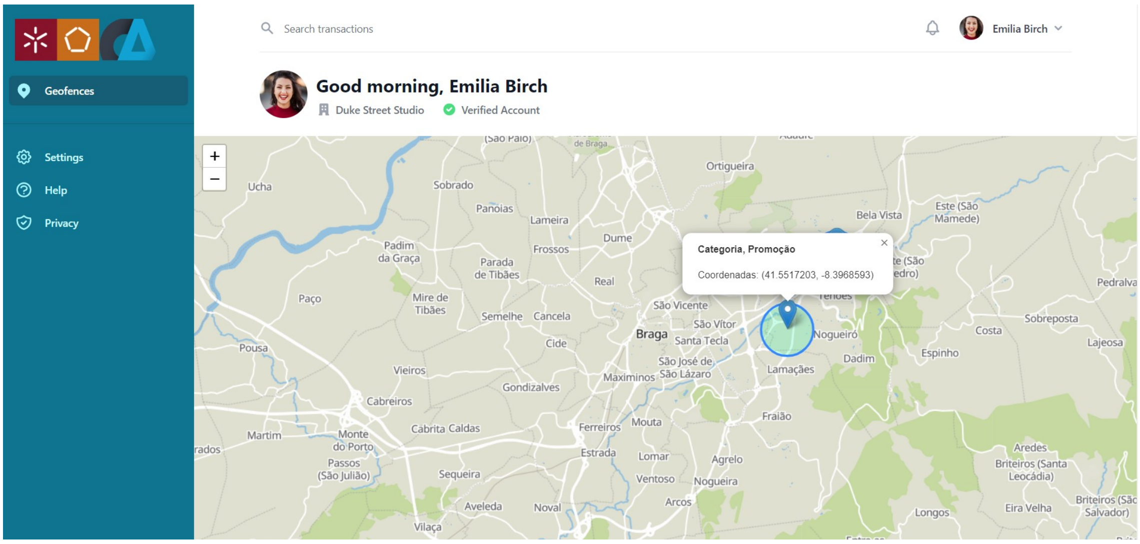


Figure 34: Geofence details

### 3.2.4 SWOT Analysis

An intenal and external SWOT analysis was performed to assess the developed system's internal strengths and weaknesses, as well as the external opportunities and threats.

- **Strengths**
  - Innovative technology for smart cities: The use of geofencing technology is a unique and innovative approach to smart city solutions, which can give the platform a competitive edge.
  - Positive impact on citizens: The notifications sent by the platform can have a significant positive impact on citizens' daily routines and improve their quality of life, which can enhance the platform's reputation and increase its usage.
  - Collaboration potential: Since the platform can be used to test and validate other research groups' work, it has great potential for collaboration with other research projects and organizations.
- **Weaknesses**
  - Dependence on data collection and analysis: The effectiveness of the platform is heavily reliant on the quality and accuracy of the data collected and analysed. Any issues in either of these components could significantly affect the platform's overall performance.
  - Potential privacy concerns: The use of location-based data and geofencing technology may raise privacy concerns among citizens, which could negatively affect the platform's adoption and usage.

  - Limited scope: The platform's functionality is currently limited to managing and storing geofences and sending notifications. The platform could benefit from the addition of more features and services to expand its capabilities and attract more users.

- **Opportunities**
  - Expansion into other cities: The platform could potentially be expanded into other cities, both nationally and internationally. This could provide new revenue streams and expand the platform's user base.
  - Integration with other smart city solutions: The platform could be integrated with other smart city solutions such as traffic management systems, waste management, and public transportation systems to provide an integrated solution to city management.
  - Monetization: The platform could be monetized through advertising, subscriptions, or partnerships with businesses and organizations that offer services related to the platform's notifications.

- **Threats**
  - Competition: The market for smart city solutions is highly competitive, and there may be other established companies that offer similar or better solutions, which could limit the platform's adoption and usage.
  - Regulatory changes: Changes in regulations around data privacy or location-based services could potentially limit the platform's use or introduce new legal challenges.
  - Technical challenges: The implementation of geofencing technology may pose technical challenges, such as connectivity issues or compatibility with different devices, which could affect the platform's usability and reliability.

### 3.2.5 Conclusions

The development of this project was driven by the wish to make Guimarães a smart city and improve the quality of life for its citizens. The project is ongoing, and the city continues to explore new ways of leveraging IoT technologies to make the city more efficient and improve the lives of its citizens.

The case study demonstrated the feasibility of implementing a crowdsensing platform for smart cities using geofencing technology in Guimarães. The platform's components have been successfully developed, creating an interface that allows external applications to create, edit, and delete geofences, a mobile application that allows citizens to interact with the platform, and a web application that allows the platform administrators to manage the geofences created.

However, this case study didn't cover the development of the data collection and analysis components, which are crucial for creating useful insights and identifying patterns. Moreover, improvements can be made to the platform's usability and scalability. For instance, the mobile application's user interface could be expanded to include more features, thus turning it into an application for citizens to interact with city services. Additionally, the platform's scalability should be tested under high-load conditions to ensure that it can handle many users and data.

Overall, the case study demonstrated the potential of using geofencing technology for smart cities and provided a foundation for future research. It showed that a crowdsensing platform can be developed and implemented to improve citizens' lives by providing relevant information that can help them adjust their daily routines.

## 3.3 Data Platform for Real-time Insights in Healthcare

### 3.3.1 Introduction

The healthcare industry is constantly evolving and advancing as the demand for quality healthcare services increases. A part of this evolution are IT solutions that aim to improve the quality of care provided to patients. From the digitalization of patient records and eletronic health records, which allowed for better organization and accessibility of medical data, to telemedicine, enabling patients to receive medical care remotely, IT solutions have been instrumental in improving the quality of care provided to patients. Furthermore, the use of IoT devices in healthcare has enabled the collection of data that can be used to check patients' health and provide real-time support to patients. Thus, data and its processing are crucial to healthcare, as they enable healthcare professionals to make informed decisions and offer timely, effective care. However, the healthcare industry is still facing several challenges in the adoption of IT solutions, such as the lack of interoperability between systems and lack of data quality (M. Miranda et al., 2009).

The Centro Hospitalar Universitário de Santo António (CHUSA) is a healthcare institution that provides healthcare services to the population of the city of Porto, Portugal. The institution is the result of the merger of the Centro Hospitalar Universitário do Porto and the Hospital de Magalhães Lemos, a psychiatric hospital. The center has frequent collaborations with the University of Minho, with the aim of developing and exploring new technologies and methodologies that can be applied to the healthcare industry. In particular, the Hospital de Santo António, one of the largest hospitals in the region, was a key partner in the development of AIDA (Agency for Integration, Dissemination and Archiving of Medical Information), an inteorperability platform that enables the integration of data from different sources and systems (Peixoto et al., 2012). AIDA is an essential component of the hospital's information system, integrating data from different sources and systems and providing a complete view of the patient's health record. Research projects regarding interoperability and data quality have been performed, as well as machine leaning and knowledge extraction in medical data.

After recognizing the importance of real-time data processing in decision-making, the health institution has undertaken a study that focuses on developing an architecture for the processing and storage of data in real-time to support decision-making. Patient monitors telemetry was chosen as a starting point for the case study, as it is a source of data that is constantly being generated and is crucial to patient care. Thus, the aim of the case study is to identify methods to process data from patient monitors in real-time, develop new approaches for aggregation, extraction, transformation, and loading of information in real-time, create a real-time analytics dashboard that incorporates an overview of the platform and its impact on decision-making, and establish automatic alerting to medical professionals when patient parameters are outliers.

The case study also aims to develop an architecture that will support the persistence of data for a sensorial ETL system through stateful data streaming. This will enable the real-time processing of patient monitor telemetry, allowing for immediate support to be given to patients and informed decisions to be made promptly. The architecture will also need to support communication protocols based on semantic interoperability that are sustained on the HL7 standard. This will enable other external applications to access and interact with the developed solution, increasing its usefulness and impact.

### 3.3.2 Methods

This case study was focused on using patient monitor telemetry as a source of data to develop a platform that can offer real-time insights in healthcare units. This data is crucial to patient care, as it provides healthcare professionals with information regarding the patient's health status, and was chosen as a starting point for the development of the platform. The Hospital de Santo António's care units are equipped with patient monitors from Mindray, a Chinese medical equipment manufacturer. These devices are capable of collecting and transmitting patient's vital signs in real-time to a centralized gateway that is connected to the hospital's network. This gateway stores the data to a Microsoft SQL database using the HL7 protocol, as well as a proprietary data format developed by Mindray. For the development of dashboards that can provide an overview of the heathcare unit, and alert medical professionals when patient parameters are outliers or critical, it is of importance to develop a solution that can retrieve the data from data sources as soon as they are available and process it in real-time.

This case study is composed of several components. The initial component are the patient monitors, which are responsible for collecting the patient's vital signs. Those vital signs are then transmitted to Mindray's eGateway, which then stores the data to a Microsoft SQL database. In addition, data from the hospital's EHR will also be used to give a complete view of the patient's health record. The first component to be built is the data processer, responsible for retrieving the data from the database as soon as vital signs are stored and providing a channel for alerts for a web application and external applications. The second component to develop is a comprehensive and user-friendly web application that not only offers an overview of the healthcare unit but also includes essential information such as the number of patients admitted and their vital signs measurement. This intuitive web application will give a detailed analysis of the healthcare units, showcasing real-time updates on patients and their constantly changing vital signs. Data integration and aggregation will be performed to ensure that the data is accurate and reliable. Moreover, this application will be equipped with an intelligent alert system. If any patient's vital signs reach critical levels, an instantaneous and highly visible alert will be generated and broadcast to all online users. This notification mechanism will guarantee that medical staff and relevant personnel are informed, enabling them to take immediate action and provide the necessary medical intervention. Figure 35 shows the architecture of the case study.

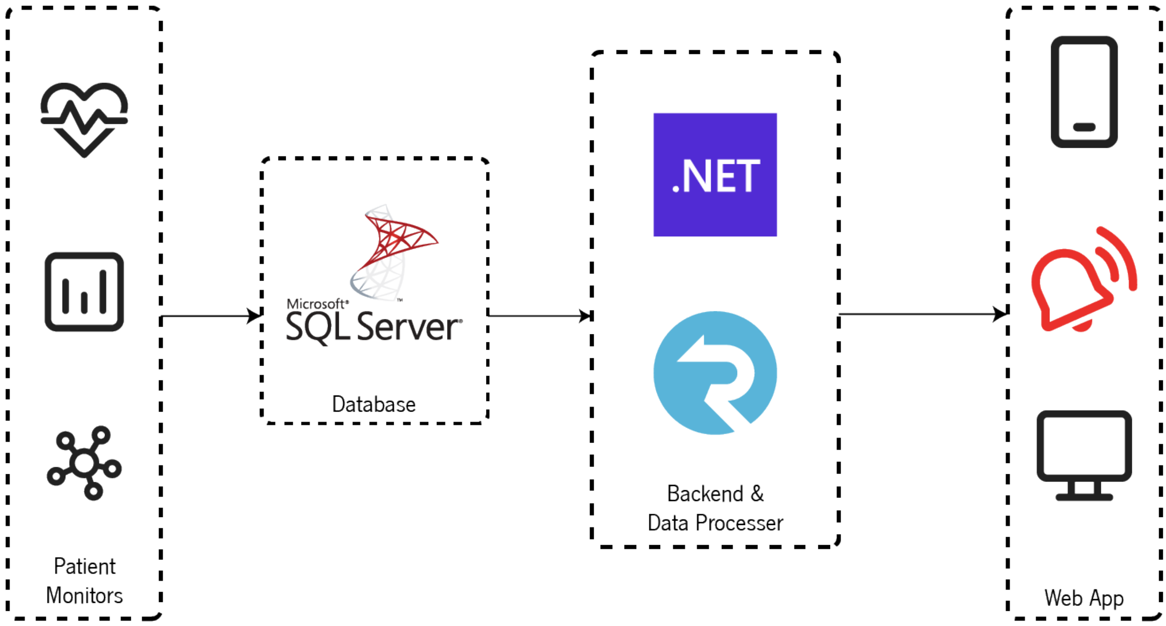


Figure 35: Architecture of the platform

### 3.3.3 Results

The data processer was programmed to continuously scan the patient monitor's database and retrieve new entries, allowing for low-latency alerts to be triggered whenever a patient's vital signs fall outside a predetermined range of normalities. This processer was developed using .NET, a cross-platform framework for building applications, as it includes a library for acessing table record change notifications from Microsoft SQL databases. This library was imperative in the development of a system that can offer medical professionals with real-time updates on their patients' statuses. In addition, the data processer also includes a SignalR hub, which is responsible for sending alerts to the web application and external applications. SignalR is an open-source library for adding real-time functionality to web applications, enabling server-side code to push content to clients instantly (Microsoft, 2023).

In addition to the data processer, a central component of the system is the web application that provides a comprehensive overview of the healthcare units. The application uses React, a JavaScript library for building user interfaces, to create a clean and interactive interface for displaying patient data.

Through this web-based application, healthcare professionals can quickly and efficiently monitor patient statistics. Figure 36 shows the dashboard of the developed web application.

The dashboard shown in Figure 36 presents an intuitive layout that allows healthcare professionals to access and analyze vital patient information. The web application offers various features that enhance usability and ease the monitoring process.

The key feature is the ability of offering real-time updates, ensuring that healthcare professionals have access to the most current patient data on any device. This dynamic aspect of the dashboard is crucial for making timely and informed medical decisions, as it allows immediate responses to patients' changing conditions.

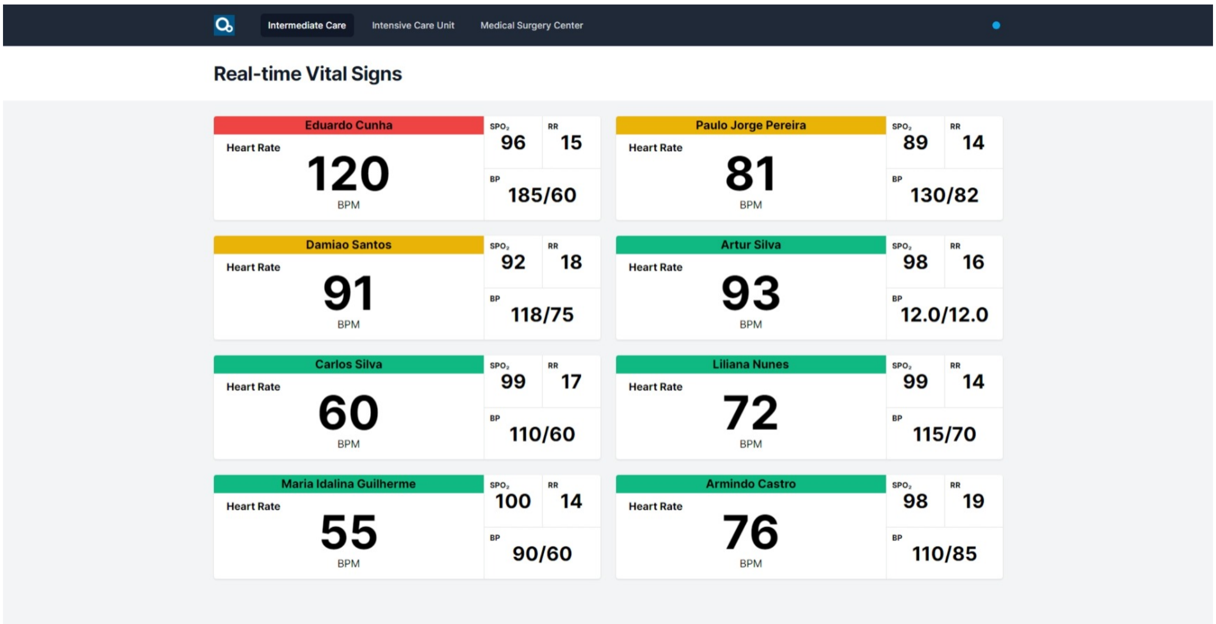


Figure 36: Vital signs dashboard

Upon reaching critical levels, the patient's signs are intelligently detected by the system, triggering the generation of the alert. This ensures that healthcare professionals are notified, allowing them to take immediate action and provide the necessary care to the patient.

The alert, depicted in Figure 37, is displayed prominently on the web application, capturing the attention of the healthcare professionals instantly. Its modal and vibrant colors contribute to its visibility.

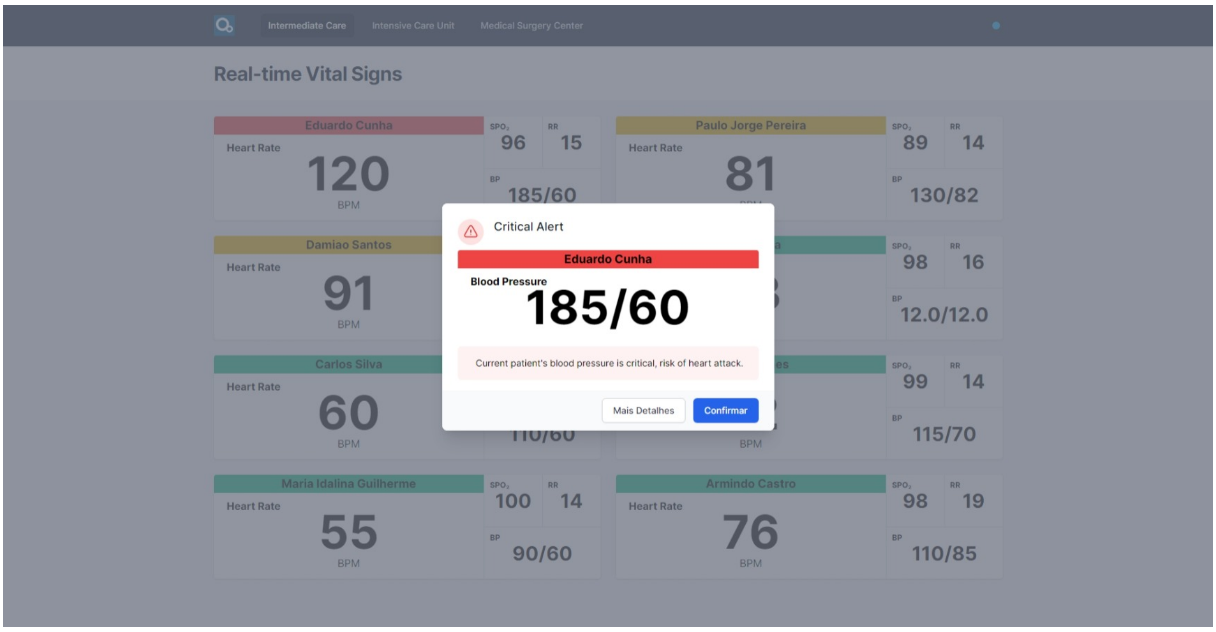


Figure 37: Alert for a patient's vital signs reaching critical levels

Overall, the developed platform is a powerful tool that combines the functionality of a data processor with a user-friendly interface to give healthcare professionals with a comprehensive overview of patient health. By harnessing the real-time information provided by patient monitors, it offers an optimized and efficient approach to healthcare data management, improving patient outcomes and healthcare delivery. With additional data sources and functionalities, this platform can be further developed to become a complete solution for healthcare data management.

### 3.3.4 SWOT Analysis

The final case study of this dissertation project was designed to address a critical need in the healthcare industry: the real-time processing of patient monitor data. To achieve this goal, a real-time analytics dashboard was created to offer medical professionals with a comprehensive overview of the platform and its impact on decision-making.

However, before the solution could be finalized and implemented, it was essential that a thorough SWOT analysis was undertaken. By understanding the potential risks and challenges associated with the solution, it is possible to develop a more robust and effective approach that would maximize its impact and benefit to the healthcare industry.

- **Strengths**
  - Real-time data processing allows for immediate decision-making, improving the quality of care provided to patients.

- The Hospital de Santo António is a leading healthcare institution in Porto, with a focus on using IT solutions to enhance healthcare services.
- The development of an interoperability platform, AIDA, demonstrates the hospital's commitment to improving data quality and access.
- The use of patient monitors and telemetry as the case study for the study is relevant to healthcare and can have a significant impact on patient care.

* **Weaknesses**
  - The lack of interoperability between systems and data quality issues remain a challenge in healthcare IT adoption.
  - The study is limited to patient monitors and may not address other sources of data that can affect patient care.
  - The study is conducted within the Hospital de Santo António, and its effectiveness in other healthcare institutions may vary.
  - The system's success is highly dependent on the accuracy and timeliness of vital sign data collected from patient monitors
* **Opportunities**
  - Real-time data processing can be applied to other areas of healthcare, such as medical imaging or electronic health records.
  - The development of a real-time analytics dashboard can increase transparency and improve communication among healthcare professionals.
  - The publication of an anonymized repository can contribute to the scientific community's development of new methodologies and technologies in healthcare IT.
  - The establishment of automatic alerting to medical professionals can improve response times and reduce adverse patient outcomes.
* **Threats**
  - The implementation of real-time data processing solutions may be costly and require significant investments in IT infrastructure and personnel training.
  - The regulatory environment surrounding healthcare data, such as GDPR, can hinder the development and implementation of IT solutions.
  - The issue of cybersecurity remains a significant concern, especially in healthcare, where patient data privacy is of utmost importance.
  - Competing technologies and IT solutions in healthcare may pose challenges to the adoption and success of real-time data processing systems.

### 3.3.5 Conclusions

The need for real-time processing of data from patient monitors is becoming increasingly important within the healthcare industry. Immediate support can make a critical difference in patient outcomes, and the ability to make informed decisions can mean the difference between life and death. However, this can also present a challenge as the amount of data generated can be overwhelming, and processing it in real-time can be complex and resource-intensive. This makes the development of an architecture for the processing and storage of data in real-time for decision support a crucial and necessary endeavor.

There are several challenges that need to be addressed in the development of an architecture for real-time data processing for decision-making. One of the primary challenges is ensuring that the data is accurate and reliable. Patient telemetry data can be highly variable, and errors can occur in acquisition and transmission. Therefore, it is important to establish protocols to ensure data accuracy and reliability.

Another challenge is developing algorithms that can extract, aggregate, transform, and load the data in real-time. This is necessary for real-time analytics dashboards that can provide an overview of the healthcare units and its impact on decision-making. Developing these algorithms can be complex and resource-intensive, and careful consideration needs to be given to ensure that the algorithms are accurate and reliable.

In conclusion, the development of an architecture for the processing and storage of data in real-time for decision support is vital to meet the demands of the healthcare industry, and this case study about patient monitors telemetry is an excellent example of this emerging field. The case study undertaken by the Centro Hospitalar Universitário de Santo António (CHUSA) in collaboration with the University of Minho is an opportunity to discover the main obstacles and necessities and identify methods to process data in real-time. The skills and knowledge gained from this case study can be applied to other areas of healthcare where real-time data processing and decision-making are crucial, and can lead to better patient outcomes and quality healthcare services.

# 4

# Final Considerations

The final chapter of this dissertation aims to summarise and present the main conclusions and contributions obtained through the development of software solutions for real-time decision support and insights in several domains. In addition, proposals for future work to improve the developed platforms and its applications are presented.

The following sections present a discussion of the results obtained in the case studies presented in this dissertation. Section 4.2 delves into a thorough examination of the research's key objectives, presenting a comprehensive discussion, and outlining how the results of the investigation have contributed to the fields of study. In Section 4.3, the perspectives for future work in this research topic are presented. Potential avenues for future research on the subject and insights on where further research could be directed to benefit the field are also discussed.

## 4.1 Discussion

In this section, a discussion of the results obtained in the case studies presented in this dissertation is presented. Each case study presents a different domain and a different set of topics and challenges. For example, the case study presented in Section 3.1 is focused on the manufacturing domain and the challenges of data integration and semantic interoperability. Section 3.2 presents a case study in the smart cities domain, where the challenges are related to the integration of data from different sources and the need for real-time insighs. Finally, Section 3.3 presents a case study in the healthcare domain, aiming to tackle the obstacles of data quality and privacy concerns. For a set of research questions, a case study is presented to answer them and to validate the proposed solutions.

Although the case studies are from different domains, they share the same goal: to provide real-time decision support and insights. Thus, it is still possible for research questions to be answered by more than a case study, if they're adapted to the domain in question. For a better understanding of the results, Table 4 presents a comparison of the case studies with their ability to answer the defined research questions.

Table 4: Comparion of the case studies with their ability to answer the research questions.

| | Manufacturing Case Study | Smart Cities Case Study | Healthcare Case Study |
|---|---|---|---|
| Manufacturing industry needs and challenges, and role of real-time sensor data in workflow optimization | ★★★★★ | ★★★ | ★★★★★ |
| Extract, process, and load raw data from multiple sensors considering their context | ★★★★★ | ★★★★★ | ★★★★★ |
| Maximizing production and minimizing costs through real-time sensor data processing | ★★★★★ | ★★★ | ★★★★ |
| Data types for real-time decision support processing and storage | ★★★★★ | ★★★★★ | ★★★★★ |
| How can real-time sensor data improve Smart Cities for a sustainable and efficient environment | ★★★★ | ★★★★★ | ★★★ |
| Use cases for real-time analytics dashboard, increasing user productivity, and optimizing CHUSA's resources | ★★★ | ★★★★★ | ★★★★★ |
| How does real-time sensor data enhance patient care? | ★★★ | ★★★★ | ★★★★★ |
| Can interoperability and ontologies develop new communication protocols for real-time distribution? | ★★★★★ | ★★★★ | ★★★★★ |

### 4.1.1 Case Study I: Bosch and the Factory of the Future

The developed platform is designed to be a flexible and scalable solution for the manufacturing industry, allowing the extraction and analysis of real-time data from multiple sensors. With the development of this platform, Bosch Car Multimedia can now have an overview of all production lines, allowing them to identify and solve anomalies as they occur. The architecture of the platform is designed to be extensible, adapting to the needs of the manufacturing industry. Furthermore, the platform supports the implementation of machine learning approaches, and the development of ontologies allows for the creation of a knowledge base for factory data. Thus, the research questions that were initially posed can now be answered, based on the findings obtained in this use.

*What are the manufacturing industry's main necessities and obstacles, and how can the extraction and analysis of real-time sensorial data help optimize its workflow and reduce costs?*

During the development of this project, the main necessities for the manufacturing industry were elicitated, namely their efficiency, productivity, and quality, as well as safety and sustainability. However, some of their main obstacles were also identified, including operataional downtime and maintenance costs, human error, and ineficcient supply chain management. The implementation of real-time sensorial data extraction and analysis tools can be used to optimize the manufacturing workflow and costs, imrproving the quality of the products and reducing waste. Several case studies were developed to showcase the potential of the platform, focusing on energy consumption, reuse of materials, and waste reduction.

*How to extract the untreated data from multiple sensors, process said data, and load it to relevant repositories having into account their context?*

To sucessfully extract untreated data from multiple sensors, process said data, and load it to relevant repositories, a structured approach to data management is required. First, all the sensors that generate data must be identified, and the metadata they produce must be determined. This will help in deciding how to process and store the data. Then, the raw data must be collected from all the sensors and stored in a central repository. It may be in different formats and may need to be converted to a common format for processing. Afterwards, the data must be cleaned by removing any redundant or corrupted data, and normalized by scaling and standardizing it. Additionally, any outliers or missing data points must be identified and handled appropriately. These steps were performed in the developed data standardizer, which is responsible for collecting and processing the data.

The preprocessed data must be processed by applying statistical and analytical techniques to derive insights and identify patterns. This includes data transformation, filtering, and feature selection. The data must also be contextualized by considering the time, location, and environmental conditions in which it was collected.

In addition, the data must be integrated with other data sources to provide a comprehensive view of the information. Finally, the processed and contextualized data must be stored in relevant repositories, such as a data warehouse, data lake, or data mart. This data must be accessible and can be used for further analysis or data modeling.

*How can real-time sensor data processing deliver new opportunities to optimize production and reduce costs?*

Real-time sensor data processing can provide manufacturers with valuable insights that can help them optimize production, reduce costs, and improve efficiency. An example of this is the use of real-time data to keep a check on equipment and processes, and identify potential problems before they occur. Maintenance can be scheduled before a problem occurs, helping manufacturers reduce downtime and maintenance costs. Sensors can also be used to monitor the quality of the products being manufactured, and identify any defects as they occur. This can help manufacturers improve the quality of their products and reduce waste. Finally, real-time data can be used to optimize the supply chain, by tracking inventory levels and automatically reorder materials when inventory levels fall below a certain threshold.

In addition, other research questions can also be answered based on the findings obtained in this case study, namely:

*What types of data can be effectively processed and stored in real-time on a platform for decision support?*

The data standardizer developed in this project is able of processing and storing various types of data in real-time. These include sensor data from Bosch XDKs deployed around production lines, gathering telemetry data from various devices across production lines, including temperature, humidity, pressure, acceleration, magnetometer, gyroscope, and light sensor data. In addition, the platform processes and store operational data related to factory processes, logistics, production process, industrial operations management, and business intelligence, while the staging data warehouse serves as a repository for processed and transformed data, allowing for historical analysis and querying of past events. Finally, the platform has enabled additional use cases, as described in Subsection 3.1.3.3,

Thus, the platform can effectively process and store various types of sensor data, telemetry data, operational data, historical data, and data related to quality control, production processes, industrial operations management, and business intelligence in real-time for decision support.

*Can interoperability and ontologies help develop new communication protocols for distributed real-time communication protocols*

Interoperability and ontologies have been proven to help develop new communication protocols for distributed real-time communication protocols, as demonstrated in this study case.

The platform that was developed includes Complex Event Processing (CEP) for querying and processing data before storing it, a web application for data representation, and ontologies for effective communication. The CEP system allows for real-time data processing and analysis, while ontologies give a standardized framework for organizing and interpreting complex events.

By adopting interoperable technologies, such as Solace PubSub and Apache Spark, and incorporating ontologies into their system, Bosch can optimize their production lines, reduce downtime, respond to changing market demands, and improve their competitive edge in the market. Therefore, interoperability and ontologies play a crucial role in developing new communication protocols for distributed real-time communication protocols in this context.

### 4.1.2 Case Study II: Crowdsensing on Smart Cities using Geofenced Notifications

The development of this project was a collaboration between the city of Guimarães, the University of Minho, aiming to improve the quality of life for citizens by providing alerts and notifications that are relevant for their location.

The first step in the development of the project was the identification of areas where sensors were deployed to collect relevant data. This was followed by the development of data collection and analysis processes, which were designed to identify patterns and discover insights.

While data collection and analysis were being performed, the development of a geofencing platform had already begun. The platform was aimed at automatically creating geofences based on past insights and sending relevant notifications to citizens. The notifications could be about traffic congestion, available parking spaces, air pollution and other city conditions.

This case study was responsible for the management and storage of geofences, and the development of a mobile application that allows citizens to interact with the platform. The platform was designed to be scalable and can be used to test and validate research on data collection and analysis, and the application of geofencing technology. Thus, the research questions that were initially posed can now be answered, based on the findings obtained in this use.

*What types of data can be effectively processed and stored in real-time on a platform for decision support?*

Real-time platforms for decision support are designed to process and analyse large volumes of data in real-time, allowing businesses and organizations to make informed decisions. In this case study, data from sensors installed in public transportation, shared bicycles and scooters, street lighting, energy production and consumption, air quality, and traffic flow was collected and analyzed. This data was in multiple types and formats, including structured, time-series and geospatial data. In addition, this data could either streaming or batch data. Other types of data that can be effectively processed and stored in real-time on a platform for decision support include unstructured data (such as text data from social media or video feeds from surveillance cameras), streaming data and machine data.

Thus, real-time platforms for decision support are designed to handle diverse types of data, as long as they're configured and optimized for the specific case study.

*How can real-time sensor data processing deliver new opportunities to Smart Cities and contribute to the development of a more sustainable and efficient urban environment?*

During the development of this case study, it was found that real-time sensor data processing can deliver new opportunities to smart cities and contribute to the development of a more sustainable and efficient urban environment in several ways, including monitoring and control, decision-making, improved public services, and citizen engagement. By using real-time approaches, cities can check and control various systems like traffic, energy, and water supply. This data can be used to optimize resource allocations, identify and resolve issues, and reduce waste and inefficiencies. In addition, citizens can be better informed of local conditions and events, and can participate in community initiatives. This can lead to improved public services and increased citizen engagement.

Furthermore, this case study can be used to answer other research questions that were posed for this dissertation, with some adaptations:

*How to extract the untreated data from multiple sensors, process said data, and load it to relevant repositories having into account their context?*

The first step in the development of this case study was identifying all areas where sensors were deployed to collect relevant data. Multiple sensors were strategically placed in Guimarães, such as those deployed in public transportation, shared bicycles, and scooters. Note that street lighting, energy production and consumption, air quality, and traffic flow sensors were already deployed to collect data. This data includes information about key performance indicators.

The next step was the development of data collection and analysis processes, to extract useful insights and identify patterns. This analysis and classification process involves statistical techniques, machine learning algorithms, and data processing methods to derive meaningful information from the raw data.

Based on the insights obtained from the data analysis, geofences are automatically created in specific locations that are relevant to citizens. For example, this can include areas with high traffic congestion or areas prone to natural disasters. Geofencing technology, which uses positioning systems to create geographic boundaries, is used to define these areas. The created geofences are designed to consider the context of the relevant repositories, meaning that the geofences are tailored to the specific needs and goals of the smart city environment in Guimarães.

Finally, the extracted and processed data, along with the created geofences, are integrated into the crowdsensing platform for smart cities. This integration involved the development of web services, a mobile application, and a backoffice to allow citizens, administrators, and external applications to interact with the platform and make use of the collected data and geofences.

*What are the case studies for a real-time analytics dashboard, increasing user productivity, as well as optimising the resources of Guimarães?*

The developed platform includes a web application that allows administrators to monitor and manage the created geofences. Futhermore, users can be aware of how many users that have the application installed are currently inside a geofence. This information can be used to optimize resource allocation and improve the efficiency of the city. These resources include public transportation, waste management, energy consumption, and citizen safety.

For example, by collecting data on traffic flow, congestion, and road conditions, a real-time analytics dashboard can help optimize traffic management strategies. This includes identifying bottlenecks, suggesting alternate routes, and providing real-time updates to drivers by delivering intelligent notifications to their smartphones. This can lead to reduced congestion, shorter travel times, and improved overall road safety.

Another use case for the platform is its ability to integrate data from various sensors and devices to enhance citizen safety. For example, the dashboard can watch air quality, noise levels, and weather conditions to provide real-time alerts and recommendations to citizens. In case of emergencies, such as natural disasters or accidents, municipalities can send immediate notifications to citizens, guiding them to safety and providing assistance.

Finally, the platform can be used to check and manage public transportation systems. A real-time analytics system can gather data from sensors placed on public transportation vehicles, such as buses and trains, to track their location and watch their performance. This data can be analyzed in real-time to optimize routes, improve scheduling, and give insights for resource allocation.

Thus, the developed real-time analytics dashboard deployed in Guimarães can improve user productivity by providing timely and relevant information, optimize the allocation of resources, and create a more efficient and sustainable living environment for its citizens.

*How can real-time sensor data processing deliver new opportunities to optimize citizen's quality of care?*

Real-time sensor data processing in the context of smart cities can deliver new opportunities to optimize citizens' quality of care in several ways. By collecting data through various sensors and devices, including smartphones, smart cities can gather information about the environment such as traffic congestion, air quality, and noise levels. This data can be analyzed and used to develop strategies to improve the quality of life for citizens.

In the specific case of Guimarães, aiming to become a smart city, a crowdsensing platform using geofencing technology was proposed. This platform aimed to automatically create geofences based on past insights and send relevant notifications to citizens. By providing citizens with real-time information about traffic congestion, available parking spaces, and natural disasters, the platform can significantly improve citizens' quality of care and enhance their overall living experience.

### 4.1.3 Case Study III: Data Platform for Real-time Insights in Healthcare

The healthcare industry is continuously seeking ways to improve the quality of care provided to patients, and one of the key ways to achieve this is through the adoption of IT solutions. These solutions have proven to be effective in improving patient outcomes and operational efficiency, by providing better access to information and enabling more effective communication between healthcare providers. Providing real-time insights into patient data is the next step in this evolution, as it allows for more timely interventions and better decision-making. However, the healthcare industry still faces challenges, such as the interoperability of systems and data quality. These challenges must be addressed before real-time analytics can be fully implemented in healthcare settings.

This case study aims to identify the value of real-time analytics in healthcare, and address the challenges faced by the healthcare industry in adopting these tools. The study focuses on patient monitors telemetry, as it is one of the most widely used technologies in healthcare settings, and the data it generates is critical for patient care. The study also aims to develop an architecture to support the persistence of data for a sensorial Extract, Transform, and Load (ETL) system. This architecture will be used to store vital signs from patient monitors, and will be integrated with existing Centro Hospitalar Universitário de Santo António (CHUSA) systems. Finally, the study aims to develop an analytics dashboard to aid decision-making, by providing real-time insights into patient data.

The research questions addressed in this case study and their answers are presented below:

*What are the use cases for a real-time analytics dashboard, increasing user productivity, as well as optimising CHUSA's resources?*

The developed real-time analytics dashboard can be used to monitor patient vital signs, provide alerts when abnormal values are detected, and offer insights into trends and patterns in patient data. This will help health professionals make better-informed decisions about patient care, and improve the quality of care provided to patients. Although this platform now only supports patient monitors telemetry, it can be extended to support other types of data sources, for example, EHR systems, to give a more comprehensive view of patient health, and additional sources of data for analysis.

Furthermore, these types of dashboards can be used to monitor the performance of the hospital as a whole, by tracking key performance indicators such as patient arrival and treatment timings, resource usage, and staff productivity. Regarding human resources, the dashboard can be used to monitor the workload of staff and highlight areas that require additional support, ensuring that productivity levels are maximized. Finally, health institutions can track the usage of resources such as hospital beds and medical equipment to optimize their deployment and reduce instances of underutilization, overusage, and waste. This will allow the hospital to identify areas for improvement and take proactive measures to address them, before they escalate into larger, systemic issues.

Overall, the use of a real-time analytics dashboards can increase user productivity by providing timely insights and actionable data, as well as optimize the hospital's resources by improving patient flow and staff productivity, with the ultimate goal of improving patient care.

*How can real-time sensor data processing deliver new opportunities to optimize patient's quality of care?*

Real-time sensor data processing can deliver new opportunities to optimize patient's quality of care in multiple ways, by providing insights into patient health and behavior that were previously unavailable, tailored to the specific needs of each patient. Sensor data has been used to detect health issues at an early stage, even before the patient exhibits symptoms, allowing for early interventions that can prevent them issue from progressing to a more severe stage. In addition, sensor data can be used to monitor patients remotely, reducing the need for frequent visits to healthcare facilities. This is particularly beneficial for patients with chronic conditions who require ongoing monitoring. With the help of real-time analytics and machine learning algorithms, sensor data can be used to predict future health issues, allowing healthcare providers to take proactive measures to prevent them from occurring. Finally, sensor data can be used to monitor patient behavior and identify patterns that may indicate a change in their health status.

*Can semantic interoperability and ontologies help develop new communication protocols for distributed real-time communication protocols?*

Semantic interoperability and ontologies can help in the development of new communication protocols for distributed real-time communication protocols. By using ontologies in the design of new communication protocols, developers can ensure that all systems involved in real-time communication share a common understanding of the concepts and data being exchanged or transmitted. As shown by all the case studies presented in this dissertation, all systems involved in real-time communication must be able to understand each other's data and concepts to exchange information effectively. While the first case study has shown that ontologies can be used to enhance semantic interoperability between different systems, this case study has shown the importance of ontologies in the implementation of communication protocols. The developed platform integrates with existing CHUSA systems, which use different communication protocols and data formats. This means that the platform must be able to understand the data and concepts used by these systems, and offer a common vocabulary for communication. By using a standardized vocabulary and a shared understanding of data and concepts, developers can reduce the risk of miscommunication, errors, and security breaches in real-time communication. To conclude, semantic interoperability and ontologies are essential tools for the development of new communication protocols for distributed real-time communication protocols. They enable seamless integration, efficient communication, and improved security, which are crucial in today's interconnected world.

In addition to the research questions presented above, other research questions were also addressed in this case study:

*What are the healthcare's main necessities and obstacles, and how can the extraction and analysis of real-time sensorial data help optimise its workflow and reduce costs?*

The healthcare industry is constantly evolving and advancing due to the increasing demand for quality healthcare services. Technological solutions have played a crucial role in improving the quality of care provided to patients, from digitalization of patient records to telemedicine. In addition, the development of IoT devices in healthcare has enabled the collection of real-time sensorial data, which can be used to watch patients' health and provide timely support. However, the healthcare industry still faces challenges in adopting IT solutions, such as lack of interoperability and data quality issues.

CHUSA is a healthcare institution in Porto, Portugal, which aims to develop and explore new technologies and methodologies for the healthcare industry. Their research focueses on interability from different sources and systems, as well as the development of new technologies to improve patient care. Thus, the health institution has undertaken a study focused on developing an architecture for real-time processing and storage of patient monitor telemetry data to support decision-making.

This study has focused on using patient monitor telemetry data as a source to develop a platform that offers real-time insights in healthcare units. These insights offer timely information about patient health status, allowing for more effective decision-making. The build platform serves as a framework for future research and development in the healthcare industry, as it can be extended to support other types of data sources and applications.

*How to extract the untreated data from multiple sensors, process said data, and load it to relevant repositories having into account their context?*

The first step in developing a platform for real-time data processing and storage is to identify the data sources and repositories. In this case, the patient monitor telemetry data from Mindray devices and data from the hospital's EHR are the primary data sources. It is important to understand the structure and format of the data from these sources to effectively extract and process it.

Next, a data retrieval process needs to be developed to extract the data from the patient monitors and the EHR database. This process should be designed to retrieve the data as soon as it is available, ensuring real-time processing. Techniques for processing the retrieved data should be implemented. This may include data cleaning, filtering, aggregation, and transformation to ensure data quality and prepare it for further analysis and loading into relevant repositories. Furthermore, the solution should include the design of relevant repositories to store the processed data. Considering the context of the domain and the research goals, these repositories should support interoperability and be compatible with standards.

The data processing and loading process should consider the context of the healthcare industry, including patient health records, real-time analytics, and decision-making support. This can be achieved by integrating the processed data with existing systems, such as AIDA, and developing intelligent alert systems based on outlier detection. Finally, the solution should be evaluated and tested to ensure its effectiveness, usability, and reliability. This evaluation should consider factors such as data accuracy, processing speed, system performance, and user satisfaction.

By following these methods, it was possible to extract untreated data from multiple sensors, process them, and load it into relevant repositories while considering the context of the healthcare industry and the specific goals of the research study described in the given context.

*What types of data can be effectively processed and stored in real-time on a platform for decision support?*

The initial case study involved patient monitor telemetry as a source of data, which provides vital signs and health status information about patients. This includes data such as heart rate, blood pressure, respiratory rate, oxygen saturation level, temperature, and other physiological measurements collected from patient monitors. However, for the platform to be effective, it should be able to process and store other types of data as well. These include patient health records, data from external applications, and aggregated and processed data. The platform should integrate data from EHR systems to give a complete view of the patient's health record, including medical history, previous diagnoses, treatments, medications, allergies, and other relevant information.

The platform should support communication protocols based on semantic interoperability, such as the HL7 protocol and the FHIR standard, to enable integration with external applications. This could include data from telemedicine solutions, wearable devices, or other IoT devices used for remote patient monitoring. Finally, the final product should also be able to aggregate and process data from multiple sources to give meaningful insights and patterns. This could include data analytics, predictive modeling, machine learning algorithms, and other techniques to extract valuable information from the collected data.

## 4.2 Main Achievements

The purpose of this research study was to develop and explore a new generation of tools in the field of ITs in order to gather and store large quantities of data, as well as the ability to process that data in real-time, enabling new analytical solutions and artificial intelligence approaches. Therefore, the research question that was initially posed to address these concerns will be answered below.

**How can real-time information processing solutions be used to increase efficiency of a wide range of processes, allow for a more representative and accurate overview of current operations, and decrease communication gaps?**

Real-time information processing techonologies have become increasingly popular due to their ability to improve the efficiency of various processes. These solutions can be used to collect data in real-time, analyze it, and offer insights that can be used to optimize operations. These solutions can be paired with big data technologies, allowing for the collection and analysis of large amounts of data.

However, the implementation of these solutions is a complex process that requires a lot of time and effort, as it involves integrating various technologies and systems, as well as ensuring data accuracy and security. As such, careful planning, collaboration, and communication are essential to successfully implement real-time information processing platforms. This includes identifying the specific needs and goals of an organization, selecting the most appropriate technology, training staff, and continuously monitoring and evaluating the system's performance. Although these processes may be challenging, the benefits have been proven to be significant and can lead to improved efficiency, cost savings, and better outcomes for businesses and individuals.

To prove the potential of these technologies, three case studies were presented in this dissertation, each of which focused on a different domain. The first case study focused on the manufacturing industry, and the second on the smart cities domain, while the third focused on a healthcare perspective.

All three case studies presented in this dissertation have been successfully executed, and the results obtained have been positive. They have all been able to show the ability to gather and store large quantities of data from sensors, IoT devices, and other data producers, and the ability to process that data in real-time, enabling new analytical solutions and artificial intelligence approaches. In addition, the results obtained from the case studies have also demonstrated the ability to increase the efficiency of a wide range of processes across multiple domains, allowing for a more representative and accurate overview of current operations, and decrease communication gaps.

In addition, new methodologies and approaches have been developed to quantify the impact and effectiveness of these technologies. These methodologies should consider factors such as data accuracy, processing speed, and the ability to give actionable insights timely. Similarly, these methodologies must ensure that the data colleted is of high quality, and that it is stored securely. Thus, the approaches chosen should be tailored to the specific needs of the organization or industry, and should consider the types of data being collected, their sources, and the processing capabilities of the system.

The first case study presented in this dissertation was developed in collaboration with Bosch Car Multimedia Portugal, and focused on the manufacturing industry. An Extract, Transform, and Load (ETL) system that gathers and analyzes data from various sources was created, such as sensors or Internet of Things (IoT) devices. This system converts the data into a standardized format, which provides valuable insights about the manufacturing process. With access to up-to-date information, the company can identify and resolve issues instead of allowing them to escalate into major problems. In addition, new communication protocols based on semantic interoperability sustained in the development of ontologies can further enhance the effectiveness of the sensorial ETL system. These protocols enable different systems to communicate with each other, even if they're using different languages or protocols. This allows for a seamless data exchange and integration, reducing the risk of errors and increasing the accuracy of data analysis.

The main objetive of the NORTE-86 project was to transform Guimarães into a smart city and enhance the well-being of its residents. The project is ongoing, with the city continuously searching for innovative methods to harness Internet of Things (IoT) technologies to optimize city operations and improve citizen experiences.

The case study successfully demonstrated the practicality of deploying a crowdsensing platform for smart cities using geofencing technology. This involved the successful development of the platform's components, including an intuitive interface that facilitates the creation, modification, and removal of geofences by external applications. Additionally, a mobile application was created to enable citizens to interact with the platform, while a web application was designed to allow platform administrators to manage the created geofences. The value of this platform lies in empowering citizens with real-time information that can help them make better-informed decisions, improving their quality of life. The insights obtained by researchers can be used to create geofences that are customized based on past events, making it easier for citizens to receive relevant notifications about the changing city conditions.

Finally, the third case study, conducted in collaboration with the Centro Hospitalar Universitário de Santo António (CHUSA), aimed to develop and implement a real-time data processing and storage architecture for decision support in healthcare. The study focused on patient monitors telemetry and sought to identify the main challenges faced by the health institution, develop new methods for real-time data extraction and transformation, and create an analytics dashboard to aid decision-making. The study also developed an architecture to support the persistence of data for the sensorial Extract, Transform, and Load (ETL) system through stateful data streaming and established automatic alerting to medical professionals when patient parameters are outliers. Overall, the study's contributions offer valuable insights into the implementation of real-time information processing solutions in healthcare and demonstrate their potential to improve patient outcomes and enhance decision-making capabilities.

To conclude, the delivery of multiple case studies has been a significant achievement of this dissertation, as it has demonstrated the practical application and successful implementation of the developed solutions in real-life scenarios. Notably, the project has achieved remarkable accomplishments in intelligent failure prediction and avoidance, data-driven process optimization, geofence creation and management, and patient telemetry. These achievements highlight the versatility and adaptability of the developed solutions for various industries and fields.

## 4.3 Future Work

Although this dissertation has made contributions to the field of real-time information processing, there are several areas for future work that could further improve the effectiveness and impact of these technologies. First and foremost, more research is needed to develop methodologies for evaluating the effectiveness and impact of real-time information processing solutions. These methodologies must consider the contextual factors that influence the performance of these solutions, as well as the various data sources and processing capabilities of the system. This will enable organizations to set realistic expectations and optimize their real-time information processing platforms to meet their specific needs.

Secondly, there is a need for continued development and advancement of technologies that enable semantic interoperability of data. Ontologies and other technologies that ease data integration and exchange between different systems and applications are crucial for ensuring the accuracy and usefulness of the data being collected and processed. Thus, more research is needed to develop and refine these technologies and ensure that they're widely adopted and implemented in various industries and domains.

In addition, although the initial and second case studies were from research projects that have been finished, the case study presented in Section 3.3 is still ongoing. Therefore, there is still room for improvement in the platform developed, especially in adding new data sources and integrations with other systems. For example, the platform could be integrated with the Electronic Health Record (EHR) used at CHUSA, allowing for the sharing of patient information, test results, and notes, enabling seamless communication and coordination among members of the healthcare team. This promotes a holistic and cohesive approach to patient care, enhancing overall efficiency and quality of treatment. Furthermore, the alerting system could be improved by offering customizable settings for each user, allowing healthcare professionals to tailor the notifications according to their preferences and specific needs. They would be able to choose the types of alerts they wish to receive, set the critical thresholds for different vital signs, and even adjust the frequency of the notifications based on their workload and availability. This flexibility ensures that healthcare professionals receive alerts that are relevant and manageable, minimizing the chances of unnecessary interruptions or overlooking critical situations.

Finally, more work is needed to address the privacy considerations that arise from collecting and processing large amounts of data in real-time. Organizations must ensure that they're collecting and processing data in an ethical and legal manner, and that the privacy and security of individuals' data are adequately protected. Guidelines and best practices for ensuring the ethical use of these technologies must be developed, and organizations must be held accountable for their actions.

In conclusion, this dissertation has demonstrated the potential of real-time information processing technologies to improve efficiency, offer real-time insights, and enhance decision-making capabilities in various industries and domains. However, there is still much work to be done to optimize and refine these solutions and address the ethical and privacy concerns associated with their use. Future research in these areas will be crucial for ensuring the continued development and effectiveness of these solutions.

# Bibliography


Abualkishik, A. (2019). Hadoop And Big Data Challenges. *Journal of Theoretical and Applied Information Technology*, *97*, 3488. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069644434%7B%5C&%7DpartnerID=40%7B%5C&%7Dmd5=88e8d3050a994a5da8c3bf7a6e662609 (cit. on pp. xi, 18, 21).

Afreen Banu, E., & Rajamani, V. (2023). Design of Online Vitals Monitor by Integrating Big Data and IoT. *Computer Systems Science and Engineering*, *44*(3), 2469–2487. https://doi.org/10.32604/csse.2023.021332 (cit. on pp. 2, 46)
looks okay.

AIMultiple. (2020, August). Complex Event Processing: How it works, Use cases & Tools. Retrieved 2020-09-16, from https://research.aimultiple.com/cep/ (cit. on p. 30).

Alahmar, A., & Benlamri, R. (2020). SNOMED CT-Based Standardized e-Clinical Pathways for Enabling Big Data Analytics in Healthcare. *IEEE Access*, *8*, 1–1. https://doi.org/10.1109/ACCESS.2020.2994286 (cit. on p. 50).

Alharbi, A., Alosaimi, W., Sahal, R., & Saleh, H. (2021). Real-Time System Prediction for Heart Rate Using Deep Learning and Stream Processing Platforms (A. M. Khalil, Ed.). *Complexity*, *2021*, 1–9. https://doi.org/10.1155/2021/5535734 (cit. on pp. 2, 46)
yes!

Alves, D. S., Maranhão, P. A., Pereira, A. M., Bacelar-Silva, G. M., Silva-Costa, T., Beale, T. W., & Cruz-Correia, R. J. (2019). Can openEHR Represent the Clinical Concepts of an Obstetric-Specific EHR - ObsCare Software? *Studies in health technology and informatics*, *264*, 773–777. https://doi.org/10.3233/SHTI190328 (cit. on p. 51).

Amaxilatis, D., Mylonas, G., Diez, L., Theodoridis, E., Gutiérrez, V., & Muñoz, L. (2018). Managing pervasive sensing campaigns via an experimentation-as-a-service platform for smart cities. *Sensors (Switzerland)*, *18*. https://doi.org/10.3390/s18072125 (cit. on pp. 34, 38, 39).

Analide, C., Novais, P., Machado, J., & Neves, J. (2011). Quality of Knowledge in Virtual Entities. *Virtual Technologies*, 1360–1367. https://doi.org/10.4018/978-1-59904-955-7.ch084 (cit. on p. 48).

Ande, R., Adebisi, B., Hammoudeh, M., & Saleem, J. (2020). Internet of things: Evolution and technologies from a security perspective. *Sustainable Cities and Society*, *54*. https://doi.org/10.1016/j.scs.2019.101728 (cit. on p. 34).

Arvind. (2012, April). Apache Sqoop Graduates from Incubator. Retrieved 2021-01-07, from https://blogs.apache.org/sqoop/entry/apache%7B%5C_%7Dsqoop%7B%5C_%7Dgraduates%7B%5C_%7Dfrom%7B%5C_%7Dincubator (cit. on p. 23).

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The Case Research Strategy in Studies of Information Systems. *MIS Quarterly*, *11*(3), 369. https://doi.org/10.2307/248684 (cit. on p. 7).

Better. (2021). Client Stories - Ministry of Health of Republic of Slovenia. Retrieved 2021-02-12, from https://www.better.care/client-stories/ministry-of-health-of-republic-of-slovenia (cit. on p. 51).

Brandão, A., Pereira, E., Esteves, M., Portela, F., Santos, M. F., Abelha, A., & Machado, J. (2016). A benchmarking analysis of open-source business intelligence tools in healthcare environments. *Information (Switzerland)*, *7*(4), 1–16. https://doi.org/10.3390/info7040057 (cit. on pp. 5, 51) done.

Brito, C., Esteves, M., Peixoto, H., Abelha, A., & Machado, J. (2019). A data mining approach to classify serum creatinine values in patients undergoing continuous ambulatory peritoneal dialysis. *Wireless Networks*, *4*. https://doi.org/10.1007/s11276-018-01905-4 (cit. on p. 49)
done.

Brown, E. (2014). *Web Development with Node and Express: Leveraging the JavaScript Stack*. O'Reilly Media. https://books.google.pt/books?id=HsLvAwAAQBAJ (cit. on p. 57).

Capponi, A., Fiandrino, C., Kantarci, B., Foschini, L., Kliazovich, D., & Bouvry, P. (2019). A survey on mobile crowdsensing systems: Challenges, solutions, and opportunities. *IEEE Communications Surveys and Tutorials*, *21*, 2419–2465. https://doi.org/10.1109/COMST.2019.2914030 (cit. on pp. 38–40).

Cardoso, L., Marins, F., Portela, F., Santos, M., Abelha, A., & Machado, J. (2014). The next generation of interoperability agents in healthcare. *International journal of environmental research and public health*, *11*(5), 5349–71. https://doi.org/10.3390/ijerph110505349 (cit. on pp. 49, 50).

Chaudhuri, S., & Dayal, U. (1997). An Overview of Data Warehousing and OLAP Technology. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *26*(1), 65–74. https://doi.org/10.1145/248603.248616 (cit. on p. 17).

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, *54*(8), 88–98. https://doi.org/10.1145/1978542.1978562 (cit. on pp. 16, 17, 21).

Cheng, G., Guo, Y., Chen, Y., & Qin, Y. (2019). Designating city-wide collaborative geofence sites for renting and returning dock-less shared bikes. *IEEE Access*, *7*, 35596–35605. https://doi.org/10.1109/ACCESS.2019.2903521 (cit. on pp. 34, 35).

Chooprayoon, V., & Che, C. (2010, February). TECTAM: An Approach to Study Technology Acceptance Model (TAM) in Gaining Knowledge on the Adoption and Use of E-Commerce/E-Business Technology among Small and Medium Enterprises in Thailand. In *E-commerce* (pp. 30–38). InTech. https://doi.org/10.5772/8907 (cit. on pp. 8, 10).

contributors, W. (2019). SWOT analysis. Retrieved 2019-03-09, from https://en.wikipedia.org/w/index.php?title=SWOT%7B%5C_%7Danalysis%7B%5C&%7Doldid=885318472 (cit. on p. 8).

Corporation, S. (2020, November). https://docs.solace.com/Solace-PubSub-Platform.htm (cit. on p. 23).

Databricks. (2020). What is Complex Event Processing? Retrieved 2020-09-16, from https://databricks.com/glossary/complex-event-processing (cit. on p. 30).

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008 (cit. on pp. 9, 10).

Demir, K. A., Döven, G., & Sezen, B. (2019). Industry 5.0 and Human-Robot Co-working. *Procedia Computer Science*, *158*, 688–695. https://doi.org/10.1016/j.procs.2019.09.104 (cit. on pp. 2, 51).

Doyle-Kent, M., & Kopacek, P. (2020). Industry 5.0: Is the Manufacturing Industry on the Cusp of a New Revolution? *Lecture Notes in Mechanical Engineering*, (October), 432–441. https://doi.org/10.1007/978-3-030-31343-2_38 (cit. on pp. 2, 52).

Duarte, J., Portela, C. F., Abelha, A., Machado, J., & Santos, M. F. (2011). Electronic Health Record in Dermatology Service. In M. M. Cruz-Cunha, J. Varajão, P. Powell, & R. Martinho (Eds.), *Enterprise information systems* (pp. 156–164). Springer Berlin Heidelberg. (Cit. on p. 49).
done.

Duarte, J., Salazar, M., Quintas, C., Santos, M., Neves, J., Abelha, A., & Machado, J. (2010). Data quality evaluation of electronic health records in the hospital admission process. *Proceedings - 9th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2010*, 201–206. https://doi.org/10.1109/ICIS.2010.97 (cit. on p. 48)
done.

EDUCBA. (2020). Introduction to Apache Flume. Retrieved 2021-01-06, from https://www.educba.com/apache-flume/ (cit. on p. 22).

El-Ganainy, N. O., Balasingham, I., Halvorsen, P. S., & Rosseland, L. A. (2020). A New Real Time Clinical Decision Support System Using Machine Learning for Critical Care Units. *IEEE Access*, *8*, 185676–185687. https://doi.org/10.1109/ACCESS.2020.3030031 (cit. on pp. 45, 46).

Elhadad, A., Alanazi, F., Taloba, A. I., & Abozeid, A. (2022). Fog Computing Service in the Healthcare Monitoring System for Managing the Real-Time Notification. (M. Elhoseny, Ed.). *Journal of healthcare engineering*, *2022*, 5337733. https://doi.org/10.1155/2022/5337733 (cit. on p. 53)
maybe?

Fernandes, B., Neves, J., & Analide, C. (2020). Safecity: A platform for safer and smarter cities. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture*

*Notes in Bioinformatics)*, *12092 LNAI*, 412–416. https://doi.org/10.1007/978-3-030-49778-1_37 (cit. on p. 35).

Ferreira, D. (2018). *Interoperability and Knowledge Discovery in Healthcare* (Publication No. September) [Doctoral dissertation, Universidade do Minho]. (Cit. on p. 50).

Firouzi, F., Farahani, B., & Marinšek, A. (2022). The convergence and interplay of edge, fog, and cloud in the AI-driven Internet of Things (IoT). *Information Systems*, *107*, 101840. https://doi.org/10.1016/j.is.2021.101840 (cit. on p. 53).

Foschini, L., Martuscelli, G., Montanari, R., & Solimando, M. (2021). Edge-enabled mobile crowdsensing to support effective rewarding for data collection in pandemic events. *Journal of Grid Computing*, *19*. https://doi.org/10.1007/s10723-021-09569-9 (cit. on pp. 35, 38–40).

Foundation, T. A. S. (2020a). Flume 1.9.0 user guide. https://flume.apache.org/FlumeUserGuide.html (cit. on pp. xi, 21, 22).

Foundation, T. A. S. (2020b). Welcome to Apache Flume. Retrieved 2021-01-06, from https://flume.apache.org/ (cit. on p. 21).

Frampton, M. (2015). *Mastering Apache Spark*. Packt Publishing. https://books.google.pt/books?id=ENZOCwAAQBAJ (cit. on pp. 27, 28).

Gavin, L. (2019). *What Is Data Engineering?* https://www.oreilly.com/library/view/what-is-data/9781492075578/ (cit. on p. 18).

Gillham, B. (2010). *Case Study Research Methods*. Bloomsbury Publishing. https://books.google.pt/books?id=F0gSBwAAQBAJ (cit. on p. 7).

Gkoulalas-Divanis, A., & Bettini, C. (2018, October). *Handbook of mobile data privacy*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98161-1 (cit. on p. 18).

Google. (2022). *Geofencing api*. https://developers.google.com/location-context/geofencing (cit. on p. 31).

Gorelik, A. (2019). *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science* (A. Oram, Ed.; First Edit). O'Reilly Media, Inc. https://www.oreilly.com/library/view/the-enterprise-big/9781491931547/ (cit. on pp. 18, 19).

Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. SAGE Publications Ltd. (Cit. on p. 42).

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). Prisma2020: An r package and shiny app for producing prisma 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis [https://doi.org/10.1002/cl2.1230]. *Campbell Systematic Reviews*, *18*, e1230. https://doi.org/https://doi.org/10.1002/cl2.1230 (cit. on pp. xi, 45).

Haddaway, N. R., & Macura, B. (2018). The role of reporting standards in producing robust literature reviews. *Nature Climate Change*, *8*(6), 444–447. https://doi.org/10.1038/s41558-018-0180-3 (cit. on p. 19).

Hak, F., Guimaraes, T., & Santos, M. (2022). Towards effective clinical decision support systems: A systematic review [Publisher: Public Library of Science]. *PLoS ONE*, *17*(8). https://doi.org/10.1371/journal.pone.0272846 (cit. on p. 41).

Hak, F., Oliveira, D., Abreu, N., Leuschner, P., Abelha, A., & Santos, M. (2020). An OpenEHR adoption in a portuguese healthcare facility [Journal Abbreviation: Procedia Comput. Sci.]. In Shakshuki E. & Yasar A.-U.-H. (Eds.), *Procedia comput. sci.* (pp. 1047–1052, Vol. 170). Elsevier B.V. https://doi.org/10.1016/j.procs.2020.03.075 (cit. on p. 51).

Hak, F., Guimarães, T., Abelha, A., & Santos, M. (2020). Clinical Decision Support Using Open Data. In *Advances in intelligent systems and computing* (pp. 484–492, Vol. 1161 AISC). https://doi.org/10.1007/978-3-030-45697-9_47 (cit. on p. 51).

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining*. Elsevier. https://doi.org/10.1016/C2009-0-61819-5 (cit. on pp. 15, 18).

Hassan, F., E., M., & Sahal, R. (2020). Real-Time Healthcare Monitoring System using Online Machine Learning and Spark Streaming. *International Journal of Advanced Computer Science and Applications*, *11*(9), 650–658. https://doi.org/10.14569/IJACSA.2020.0110977 (cit. on pp. 2, 47) okay.

Hassan, Q. (Ed.). (2018, May). *Internet of Things A to Z*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119456735 (cit. on p. 3).

Hay, G., & Castilla, G. (2006). Object-based image analysis: Strengths, weaknesses, opportunities and threats (SWOT). *Proceedings of 1st International Conference on Object-based Image Analysis (OBIA 2006)*, *XXXVI-4/C4*(May 2014), 3. https://www.isprs.org/proceedings/XXXVI/4-C42/ (cit. on pp. 8, 9).

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0 (cit. on p. 49).

Hevner, A. R., & Chatterjee, S. (2010). *Design Research in Information Systems: Theory and Practice* (Vol. 2). Springer, New York, NY. https://doi.org/10.1007/978-1-4419-6108-2 (cit. on p. 6). 1-350.

HL7 International. (2023). HL7 Backgrounder Brief. Retrieved 2023-03-16, from http://www.hl7.org/newsroom/HL7backgrounderbrief.cfm (cit. on p. 50).

Hoffman, S. (2013). *Apache flume: Distributed log collection for hadoop*. Packt Publishing Ltd. (Cit. on p. 21).

Hu, J., Yang, K., Wang, K., & Zhang, K. (2020). A blockchain-based reward mechanism for mobile crowdsensing. *IEEE Transactions on Computational Social Systems*, *7*, 178–191. https://doi.org/10.1109/TCSS.2019.2956629 (cit. on pp. 36, 38, 40).

Inmon, W. H. (2002). *Building the Data Warehouse* (R. Ipsen, Ed.; Third Edit). John Wiley & Sons, Inc. (Cit. on pp. 17, 18).

Ismagilova, E., Hughes, L., Dwivedi, Y. K., & Raman, K. R. (2019). Smart cities: Advances in research — an information systems perspective. *International Journal of Information Management*, *47*, 88–100. https://doi.org/10.1016/j.ijinfomgt.2019.01.004 (cit. on pp. 31, 37, 39).

JBoss Community Documentation. (2020). Complex Event Processing. Retrieved 2020-11-19, from https://docs.jboss.org/drools/release/6.2.0.CR3/drools-docs/html/DroolsComplexEventProcessingChapter.html%7B%5C#%7Dd0e10678%7B%5C%%7D5C (cit. on p. 30).

John, T., & Misram, P. (2017). *Data Lake for Enterprises*. Packt Publishing Ltd. https://www.oreilly.com/library/view/data-lake-for/9781787281349/ (cit. on p. 22).

Kale, S., Tamakuwala, H., Vijayakumar, V., Yang, L., & Rawal Kshatriya, B. S. (2020). Big Data in Healthcare: Challenges and Promise. In *Smart innovation, systems and technologies* (pp. 3–17, Vol. 164). https://doi.org/10.1007/978-981-32-9889-7_1 (cit. on p. 19).

Kendig, C. E. (2016). What is Proof of Concept Research and how does it Generate Epistemic and Ethical Categories for Future Scientific Practice? *Science and Engineering Ethics*, *22*(3), 735–753. https://doi.org/10.1007/s11948-015-9654-0 (cit. on p. 8).

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit* (Wiley, Ed.; Third Edit, Vol. 3). John Wiley & Sons. (Cit. on pp. 16–20, 48, 49).

Kirimtat, A., Krejcar, O., Kertesz, A., & Tasgetiren, M. F. (2020). Future trends and current state of smart city concepts: A survey. *IEEE Access*, *8*, 86448–86467. https://doi.org/10.1109/ACCESS.2020.2992441 (cit. on pp. 2, 31, 36).

KnowledgeHut. (2019, August). Apache Spark Pros and Cons. Retrieved 2021-01-07, from https://www.knowledgehut.com/blog/big-data/apache-spark-advantages-disadvantages (cit. on p. 28).

Kousiouris, G., Akbar, A., Sancho, J., Ta-shma, P., Psychas, A., Kyriazis, D., & Varvarigou, T. (2018). An integrated information lifecycle management framework for exploiting social network data to identify dynamic large crowd concentration events in smart cities applications. *Future Generation Computer Systems*, *78*, 516–530. https://doi.org/10.1016/j.future.2017.07.026 (cit. on p. 35).

Kuechler, B., & Petter, S. (2017). Design Science Research in Information Systems. *Design Science Research in Information Systems:* 8–10. https://doi.org/10.1007/978-3-642-29863-9 (cit. on p. 6)
Way too esoteric for me! 1-66.

Kunigk, J., Buss, I., Wilkinson, P., & George, L. (2018). *Architecting Modern Data Platforms* (N. Tache & M. Cronin, Eds.; First Edit). O'Reilly Media, Inc. (Cit. on p. 19).

Lalmi, F., & Adala, L. (2021). Big Data for Healthcare: Opportunities and Challenges. In *Studies in computational intelligence* (pp. 217–229, Vol. 935). https://doi.org/10.1007/978-3-030-62796-6_12 (cit. on p. 18).

Lasi, H., Fettke, P., Kemper, H. G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business and Information Systems Engineering*, *6*(4), 239–242. https://doi.org/10.1007/s12599-014-0334-4 (cit. on p. 2).

Li, B., & Tsui, R. (2020). How to Improve the Reuse of Clinical Data– openEHR and OMOP CDM. *Journal of Physics: Conference Series*, *1624*(3), 032041. https://doi.org/10.1088/1742-6596/1624/3/032041 (cit. on p. 51).

Liu, G., Zhu, W., Saunders, C., Gao, F., & Yu, Y. (2015). Real-time Complex Event Processing and Analytics for Smart Grid. *Procedia Computer Science*, *61*, 113–119. https://doi.org/10.1016/j.procs.2015.09.169 (cit. on p. 30).

Lourenço, J. M. (2021). *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. ii).

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. *Journal of Industrial Information Integration*, *6*, 1–10. https://doi.org/10.1016/j.jii.2017.04.005 (cit. on p. 52).

Luraschi, J., Kuo, K., & Ruiz,
bibinitperiod E. (2918). *Mastering Spark with R The Complete Guide to Large-Scale Analysis and Modeling* (First Rele). O'Reilly Media, Inc. (Cit. on p. 18).

Maas, G., & Garillot, F. (2019). *Stream Processing with Apache Spark* (First Rele). O'Reilly Media, Inc. (Cit. on p. 19).

March, S., & Storey, V. (2008). Design Science in the Information Systems Discipline: An Introduction to the Special Issue on Design Science Research. *MIS Quarterly*, *32*(4), 725. https://doi.org/10.2307/25148869 (cit. on p. 6).

Martinez, G. (2016). *Research Methods for Business Students* (Seventh ed). Pearson Education Limited. www.pearson.com/uk. (cit. on p. 5).

Mavriki, P., & Karyda, M. (2020). Big Data Analytics in Healthcare Applications: Privacy Implications for Individuals and Groups and Mitigation Strategies. In *Lecture notes in business information processing* (pp. 526–540, Vol. 402). https://doi.org/10.1007/978-3-030-63396-7_35 (cit. on p. 19).

Meredith, J. (2021, February). What is openEHR and why is it important? | Digital Health Wales. Retrieved 2021-02-03, from https://digitalhealth.wales/news/what-openehr-and-why-it-important (cit. on p. 51).

Microsoft. (2023, February 14). *Overview of ASP.NET core SignalR* [Https://learn.microsoft.com/en-us/aspnet/core/signalr/introduction?view=aspnetcore-7.0]. (Cit. on p. 80).

Mind Tools Content Team. (2019). SWOT Analysis (Strengths, Weaknesses Opportunities, Threats). Retrieved 2019-03-10, from https://www.mindtools.com/pages/article/newTMC%7B%5C_%7D05.htm (cit. on p. 8).

Miranda, M., Duarte, J., Abelha, A., Machado, J., & Neves, J. (2009). Interoperability and Healthcare. *European Simulation and Modelling Conference 2009*, 205–212. http://hdl.handle.net/1822/11288 (cit. on pp. 49, 50, 78)
done.

Miranda, R., Ramos, V., Ribeiro, E., Rodrigues, C., Silva, A., Durães, D., Analide, C., Abelha, A., & Machado, J. (2022). Crowdsensing on smart cities: A systematic review [Series Title: Lecture Notes in Computer Science]. In A. C. Bicharra Garcia, M. Ferro, & J. C. Rodríguez Ribón (Eds.), *Advances in artificial intelligence – IBERAMIA 2022* (pp. 103–106, Vol. 13788). Springer International Publishing. https://doi.org/10.1007/978-3-031-22419-5_9 (cit. on pp. xiii, 39).

Mitchell, G. (2019). SWOT Analysis Definition. Retrieved 2019-03-09, from https://www.investopedia.com/terms/s/swot.asp (cit. on pp. xi, 8, 9).

Mujahid Ghouri, A., Mani, V., Jiao, Z., Venkatesh, V. G., Shi, Y., & Kamble, S. S. (2021). An empirical study of real-time information-receiving using industry 4.0 technologies in downstream operations. *Technological Forecasting and Social Change*, *165*(December 2020). https://doi.org/10.1016/j.techfore.2020.120551 (cit. on p. 52).

Nahavandi, S. (2019). Industry 5.0-a human-centric solution. *Sustainability (Switzerland)*, *11*(16). https://doi.org/10.3390/su11164371 (cit. on p. 2).

Neves, J. J., Machado, J. J., Ribeiro, J., Abelha, A., Vicente, H., Ferraz, F., Neves, J. J., Machado, J. J., Esteves, M., Sampaio, L., Ferraz, F., Abelha, A., Machado, J. J., Machado, J. J., Neves, J. J., Ribeiro, J., & Sampaio, L. (2018). A Deep-Big Data Approach to Health Care in the AI Age. *Mobile Networks and Applications*, *23*(4), 1123–1128. https://doi.org/10.1007/s11036-018-1071-6 (cit. on p. 49)
done.

Nižetić, S., Šolić, P., González-de-Artaza, D. L.-I., & Patrono, L. (2020). Internet of things (iot): Opportunities, issues and challenges towards a smart and sustainable future. *Journal of Cleaner Production*, *274*. https://doi.org/10.1016/j.jclepro.2020.122877 (cit. on pp. 2, 31, 37, 38).

Oliveira, D. (2018). Knowledge Engineering for Interoperable Electronic Health Records - Pre Thesis. (March) (cit. on p. 50).

Oliveira, D., Miranda, R., Abreu, N., Leuschner, P., Abelha, A., Santos, M., & Machado, J. (2020). Management of a pandemic based on an openEHR approach. *Procedia Computer Science*, *177*, 522–527. https://doi.org/10.1016/j.procs.2020.10.072 (cit. on p. 51).

Oliveira, D., Miranda, R., Hak, F., Abreu, N., Leuschner, P., Abelha, A., & Machado, J. (2021). Steps towards an healthcare information model based on openEHR. *Procedia Computer Science*, *184*(2019), 893–898. https://doi.org/10.1016/j.procs.2021.04.015 (cit. on pp. 47, 49, 51).

Oliveira, D., Miranda, R., Leuschner, P., Abreu, N., Santos, M. F., Abelha, A., & Machado, J. (2021). OpenEHR modeling: improving clinical records during the COVID-19 pandemic. *Health and Technology*, *11*(5), 1109–1118. https://doi.org/10.1007/s12553-021-00556-4 (cit. on p. 51).

openEHR Foundation. (2021). What is openEHR? Retrieved 2021-03-26, from https://www.openehr.org/about/what%7B%5C_%7Dis%7B%5C_%7Dopenehr (cit. on p. 51).

Oracle. (2023). Data Warehouse Reference Architecture. Retrieved 2023-03-27, from https://docs.oracle.com/database/121/TDPDW/tdpdw%7B%5C_%7Darch.htm%7B%5C#%7DTDPDW380 (cit. on p. 18).

Oztemel, E., & Gursev, S. (2020). Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing*, *31*(1), 127–182. https://doi.org/10.1007/s10845-018-1433-8 (cit. on p. 2).

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021a). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71 (cit. on pp. xi, 32, 34).

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021b). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/BMJ.N71 (cit. on p. 42).

Pánek, J. (2019). Emotional maps: Participatory crowdsourcing of citizens' perceptions of their urban environment. *Cartographic Perspectives*, *2018*, 17–29. https://doi.org/10.14714/CP91.1419 (cit. on pp. 35, 38, 39).

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 15. https://doi.org/10.2753/MIS0742-1222240302 (cit. on pp. xi, 6)
45-77.

Peixoto, H., Santos, M., Abelha, A., & Machado, J. (2012). Intelligence in interoperability with AIDA. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7661 LNAI*, 264–273. https://doi.org/10.1007/978-3-642-34624-8_31 (cit. on p. 78).

Picaut, J., Fortin, N., Bocher, E., Petit, G., Aumond, P., & Guillaume, G. (2019). An open-science crowdsourcing approach for producing community noise maps using smartphones. *Building and Environment*, *148*, 20–33. https://doi.org/10.1016/j.buildenv.2018.10.049 (cit. on p. 36).

Pilloni, V. (2018). How data will transform industrial processes: Crowdsensing, crowdsourcing and big data as pillars of industry 4.0. *Future Internet*, *10*. https://doi.org/10.3390/fi10030024 (cit. on pp. 36, 37).

Rai, B. (2022). *Security Issues and Solutions for Healthcare Informatics*. https://doi.org/10.1007/978-3-030-85559-8_12 (cit. on p. 18).

Raja, R., Mukherjee, I., & Sarkar, B. K. (2020). A Systematic Review of Healthcare Big Data. *Scientific Programming*, *2020*, 1–15. https://doi.org/10.1155/2020/5471849 (cit. on p. 19)
systematic review.

Robert Bosch GmbH. (2020). Industry 4.0. Retrieved 2020-06-01, from https://www.bosch.com/products-and-services/connected-products-and-services/industry-4-0/ (cit. on p. 2).

Roman, C., Liao, R., Ball, P., Ou, S., & Heaver, M. D. (2018). Detecting on-street parking spaces in smart cities: Performance evaluation of fixed and mobile sensing systems. *IEEE Transactions on Intelligent Transportation Systems*, *19*, 2234–2245. https://doi.org/10.1109/TITS.2018.2804169 (cit. on p. 35).

Runsewe, O., & Samaan, N. (2021). Cloud resource scaling for time-bounded and unbounded big data streaming applications. *IEEE Transactions on Cloud Computing*, *9*(2), 504–517. https://doi.org/10.1109/TCC.2018.2876242 (cit. on p. 53).

Saleh, H., Younis, E. M. G., Sahal, R., & Ali, A. A. (2021). Predicting Systolic Blood Pressure in Real-Time Using Streaming Data and Deep Learning. *Mobile Networks and Applications*, *26*(1), 326–335. https://doi.org/10.1007/s11036-020-01645-w (cit. on p. 46)
gotta make in query.

Salesforce. (2020). Meet the Three Industrial Revolutions Unit. Retrieved 2020-09-07, from https://trailhead.salesforce.com/en/content/learn/modules/learn-about-the-fourth-industrial-revolution/meet-the-three-industrial-revolutions (cit. on p. 52).

Shafqat, S., Kishwer, S., Rasool, R. U., Qadir, J., Amjad, T., & Ahmad, H. F. (2020). Big data analytics enhanced healthcare systems: a review. *The Journal of Supercomputing*, *76*(3), 1754–1799. https://doi.org/10.1007/s11227-017-2222-4 (cit. on p. 51).

Shahrour, I., & Xie, X. (2021). Role of internet of things (iot) and crowdsourcing in smart city projects. *Smart Cities*, *4*, 1276–1292. https://doi.org/10.3390/smartcities4040068 (cit. on pp. 36, 38).

Shehab, R., Taher, M., & Mohamed, H. (2021). Live Big Data Analytics Resource Management Techniques in Fog Computing for TeleHealth Applications. *Jordanian Journal of Computers and Information Technology*, *7*(1), 89. https://doi.org/10.5455/jjcit.71-1605864596 (cit. on p. 46)
From Duplicate 1 (Live Big Data Analytics Resource Management Techniques in Fog Computing for TeleHealth Applications - Shehab, Ragaa; Taher, Mohamed; Mohamed, Hoda) maybe From Duplicate 2 (Live Big Data Analytics Resource Management Techniques in Fog Computing for TeleHealth Applications - Shehab, Ragaa; Taher, Mohamed; Mohamed, Hoda) seems like they use YARN for ecg.

Shirer, M., & Goepfert, J. (2019, April). IDC Forecasts Revenues for Big Data and Business Analytics Solutions Will Reach $189.1 Billion This Year with Double-Digit Annual Growth Through 2022. Retrieved 2020-05-28, from https://www.idc.com/getdoc.jsp?containerId=prUS44998419 (cit. on p. 15).

Singaram, M., & Jain, P. (2018). What is the Difference between Proof of Concept and Prototype? Retrieved 2019-05-20, from https://www.entrepreneur.com/article/307454 (cit. on p. 8).

Solace. (2021a). PubSub+ Event Broker. Retrieved 2021-01-07, from https://solace.com/products/event-broker/ (cit. on p. 24).

Solace. (2021b). PubSub+ Platform. Retrieved 2021-01-07, from https://solace.com/products/platform/ (cit. on pp. xi, 24).

Sousa, R., Miranda, R., Moreira, A., Alves, C., Lori, N., & Machado, J. (2021). Software Tools for Conducting Real-Time Information Processing and Visualization in Industry: An Up-to-Date Review. *Applied Sciences*, *11*(11), 4800. https://doi.org/10.3390/app11114800 (cit. on pp. 2, 19)
...no comment required.

Technologies, M. (2018, March). What is power bi architecture? a complete tutorial. https://mindmajix.com/power-bi-architecture (cit. on pp. xi, 29).

Techopedia. (2018, January). Semantics. Retrieved 2023-03-10, from https://www.techopedia.com/definition/687/semantics-computing (cit. on pp. 3, 47).

Techopedia. (2019). Definition - What does Proof of Concept (POC) mean? Retrieved 2019-05-20, from https://www.techopedia.com/definition/4066/proof-of-concept-poc (cit. on p. 8).

Thames, L., & Schaefer, D. (2016). Software-defined Cloud Manufacturing for Industry 4.0. *Procedia CIRP*, *52*, 12–17. https://doi.org/10.1016/j.procir.2016.07.041 (cit. on p. 53).

Thanh, L. N. T., Phien, N. N., Nguyen, T. A., Vo, H. K., Luong, H. H., Anh, T. D., Tuan, K. N. H., & Son, H. X. (2021). IoHT-MBA: An Internet of Healthcare Things (IoHT) Platform based on Microservice and Brokerless Architecture. *International Journal of Advanced Computer Science and Applications*, *12*(7), 594–601. https://doi.org/10.14569/IJACSA.2021.0120768 (cit. on pp. 46, 52)
looks promising.

The Apache Software Foundation. (2019). Sqoop. Retrieved 2021-01-06, from http://sqoop.apache.org/ (cit. on p. 23).

The Apache Software Foundation. (2021). Spark Overview. Retrieved 2021-01-07, from http://spark.apache.org/docs/latest/index.html (cit. on p. 27).

The openEHR Foundation. (2015, December). openEHR Architecture Overview. Retrieved 2023-03-20, from https://specifications.openehr.org/releases/BASE/Release-1.0.3/architecture%7B%5C_%7Doverview.html (cit. on pp. xi, 51, 52).

Venkatesh, V. (2000a). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. *Information Systems Research*, *11*(4), 342–365. https://doi.org/10.1287/isre.11.4.342.11872 (cit. on p. 10).

Venkatesh, V. (2000b). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. *Information Systems Research*, *11*(4), 342–365. https://doi.org/10.1287/isre.11.4.342.11872 (cit. on p. 10).

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x (cit. on p. 10).

Verma, R., & Bhatt, R. (2021). Security in Big Data Health Care System. *2021 Sixth International Conference on Image Information Processing (ICIIP)*, *2021-Novem*, 551–555. https://doi.org/10.1109/ICIIP53038.2021.9702663 (cit. on p. 19).

Vohra, D. (2016a). Apache flume. In *Practical hadoop ecosystem* (pp. 287–300). Springer. (Cit. on p. 21).

Vohra, D. (2016b). *Practical Hadoop Ecosystem* (Springer, Ed.; 1st). Apress. https://doi.org/10.1007/978-1-4842-2199-0 (cit. on p. 20).

Vohra, D. (2016c). Using apache sqoop. In *Pro docker* (pp. 151–183). Springer. (Cit. on p. 23).

Wang, L., & Alexander, C. A. (2019). Big Data Analytics in Healthcare Systems. *International Journal of Mathematical, Engineering and Management Sciences*, *4*(1), 17–26. https://doi.org/10.33889/IJMEMS.2019.4.1-002 (cit. on pp. 36, 37, 39)
yes but. 2019.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099 (cit. on p. 48).

Wasson, M. (2013). Single-Page Applications: Build Modern, Responsive Web Apps with ASP.NET. Retrieved 2019-05-31, from https://msdn.microsoft.com/en-us/magazine/dn463786.aspx (cit. on p. 74).

Wilson, E. J. (1994). The Relative Importance of Supplier Selection Criteria: A Review and Update. *International Journal of Purchasing and Materials Management*, *30*(2), 34–41. https://doi.org/10.1111/j.1745-493X.1994.tb00195.x (cit. on p. 19).

Yang, M., Zhu, T., Liang, K., Zhou, W., & Deng, R. H. (2019). A blockchain-based location privacy-preserving crowdsensing system. *Future Generation Computer Systems*, *94*, 408–418. https://doi.org/10.1016/j.future.2018.11.046 (cit. on pp. 36, 38).

Yin, R. (2018). *Case Study Research and Applications* (K. DeRosa, Ed.; Sixth Edit, Vol. 53). SAGE Publications, Inc. https://us.sagepub.com/en-us/nam/case-study-research-and-applications/book250150%7B%5C#%7Ddescription (cit. on p. 7).

Zikopoulos, P., Eaton, C., DeRoos, D., Deutsch, T., & Lapis, G. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (1st). McGraw-Hill Osborne Media. http://it.toolbox.com/blogs/db2luw. (cit. on pp. 16, 20).

# A
# Publications

In this appendix, the list of scientific papers published in peer-reviewed conferences and journals during the course of this doctoral theis is presented.

## A.1 Management of a pandemic based on an openEHR approach

**DOI:** 10.1016/j.procs.2020.10.072

**Authors:** Daniela Oliveira, Rui Miranda, Nuno Abreu, Pedro Leuschner, António Abelha, Manuel Santos, José Machado

**Title:** Management of a pandemic based on an openEHR approach

**Publication:** Procedia Computer Science

**Conference Proceedings:** 11th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, EUSPN 2020

**Year of Publication:** 2020

**Abstract:** The COVID-19 pandemic has collapsed several national health systems, due to the lack of healthcare professionals and exhaustion of those employed, as well as the lack of interoperability and capacity to restructure their informatic systems. Therefore, the restructuring of institutions at all levels is essential, mainly at the level of their Information Systems. When the COVID-19 pandemic had spread to Portugal in March 2020, with a breakout on the northern region, the Centro Hospitalar Universitário do Porto (CHUP) healthcare institution had felt the need to develop and integrate a new approach based on

the openEHR standard to interoperate with the institution's existing information systems, with the aim of responding quickly to the pandemic's evolution.

**Keywords:** COVID-19; Electronic health record; Healthcare information systems; Interoperability; OpenEHR

**State of Publication:** Published

## A.2 Health professional's decision-making based on multichannel interaction services

**DOI:** 10.1016/j.procs.2021.03.112

**Authors:** Ailton Moreira, Rui Miramda, Manuel Filipe Santos

**Title:** Health professional's decision-making based on multichannel interaction services

**Publication:** Procedia Computer Science

**Conference Proceedings:** 2th International Conference on Ambient Systems, Networks and Technologies, ANT 2021 / 4th International Conference on Emerging Data and Industry 4.0, EDI40 2021 / Affiliated Workshops

**Year of Publication:** 2021

**Abstract:** The COVID-19 pandemic has reinforced the importance and impact of telemedicine and multichannel interactions in healthcare services provided to patients. Health professionals are in turn increasingly dependent on patient data collected through multichannel interactions to make their clinical decisions. This article intends to present a brief analysis from the viewpoint of health professionals regarding the use of technologies in telemedicine and multichannel interactions to support decision making, basing on the analysis of clinical data of patients collected in a telemedicine environment. These technologies have numerous advantages for healthcare professionals and patients, but there are also some obstacles and gaps inherent that need to be overcome. Furthermore, health professionals can perform a more detailed analysis of patient data before taking any decision, as this practice promotes data collection to facilitate the decision-making process of health professionals.

**Keywords:** Decision-making; Health professionals; Multichannel interaction services; Telemedicine

**State of Publication:** Published

## A.3 Steps towards an healthcare information model based on openEHR

**DOI:** 10.1016/j.procs.2021.04.015

**Authors:** Daniela Oliveira, Rui Miranda, Francini Hak, Nuno Abreu, Pedro Leuschner, António Abelha, José Machado

**Title:** Data Platforms for Real-time Insights in Healthcare: Systematic Review

**Publication:** Procedia Computer Science

**Conference Proceedings:** 12th International Conference on Ambient Systems, Networks and Technologies, ANT 2021 / 4th International Conference on Emerging Data and Industry 4.0, EDI40 2021 / Affiliated Workshops

**Year of Publication:** 2021

**Abstract:** During COVID-19 pandemic crisis, healthcare institutions globally were experiencing a VUCA - Volatile, Uncertain, Complex, and Ambiguous - environment. Efficient clinical and administrative management had never been so emergent. To achieve this goal, different components of the Healthcare Information System (HIS) must cooperate and interoperate flawlessly. Data standardization is a necessary step towards normalization and interoperability between existing Legacy Systems (LSs), and provides for longitudinal, highly reliable and persistent Electronic Health Records (EHRs). The openEHR standard was chosen for its overall dual domain architecture, where the more dynamic clinical information model may evolve independently from the relatively stable Reference Model (RM). Its Information Model (IM) comprises demographic, administrative and clinical systems. Critical clinical terms have been aligned to the FHIR HL7 standard, as to further support interoperability.

**Keywords:** Clinical information model; Electronic health record; Healthcare information systems; OpenEHR; Reference model

**State of Publication:** Published

## A.4 Software tools for conducting real-time information processing and visualization in industry: An up-to-date review

**DOI:** 10.3390/app11114800

**Authors:** Regina Sousa, Rui Miranda, Ailton Moreira, Carlos Alves, Nicolas Lori, José Machado

**Title:** Software tools for conducting real-time information processing and visualization in industry: An up-to-date review

**Publication:** Applied Sciences (Switzerland)

**Year of Publication:** 2021

**Abstract:** The processing of information in real-time (through the processing of complex events) has become an essential task for the optimal functioning of manufacturing plants. Only in this way can artificial intelligence, data extraction, and even business intelligence techniques be applied, and the data produced daily be used in a beneficent way, enhancing automation processes and improving service delivery. Therefore, professionals and researchers need a wide range of tools to extract, transform, and load data in real-time efficiently. Additionally, the same tool supports or at least facilitates the visualization of this data intuitively and interactively. The review presented in this document aims to provide an up-to-date review of the various tools available to perform these tasks. Of the selected tools, a brief description of how they work, as well as the advantages and disadvantages of their use, will be presented. Furthermore, a critical analysis of overall operation and performance will be presented. Finally, a hybrid architecture that aims to synergize all tools and technologies is presented and discussed.

**Keywords:** Complex Event Processing (CEP); Data visualization; Kafka; Power BI; Real time information processing; Spark; Tableau

**State of Publication:** Published

## A.5 OpenEHR modeling: improving clinical records during the COVID-19 pandemic

**DOI:** 10.1007/s12553-021-00556-4

**Authors:** Daniela Oliveira, Rui Miranda, Pedro Leuschner, Nuno Abreu, António Abelha, Manuel Santos, José Machado

**Title:** OpenEHR modeling: improving clinical records during the COVID-19 pandemic

**Publication:** Health and Technology

**Journal:** Health and Technology

**Year of Publication:** 2021

**Abstract:** The COVID-19 pandemic had put pressure on various national healthcare systems, due to the lack of health professionals and exhaustion of those avaliable, as well as lack of interoperability and inability to restructure their IT systems. Therefore, the restructuring of institutions at all levels is essential, especially at the level of their information systems. Furthermore, the COVID-19 pandemic had arrived in Portugal at March 2020, with a breakout on the northern region. In order to quickly respond to the pandemic, the CHUP healthcare institution, known as a research center, has embraced the challenge of developing and integrating a new approach based on the openEHR standard to interoperate with the institution's existing information and its systems. An openEHR clinical modelling methodology was outlined and adopted, followed by a survey of daily clinical and technical requirements. With the arrival of the virus in Portugal, the CHUP institution has undergone through constant changes in their working methodologies as well as their openEHR modelling. As a result, an openEHR patient care workflow for COVID-19 was developed.

**Keywords:** Archetypes; COVID-19; Electronic health record; Healthcare information systems; Interoperability; OpenEHR; Templates

**State of Publication:** Published

## A.6 Crowdsensing on Smart Cities: A Systematic Review

**DOI:** 10.1007/978-3-031-22419-5$_9$

**Authors:** Rui Miranda, Vasco Ramos, Eduarda Ribeiro, Carla Rodrigues, António Silva, Dalila Durães, César Analide, António Abelha, José Machado

**Title:** Crowdsensing on Smart Cities: A Systematic Review

**Proceedings:** Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

**Conference:** 17th Ibero-American Conference on Artificial Intelligence, IBERAMIA 2022

**Year of Publication:** 2023

**Abstract:** With the rise of the internet and the Internet of Things (IoT), the concept of a Smart City began to materialise. Crowdsensing is the process of using portable sensing devices to gather information about people's surroundings. Furthermore, the research on these domains has pivoted from solely technology concepts to now including how they improve the quality of life of citizens and their utility. This paper presents a systematic review aiming to identify the role and purpose of crowdsensing and the improvement of citizens' lives in a smart city through this technique. Using the Scopus citation and abstract database, six papers were picked out as relevant for discussion in the review.

**Keywords:** Crowdsensing; Internet of Things; Smart cities; Smart notifications

**State of Publication:** Published

## A.7 Data Platforms for Real-time Insights in Healthcare: Systematic Review

**DOI:** 10.1016/j.procs.2023.03.110

**Authors:** Rui Miranda, Carlos Alves, António Abelha, José Machado

**Title:** Data Platforms for Real-time Insights in Healthcare: Systematic Review

**Publication:** Procedia Computer Science

**Conference Proceedings:** The 14th International Conference on Ambient Systems, Networks and Technologies Networks (ANT 2022) and The 6th International Conference on Emerging Data and Industry 4.0 (EDI40)

**Year of Publication:** 2023

**Abstract:** The ever-growing usage and popularity of Internet of Things devices, coupled with Big Data technologies and machine learning algorithms, have allowed for data engineers to explore new opportunities in healthcare and continuous care. Furthermore, there is a need to reduce the gap on time from when information is created to when actions and insights can be offered. However, a challenge in implementing a large-scale data processing architecture is deciding which tools are appropriate, and how to apply them in the best way possible. For example, streaming systems are now mature enough that hospitals worldwide can use their extremely large datasets, along with data producers, to predict and influence future events. Thus, the main objective of this systematic review is to identify the state-of-the-art in data platforms on healthcare that allow the creation of metrics and actions in real-time. The PRISMA guideline for reporting systematic reviews was implemented to deliver a transparent and consistent report, validating the technological advances in a critical sector. Multiple pertinent articles and papers were retrieved from the Scopus abstract and citation database on May 13, 2022, using several relevant keywords to identify potentially relevant documents published from January 2020 onward. These documents must have already been published in English and been already published, and accessible through the B-ON consortium that allows Portuguese students to legally download from most publishers. Over seven studies have been selected for deeper discussion based on their relevance and impact for this review, showcasing their main objectives, data sources, and tools used, as well as their approaches for interoperability and support of machine learning algorithms for decision support. In closing, the collected articles have shown that while Big Data is currently in use at health institutions of all sizes, the ability of processing large amounts of data from sensors and events, and notifying stakeholders as quickly as possible is still in its infancy.

**Keywords:** Data Engineering; Business Intelligence; Big Data; Machine Learning; Streaming Systems; Data Lake; Data Mining; Event-Driven Microservices

**State of Publication:** Published