Contents

List of Figures

List of Tables

Chapter 1

Introduction

Health has played an important role in human history, helping civilization, behind the curtains, to evolve into the society of today[56]. Recently, the healthcare sector has witnessed the development of a wide range of IoT devices and applications[44]. And a new field has been unlocked: Healthcare information technology (HIT).

1.1 Background

Healthcare information technology (HIT) has been defined as "the application of information processing involving both computer hardware and software that deals with the storage, retrieval, sharing, and use of healthcare information, data, and knowledge for communication and decision making[31]" where the oil of it is the Medical Informatics as Morris F Collen defines it: "Medical informatics is the application of computer technology to all fields of medicine-medical care, medical teaching and medical research"; in other words The medical informatics is the foundation for understanding and practice of the up-to-day medicine. Its basic tool is the computer, the subject of studying and the means by which the aspects and achievement in the new knowledge in studying of a man, his health and disease and functioning of the total health activities is performed[41].

Medical informatics as a discipline is still young, in particular when you compare it with other medical disciplines. However, within the past decades, societies in general, and medicine and healthcare in particular, have tremendously changed by the adoption of health information technology. This change has significantly impacted the healthcare field as well[36]. As a result, health information technology improves patient's safety by reducing medication errors, reducing adverse drug reactions, and improving compliance to practice

guidelines. There should be no doubt that health information technology is an important tool for improving healthcare quality and safety[24].

1.2 Problem & motivation

With the progress of health information technology the healthcare data is increasingly digitized and, like in most other industries, data is growing in Velocity, Volume and Value. According to Statista[7], the amount of global healthcare data is expected to increase dramatically by the year 2020. Early estimates from 2013 suggest that there were about 153 exabytes of healthcare data generated in that year. However, projections indicate that there could be as much as 2,314 exabytes of new data generated in 2020 (Figure 1.1).

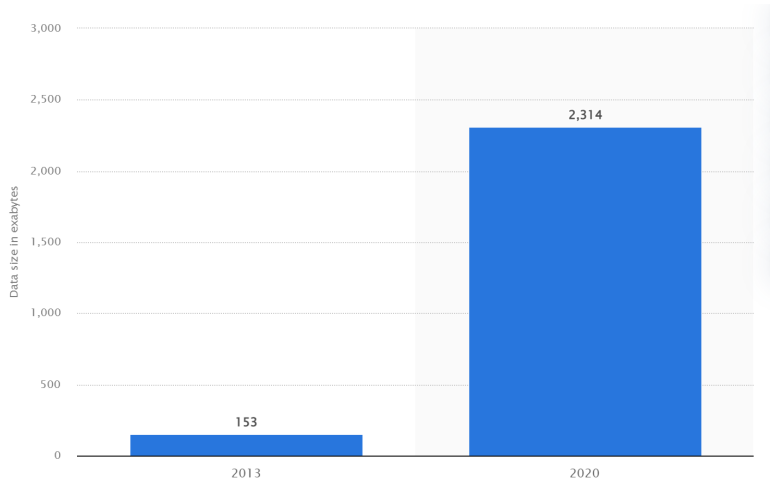

Figure 1.1: Total amount of global healthcare data generated in 2013 and a projection for 2020* (in exabytes).

Health Data Management is the practice of making sense of this data and managing it to the benefit of healthcare organizations, practitioners, and ultimately patient well being and health, It enables the integration and analysis of medical data to make patient care more efficient, and extract insights that can improve medical outcomes, while protecting the security and privacy of the data. In the past forty years, medical data began a transition from purely paper-based tracking to digitized information. Even today, many types of medical data have yet to be digitized, or have not yet been integrated into Health Data Management systems. Some of the important challenges facing health data professionals today are[6]:

- ***Fragmented data:*** medical data can be structured data in spreadsheets or databases, images or video files, digital documents, scanned paper documents, or may be stored in specialized formats such as the DICOM format used for MRI scans. Data is widely duplicated, collected multiple times and stored in different versions by healthcare providers, public health organizations, insurance bodies, pharmacies, and patients themselves. There is no one source of truth for information on patient well being.

- ***Changes to data:*** medical data constantly changes as do the names, professions, locations and conditions of patients and physicians. Patients undergo numerous tests and are administered many types of treatment over the years, and the treatments and medications themselves evolve over time. New types of medical treatment, such as telehealth models, create new types of data.

- ***Regulations and compliance:*** medical data is sensitive and must adhere to government regulations, such as the USA Health Insurance Portability and Accountability Act (HIPAA). Data discovery challenges and poor data quality make it much more difficult to perform the required audits and meet regulatory requirements and limits the diversity of data healthcare providers can use for the benefit of patients.

1.3 Purpose & delimitations

The goal of this work is to design a visualization system for medical data in a conventional common language between the different health actors.

We intend to achieve our goal by designing a warehouse system that brings data from different sources together, and structures it using the eXtensible Markup Language (XML).
First, we created a data warehouse that takes care of the data integration.
Second, we integrated the focus data in a structured form using XML.
Third, we designed a visual presentation after the data processing and formatting.

1.4 Document structure

This document is presented in 6 chapters, starting with the chapter1: Introduction, in which we present a bit of background of the topic and then delve into

formally defining the problem we intend to tackle, followed by a brief description of what lies within and beyond the scope of this work.

In Chapter 2: Health sector & Data, we present the healthcare environment including the principal actors, activities, and the type of data generated from each one. Then we presented the Medical records management and its various electronic types and how important security is to them.

In chapter 3: Information Visualization, we present its definition, then we explain the visualization pipeline: how it deals with data, then we move to the data warehouse and its data integration approaches.

Next comes chapter 4: Contribution & Discussion, we talk about the work objectives then we go through several related work reviews and then present the proposed work. Chapter 5: Implementation, a chapter about the implementation of the system, which tools are introduced and results are displayed using screenshots and diagrams.

Finally, in chapter 5: Conclusion & Future work, the results and insights gained through the journey of making the proposed solution, few conclusions drawn and perspectives on what could be enhanced moving forward with this project.

Chapter 2

Health sector & Data

In this chapter, we provide an overview of data from the health sector and the actors involved in it, by presenting the components of a health system and the types of data generated by each activity. In the second part, we will move on to the multiple electronic records used in the healthcare.

2.1 Healthcare Environment

Healthcare is a multi-dimensional system established with the sole aim for the prevention, diagnosis, and treatment of health-related issues or impairments in humans. There is three components of a healthcare system[32]:

- The health professionals (physicians or nurses): belong to various health sectors like dentistry, medicine, midwifery, nursing, psychology, physiotherapy, and many others.
- Health facilities (clinics, hospitals for delivering medicines and other diagnosis or treatment technologies).
- Financing institution supporting the former two.

Healthcare is required at several levels depending on the urgency of situation:

1. ***Primary care:*** Professionals serve it as the first point of consultation.
2. ***Secondary care:*** acute care requiring skilled professionals.
3. ***Tertiary care:*** advanced medical investigation and treatment.
4. ***Quaternary care:*** highly uncommon diagnostic or surgical procedures.

At all these levels, the health professionals are responsible for different kind of information such as a patient's medical history (diagnosis and prescriptions related data), medical and clinical data (like data from imaging and laboratory examinations), and other private or personal medical data

Regardless of what form it takes, data has the potential to tell stories, identify cost savings and efficiencies, new connections and opportunities, and enable improved understanding of the past to shape a better future[57].

The term "big data" has become a buzzword in recent years, with its usage frequency having doubled each year in the last few years according to common search engines (Figure 2.1).

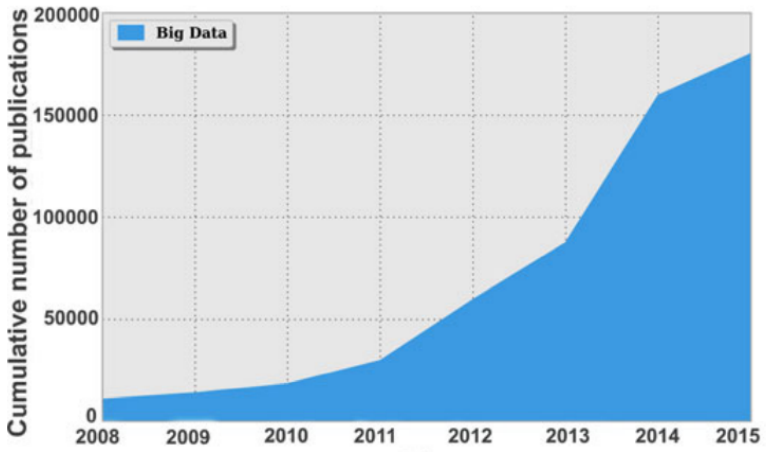

Figure 2.1: Cumulative number of publications referring to "big data" indexed by Google Scholar.

Big data is a vague term with a definition that is not universally agreed upon. A definition by Demchenko et al[33] who define Big Data by five V's: Volume, Velocity, Variety, Veracity, and Value. Volume pertains to vast amounts of data, Velocity applies to the high pace at which new data is generated, Variety pertains to the level of complexity of the data, Veracity measures the genuineness of data, and Value evaluates how good the quality of the data is in reference to the intended results.

If we trace the relationship between the use of the term big data per health research , we can easily infer the growth of medical informatics (Figure 2.2). Big data in health is concerned with meaningful datasets that are too big, too fast, and too complex for healthcare providers to process and interpret with existing tools[27].

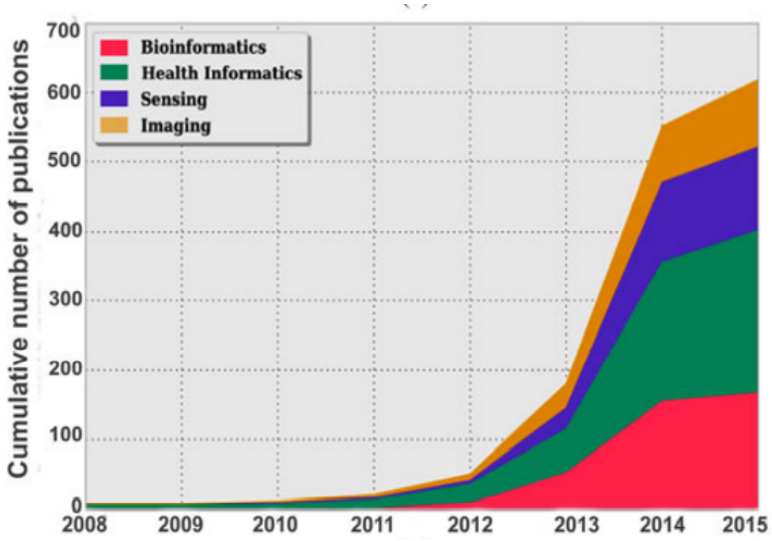

Figure 2.2: Cumulative number of publications per health research area referring to "big data," as indexed in IEEE Xplore, ACM Digital library, PubMed (National Library of Medicine, Bethesda, MD), Web of Science, and Scopus.

There are numerous current areas of research within the field of Health Informatics, including Bioinformatics, Image Informatics (e.g. Neuroinformatic), Clinical Informatics, Public Health Informatics, and also Translational BioInformatics (TBI). Research done in Health Informatics (as in all its subfields) can range from data acquisition, retrieval, storage, analytics employing data mining techniques, and so on.

Data gathered for Health Informatics research does exhibit many of these qualities. Big Volume comes from large amounts of records stored for patients: for example, in some datasets each instance is quite large (e.g. datasets using MRI images or gene microarrays for each patient), while others have a large pool with which to gather data (such as social media data gathered from a population). Big Velocity occurs when new data is coming in at high speeds, which can be seen when trying to monitor real-time events whether that be monitoring a patient's current condition through medical sensors or attempting to track an epidemic through multitudes of incoming web posts (such as from Twitter). Big Variety pertains to datasets with a large amount of varying types of independent attributes, datasets that are gathered from many sources (e.g. search query data comes from many different age groups that use a search engine), or any dataset that is complex and thus needs to be seen at many levels of data throughout Health Informatics.

Schematically, several health-related activities can be distinguished[40]:

- Pre-admission, admission and administrative discharge activities.
- T2A invoicing or valuation activities.
- Care activities in the accommodation service.
- Activities in operating theaters and technical platforms.
- Laboratory activities.
- Imaging activities.

For each activity, different types of data are generated. In France and most developed countries, the following data is collected and digitally available (in chronological order)[1, 30]:

1. **Administrative data** related to patient movements (identity, dates, places, etc.), demographic (age, sex, place of residence, etc.) and insurance (health coverage, etc.).
2. **Results of biological analyses**, generally taken by nurses and analyzed by professionals or by robots.
3. **Medical data** produced automatically by autonomous medical devices. These devices can be implantable or external.
4. **Data relating to the drugs administered to the patient**, generally by nurses or doctors, possibly as part of a diagnostic or therapeutic procedure.
5. **Data relating to medical devices** implanted in the patient during surgery.
6. **Data relating to medical procedures**, whether diagnostic or therapeutic. These data are generally coded by the producer, sometimes by the machine which produces them.
7. **Comments** in free text, possibly formalized in letters or reports.
8. **Medical diagnoses**, coded a posteriori by the doctors who treated the patient, or by specialized technicians reading the letters[35].

This data can be structured (which can be used directly by an algorithm) or unstructured (they are stored without a predefined format, such as the text of reports or medical letters, and are interpreted by humans). Machines generally produce raw structured information (eg medical biology measurements), while healthcare professionals exchange unstructured information with high interpretative value (eg a diagnosis). Medical records are created by aggregating information from different sources: (Table 2.1) gives an overview of this data[40].

Category	Nature	Structure
Administrative data	Values, Text	Structured
Communications between caregivers (Transmissions and medical observations, medical letters, reports)	Text	Unstructured
Data managed by centralized pharmacies	Values	Structured
Medical biology results	Values, Text	Structured
Data from monitoring devices	Values	Structured
Image Data Images	Images, Text	Unstructured
PMSI data and codes	Codes, Values	Structured

Table 2.1: Origin, nature and structure of medical record data.

2.2 Medical records management

Self-tracking and documenting information about aspects of one's personal and daily life has a long history. It is an effective method which helps us to learn more about ourselves, rather than depending on our limited memory[25].

The medical record is a multifunctional document that is used to communicate and document critical information about patients' medical care among health care professionals. Comprehensive medical records are a cornerstone in the quality and efficiency of patient care during the hospitalization and in subsequent follow-up visits, as they can provide a complete and accurate chronology of treatments, patient results and future plans for care[55], it involves many kinds of records, including patient charts, x-rays, images, scans, and even emails. Additionally, it involves making sure all of these items are accessible, safe, and secure. There are multiple electronic records used in the healthcare.

2.2.1 Electronic medical record

Electronic medical record (EMR) systems, defined as "an electronic record of health-related information on an individual that can be created, gathered, managed, and consulted by authorized clinicians and staff within one health care organization," have the potential to provide substantial benefits to physicians, clinic practices, and health care organizations.

It is a digital version of the paper medical record that has been used for years and it will contain the patient's medical and surgical history, allergy information, treatment history, current, and past prescriptions, and other pertinent information that can be used in making future medical decisions[19]. These systems can facilitate workflow and improve the quality of patient care and patient safety[5].

2.2.2 Electronic health record

An Electronic Health Record (EHR) is an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the key administrative clinical data relevant to that persons care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports. The EHR automates access to information and has the potential to streamline the clinician's workflow. It also has the ability to support other care-related activities directly or indirectly through various interfaces, including evidence-based decision support, quality management, and outcomes reporting.

EHRs are the next step in the continued progress of healthcare that can strengthen the relationship between patients and clinicians. The data, and the timeliness and availability of it, will enable providers to make better decisions and provide better care[4].

2.2.3 The Difference Between EMR & EHR

Both the EMR and EHR contain electronic versions of a patient's medical history. Most of the information in an EMR goes into an EHR.

The EMR can contain medical history, diagnoses, medications, immunizations and dates, allergies, etc. Often, a patient needs to ask for a printed copy of an EMR to share with another medical provider.

The EHR contains similar details as the EMR, but also other relevant data like information from wearable devices, demographics, and insurance information. It can also contain lab data and imaging reports that come from other

offices or practices. Assuming the software is compatible, other offices and practices can access the information within an EHR to help coordinate care and make clinical decisions[9] (Table 2.2) gives an overview of this difference:

EMR (electronic medical record)	**EHR (electronic health record)**
Medical and clinical data gathered in one provider's office	Medical and clinical data gathered from many providers' offices and hospitals
Narrower view	Broader view
Digital version of a paper chart in one office	Digital version of varied health information
Not designed for sharing	Designed for sharing outside of an individual medical practice
Providers use mainly for diagnosis and treatment	Providers have access to many diagnostic tools to make decisions

Table 2.2: EMR vs. EHR: Similarities and Differences.

2.2.4 Personal health record

Electronic Personal Health Records (ePHRs) are a representation of health records connected to the care of a patient and are managed by the patient[34], unlike EHRs, which are managed by health care providers. ePHRs allow healthcare consumers the luxury of deciding which health information to share with healthcare providers[50]. Ozok et al.[11] defined ePHR systems as patient centric, multi-functional, health management systems developed for managing and storing lifelong personal health information for various purposes from chronic to critical, medical and preventive care[26].

The information in an EHR is keyed in by healthcare providers and is only accessible to healthcare providers. In addition, an EHR might only contain information from a single healthcare provider. On the other hand, an individual will retain control of their own ePHR, which might encompass health information from different sources, such as various healthcare providers, as well as from the patient, as integrated ePHRs have the capability to incorporate data from different sources. Thus, at any one time, there may be various EHRs for one person but only one ePHR[26].

2.3 Security in medical records management

The actual technology of an electronic medical record seems to be falling into place. As the world moves more toward the use of telemedicine to preclude

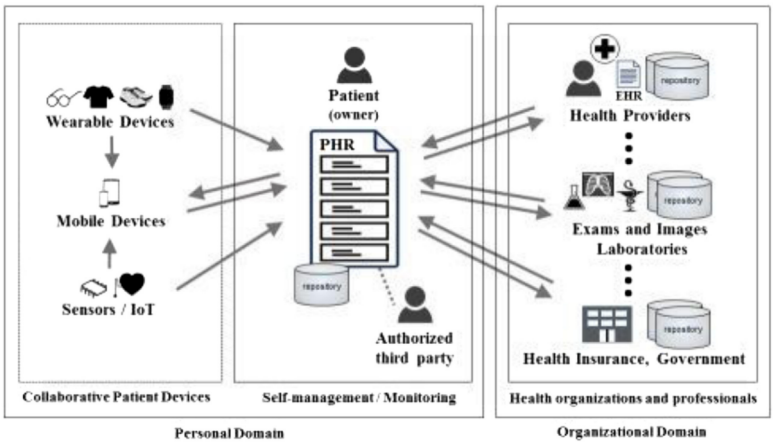

Figure 2.3: Difference between an EHR and ePHR (Taken from [26]).

the movement of patients to more advanced facilities, the need for a fully functioning medical record is paramount. One of the important ethical issues in electronic records management involves privacy. Privacy is the "claim of individuals to be left alone, free from surveillance or interference from other individuals, organizations, including state" (Laudon and Laudon 2005:159), privacy deals with the collection and use or misuse of data. Data is constantly being collected and stored on each of us. This data is often distributed over easily accessed networks and without our knowledge or consent[45].

There are a number of security dilemmas in electronic records management. There can be illegal access and use of records, data alteration and destruction (Stair and Reynolds 2006: 583).

Typical cross-organisational e-health applications are[46]:

- sharing of patient records among different healthcare professionals;
- access to distributed EHRs any place and any time;
- on-line teleconsultation, telemonitoring and assistance;
- patient—doctor consultation services;
- patients' access to their own EHRs.

That is why Electronic health records management attracts significant international interest and sets the scenery for the establishment of a distributed, coalition-based, security policy enhanced records exchange framework among different medical domains. Several European projects have proposed candidate solutions for secure inter-operations between medical domains. In the

HARP project, security profiles related to access rights are dynamically downloaded to the client side. The MEDITRAV EUproject attempts to overcome national or linguistic barriers by adopting the solution of a multilingual portable personal record. These approaches, pose mainly their research effort on the security requirements for effective electronic health record management, still they confront mainly to stable infrastructures[29].

2.4 Conclusion

We have seen in this chapter the principal components of the healthcare sector from its principal actors through its activities to the data generated from each activity. Then we moved to the different medical electronic records management types, and we presented the differences between them. Finally we moved on to the importance of the security field in the treatment of the patient data and medical data, then we did a quick overview of some of the projects implemented in this regard.
In the next chapter we will present the Information visualization field and its applications.

Chapter 3

Information Visualization (infoVis)

Human mind is very visual, following Williams et al., visualization is “a cognitive process performed by humans in forming a mental image of a domain space. In computer and information science it is, more specifically, the visual representation of a domain space using graphics, images, animated sequences, and sound augmentation to present the data, structure, and dynamic behavior of large, complex data sets that represent systems, events, processes, objects, and concepts”[54]. The (Figure 3.1) below presents the Florence Nightingale’s ’Rose diagram’ published in 1858 showing the reduction in the number of deaths in military hospitals in Scutari arising from the changes she instituted[49]
.

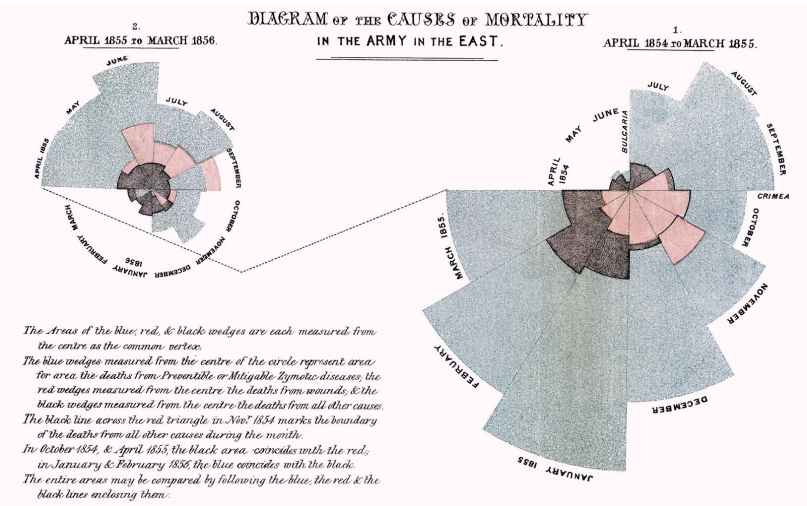

Figure 3.1: Rose Diagram

Data visualization involves presenting data in graphical or pictorial form which makes the information easy to understand. It helps to explain facts and determine courses of action. It will benefit any field of study that requires innovative ways of presenting large, complex information[49].

Traditionally, data visualization has been structured along two main fields: scientific visualization and information visualization. A third, newer field, called visual analytics has emerged in the past several years, as a bridge between and also an extension of the former two fields[51]. In this paper we will focus mainly on the information visualization field.

3.1 Definition

Information visualization (InfoVis) is the practice of representing data in a meaningful, visual way that users can interpret and easily comprehend, it is a research area that aims to aid users in exploring, understanding, and analyzing data through progressive, iterative visual exploration. With the boom in big data analytics, InfoVis is being widely used in a variety of data analysis applications in different domains, ranging from finance to sports to politics[39].

Information visualizations are often created with an audience in mind and designed to display certain important information that they need to understand. With an idea of how the visualization will be used, using multiple tools (Column chart, Bar graph, Network graph, Stacked bar graph, Histogram, Line chart, Pie chart, Box plot, Bubble chart, Dual-axis chart,...)3.2 that can help users compare different values, show the bigger picture, track trends in the data, and understand different relationships between variables[16]. These tools follow the model of the visualization pipeline.

Figure 3.2: Information visualization tools

3.2 Visualization pipeline

A visualization pipeline embodies a dataflow network in which computation is described as a collection of executable modules that are connected in a directed graph representing how data moves between modules. In a basic pipeline (Figure 3.3), there are three types of modules: sources, filters, and sinks. A source module produces data that it makes available through an output. File readers and synthetic data generators are typical source modules. A sink module accepts data through an input and performs an operation with no further result (as far as the pipeline is concerned). Typical sinks are file writers and rendering modules that provide images to a user interface. A filter module has at least one input from which it transforms data and provides results through at least one output[43].

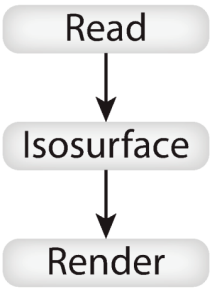

Figure 3.3: A simple visualization pipeline.

As science progresses, this model has been detailed, Figure 3.4 provides an overview of the infoVis pipeline. It has five main modules: Data Analysis, Filtering, Mapping, Rendering, Image data, explained as follows:

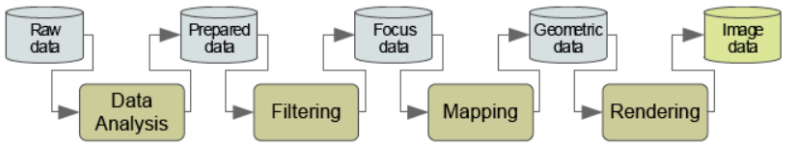

Figure 3.4: A visualization pipeline describes the process of creating visual representations of data.

1. ***Raw data:*** First, we have to import the data. This implies finding a representation of the original information we want to investigate in terms of a data set. Practically, importing data means choosing a specific dataset

implementation and converting the original information to the representation implied by the chosen dataset in order to turn this data into information using Data analysis.

2. ***Data Analysis:*** Is the process of bringing order and structure to collected data. mostly using data warehouse systems (3.3), It turns data into information that teams can use. Analysis is done using systematic methods to look for trends, groupings, or other relationships between different types of data[3], following this process:

    - ***Data Requirements Specification:*** The data required for analysis is based on a question or an experiment. Based on the requirements of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data may be numerical or categorical[2].
    - ***Data Collection:*** Guided by the requirements identified, Data can be collected through several sources, including online sources, computers, personnel, and sources from the community.
    - ***Data processing:*** The data that is collected must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (known as structured data) for further analysis, often through the use of spreadsheet or statistical software[21].
    - ***Data Cleaning:*** The processed and organized data may be incomplete, contain duplicates, or contain errors. Data Cleaning is the process of preventing and correcting these errors[2].
    - ***Perform data analysis:*** One of the last steps in the data analysis process is analyzing and manipulating the data. This can be done in a variety of ways depending on the cleaned data nature[18].

    The data analysis step produces the prepared data.

3. ***Filtering:*** Data filtering is the process of choosing a smaller part of your data set and using that subset for viewing or analysis[15], this portion of data called focus data.

4. ***Mapping:*** Focus data are mapped to geometric primitives (e.g., points, lines) and their attributes (e.g., color, position, size); most critical step for achieving expressiveness and effectiveness.

5. ***Rendering:*** The rendering operation is the final step of the visualization process, rendering takes the geometric data created by the mapping operation and transforms it to an image data.

3.3 Data Warehouse Systems

The concept of "data warehousing" arose in the mid 1980s with the intention to support huge information analysis and management reporting[53]. Data warehouse was defined According to Bill Inmon a “subject-oriented, integrated, time variant and non-volatile collection of data in support of management’s decision making process”[37].

According to Ralph Kimball "a data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making"[38].

There are three major areas in the data warehouse architecture as following:

- ***Data Acquisition:*** This step covers the process of extracting data from the multi sources, moving all extracted data to the stage and preparing the data for loading into the repository. The two main architectural components of this area are the source data and the data store, which is where all the extracted data is gathered and prepared for loading into the data warehouse.

- ***Data processing and storage:*** This stage covers all preparations and analysis that take place on our data from data cleaning to data processing to filtering till the rendering step (the steps mentioned in section3.2). At this stage, the data goes through a series of transformations to extract the focus data into clear and clean formats to build a sort of common language.

- ***Information visualization:*** This step focuses mainly on the visualization part, it makes it easy for the users to access the information directly from the data warehouse.

3.4 Data Integration Approaches

Data integration is the most tedious and time-consuming step in setting up a decision-making information system. During this step, the data is transformed and filtered to represent a homogeneous, common and stable source of information. The performance of the SID is closely linked to the quality of data integration. It should be noted that the data integration step is not limited to the decision-making domain. It is more general and can be applied for different needs: bringing together and requesting several operational information systems, communicating applications that have been made in silos (independently of each other), etc.

Several approaches have been developed depending on the integration needs. We present the most used approaches[28]:

- ***Extract Transform and Load (ETL):***
  This is the most used approach in setting up a data warehouse. In this approach, the integration is done in three steps:
  - Extracting data from sources.
  - Data transformation, which consists of cleaning and aggregating data to integrate them into a predefined schema.
  - Loading data into the target (the data warehouse).
- ***Enterprise Information Integration(EII):***
  In the EII approach, no physical integration is performed. Heterogeneous data sources are consolidated using a virtual database, transparent to applications using the data. The virtual database provides a unified view of data. Users send their request directly to the database. The query is then broken down into sub-queries that will be sent to the respective sources. The answers are assembled into a final result.
- ***Enterprise Application Integration(EAI):***
  In order to connect applications built in different environments and with different technologies, the EAI approach is based on application integration and sharing of their data using web services (SOA architecture). This approach allows for real-time communication. It is also used to feed data warehouses. This approach does not replace ETL.

Criteria	**ETL**	**EII**	**EAI**
Data flow	Unidirectional	Bidirectional	Bidirectional
Latency	Daily to monthly	Real time	Real time
Data transformation	Big capacity	Medium capacity	Low capacity
Context of use	-Consolidation of a large amount of data. -Complex transformations.	-Link an existing warehouse with specific data sources. -Source data volatile and accessible using simple queries.	-Sources not directly accessible. -Simple queries.

Table 3.1: Comparison between different integration approaches.

3.5 Conclusion

In this chapter, we went through the definition of Information visualization, presenting the visualization pipeline process, then moving to the data warehouse and its different data integration approaches.
In the next chapter, we present our contribution based on the different concepts we have already seen in the previous chapters.

Chapter 4

Contribution & Discussion

In this chapter we propose our data warehouse and visualization system for medical data that aims to manage the uncommon data from the various sources using a conventional common language between the different health actors.

4.1 Work Objectives

The aim of this work is to propose a data warehouse system that organizes data coming from different sources by producing a common language. We are interested in the step between Rendering and Data image in the pipeline process(3.4). After reviewing and analyzing similar proposed works, we created our own visualization system for medical data in a conventional common language.

4.2 Literature & Related works review

We present in this section architecture for healthcare data warehouses and solutions that attempt to integrate infoVis into medical data and medical structures which could be used by executive managers, doctors, physicians and other health professionals to support the healthcare process. Medical data existing today in multiple sources with different formats makes it necessary to have certain data integration techniques. A healthcare data warehouse is therefore needed to integrate the different data sources into a central data repository and analyze this data.

- ***A Healthcare Data Warehouse for Cancer Diseases:*** Dr.Osama E.Sheta and Ahmed Nour Eldeen discussed in their paper[48] the implementa-

tion of a healthcare data warehouse for cancer diseases, they proposed two stages approach for the building cancer data warehouse:

1. *Business Analysis:* Consist of business process analysis and business requirement analysis (Figure4.1).

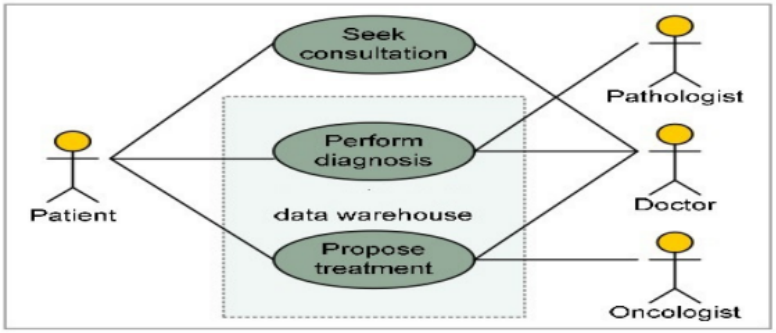

Figure 4.1: Cancer data warehouse use case diagram.

2. *Architecture Design:* Data is imported from several sources and transformed within a staging area before it is integrated and stored in the production data warehouse for further analysis (Figure4.2).

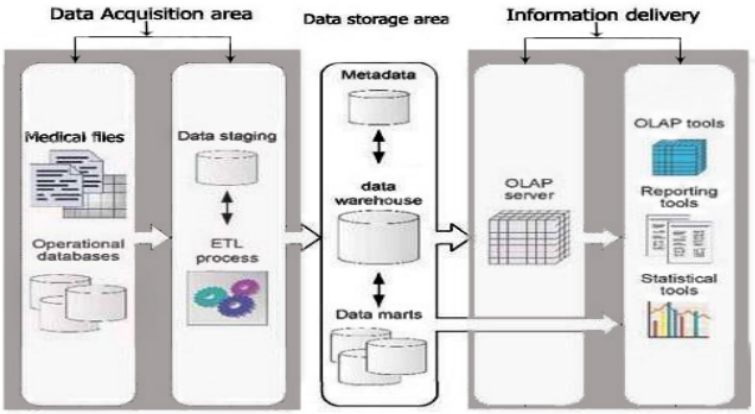

Figure 4.2: Cancer data warehouse Architecture Taken from the source.

- ***Data Warehouse Framework in Pharmaceutical Sector:*** In this paper[22] authors proposed a data warehouse framework to enhance decisions of distribution systems in pharmaceutical companies to decrease the medicine industry cost and increase productivity. The framework can be described

in four phases shown in (Figure 4.3). Phase one consists of a data preparation phase which has four steps (data collection, building DBs, DWH and data cleaning). Phase two consists of training phase which is applying time series to three types of Neural Networks techniques (levenberg marquardt, Bayesian regularized, and Scaled conjugate gradient).Phase three is testing the performance based on mean square error (MSE). Phase four consists of evaluating the performance of the best prediction model.

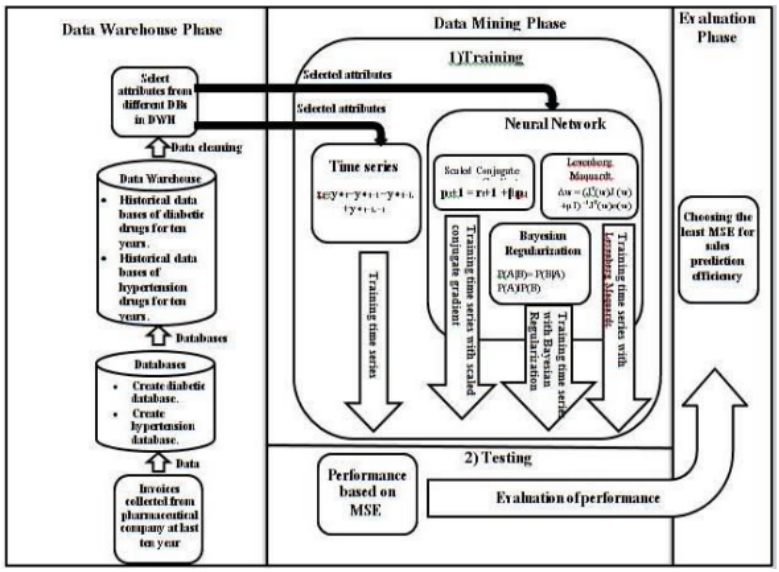

Figure 4.3: The Proposed Framework of Sales Prediction.

- ***Big Bata Warehouse Based On Hadoop Architecture:*** In this paper[47] entitled "Medical Big Data Warehouse: Architecture and System Design, a Case Study: Improving Healthcare Resources Distribution" authors proposed a system architecture and a conceptual data model for a MBDW (Medical Big Data Warehouse), and then offer a solution to overcome both the growing of fact table size and the lack of primary and foreign keys in the framework Apache Hive required in the conceptual data model. This solution is based on nested partitioning according to the dimension tables keys, then applying their solution to implement a MBDW to improve medical resources distribution for the health sector in the Bejaia region (in Algeria).

The overall architecture is depicted in (Figure 4.4). It is a scalable, reliable, and distributed architecture to extract, store, analyze, and visualize healthcare data extracted from various resources HIS (Hospitals Information systems).

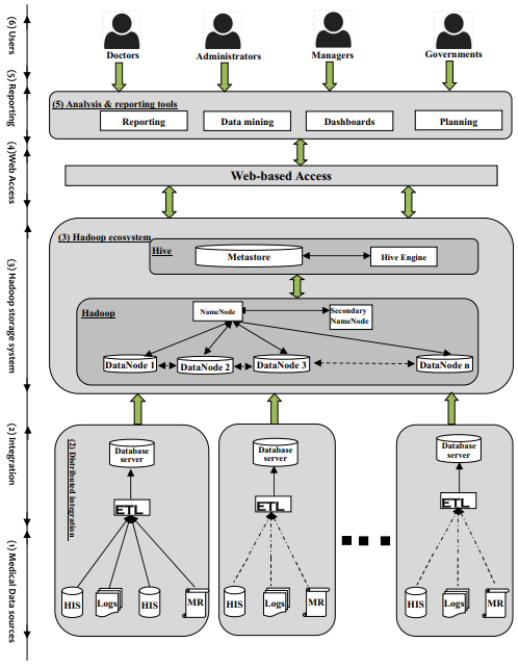

Figure 4.4: Hadoop-based system architecture of medical big data warehousing.

4.3 Proposed Solution

As mentioned in the project description, the presence of a visualization presentation in the medical sector is necessary. The proposed solution was to create a system that organizes the data and structures it, we used for that the eXtensible Markup Language (XML). We mainly focused on the "Rendering" step in the visualization pipeline(3.4), and we worked on the medical data that are mentioned in the Table2.1.

4.3.1 Why XML?

The rise of XML (eXtensible Markup Language) in patient care has been driven by the needs for communication among health professionals and between health-care organizations such as hospitals and health insurance companies.The main advantage of XML is its flexibility, as it allows creators to describe any content easily by generating their own tags[52]. Some of XML's features are[23]:

- The XML and DTD files are human readable and thus can be easily edited by people with only a few computer skills. Updating a data model is, therefore, straightforward (at least from a technical point of view).
- XML is Internet-oriented and has very rich capabilities for linking data; this can be used for interconnecting databases.
- XML provides an open framework for defining standard specifications. This is an important point because medical informatics clearly lacks standardization. For example, querying on multiple molecular biology databases could be greatly facilitated if each database would offer an XML view of their content.

On the other hand, XML has some weaknesses:

- The overhead of a text based format in data parsing, storage and transmission needs to be evaluated before adopting XML as a general solution. However, a text format means that the source code can be read and edited with any text editor.
- It is not clear whether XML satisfactorily addresses the problems of technological scalability. Indeed if XML data are stored in flat files, queries on XML files will not scale because XML in itself does not provide scalable facilities such as indexing or data clustering. This means that parsing should be done on the fly which leads to poor performances. One solution could be to have query optimizations done externally for example using a DataBase Management System (DBMS).

At this point the question to be answered is whether the pros prevail over the cons, for this reason Frederic Achard and al[23] have provided a comparison between XML and some of the most popular solutions that are used for the management and exchange of bioinformatics data summarized in the (Table4.1), each one is rated with one to four stars for different criteria: the higher the number of stars, the better the solution with regard to the criteria.

Criteria	**XML**	**Field/ value**	**ASN.1**	**CORBA**	**Java RMI**	**OODBMS**
Model expressiveness	**	*	***	***	***	****
Constraints	**	*	*	**	***	****
Self-descriptive	yes	no	yes	yes	yes	yes
Query language	soon[a]	no	no	soon[b]	no	yes
Flexibility	****	*	***	***	***	****
Simplicity	****	****	***	*	**	**
Scalability	**	*	**	***	***	****
Interoperability	****	*	**	****	****	***

Table 4.1: Summary of comparison of different alternatives to XML.

They conclude that the use of XML as an intermediate medium would be really efficient only if all databases share common or very similar DTDs. Whatever language is used, it is always difficult to find an agreement on a common semantics, and when one is found, it is often revised. However, XML would be an excellent candidate for this role because of its flexibility.

4.3.2 Architecture

We focused on extracting the following type of data (each head represents an XML node):

- **Patient:** That presents the administrative information about the patient x, including his full name, gender, date of birth, address.
- **Drug:** Presents the list of drugs in the patient's prescription, it includes: the drug's name, its dosage, strength usually mentioned by the doctor, the quantity, and its type(liquid or table).
- **Diagnosis:** Includes the doctor's name, the diseases name, and the observations taken by the medical actor in charge, it also contains the name of the prescribed medications and the date of diagnosis.

- **LabResults:** It presents the medical biology results that includes the analysis's name,the patient rate saved, the gender and the age for the comparison (a predefined high/low rate is defined corresponding to each analysis), attached with the laborator name.
- **ImagesData:** Present the data of the image obtained by the patient, it includes the image name and the image itself, the day it was taken, and the patient's national ID number.

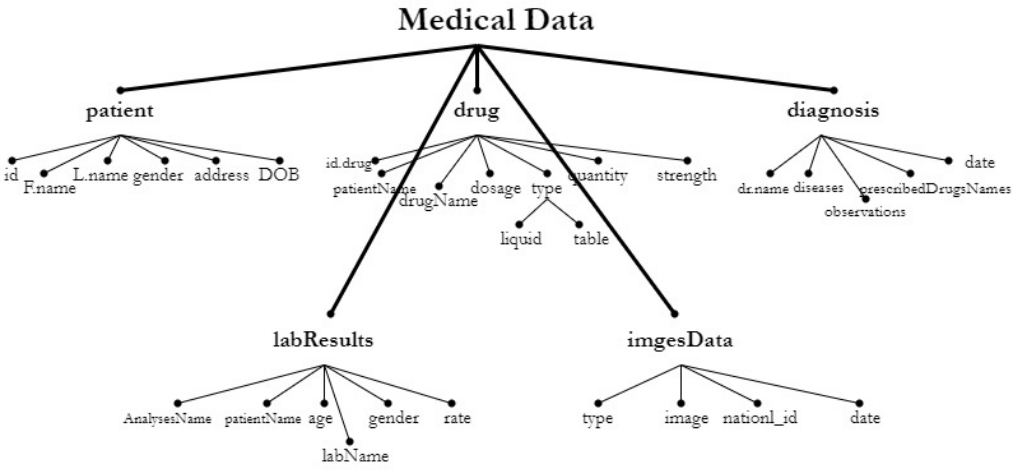

Figure 4.5: The corresponding tree of XML schema.

4.3.3 Data Integration & Processing

Our main objective was to create a visualization system for that: we collected medical data from different sources that represent various medical actors, and we process them following the visualization process mentioned before (3.4), we used Talend open studio (5.1.2) to handle this step.

The resulting data will be used as a source to feed all the components of the digital marketing reporting applications that are built in Tableau Software (5.1.3).

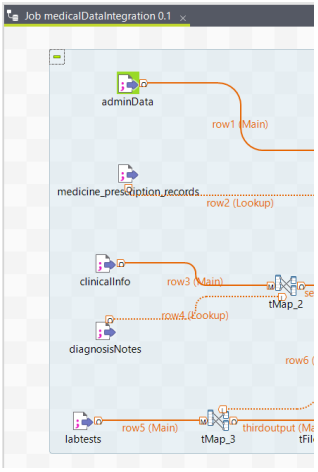

Figure 4.6: Input data.

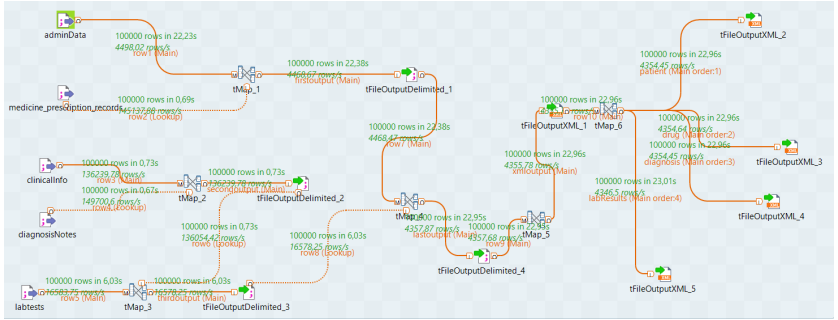

Figure 4.7: Talend job execution.

Chapter 5

Implementation

In this chapter, the process of implementation will be covered, starting by the architecture and the tools to the results and how everything fits together.

5.1 Tools

5.1.1 Python

Python[14] is a programming language that can be used in many contexts and is suitable for any type of use thanks to specialized libraries. However, it is particularly used as a scripting language to automate simple but tedious tasks. It is also used as a prototype development language when a functional application is needed before optimizing it with a lower level language. It is particularly widespread in the scientific world, and has many libraries optimized for numerical calculations[12].

Pandas: Pandas[10] is a library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical arrays and time series. Pandas is free software under the BSD license.

The main data structures offered by Pandas are series (to store data according to one dimension - size according to an index), DataFrames (to store data according to 2 dimensions - rows and columns), Panels (to represent data according to 3 dimensions, 4D Panels or Data Frames with hierarchical indexes also called Multi Index (to represent data according to more than 3 dimensions - hypercube))[20].

Figure 5.1: Python et Pandas Logo.

5.1.2 Talend

We used Talend Open Studio to create and develop
ETL processes. It is a tool based on Java and with an interface derived from that of Eclipse (Figure 5.2). It allows you to design ETL processes visually, and offers more than nine hundred components (the following list is not exhaustive)[13]:

- Connect to different data sources for reading and writing:
  - Flat files, .xml, .csv, xls, etc.
  - Relational databases (Postgresql, MsSql, etc) and Nosql.
- To manipulate the data, namely:
  - Filter them.
  - Apply aggregate functions on them.
  - Sort them.
- to organize data flows.

These components are then assembled as needed to design ETL processes[42].

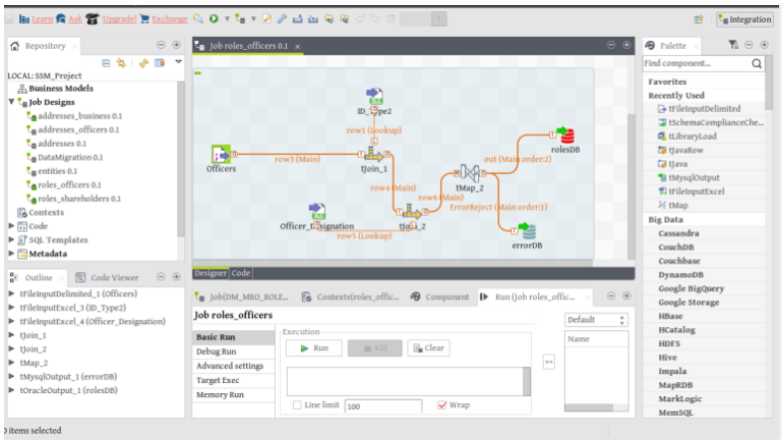

Figure 5.2: Talend Interface.

5.1.3 Tableau

Tableau is an excellent data visualization and business intelligence tool used for reporting and analyzing vast volumes of data. It helps users create different charts, graphs, maps, dashboards, and stories for visualizing and analyzing data, to help in making business decisions. Tableau supports powerful data discovery and exploration that enables users to answer important questions in seconds, it can connect to several data sources that other BI tools do not support. Tableau enables users to create reports by joining and blending different datasets and it supports a centralized location to manage all published data sources within an organization[17].

Figure 5.3: Tableau Logo.

5.1.4 Highcharts

Highcharts is a software library for charting written in pure JavaScript meant to enhance web applications by adding interactive charting capability. It has all the tools needed to create reliable and secure data visualizations by providing a wide variety of charts. For example, line charts, spline charts, area charts, bar charts, pie charts and so on.They offer wrappers for the most popular programming languages (.Net, PHP, Python, R, Java) as well as iOS and Android, and frameworks like Angular, Vue, and React[8].

Figure 5.4: Highcharts Logo.

5.2 Result

To validate the proposed solution, we visualize the data using various technologies highlighted in previous sections, in this section we will go through the results that were obtained by this investigatory effort.

5.2.1 Processed Data Result

After the data cleansing using python and its libraries, and the data processing using Talend, we got these results:

Chapter 6

Conclusion & Future work

Bibliography

[1] ATIH : Agence technique de l'information sur l'hospitalisation. https://atih.sante.fr/.

[2] Data Analysis - Process. https://www.tutorialspoint.com/excel_data_analysis/data_analysis_pr

[3] Data Analysis, Visualization and Interpretation | MEALD Pro Starter. https://mealdprostarter.org/n-data-analysis-visualization-and-interpretation/.

[4] Electronic Health Records | CMS. https://www.cms.gov/Medicare/E-Health/EHealthRecords.

[5] Electronic Medical Record Systems | Digital Healthcare Research. https://digital.ahrq.gov/electronic-medical-record-systems.

[6] Health Data Management: Benefits, Challenges and Storage.

[7] Healthcare data volume globally 2020 forecast. https://www.statista.com/statistics/1037970/global-healthcare-data-volume/.

[8] Interactive javascript charts library.

[9] The Key to Maintaining Medical Records | Smartsheet. https://www.smartsheet.com/medical-records-management.

[10] Pandas - Python Data Analysis Library. https://pandas.pydata.org/.

[11] Paper-based versus computer-based records in the emergency department: Staff preferences, expectations, and concerns - Haleh Ayatollahi, Peter A. Bath, Steve Goodacre, 2009. https://journals.sagepub.com/doi/10.1177/1460458209337433.

[12] Python (langage) — Wikipédia. https://fr.wikipedia.org/wiki/Python_(langage).

[13] Talend Open Studio for Data Integration. https://www.next-decision.fr/editeurs-bi/etl/talend-open-studio.

[14] Welcome to Python.org. https://www.python.org/.

[15] What is Data Filtering? - Displayr. https://www.displayr.com/what-is-data-filtering/.

[16] What is Information Visualization? https://www.tibco.com/reference-center/what-is-information-visualization.

[17] What is Tableau: The Ultimate Guide To Know All About Tableau in 2021. https://www.simplilearn.com/tutorials/tableau-tutorial/what-is-tableau.

[18] What Is the Data Analysis Process? 5 Key Steps to Follow. https://www.g2.com/articles/data-analysis-process.

[19] Wheel | What are Electronic Medical Records. https://www.wheel.com/companies-blog/what-are-electronic-medical-records.

[20] Pandas. *Wikipédia*, June 2020.

[21] Data analysis. *Wikipedia*, May 2022.

[22] Noura Mahmoud Abd Elazeem, Nevine Makram Labib, and Aliaa Kamal Abdella. A proposed data warehouse framework to enhance decisions of distribution system in pharmaceutical sector. *Egyptian Computer Science Journal*, 43(2):43–60, 2019.

[23] Frederic Achard, Guy Vaysseix, and Emmanuel Barillot. Xml, bioinformatics and data integration. *Bioinformatics*, 17(2):115–125, 2001.

[24] Yasser K Alotaibi and Frank Federico. The impact of health information technology on patient safety. *Saudi medical journal*, 38(12):1173, 2017.

[25] Majedah Mohammad Alrehiely. Evaluating Different Visualization Designs for Multivariate Personal Health Data. page 461.

[26] Yaser A Alsahafi and Valerie Gay. An overview of electronic personal health records. *Health Policy and Technology*, 7(4):427–432, December 2018.

[27] Javier Andreu-Perez, Carmen CY Poon, Robert D Merrifield, Stephen TC Wong, and Guang-Zhong Yang. Big data for health. *IEEE journal of biomedical and health informatics*, 19(4):1193–1208, 2015.

[28] Sarah NAIT BAHLOUL. Les entrepôts de données pour le décisionnel: Concepts et notions de base. pages 1–73, 2019.

[29] Petros Belsis, Apostolos Malatras, Stefanos Gritzalis, Christos Skourlas, and Ioannis Chalaris. *Pervasive Secure Electronic Healthcare Records Management.* January 2005.

[30] SAS bOracle France. Detection of adverse drug events: proposal of a data model. *Detection and Prevention of Adverse Drug Events: Information Technologies and Human Factors*, 148:63, 2009.

[31] D Brailer. The decade of health information technology. *HHS Report, July*, 21, 2004.

[32] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):1–25, 2019.

[33] Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, and Cees De Laat. Addressing big data challenges for scientific data infrastructure. In *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, pages 614–617. IEEE, 2012.

[34] George Demiris, Lawrence B. Afrin, Stuart Speedie, Karen L. Courtney, Manu Sondhi, Vivian Vimarlund, Christian Lovis, William Goossen, and Cecil Lynch. Patient-centered Applications: Use of Information Technology to Promote Disease Management and Wellness. A White Paper by the AMIA Knowledge in Motion Working Group. *Journal of the American Medical Informatics Association*, 15(1):8–13, January 2008.

[35] Emmanuel Chazard. Réutilisation et fouille de données massives de santé produites en routine au cours du soin. page 174, July 2017.

[36] Reinhold Haux. Medical informatics: past, present, future. *International journal of medical informatics*, 79(9):599–610, 2010.

[37] William H. Inmon. *Building the Data Warehouse*. Wiley, Indianapolis, Ind, 4th ed edition, 2005.

[38] Ralph Kimball and Joe Caserta. *The Data WarehouseETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* John Wiley & Sons, Inc., Hoboken, 2011.

[39] Shixia Liu, Weiwei Cui, Yingcai Wu, and Mengchen Liu. A survey on information visualization: Recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, December 2014.

[40] Par Niels Martignene. visualisation unifiée de données cliniques temporelles, hétérogènes, hiérarchiques. page 66.

[41] Izet Masic. The history and new trends of medical informatics. *Donald School J Ultrasound Obstet Gynecol*, 7(3):301–302, 2013.

[42] Benmohamed Aek El Mehdi. Intégration continue dans un projet decisionnel au sein de la cnl. pages 1–73, 2018.

[43] Kenneth Moreland. A Survey of Visualization Pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):367–378, March 2013.

[44] Somayeh Nasiri, Farahnaz Sadoughi, Mohammad Hesam Tadayon, and Afsaneh Dehnad. Security requirements of internet of things-based healthcare system: a survey study. *Acta Informatica Medica*, 27(4):253, 2019.

[45] Mpho Ngoepe, Lebohang Mokoena, and Patrick Ngulube. Security, privacy and ethics in electronic records management in the South African public sector. *ESARBICA Journal: Journal of the Eastern and Southern Africa Regional Branch of the International Council on Archives*, 29, March 2011.

[46] Pekka Ruotsalainen. A cross-platform model for secure Electronic Health Record communication. *International journal of medical informatics*, 73:291–5, April 2004.

[47] Abderrazak Sebaa, Fatima Chikh, Amina Nouicer, and AbdelKamel Tari. Medical big data warehouse: Architecture and system design, a case study: Improving healthcare resources distribution. *Journal of medical systems*, 42(4):1–16, 2018.

[48] Osama El-Sayed Sheta and Ahmed Nour Eldeen. Building a health care data warehouse for cancer diseases. *International Journal of Database Management Systems*, 4(5):39–46, October 2012.

[49] Robert Spence. *Information visualization*, volume 1. Springer, 2001.

[50] Paul C. Tang, Joan S. Ash, David W. Bates, J. Marc Overhage, and Daniel Z. Sands. Personal Health Records: Definitions, Benefits, and Strategies for

Overcoming Barriers to Adoption. *Journal of the American Medical Informatics Association*, 13(2):121–126, March 2006.

[51] Alexandru Telea. *Data Visualization: Principles and Practice*. January 2008.

[52] Pham Thi Thu Thuy, Young-Koo Lee, and Sungyoung Lee. S-trans: Semantic transformation of xml healthcare data into owl ontology. *Knowledge-Based Systems*, 35:349–356, 2012.

[53] Teh Wah and Ong Sim. Development of a data warehouse for Lymphoma cancer diagnosis and treatment decision support. 6, March 2009.

[54] James G. Williams and And Others. Visualization. *Annual Review of Information Science and Technology (ARIST)*, 30:161–207, 1995.

[55] REX Wong and Elizabeth H Bradley. Developing patient registration and medical records management system in ethiopia. *International journal for quality in health care*, 21(4):253–258, 2009.

[56] Jian Xu, Laiwen Wei, Wei Wu, Andi Wang, Yu Zhang, and Fucai Zhou. Privacy-preserving data integrity verification by using lightweight streaming authenticated data structures for healthcare cyber–physical system. *Future Generation Computer Systems*, 108:1287–1296, 2020.

[57] Sonja Zillner, Tilman Becker, Ricard Munné, Kazim Hussain, Sebnem Rusitschka, Helen Lippell, Edward Curry, and Adegboyega Ojo. *Big Data-Driven Innovation in Industrial Sectors*, pages 169–178. Springer International Publishing, Cham, 2016.