CLINICAL DATA WAREHOUSE: A REVIEW

Alaa Alaa Khalaf Hamoud[1]

[1] College of Computer Science and Information Technology / University of Basrah. Basrah Iraq
Alaak7alaf@gmail.com

Ali Salah Hashim[2]

[2] College of Computer Science and Information Technology / University of Basrah, Basrah, Iraq.
alishashim2009@gmail.com

Wid Akeel Awadh [3]

[3] College of Computer Science and Information Technology / University of Basrah, Basrah, Iraq.
umzainali@gmail.com

***Abstract* - Clinical decisions are crucial because they are related to human lives. Thus, managers and decision makers in the clinical environment seek new solutions that can support their decisions. A clinical data warehouse (CDW) is an important solution that is used to achieve clinical stakeholders' goals by merging heterogeneous data sources in a central repository and using this repository to find answers related to the strategic clinical domain, thereby supporting clinical decisions. CDW implementation faces numerous obstacles, starting with the data sources and ending with the tools that view the clinical information. This paper presents a systematic overview of purpose of CDWs as well as the characteristics; requirements; data sources; extract, transform and load (ETL) process; security and privacy concerns; design approach; architecture; and challenges and difficulties related to implementing a successful CDW. PubMed and Google Scholar are used to find papers related to CDW. Among the total of 784 papers, only 42 are included in the literature review. These papers are classified based on five perspectives, namely methodology, data, system, ETL tool and purpose, to find insights related to aspects of CDW. This review can contribute answers to questions related to CDW and provide recommendations for implementing a successful CDW.**

Index Terms - Clinical Data Warehouse, Data Warehouse, ETL, Clinical Operational Systems, Electronic Medical Records.

I. Introduction

Interest in medical systems should be considered a priority because all stakeholders in the medical environment aim to provide the best services for patients and find the best platforms for decision making. Recently, clinical data have been used for new objectives aside from clinical purposes, such as research, treatment enhancement and critical decision making [1]. Clinical organizations are searching for new technologies to find relationships between uncorrelated clinical records, such as a patient's history, treatment, diagnosis, physician's notes, hospital records and personal information [2].

The costs of medicines and treatments are constantly increasing; thus, finding the tools and systems that reduce these costs is a goal of all medical institutions. A clinical data warehouse (CDW) is regarded as the best approach to achieve this goal. A decision based on a false or an incorrect data may lead to disastrous results rather than support decisions [3, 4].

A data warehouse (DW) is one of the most important platforms that help stakeholders in various disciplines make decisions. Data in the DW are integrated and modelled in multidimensional form, thereby making visualization and analysis fast and easy [5]. The types of data stored in DW should enable stakeholders and institutions to obtain high-quality results that support critical decisions [6].

The DW processes data from operational data storage systems. This process requires tools and hardware components to ensure safe storage and efficient analysis of large data that institutions, organizations, researchers and others need in making strategic and operational decisions. The DW is not only an instrument used in transferring data but is also a tool in consolidating, analyzing, querying and presenting information. The success of DW in many fields has encouraged clinical institutions to adopt it as a platform for research, management, analysis and decision making [7-9].

As a new approach of DW, CDW can enhance the quality of medical decisions and online data processing. CDW can serve as a basis for reporting, studying, planning and supporting clinical research. Moreover, CDW simplifies data processing, analysis and improves clinical decision making. The use of CDW in biomedical research faces many challenges. The required characteristics for implementing a successful CDW have not been defined clearly because many DWs in medical institutions focus only on management [9, 10].

CDW construction is a difficult task from planning to implementation. Different clinical procedures from intensive care to treatment contain a variety of data and produce heterogeneous data [11]. The implementation process of CDW if full of obstacles start from analyzing data sources and ending with implementing access tools (OLAP, KPI, and reports console). The difficulty of detecting the proper data form and how to consolidate the different data formats is a challenge. Another challenge is handling long-term clinical data, which differs from dealing with short-term clinical data. The challenge experienced by the stakeholders is that their needs vary depending on the clinical procedures and data formats. In this paper, the following questions will be answered: What are the main objectives of CDW? What are the proper tools to implement the extract, transform and load (ETL) process? What is the proper approach to implement

doi:10.25195/2017/4424

CDW? What are the security concerns related to implementing CDW? Does CDW implementation involve data privacy concerns? What are the systematic requirements for building a successful CDW? How CDW differs from other DW types? What are the most important issues related to data that affect the implementation of CDW? Does backup required? What is the preferable ETL tool to implement CDW?

The paper is organized as follows. Section 2 describes DW and CDW in simple terms. Section 3 lists the characteristics of the CDW. Section 4 provides the possible data sources that can be used in CDW. Section 5 briefly explains the ETL process, which is considered as the base operation in CDW. Section 6 lists the challenges and difficulties of CDW implementation, and Section 7 presents a review of related literature. Finally, Section 8 provides the conclusions and recommendations.

II. CDW

The DW has been defined from various perspectives. Inmon, the inventor of the DW, defined DW as “a subject-oriented, integrated, time-variant, non-volatile data in support of management decisions”. Being subject-oriented means that only the relevant data are collected and stored to present useful information related to the subject. Integrated property describes the stored data style and format where all data types, naming conventions, encoding, data domains and measurements should be unified in standard form. Non-volatile property ensures that the data stored in DW should not change after any operational process execution, where time-variant means that the data in the DW should be historical and present (see Figure 1) [12-14].

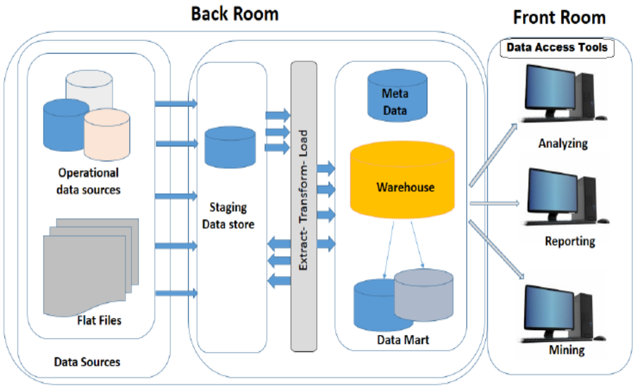

Fig 1: Architecture of DW

CDW, an emerging discipline of DW, refers to the central storage which provides access for different CDW stakeholders to utilize clinical data and knowledge so that they can analyze care situations and make critical decisions. CDW collates the data from different departments, laboratories and operational data stores into a single storage system [8, 14]. CDW processes the DW information related to hospitals and validates which data can be used for research, management, clinical practice and/or administration. CDW may be used by all healthcare stakeholders to access clinical data and obtain results in different disciplines to support decision making. The clinical data in CDW vary and differ from information related to patients’ records (such as treatments, procedures performed, vital signs, demographics, treatment costs and supplies used) to research, management and administration data. CDW is distinguished into many categories that support research, such as single-institution CDW, multi-institution CDW and research usage of CDW [9, 15].

Opinions are divided on whether CDWs should be located inside or outside hospitals or clinical departments. Accordingly, if CDWs were located outside hospitals or clinical departments, implementation would be difficult because communication is required during ETL and data integration between clinical stakeholders and the IT team. Also, if the CDW is located outside clinical departments, it may be neglected. However, if the location of the CDW is outside of hospitals, integration with non-clinical data may be easier.

Several factors are required to deal with barriers for implementing a CDW, such as data integrity, sound temporal schema design, query expressiveness, heterogeneous data integration, knowledge evolution integration, source evolution integration, traceability and guided automation [11].

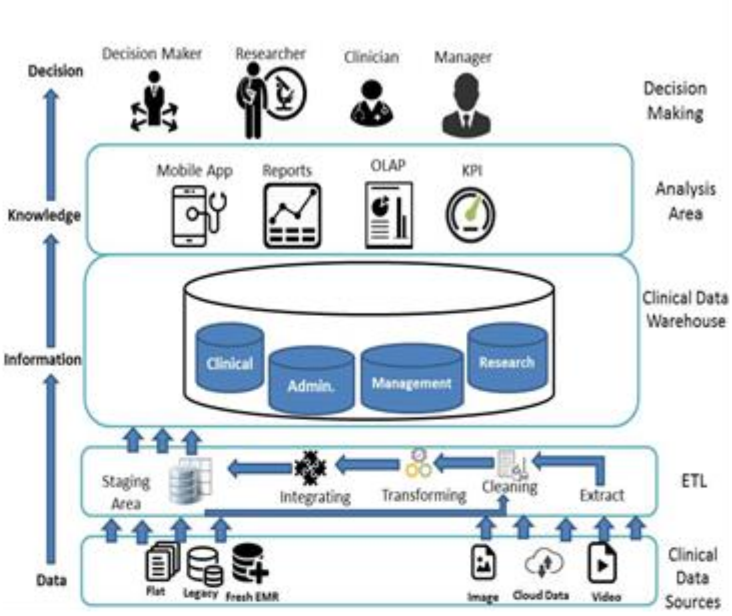

Fig 2: Architecture of CDW

Figure 2 shows a simplified architecture of the CDW, how the data go through ETL into the CDW and how the analysis tools are used by clinical stakeholders for decision making, research and management purposes. Data sources in CDW varies from other DW data sources types in many criteria such as variety and complexity of data structure, privacy concerns related to patients, different data types, duplication of clinical data, and variety of data sources platforms. Clinical data still emerging and different and complex data structures appear and these data structures need more analysis and time to add them to the existing CDW. ETL part in the most important part in CDW since it is the back room for implementing successful CDW. ETL tools still emerging and new tools are required to handle the new data and to turn the new data into useful information. The CDW type that stores the clinical information varies from enterprise DW to clinical data marts depends on the size of data and number of departments involved. The best recommended type to implement CDW is enterprise DW since it holds all the data related to the enterprise. Since it is difficult to implement clinical enterprise DW due to the different data of departments and difficult to handle all data in single schema but the benefit of enterprise clinical DW is to discover the hidden patterns in clinical data.

CDW reduces the time of collecting data and storing them in clinical operational data stores (CODS). The quality of the data is guaranteed by using CDW because the data are validated before storage in CODS [16, 17].

The benefits of using CDW are listed as follows [8-10, 16, 18, 19]:

- Helps determine the relationship between clinical data attributes, discover disease behavior, evaluate treatment procedures and increase patients' outcomes;
- Provides users with various information related to management and research fields;
- Uses the DW platform to enhance data quality and quantity, and improve query performance and business intelligence [20].
- Enhances the quality of care provided for patients;
- Uses a knowledge-based platform to make the right decisions on critical issues;
- Reduces the time spent on data collection and enhances data quality; and
- Provides a platform for timely analysis and online decision-making systems for administration, research, clinical and management systems [21].

III. CDW Characteristics

The clinical data should be collected depending on the nature and context, time and purpose of the future analysis [11]. CDW should ensure patients' privacy protection. Research design, chart presentation and data extraction are the major areas of CDW [9]. A major advantage of CDW is data quality because it is able to determine data reliability required for planning, analysis and decision making. However, data quality problems can occur in terms of accessibility, validity, freshness, relevance, completeness, consistency, reliability and integrity.

The clinical data quality is a critical issue because it affects the decision making and reliability of research. Data quality can be ensured by extracting the data that meet the needs of CDW stakeholders and storing them in a particular format. The problem of data quality occurs in various parts of data warehousing and ETL (data profiling, data staging and ETL processes) and CDW implementation (schema design and modelling) [22].

Medical ETL is sensitive to data quality and integrity because low-quality data could affect the clinical organization's income and decision-making process. The complexity of clinical data structure and diversity of medical operations requires implementation of complex ETL before the data are loaded into CDW storage. Different medical departments require various tools to connect different data sources and deal with a variety of data formats produced to apply ETLs [10, 23].

To establish and implement successful CDWs, many approaches are available such as requirements on users, information, regularity and ethics. User requirements may cause difficulties in ETL because the stakeholders have varying needs for CDW reports. Online and other ODSs and the types of data stored in them result in different user requirements. Information requirements refer to the types and costs of data used to accomplish ETL processes. ETL tools vary from open source to commercial, and have different capabilities and methods of processing specific types of datasets. Ethical and legal conditions are mandatory to maintain patients' privacy and protect their data. Patients have to be informed about the use of their data for research and how to cooperate further with data entry [24, 25].

Developing CDWs involves privacy and security constraints aside from policies related to medical data. The main issues in security and privacy are integrity, availability and confidentiality of the data to be shared between departments. Integrity means that the data should not be altered through any unauthorized action. Availability means that the data should be accessible to authorized persons any time. Confidentiality means keeping the data unreachable to unauthorized persons. Sharing patients' information between different departments and keeping confidentiality is a major challenge [26, 27].

The scenarios in keeping clinical data privacy are the following [28]:

- implementing a doctor–patient standard policy for sharing data;
- implementing data privacy restrictions during creation of the first ODS tables; and

- following government regulations to preserve patients' privacy.

In recent years, clinical systems have been attacked by hackers who have breached patients' medical files, billing and insurance records, payment details and other data. These incidents emphasize the need for CDWs to have high-security data protection.

IV. Data Sources

The data source is the foundation of CDW implementation. Clinical data sources vary depending on clinical procedures, devices and medical departments. The types of clinical data sources are the following:

- Laboratory, which represents results of laboratory tests;
- Diagnosis, which lists the details of the diagnosis process;
- Demographics, which is used to enrich the analysis of the environment data;
- Treatment, which refers to information related to treatment processes such as procedure, type and risk;
- Clinical, which represents patients' information related to their lifestyle and habits; this information can be used to improve the capabilities of data analysis [1, 10].

Table 1: CDW data compared with other domains [9]

Category	**Clinical**	**Other Domains**
Transaction	Unique	Repetitive
Data type	Mixed (text, code, number, image)	Number
Common vocabulary	Normalization required	Existing
Time value information	Significant	Not significant
External category	Essential	Not essential

Table 1 shows that the clinical data differ from that of other domains, thereby causing difficulty in the implementation process of CDW. The transaction related to the clinical domain is unique for each patient each time, whereas the transaction in other domains, such as banks and universities, is repetitive. As mentioned, the data types range from text, code, numbers, images and videos, while other domains can be implemented based on numbers only. The normalization process in the clinical data is required to remove duplications, while normalization is not required for the other domains and existing records can be depended upon. The time value information is not significant to other domains but is significant to clinical data. The external data sources and categories are essential to the clinical domain but is not essential to other domains.

Electronic health record (EHRs) that store clinical data of patients are one of the most frequently used data sources for CDW. ETL in CDW loads the raw data of electronic medical records (EMR) after extraction, cleaning, reconstructing and transforming them into CDW schema tables to make them consistent with the other clinical legacy databases. CDW can provide an analytical interface to access and assess the EHR and use the results in clinical research [9, 29].

The data quality parameters which should be ascertained are completeness, accuracy and consistency. Incomplete data from ODSs may generate additional tasks for the ETL. Data inconsistency and inaccuracy may also result in additional data preparation for the ETL. Low-quality data such as those mentioned above require new ways of handling and processing, which lead to additional efforts and costs [29].

V. ETL

ETL involves three stages of data handling: extraction, transformation and loading. Extraction is responsible for connecting the various data sources and extracting the data relevant for analysis and research. The difficulty in extraction is the existence of heterogeneous data sources that need different approaches for connecting and extracting. ETL requires specific tools to deal with heterogeneous data sources. The second stage is transformation, in which the extracted data are transformed to a specific format based on rules, functions and conditions in preparation for the next stage. The second stage ensures that the data are integrated and consolidated to facilitate the final stage. In the final stage, the data are transformed into dimensional forms and loaded into DW tables with star or snowflake schemas. The difficulty faced in the last stage of the ETL is how to handle and differentiate new data records from the existing data [7, 10, 30, 31].

VI. Issues and Challenges

The challenges and issues related to CDW are the following [1, 8-11, 16]:

- Data source independence. The independent clinical data sources with various conditions and environments that may cause different clinical systems are constructed with different storage media. Data source independence requires analysis and planning to implement the flexible ETL, which may take time and effort.
- Data availability. Availability of data across different sources depends on completeness and design. The old operational systems may work with various policies and obligations on data entry and types, which may affect the overall data accessibility. The massive increase in clinical data volume requires new setups to link the old and new data.

- Data format. The format of the clinical data ranges from text and images to videos and signals. The clinical data are also in numeric, qualitative, quantitative to image, ultrasound, sequential time, signal, protein and microarray forms [32].
- Data collection methods. The two types of data collection methods are manual and automated. Manual data collection consumes time and effort in data entry and is susceptible to errors that require cleaning. Automated data collection does not consume as much time and effort; thus, it is less susceptible to errors compared with manual data collection. Long-term clinical data related to specific diseases, such as continuous diagnosis, need a different approach compared with short-term medical data.
- Data integration tools. One of the most important challenges in CDW is implementing data integration tools. Data integration is the process of combining multiple data from uncorrelated data and from different departments in a single repository. The various clinical departments, treatment procedures, data types and attributes make the data integration process extremely difficult. The integration process involves rearranging, consolidating and integrating data in a unified form to analyze the data. Data preparation and integration time may consume 90% of the overall CDW construction, which requires efforts to analyze the data and build a solid schema.
- ETL issues. The different data formats from multiple data sources require ETL tools that can make the format flexible to enable data mining and machine learning approaches for information retrieval. Dealing with various data sources, schemas, attributes and data types is a challenging task in CDW. Handling old clinical data and transforming them into specific forms to be loaded into CDW tables require tools, scenarios and plans to merge with new data. Selecting the proper schema (whether star or snowflake) requires a large data analysis plane, which should be compatible with the resulting research reports.
- Legacy systems. Considerable time and effort have to be spent on collecting data from legacy clinical systems, but clinical data are beneficial for future research.
- Data quality. Data completeness, validity, accuracy, conformity and integrity problems should be addressed by using different solutions. Low-quality data should be refined and assessed based on specific criteria.
- Data privacy. Data extraction should ensure the patients' privacy and protection. Government policies and regulations are crucial aside from legal and ethical restrictions.
- CDW schema. Relational and dimensional data model designs are two familiar models for implementing DW. Relation model design can be used to solve data consistency and integrity problems and handle evolving volume data. Dimensional model design can be used for stable and known problems to fix end-user needs. In general, an ad-hoc architecture is preferable because the requirements vary from one department to another.
- Clinical institution standards. Lack of standards among institutions makes data gathering and integrating extremely difficult.
- Clinical stakeholders. These are all persons involved in the use of CDW, such as clinicians, physicians, researchers, doctors, managers and administrators in medical institutions. Clinical stakeholders can use CDW to improve healthcare, enhance patients' quality of living and decrease disease outbreaks by making the right strategic decisions [33].
- Analytical tools. The front end window holds different tools and approaches that use CDW to show results in the form of reports, charts and indicators. Different tools, such as online analytical processing (OLAP), can be implemented on CDW to present information such as key performance indicators [12].

VII. Literature Analysis

Searching through PubMed and Google Scholar for "clinical data warehouse" revealed 784 papers; after filtering, only 42 papers are included in this review. These papers will be classified and reviewed to answer the following questions: What are the main objectives of CDW? What are the proper tools to implement ETL? Do data privacy concerns exist? What are the other systematic requirements for building a successful CDW? The literature analysis in Table 2 classifies the papers according to five categories: methodology, system, data, ETL tool and purpose.

Table 2: Methodology and System Perspectives

Seq.	Ref.	Author	Methodology			System	
			Architecture	Design Approach	Dep.	Backup	Security
1	[34]	Nicolas		Top–Down	1H		√
2	[33]	Iain	√	Top–Down		√	√
3	[35]	John	√	Top–Down	1D		
4	[36]	Lekha	√	Top–Down			
5	[37]	Kislaya	√	Top–Down	7C	√	√
6	[38]	Christine		Top–Down			√
7	[39]	Nicolas		Top–Down			
8	[40]	Barrett		Top–Down			

9	[41]	Denis		Top–Down	1D		
10	[42]	Christoph		Top–Down			
11	[43]	Martin	√	Top–Down	1D		
12	[44]	Osama	√	Top–Down	1D		
13	[45]	Eric		Top–Down	1H		√
14	[46]	Christian		Top–Down	3D		
15	[47]	Jyoti	√	Top–Down	1C		√
16	[48]	Marleen		Top–Down	1C		
17	[49]	Alaa	√	Top–Down	1H	√	√
18	[50]	Young	√	Top–Down	1C		
19	[51]	Christopher		Top–Down			√
20	[52]	Tyler	√	Top–Down			√
21	[53]	Matthew		Top–Down	1D		
22	[54]	Marc		Top–Down			√
23	[55]	Mary					
24	[56]	Birger	√	Top–Down			√
25	[57]	Elene		Top–Down	12C	√	√
26	[58]	Hai		Top–Down			√
27	[59]	Anne		Top–Down	3D		√
28	[60]	Khan	√	Top–Down	8717C		
29	[61]	Luis		Top–Down			
30	[62]	Andrew	√	Top–Down			√
31	[63]	Alaa	√	Top–Down	1D	√	
32	[64]	Lumel		Top–Down			√
33	[65]	Dominic		Top–Down			√
34	[66]	Taxiarchis		Bottom–Up	1D		
35	[67]	Reesa		Top–Down	18D		√
36	[68]	Tanya		Top–Down	2D		√
37	[69]	Nicolas		Top–Down			√
38	[70]	David	√	Top–Down			√
39	[71]	Axel		Top–Down			√
40	[72]	Genes		Top–Down	1D		√
41	[20]	Monica		Top–Down			√
42	[73]	Jean		Top–Down	8S		√

A. Methodology

From a research perspective, concerns on design approach, architecture and number of departments are involved in CDW implementation. The research perspective provides a simple view of the entire design approach and the best methodology of the design plan to help the IT team and clinical stakeholders in understanding CDW. The methodology is divided into three criteria: architecture, design approach and departments.

i. Architecture

Architecture refers to the general structure of the CDW building process and how the CDW schematic components are connected. The architecture diagram can help in understanding the general implementation process. Many studies have demonstrated CDW architecture diagrams such as [33], [35], [36], [37], [43], [44], [47], [49], [50], [52], [60], [62], [63] and [70]. A few studies have presented their CDW schema approaches. As mentioned, the two familiar schemas are star and snowflake. The star schema consists of single fact table and tables called dimensions connected to a fact table by keys. The snowflake schema is an extended form of the star schema where many other tables are connected to the dimensions.

ii. 7.1.2 Design Approach

The two familiar design approaches are top–down and bottom–up. The top–down design approach provides the final shape of the system. This approach starts with implementing and constructing the pieces to reach the final goal. The bottom–up approach starts with dividing the large problem into small pieces of obstacles and solving each obstacle individually. Most studies used the top–down approach in designing, which saves time because the basic idea is clear and the required components are available. In the top–down approach, each team member knows the assigning task, which makes the system implementation flexible. For CDW, the proper development approach is top–down. It can be used as a systematic approach to help in decreasing integration obstacles. This approach is time consuming and difficult to implement because concept consistency is difficult to achieve for all clinical organization data. The bottom–up design approach is preferable for design, implementation and development of clinical data marts. This approach is characterized by flexibility and low implementation cost of CDW data marts, but it faces difficulty in integrating various data marts in the clinical enterprise of DW [74][75].

iii. 7.1.3 Departments

Departments refer to several clinical departments involved in the CDW implementation. Abbreviations used in clinical departments are H for hospitals, C for centers and D for hospital department or study. CDW covers data starting from one department to other hospitals and their departments and centers. A few papers mentioned the departments or data size used to implement CDW. The number of departments can provide a general view of the data sources and the data to be used to implement clinical data marts.

B. System

The CDW is a system, and two of the most important points in implementing any successful system are keeping it backed up and securing it from unauthorized access.

i. Backup

Backup is the process of keeping a copy of all data to use the image of the files in restoration when needed. Backup is a crucial process because it keeps all the data safe from loss when they are deleted or corrupted. As shown in Table 2, 5 (namely [33], [37], [49], [57] and [63]) out of 40 studies have used two systems of backing up their data or implementing a specific backup system to keep copies of all CDW data. The types of backup methods are incremental, differential, full and virtual full backup. Each method has its own capabilities and limitations. Full backup takes a snapshot of all the data while incremental backup takes a copy of the files that have been created or changed after the last backup. Differential backup stores only the new file changes after the last full backup, while virtual full backup takes a backup of all the data and synchronizes it with the original data periodically. Selecting a backup method and tool depends upon but is not comparable with data evaluability.

ii. Data Security

To safeguard CDW from unauthorized access, data security should be implemented and restrict access to specific persons in the decision making part. Access to CDW should be limited to clinical decision makers. Each authorized member to CDW should have a specific permissions to access specific part of CDW. Since CDW scope varies and cover many departments, so each department's should access the specific part that covered their needs and shows the results that support their decisions. As shown in Table 2, 25 CDWs apply security measures to prevent unwanted access. Physical protection is required in implementation, but the best solution to overcome security issues is to implement cloud storage technology, which can ensure safety, reduce costs, eliminate the need for physical protection and provide a reliable platform.

Table 3: Data, ETL Tool and Purpose Perspectives

Seq	**Ref.**	**Author**	**Data**				**ETL Tool**	**Purpose**			
			Size	**Availability**	**Privacy**	**Quality**		**Administration**	**Management**	**Research**	**Clinical**
1	[34]	Nicolas	>15m		√	√		√	√	√	√
2	[33]	Iain	>100 m	√	√	√	ODI	√	√	√	√
3	[35]	John		√		√		√	√	√	√
4	[36]	Lekha		√			SSIS	√	√	√	√
5	[37]	Kislaya	>3m	√	√	√		√	√	√	√
6	[38]	Christine	>99 m	√	√	√	i2b2	√	√	√	√
7	[39]	Nicolas					AT	√	√	√	√
8	[40]	Barrett	>4.4m	√	√	√		√	√	√	√
9	[41]	Denis	>17 thou.	√			Talend	√	√	√	√
10	[42]	Christoph	>10 m	√	√		i2b2+SQL	√	√	√	√
11	[43]	Martin	11,898	√	√	√		√	√	√	√
12	[44]	Osama				√	SSIS 2008	√	√	√	√

13	[45]	Eric	1.2 m	√	√	√	Oracle SAP	√	√	√	√
14	[46]	Christian			√	√	Talend	√	√	√	√
15	[47]	Jyoti	6.3 m	√	√	√	OBIEE	√	√	√	√
16	[48]	Marleen	75 GB	√	√		Extelligence Critical Care Export	√	√	√	√
17	[49]	Alaa	250 thou.	√	√	√	SSIS2014	√	√	√	√
18	[50]	Young	>1200	√				√	√	√	√
19	[51]	Christopher	268 m	√	√	√		√	√	√	√
20	[52]	Tyler		√	√	√	SAS	√	√	√	√
21	[53]	Matthew		√	√			√	√	√	√
22	[54]	Marc		√	√	√	Talend	√	√	√	√
23	[55]	Mary		√	√	√		√	√	√	√
24	[56]	Birger	2.17 m	√	√	√	i2b2	√	√	√	√
25	[57]	Elene		√		√		√	√	√	√
26	[58]	Hai		√	√	√		√	√	√	√
27	[59]	Anne	127m	√	√	√	Talend	√	√	√	√
28	[60]	Khan	12m	√	√	√		√	√	√	√
29	[61]	Luis	230 thou.	√			Java EE7 and Spring Framework	√	√	√	√
30	[62]	Andrew	15 m			√		√	√	√	√
31	[63]	Alaa	7 thou.	√	√	√	SSIS2014	√	√	√	√
32	[64]	Lumel	411 m	√	√	√	SQL+Python	√	√	√	√
33	[65]	Dominic		√	√	√	Kettle	√	√	√	√
34	[66]	Taxiarchis	2.7 m	√	√	√		√	√	√	√
35	[67]	Reesa	500 thou.	√	√	√	i2b2	√	√	√	√
36	[68]	Tanya				√	Oracle+SQL	√	√	√	√
37	[69]	Nicolas	2 m.	√		√		√	√	√	√
38	[70]	David		√	√	√		√	√	√	√
39	[71]	Axel		√	√	√	i2b2	√	√	√	√
40	[72]	Genes	>1b	√	√	√	Tool in Java	√	√	√	√
41	[20]	Monica	>136 m	√	√	√	i2b2	√	√	√	√
42	[73]	Jean	250G	√	√	√	Microsoft .Net 2.0	√	√	√	√

C. Data Processing

Data processing in CDW is the basic step in successful decision making. The four major components related to data processing are data size, data availability, data privacy and data quality. The size of the data involved in CDW implementation varies from thousands (thou), millions (m), billions (b) and gigabytes (G). Data size ranges from thousands to billions of records. The normal data size of DW varies from few kilobytes to a large terabyte. CDW is characterized by a large volume of clinical data composed of treatments, diagnosis records and EMRs, which are stored and processed to obtain analytical results.

i. Data Availability

Data availability means that data will still be accessible even if disastrous events occur. This perspective may depend on factors such as system security and backup, which can ensure the continuous availability of CDW. According to our research, 37 studies achieved data availability in implementing CDW, which proves the importance of this factor.

ii. Data Privacy

Data privacy should be ensured from the first step of the CDW implementation process. Data privacy refers to the protection of patients' personal information and determination of the parts that can be shared. A total of 32 studies achieved this objective in CDW implementation.

iii. Data Quality

Data quality is the measure of data usefulness. This concept refers to the data with consistency and unambiguity. As many heterogeneous data sources exist, this concept is difficult to achieve but is required. We found 34 studies that achieved this aspect in CDW implementation.

D. ETL tool

The ETL tool is necessary in CDW implementation. Selecting a license or source of ETL tool depends significantly on the project funding and nature of data sources. Only a few papers did not mention the ETL tool used in implementing CDW. i2b2 was used in six papers, SSIS in four papers, Talend also in four papers, while other papers used Java, SAS, AT and Microsoft.Net.

E. Purpose

The patient's information, financial information and medical data, including diagnoses, prescriptions, tests, medical records and nursing records were automatically updated from the operational EMR database to the DW system daily using an ETL tool [p26]. These procedures result in different data subjects. The main definition of CDW clarifies the purpose based on four goals: administration, management, clinical and research. CDW should achieve these goals to be successful. These four goals are derived from data types (ODSs) and stakeholders' needs. The four goals are correlated. Administration and management are the base purposes of building CODSs, and thus require information on managerial and administrational problems to fill gaps and enhance healthcare for patients. Clinical and research purposes are important to find hidden patterns, relationships between different attributes, disease behavior and ways to explore the clinical knowledge to support decisions.

VIII. CONCLUSIONS AND RECOMMENDATIONS

The various hardware components and software tools require complicated steps in integration to handle different data formats to produce improved information for research and decision making. The integration process should be built based on a planned approach to analyze the collected data and clean them to produce useful information. CDW is complicated because of the need for data integrity in ODS. The new ODS platforms make the ETL processes highly complicated and increase the need for new technologies.

Various clinical departments, procedures, jobs and tasks involve challenges due to new technologies, and new scenarios should be planned to adapt to these changes. Different data types are considered the first challenge in building and implementing a successful CDW. To ensure clinical data privacy, new government policies and regulations are needed. The data management in clinical ODS prior to the ETL process can influence the overall knowledge result from CDW.

One of the most important recommendations is to focus on clinical ODS and provide special courses for employees who work in data entry related to the process. Clinical ODS should not be accessed by all staff, and special security should be ensured to protect ODSs. A manual for clinical ODSs should be developed and new features should be added to enhance and reduce data collection time. Paper-based clinical records pose another challenge because the data contained in these records should be transformed to EMR, which is time consuming and laborious. Old legacy data sources should be treated carefully with a dedicated approach because they require all the ETL processes such as data cleaning, integration, transforming and loading.

The cost of building a CDW depends on the organization's needs and goals. The organization may decide to adapt low-cost or high-cost solutions by purchasing licenses of ETL, ODS, OLAP and reporting tools and software. Open-source solutions require a team to know all system requirements and methods to fix bugs. The CDW should be located inside hospitals and clinical departments, which makes CDW implementation fast and accurate, because it does not require agreements to work with clinical data sources, and it enables IT teams to obtain data-related answers directly from the clinical stakeholders.

The top–down design approach is preferred by many. This approach provides a pre-analysis of all operational data sources because it starts from analyzing all base components and goes further to all the implementing processes by integrating all heterogeneous clinical data sources. Despite the high cost and slow implementation of the top–down approach, it is the best choice for designing and implementing CDW. The bottom–up approach is preferable when CDW implementation starts from implementing CODSs, thereby making the CDW implementation process flexible and almost without obstacles. The bottom–up approach is preferable when the stakeholders decide to build separate data marts for each department.

Enhanced ETL is required to handle various data types and reduce the time spent on all ETL processes. Selecting the licenses for the ETL tool and overall project depends firstly on institutional funding. Open-source tools have proven their performance and accuracy in many studies but they need an experienced IT team to fix any bugs that may appear. A successful CDW should ensure data privacy; fulfil security and backup requirements with data availability and quality; and provide a solid clinical platform for research, administration and managerial purposes.

References

[1] Lau, E.C., et al., Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. Clinical epidemiology, 2011. 3: p. 259.

[2] Ballantyne, D.J. and M. Mulhall, Method and apparatus for electronically accessing and distributing personal health care information and services in hospitals and homes. 1999, Google Patents.

[3] Cao, H., et al. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. in AMIA Annual Symposium Proceedings. 2005. American Medical Informatics Association.

[4] Ewen, E.F., C.E. Medsker, and L.E. Dusterhoft. Data warehousing in an integrated health system: building the business case. in Proceedings of the 1st ACM international workshop on Data warehousing and OLAP. 1998. ACM.

[5] Chaudhuri, S. and U. Dayal, An overview of data warehousing and OLAP technology. ACM Sigmod record, 1997. 26(1): p. 65-74.

[6] Jaber, M.M., et al., Flexible data warehouse parameters: Toward building an integrated architecture. International Journal of Computer Theory and Engineering, 2015. 7(5): p. 349.

[7] Kimball, R. and M. Ross, The data warehouse toolkit: the complete guide to dimensional modeling. 2011: John Wiley & Sons.

[8] Sahama, T.R. and P.R. Croll. A data warehouse architecture for clinical data warehousing. in Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68. 2007. Australian Computer Society, Inc.

[9] Shin, S.-Y., W.S. Kim, and J.-H. Lee, Characteristics desired in clinical data warehouse for biomedical research. Healthcare informatics research, 2014. 20(2): p. 109-116.

[10] Mohammed, R.O. and S.A. Talab, Clinical data warehouse issues and challenges. International Journal of u-and e-Service, Science and Technology, 2014. 7(5): p. 251-262.

[11] Khnaisser, C., et al. Data warehouse design methods review: trends, challenges and future directions for the healthcare domain. in East European Conference on Advances in Databases and Information Systems. 2015. Springer.

[12] Hamoud, A.K. and T. Obaid, Using OLAP with Diseases Registry Warehouse for Clinical Decision Support. 2014.

[13] Evans, R.S., J.F. Lloyd, and L.A. Pierce. Clinical use of an enterprise data warehouse. in AMIA Annual Symposium Proceedings. 2012. American Medical Informatics Association.

[14] Stolba, N. and A.M. Tjoa, The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making. International Journal of Computer Systems Science and Engineering, 2006. 3(3): p. 143-148.

[15] Zhou, X., et al. Building clinical data warehouse for traditional Chinese medicine knowledge discovery. in 2008 International Conference on BioMedical Engineering and Informatics. 2008. IEEE.

[16] Roelofs, E., et al., Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiotherapy and Oncology, 2013. 108(1): p. 174-179.

[17] DeWitt, J.G. and P.M. Hampton, Development of a data warehouse at an academic health system: knowing a place for the first time. Academic Medicine, 2005. 80(11): p. 1019-1025.

[18] Yoo, S., et al., Electronically implemented clinical indicators based on a data warehouse in a tertiary hospital: its clinical benefit and effectiveness. International journal of medical informatics, 2014. 83(7): p. 507-516.

[19] Lieberman, M.I. and T.N. Ricciardi. The use of SNOMED© CT simplifies querying of a clinical data warehouse. in AMIA Annual Symposium Proceedings. 2003. American Medical Informatics Association.

[20] Horvath, M.M., et al., The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. Journal of biomedical informatics, 2011. 44(2): p. 266-276.

[21] Weng, C., et al. Comparing the effectiveness of a clinical registry and a clinical data warehouse for supporting clinical trial recruitment: a case study. in AMIA Annual Symposium Proceedings. 2010. American Medical Informatics Association.

[22] Tang, P.C., et al., Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. Journal of the American Medical Informatics Association, 2007. 14(1): p. 10-15.

[23] Denney, M.J., et al., Validating the extract, transform, load process used to populate a large clinical research database. International journal of medical informatics, 2016. 94: p. 271-274.

[24] Wiesenauer, M., C. Johner, and R. Röhrig, Secondary use of clinical data in healthcare providers-an overview on research, regulatory and ethical requirements. Stud Health Technol Inform, 2012. 180: p. 614-8.

[25] Schubart, J.R. and J.S. Einbinder, Evaluation of a data warehouse in an academic health sciences center. International Journal of Medical Informatics, 2000. 60(3): p. 319-333.

[26] Puppala, M., et al. Data security and privacy management in healthcare applications and clinical data warehouse environment. in Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on. 2016. IEEE.

[27] Khan, S.I. and A.S.L. Hoque. Privacy and security problems of national health data warehouse: a convenient solution for developing countries. in Networking Systems and Security (NSysS), 2016 International Conference on. 2016. IEEE.

[28] Darley, B., et al., How to Keep a Clinical Confidence: A Summary of Law & Guidance on Maintaining the Patient's Privacy. 1994.

[29] Botsis, T., et al., Secondary use of EHR: data quality issues and informatics opportunities. Summit on Translational Bioinformatics, 2010. 2010: p. 1.

[30] Kimball, R. and J. Caserta, The Data WarehouseÂ ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. 2011: John Wiley & Sons.

[31] Kimball, R., et al., The data warehouse lifecycle toolkit. 2008: John Wiley & Sons.

[32] Inmon, B., Data warehousing in a healthcare environment. The Data Administration Newsletter-TDAN. com, 2007.

[33] Karami, M., A. Rahimi, and A.H. Shahmirzadi, Clinical Data Warehouse: An Effective Tool to Create Intelligence in Disease Management. The health care manager, 2017. 36(4): p. 380-384.

[34] Garcelon, N., et al., A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. Journal of biomedical informatics, 2018. 80: p. 52-63.

[35] Chelico, J.D., et al. Designing a Clinical Data Warehouse Architecture to Support Quality Improvement Initiatives. in AMIA Annual Symposium Proceedings. 2016. American Medical Informatics Association.

[36] Narra, L., T. Sahama, and P. Stapleton. Clinical data warehousing for evidence based decision making. in MIE. 2015.

[37] Kunjan, K., et al. A Multidimensional Data Warehouse for Community Health Centers. in AMIA Annual Symposium Proceedings. 2015. American Medical Informatics Association.

[38] Turley, C.B., et al., Leveraging a statewide clinical data warehouse to expand boundaries of the learning health system. eGEMs, 2016. 4(1).

[39] Garcelon, N., et al., Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. Journal of the American Medical Informatics Association, 2016. 24(3): p. 607-613.

[40] Jones, B. and D.K. Vawdrey, Measuring Mortality Information in Clinical Data Warehouses. AMIA Summits on Translational Science Proceedings, 2015. 2015: p. 450.

[41] Delamarre, D., et al., Semantic integration of medication data into the EHOP Clinical Data Warehouse. Studies in health technology and informatics, 2015. 210: p. 702-706.

[42] Rinner, C., et al., A Clinical Data Warehouse Based on OMOP and i2b2 for Austrian Health Claims Data. Studies in health technology and informatics, 2018. 248: p. 94-99.

[43] Seneviratne, M.G., et al., Architecture and Implementation of a Clinical Research Data Warehouse for Prostate Cancer. eGEMs (Generating Evidence & Methods to improve patient outcomes), 2018. 6(1).

[44] Sheta, O.E.-S. and A.N. Eldeen, Building a health care data warehouse for cancer diseases. arXiv preprint arXiv:1211.4371, 2012.

[45] Zapletal, E., et al. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. in MedInfo. 2010. Citeseer.

[46] Karmen, C., et al. A framework for integrating heterogeneous clinical data for a disease area into a central data warehouse. in MIE. 2014.

[47] Kamal, J., et al. Information warehouse–a comprehensive informatics platform for business, clinical, and research applications. in AMIA Annual Symposium Proceedings. 2010. American Medical Informatics Association.

[48] De Mul, M., et al., Development of a clinical data warehouse from an intensive care clinical information system. Computer methods and programs in biomedicine, 2012. 105(1): p. 22-30.

[49] Hamoud, A.K. and T. Obaid, Building Data Warehouse for Diseases Registry: First step for Clinical Data Warehouse. 2013.

[50] Choi, I.Y., et al., Development of prostate cancer research database with the clinical data warehouse technology for direct linkage with electronic medical record system. Prostate international, 2013. 1(2): p. 59-64.

[51] Chute, C.G., et al., The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association, 2010. 17(2): p. 131-135.

[52] Ross, T.R., et al., The HMO research network virtual data warehouse: a public data model to support collaboration. Egems, 2014. 2(1).

[53] Krasowski, M.D., et al., Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. Journal of pathology informatics, 2015. 6.

[54] Cuggia, M., et al. Roogle: an information retrieval engine for clinical data warehouse. in MIE. 2011.

[55] Wisniewski, M.F., et al., Development of a clinical data warehouse for hospital infection control. Journal of the American Medical Informatics Association, 2003. 10(5): p. 454-462.

[56] Haarbrandt, B., E. Tute, and M. Marschollek, Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. Journal of biomedical informatics, 2016. 63: p. 277-294.

[57] Katzan, I., et al. The Knowledge Program: an innovative, comprehensive electronic data capture system and warehouse. in AMIA Annual Symposium Proceedings. 2011. American Medical Informatics Association.

[58] Hu, H., et al., DW4TR: a data warehouse for translational research. Journal of biomedical informatics, 2011. 44(6): p. 1004-1019.

[59] Jannot, A.-S., et al., The georges pompidou university hospital clinical data warehouse: a 8-years follow-up experience. International journal of medical informatics, 2017. 102: p. 21-28.

[60] Khan, S.I. and A.S.M.L. Hoque. Towards development of health data warehouse: Bangladesh perspective. in Electrical Engineering and Information Communication Technology (ICEEICT), 2015 International Conference on. 2015. IEEE.

[61] Marco-Ruiz, L., et al., Archetype-based data warehouse environment to enable the reuse of electronic health record data. International journal of medical informatics, 2015. 84(9): p. 702-714.

[62] Post, A.R., et al., The Analytic Information Warehouse (AIW): A platform for analytics using electronic health record data. Journal of biomedical informatics, 2013. 46(3): p. 410-424.

[63] Hamoud, A.K., et al., Design and Implementing Cancer Data Warehouse to Support Clinical Decisions. 2016.

[64] Waitman, L.R., et al. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. in AMIA Annual Symposium Proceedings. 2011. American Medical Informatics Association.

[65] Girardi, D., J. Dirnberger, and M. Giretzlehner, An ontology-based clinical data warehouse for scientific research. Safety in Health, 2015. 1(1): p. 6.

[66] Botsis, T., et al., Developing a multivariable prognostic model for pancreatic endocrine tumors using the clinical data warehouse resources of a single institution. Applied clinical informatics, 2010. 1(1): p. 38.

[67] Laws, R., et al., The Community Health Applied Research Network (CHARN) data warehouse: a resource for patient-centered outcomes research and quality improvement in underserved, safety net populations. EGEMS, 2014. 2(3).

[68] Podchiyska, T., et al. Managing medical vocabulary updates in a clinical data warehouse: An RxNorm case study. in AMIA Annual Symposium Proceedings. 2010. American Medical Informatics Association.

[69] Garcelon, N., et al., Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. Journal of biomedical informatics, 2017. 73: p. 51-61.

[70] Foran, D.J., et al., Roadmap to a comprehensive clinical data warehouse for precision medicine applications in oncology. Cancer informatics, 2017. 16: p. 1176935117694349.

[71] Schumacher, A., T. Rujan, and J. Hoefkens, A collaborative approach to develop a multi-omics data analytics platform for translational research. Applied & translational genomics, 2014. 3(4): p. 105-108.

[72] Genes, N., et al., Validating emergency department vital signs using a data quality engine for data warehouse. The open medical informatics journal, 2013. 7: p. 34.

[73] Couderc, J.-P. The telemetric and Holter ECG warehouse initiative (THEW): a data repository for the design, implementation and validation of ECG-related technologies. in Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2010. NIH Public Access.

[74] Han, J., J. Pei, and M. Kamber, Data mining: concepts and techniques. 2011: Elsevier.

[75] Hamoud, A. and T.A.S. Obaid, Design and Implementation Data Warehouse to Support Clinical Decisions Using OLAP and KPI. 2013, Department of Computer Science, University of Basrah.

AUTHOR PROFILE

Alaa Khalaf Hamoud is a lecturer in Computer Information Systems, University of Basrah, Iraq. He received BSc degree from Computer Science Department, University of Basrh in 2008 with first ranking college student. He also received his MSc degree from the same department with first ranking department student. He participated in (seven months) IT administration course in TU berlin-Germany. His scientific interests are data mining, data warehousing.

Ali Salah received M.sc in Information Technology from Tenaga University, (Malaysia) and B.sc in Computer Science from Basra University (Iraq). His current research interests includes cloud computing and data mining.

Wid Akeel received M.sc in Information Security (2012) from Basra University, (Iraq) and B.sc in Computer Science (2006) from Basra University (Iraq). Her current research interests includes information security, Steganography, image processing and data mining.

doi:10.25195/2017/4424