Republic of Sudan

Ministry of Higher Education and Scientific Research

University of Shendi

College of Graduate Studies

DEVELOPING OF EXTRACTION, TRANSFORMATION AND LOADING DATA WAREHOUSE TECHNIQUES FOR CANCER

A Thesis Submitted in Fulfilment of the Requirement for the
PH.D Degree in Computer Science

By
Abubaker Elrazi Osman Mohammed Ahmed

Supervisor:
Prof. Dr. Elsamani Abd Elmutalib Ahmed Abd Elmutalib

May 2015

آية

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

وَلِيَعْلَمَ الَّذِينَ أُوتُوا الْعِلْمَ أَنَّهُ الْحَقُّ مِن رَّبِّكَ فَيُؤْمِنُوا بِهِ فَتُخْبِتَ لَهُ قُلُوبُهُمْ وَإِنَّ اللَّهَ لَهَادِ الَّذِينَ آمَنُوا إِلَى صِرَاطٍ مُّسْتَقِيمٍ ﴿٥٤﴾

(الحج : 54)

DEDICATION

To my Mother

To my Father… Osman

To my Fathers… Abd Elrahim, Abd Allah and Ahmed

To my Family Noha, Mohammed and Mina

To my Brothers and Sisters

ACKNOWLEDGEMENTS

In the beginning I have to thank Allah for my success to complete this work which I hope it can be beneficial. God blessing and peace be upon our prophet Mohammed. I would like to thank and appreciate Prof. Dr. Elsamani Abd Elmutalib Ahmed the supervisor of this research who supported me fully to complete the work. I would also express my appreciation to examination committee members Assoc. Prof. Dr. Saif Eldin Fatooh Osman, Asst. Prof. Dr. Mahmoud Ali Ahmed for their valuable discussion, suggestion and comments. Also I would like to extend my gratitude to Prof. Dr. Yahia Fadl Allah Mukhtar. Furthermore thanks to Asst. Prof. Dr. Mohammed Bakri Basheer for the valuable discussion & observations which helped me too much. In addition to appreciation to information technology center family of Shendi University particularly Eng. Mohammed Qasim Elseed, Eng. Osama Taha and Mr. Hashim Abdul Allah Taha. Finally, I would like to thank my family who shouldered the burden of this study.

Abstract

The Clinical Data Warehouse (CDWH) is a result of utilizing data warehouse technology in medical field for integrating clinical data. The CDWH integrates the relevant data according to medical goals from more than one disparate data sources. However, many of medical institutions do not utilize this large volume of clinical data. Furthermore, the harnessing of this huge data produces several issues need to addressed, which include: the data integration, data quality, and Extract, Transform and Load (ETL) process. The integration of data and data quality issues are observe through the ETL process. This research developed CDWH architecture to support clinical data analysis according clinical requirements. Additionally, ETL techniques developed to integrate clinical data and improve the quality of data in CDWH. The ETL techniques are responsible for the extraction and integration of data from several sources, cleansing, transformation, mapping, and loading data into a CDWH. The experiment conducted using cancer disease dataset collected from two Radiation and Isotopes Centers –Sudan. The datasets store in different format (SQL, MySQL, and XML). The ETL techniques evaluated using four parameters; subject-oriented, integrated, non-volatile and time-variant. In addition, the result shows that all data integration and quality problems are handled; the produced data relevant, complete, integrate, unique, valid, consistent and most importantly they are in an appropriate forms for data mining. Furthermore, the clinical data warehouse provides a rich analysis environment to improve the quality of diagnosis, clinical care, treatment recommendation decision making and support research work.

الملخص

مستودع البيانات السريرية هو نتيجة لإستخدام تقانة مستودع البيانات فى المجال الطبى وذلك لتوحيد البيانات السريرية وتكاملها. نجد إن مستودع البيانات السريرية يُوحد بها البيانات المطلوبة (ذات الصلة بموضوع الدراسة) وفق الأهداف الطبية من أكثر من مصدر واحد غير متجانس البيانات. ولكن الكثير من هذه المؤسسات لا تستخدم هذا الكم من البيانات. علاوة على ذلك، نجد أن تسخير واستخدام هذا الكم الهائل من البيانات ينتج عنه عدة تحديات تتضمن: تكامل البيانات، جودة البيانات، وعملية الاستخراج، التحويل والتحميل. تمت معاجة قضايا تكامل وجودة البيانات من خلال عملية الاستخراج، التحويل والتحميل. فى هذا البحث تم تطوير معمارية لمستودع البيانات السريرية لدعم تحليل البيانات السريرية حسب الإحتياجات السريرية. بالإضافة إلي ذلك، تم تطوير تقتنيات الإستخراج، التحويل والتحميل لتُوحد البيانات السريرية وتحسين جودتها فى مستودع البيانات السريرية. بحيث نجد أن تقنيات الإستخراج ، التحويل والتحميل مسئولة عن إستخراج وتوحيد البيانات من عدة مصادر، وعن التنظيف، التحويل، رسم الخرائط وتحميل البيانات فى مستودع البيانات السريرية. إستخدمت التجربة مجموعه بيانات لمرض السرطان جمعت من مركزين للطب النووي والمعالجة بالأشعة بالسودان. ومجموعة هذه البيانات مخزنة فى نسق مختلف (SQL, XML, MySQL). تم استخدام أربعة معلمات لتقييم تقنيات الاستخراج، النقل والتحميل تضمنت:-subject oriented, integrated, non-volatile and time-variant. إضافة إلى ذلك، تظهر النتائج معالجة مشكلات تكامل كل البيانات السريريه وجودة البيانات التي تؤثر علي تصميم وتطوير أنظمة مستودع البيانات السريرية، حيث أن البيانات التي تم تخزينها في مستودع البيانات السريرية ذات الصلة بالموضوع المحدد، كاملة، متكاملة، فريدة من نوعها، صالحة، متجانسة والأهم من ذلك أنها تم تحويلها للأشكال المناسبة للتحليل واستخلاص النتائج. علاوة علي ذلك، يوفر مستودع البيانات السريريه بيئة تحليل غنية لتحسين جودة التشخيص والرعاية السريرية والعلاج للمساعدة في صنع القرار الطبي الفعال ودعم ودعم العمل البحثي.

Research Content

LIST OF TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

CDSS:	Clinical Decision Support System
CDWH	Clinical Data Ware House
DB	Database
DBMS	Database Management System
DDL	Data Definition Language
DSS	Decision Support System
DWH	Data Ware House
EMR	Electronic Medical Record
ER	Entity Relationship
ERD	Entity Relationship Diagram
ETL	Extracting, Transforming, Loading
HIS	Hospital Information System
LIS	Laboratory information system
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
PACS	Picture Archiving and Communications System
RDBMS	Relational Database Management System
RICK	Radiation and Isotopes Centre Khartoum– Sudan
RICSH	Radiation and Isotopes Hospital –Shendi– Sudan
RIS	Radiology Information System
SQL	Structured Query Language

Abstract

The Clinical Data Warehouse (CDWH) is a result of utilizing data warehouse technology in medical field for integrating clinical data. The CDWH integrates the relevant data according to medical goals from more than one disparate data sources. However, many of medical institutions do not utilize this large volume of clinical data. Furthermore, the harnessing of this huge data produces several issues need to addressed, which include: the data integration, data quality, and Extract, Transform and Load (ETL) process. The integration of data and data quality issues are observe through the ETL process. This research developed CDWH architecture to support clinical data analysis according clinical requirements. Additionally, ETL techniques developed to integrate clinical data and improve the quality of data in CDWH. The ETL techniques are responsible for the extraction and integration of data from several sources, cleansing, transformation, mapping, and loading data into a CDWH. The experiment conducted using cancer disease dataset collected from two Radiation and Isotopes Centers –Sudan. The datasets store in different format (SQL, MySQL, and XML). The ETL techniques evaluated using four parameters; subject-oriented, integrated, non-volatile and time-variant. In addition, the result shows that all data integration and quality problems are handled; the produced data relevant, complete, integrate, unique, valid, consistent and most importantly they are in an appropriate forms for data mining. Furthermore, the clinical data warehouse provides a rich analysis environment to improve the quality of diagnosis, clinical care, treatment recommendation decision making and support research work.

الملخص

مستودع البيانات السريرية هو نتيجة لإستخدام تقانة مستودع البيانات فى المجال الطبى وذلك لتوحيد البيانات السريرية وتكاملها. نجد إن مستودع البيانات السريرية يُوحد بها البيانات المطلوبة (ذات الصلة بموضوع الدراسة) وفق الأهداف الطبية من أكثر من مصدر واحد غير متجانس البيانات. ولكن الكثير من هذه المؤسسات لا تستخدم هذا الكم من البيانات. علاوة على ذلك، نجد أن تسخير واستخدام هذا الكم الهائل من البيانات ينتج عنه عدة تحديات تتضمن: تكامل البيانات، جودة البيانات، وعملية الاستخراج، التحويل والتحميل. تمت معاجة قضايا تكامل وجودة البيانات من خلال عملية الاستخراج، التحويل والتحميل. فى هذا البحث تم تطوير معمارية لمستودع البيانات السريرية لدعم تحليل البيانات السريرية حسب الإحتياجات السريرية. بالإضافة إلي ذلك، تم تطوير تقتنيات الإستخراج، التحويل والتحميل لتُوحد البيانات السريرية وتحسين جودتها فى مستودع البيانات السريرية. بحيث نجد أن تقنيات الإستخراج ، التحويل والتحميل مسئولة عن إستخراج وتوحيد البيانات من عدة مصادر، وعن التنظيف، التحويل، رسم الخرائط وتحميل البيانات فى مستودع البيانات السريرية. إستخدمت التجربة مجموعه بيانات لمرض السرطان جمعت من مركزين للطب النووي والمعالجة بالأشعة بالسودان. ومجموعة هذه البيانات مخزنة فى نسق مختلف (SQL, XML, MySQL). تم استخدام أربعة معلمات لتقييم تقنيات الاستخراج، النقل والتحميل تضمنت:-subject oriented, integrated, non-volatile and time-variant. إضافة إلى ذلك، تظهر النتائج معالجة مشكلات تكامل كل البيانات السريريه وجودة البيانات التي تؤثر علي تصميم وتطوير أنظمة مستودع البيانات السريرية، حيث أن البيانات التي تم تخزينها في مستودع البيانات السريرية ذات الصلة بالموضوع المحدد، كاملة، متكاملة، فريدة من نوعها، صالحة، متجانسة والأهم من ذلك أنها تم تحويلها للأشكال المناسبة للتحليل واستخلاص النتائج. علاوة علي ذلك، يوفر مستودع البيانات السريريه بيئة تحليل غنية لتحسين جودة التشخيص والرعاية السريرية والعلاج للمساعدة في صنع القرار الطبي الفعال ودعم ودعم العمل البحثي.

CHAPTER ONE
INTRODUCTION

This chapter describes the introduction of this research which includes: research motivation, problem statement that addressed by the research work, objectives that achieved, research contributions, research scope and research structure.

1.1 Research Motivation

Data Warehouse (DWH) today plays a major role in business in order to improve decision making; according to Bill Inmon in 1990, “data warehouse is a subject oriented, integrated, time-variant and nonvolatile collection of data in support of the management’s decisions making process” [1] , while Ralph Kimball provided a simpler definition of a data warehouse as "a copy of transaction data specifically structured for query and analysis" [2]. Therefore, DWH is defined as a central repository; integrated a collection of data extracted from more than one source (operational or transactional systems), the data have to be transformed after cleansed from any data problems, in order to optimized data for storage to support the decision making processes and provide powerful analysis and researches environment.

In the recent years the using of DWH technologies in the healthcare field is becoming an important and fertile field, and acquired the attention of researchers to benefit from the functionality, capabilities and features offer by the DWH technologies. However, the clinical data is diverse from business data and using the DWH technologies in medical field produce new issues and challenges. In medical institutions the clinical data are stored in disparate medical operational systems that have limited almost no interconnectivity between these systems. Furthermore these systems contain accumulated substantial amounts of data about patients with the associated clinical conditions and treatment details. Furthermore, data that stored in the CDWH must meet specific

requirements to improve data analysis processes to support medical decision and enhance medical research.

Consequently, the efficiency of DWH is mainly depend on ETL process, which integrate these clinical data, and presents as the major part of a DWH environment [1] [3] [4]. These process activities are highly sensitive to quality of data; poor quality of data will affect the revenue of an organization and causes low quality decision making [5] [6] [7] [8] [9]. The basically tasks of ETL process included extracting data from heterogeneous data sources, converting data into a common structure suitable for analyzing and mining, cleansing data, and loading it into the DWH [10] [11] [12]. This gives a motivation to this research on how to develop ETL techniques for extracting, cleansing, transforming, integrating and loading data processes, because these processes are a success key of a DWH to provide more effective analysis environment to support medical decision making.

1.2 Problem Statement

Cancer is a generic term for a large group of diseases (more than 100 different types of cancer) that can affect any part of the body[13]. Cancer is a leading cause of death worldwide, and more than 60% of world's total new annual cases occur in Africa, Asia and Central and South America. These regions represent for 70% of the world's cancer deaths [14]. In the recent years the cancer disease is spread in Sudan; the statistics from Radiation and Isotopes Centre Khartoum (RICK) and Shendi (RISH) showed that the numbers of new cases and deaths is increasing. Furthermore, Martel C et al in [15] predicted that from 2008 to 2030, cancer incidence will rise 75 percent globally and will double in the least developed countries. The medical institutions store a huge data about patients with the associated clinical conditions, diagnosis and treatment details. These data must meet specific requirements to improve data analysis processes, support medical decision, and enhance medical research. Using of DWH technologies in the medical field is produced new issues as the following:

1. Data collected from different hospitals which use diverse data format, data structure, and DBMS. These data is differ from business data, which it has their distinct set of characteristics that made the data integration complicated.
2. The clinical data is complex and rise several issues for instance; poorly characterized mathematically, difficult data type for mining, and difficult to determine hidden relationships.
3. Using DWH technologies in the medical field has produced new challenges, e.g. medical data required; more powerful data model construct than conventional approaches, and strong techniques to integrate clinical data.

4. The CDWH aim to integrate, utilizing the clinical data volumes, discovery hidden relationship in data, evaluate the performance of using of different treatments protocols, and provide information in areas ranging from research to management.
5. Data quality is an important issue of CDWH development and implementation life cycle. Moreover, the data quality determines the reliability of data for analysis, and making decisions.

The integration process performs through ETL process, which it deals with the extraction, cleansing, transformation and loading of data from sources into clinical data warehouse. Additionally, the ETL process is one of the major critical issues in the process of CDWH development to facilitate the mining and understanding of complex medical characteristics and trends of cancer disease management. The ETL integrated, and store in clinical data warehouse with good quality guarantees.

1.3 Research Questions

In order to develop integration technique and achieve quality of data in CDWH, there are five research questions considered in this research:

1. How the medical purpose and requirements are determined in proper ways?
2. What are the extraction, cleansing, transformation and loading (ETL) problems that affect the quality of data?
3. How the ETL Techniques can be used to integrate heterogeneous clinical data, cleanse, convert the clinical data into format that is conducive to effective data analysis, map and load the medical data into CDWH according medical requirements?
4. How to handle the data problems that may affect the quality of data during ETL process.
5. How the CDWH architecture can be designed to implement CDWH and data storage services that provide data analysis and mining.

1.4 Research Objectives

This research aims to integrate the data from several data sources and improve the data quality at clinical datasets, in order to support the medical decision making and provides powerful analysis and researches environment. The objectives of this research work are:

1. Building integration techniques to extract, cleanse, transform, and load heterogeneous clinical data from several data source into CDWH.
2. Designing and developing CDWH architecture based on the medical needs.
3. Utilizing OLAP and data mining technologies to mine these data and assist to understand complexity of the medical characteristics and trends of cancer as well as support decision making.

1.5 Research Contributions

This research presents a ETL techniques to integrate heterogeneous clinical data from several medical data sources into clinical data warehouse, as well as all related works to integration clinical data issues, data quality issues, ETL techniques, and clinical data warehouse architecture and technique. This section lists the contributions of this thesis:

1. **Analysis of relevant work:** Determine the data quality criteria and integration issues in order to ensure high quality data in CDWH. These issues include data warehousing, clinical data warehouse issues and challenges, data warehouse architecture, ETL processes and clinical data quality and clinical data integration.
2. **ETL techniques**: Four integration techniques (Extraction, Cleansing, Transformation, and Loading) are implementing to improve the quality of data in CDWH, by investigating how the data integration techniques efficiency handles data integration and quality problems, presenting how the four key activities; extraction, cleansing, transforming, and loading, are performed.
3. **CDWH technique:** the third contribution is designing and implementing of clinical data warehouse technique. The clinical data warehouse stores the integrated clinical data from various sources to improved medical decision making and provide data for clinical research.
4. **Development of information systems:** The fourth contribution, two running high quality systems for cancer information and early detection of breast cancer are implemented at Mc Nimer Hospital-Shendi-Sudan.

1.6 Research Scope

This research work aims to develop ETL techniques to integrate clinical data into clinical data warehouse, which contains clinical data about cancer's patients. Furthermore, this research work only involves the related data warehousing processes to process the data, such as Data Extraction, Transformation, Loading, Cleansing, Integration and querying etc.

Additionally, this research work aim to develop clinical data warehouse to store the clinical with quality grantee from staging area. There are some issues are not scope of this research. These issues include: Security, Privacy and data transfer time (Bandwidth).

1.7 Research Structure

Chapter One: Provides details about the research motivation, problem statement, research objectives, research contribution, research scope and research structure.

Chapter Two: Presents the literature review that relevant to the study. This chapter provides overview of previous work efforts on the CDWH that are include, data warehousing, clinical data warehouse issues and challenges, data warehouse architecture, and clinical data quality and clinical data integration.

Chapter Three: Describes the research methodology. The methodology consists of four stages; literature review, design and development of ETL techniques, design and development of clinical data warehouse, and study results. Each stage is explained in detail.

Chapter Four: Presents and discusses the design and develop of ETL techniques (extraction, transformation, cleansing, and loading). The proposed ETL techniques consist of medical analysis, physical development, implementation and evaluation.

Chapter Five: Presents and discusses the proposed CDWH design and development technique. The proposed CDWH technology consists of: the medical Analysis, CDWH architectural selection, CDWH schemes model creation, population CDWH and data storage services, physical development, and CDWH evaluation.

Chapter Six: Provide a discussion of result analysis. This chapter contains a discussion and results.

Chapter Seven: Concludes the research work by providing information on the research process, its benefits to CDWH quality and recommendations and future work.

CHAPTER TWO
LITERATURE REVIEW

This chapter reviewed the literature according to what data quality and integration needs in order to ensuring high quality data in CDWH to support medical decision making and improve medical research. This chapter provides overview of previous work efforts on the CDWH that are relevant to this study, data warehousing, clinical data warehouse issues and challenges, data warehouse architecture, and clinical data quality and clinical data integration.

2.1. Data Warehousing

The first Data Warehouses (DWH) technology developed in the 1980s as a response to the lack of information provided by several online application systems that were being built, and they were rarely integrated with each other [16]. The DWH approach integrates data from the operational systems into one common data source, known as the DWH, which is structured for intelligent query data analysis purposes [17] [2] [18] [19]. The most popular definition is Inmon definition [1]; "A DWH is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process ". The meanings of the key terms are defined bellow:

1. Subject-oriented: means that all relevant data about a subject is gathered and stored as a single set in a useful format. Information is presented according to specific subjects or areas of interest.
2. Integrated: refers to data being stored in a globally acceptable fashion with consistent naming conventions, measurements, encoding structures, and physical attributes, even when the underlying operational systems store the data differently.
3. Non-volatile: means stable information that does not change each time an operational process is executed. Information is consistent regardless of when the DWH is accessed.

4. Time-variant: means that the DWH contains a history of the subject, as well as current information. DWH's data represents long-term data from five to ten years.

The DWH technology continued providing function and speed for decision making and researches. Therefore, the DWH is a data structure that is optimized for distribution, mass storage and complex query processing, which collects and stores integrated sets of historical data from multiple operational systems and integrates them to one or more Data Marts [20]. Furthermore, the DWH integrates data from more than one operational system, which exist on one or several organizations.

The integration process in DWH includes three steps: A) developing a unified model that can accommodate all information various single databases, B) transferring the data into developed model before loading them into the DWH, and C) extracting the data from the source databases and integrating it in one environment, to provide access to obtain the required knowledge which the individual sources cannot provide it [21] [22] [23] [24] [25] [26] [27]. These integration processes require a set of hardware and software components that can be used to get better analysis of massive data as well as making better decisions making and research. Furthermore, the integration process requires architecture and techniques to collect, analyze, cleanse and present information [28] [20]. According to McFadden et al. [29] and Gupta et al. [30], the DWH principle is supported by the following factors: 1) A DWH centralizes and integrates data from disparate operational systems and makes them readily available to allow quick decisions on historical data, 2) A properly designed DWH integrates data across the enterprise improving data quality and consistency, 3) Managing and controlling business enable managers and end users to understand the business and make decisions, and 4) A separate data warehouse eliminates much of

the contention for resources that result when informational applications are confounded with operational processes.

In addition to, there are two paradigms in the DWH field; the first one is Bill Inmon's paradigm where it describes the DWH as one part of the overall business intelligence system. In Bill Inmon's paradigm any enterprise has one DWH, and data marts source their information from the DWH. The DWH stores information in 3rd normal form. The Bill Inmon's paradigm proposed the snowflake schema, where some dimension tables are normalized, thereby further splitting the data into additional tables [31] [1]. The second paradigm, is Ralph Kimball's paradigm where it describes the DWH as the conglomerate of all data marts within the enterprise. The Ralph Kimball's paradigm proposed star schema model, to represent dimensional data model [32] [33]. There is no right or wrong between these two ideas, as they represent different data warehousing philosophies. In reality, the DWH systems in most enterprises are closer to the Ralph Kimball's idea. This is because most DWHs started out as a departmental effort hence originated as a data mart.

The DWHs technology provides numerous benefits to an organization, that described on multiple levels, such as time-saving for users, improving quantity and quality of information, enhancing the business intelligence and query performance, and supporting the decision-making by enabling quick and efficient access to information from legacy systems and linkage to multiple operational data sources [34] [20].

On the other hand, clinical fields are indeed becoming a very attractive research domain for computer science in general and DWH in particular. However, some business areas are in need of more complex data structures, one such area is medical field, where clinical data about

large patient populations is experiencing a larger fluctuation of incoming data, due to the greater number of requirements that the data is used for [35]. The CDWH is integrating medical data from various operational medical and administrative systems into DWH to provide quality and research purposes in a timely, efficient and secure way [36] [37]. Therefore, the DWH is a technical solution for immense data storage, management and processing [1]. Furthermore, the CDWH supports financial analysis [38], disease control [39], adverse drug events [40], laboratory test data analysis [41], and clinical decision process, and medical research [20] [42]. Additionally, the CDWH facilitates efficient storage, enhances timely analysis and increases the quality of real time decision making processes [1] [43] [44] [45] [46].

As a result of the previous discussion the DWHs and CDWHs are not different so much; they use the same technology [29]. However everything gets more complex and complicated with CDWHs [47]. Table 2-1, summarize that CDWH is different from a commercial DWH in the 4 aspects below [48].

Table (2-1): Differences between DWHs and CDWHs [48]

Aspect	Data Warehouse	Clinical Data Warehouse
Usage of data stored in the DWH	To identity patterns from the enormous amount of data in the operational database for better management decisions.	To validate assumptions, find indicators, descriptors and risk factors in order to understand and characterize the complex medical data and trends.
Data types support	Deals with simple data types: • String, text • Numeric, decimal • Boolean	Supports both simple conventional data types a advanced data types that suit to medical data specificity: • Advanced temporal support; • Advanced classification structure; • Continuously valued data; • Image data, e.g. X-ray, electro-cardiogram.
Semantic of the data	The semantic of the data is clear and explicit.	The semantic of the data is implicit, and special tools or features are often required for data understanding and semantic derivation.
Data processing method	The Extract, Load and Transform (ETL) processes are simple and straight forward.	Complex algorithms based on signal processing, pattern recognition, statistical methodologies, are often required to extract and transform the raw data into relevant information as well as to validate them.

2.2. Clinical Data Warehouse Issues and Challenges

The CDWH is a result of utilizing DWH technology in medical field to provide rich analysis environment. The usage of CDWH technology aims to integrating information from various sources, reducing administrative costs and improving the quality of healthcare [49] [50]. The medical data is a central to both effective health care and to management process. Therefore, CDWH facilitates the processes of determining the relationships in clinical data, discovering disease trends, evaluating the performance of different treatments protocols used, support measuring and improving patient outcome, and providing information to users in areas ranging from research to management [51] [52] [53]. The success of the CDWH depends critically on the collection and integrating data from various medical data sources, analysis, and information retrieval about the effectiveness of treatments, the accuracy of diagnoses, and other medical information [43] [54].

The process of gathering medical data from medical operation systems is complex, time-consuming and labor intensive [55] [20]. Furthermore, this data is of a sensitive nature, complexity of medical data, diverse storage formats security and privacy [49] [56]. The medical data contain the data related to patient care including specific demographics, input and output data recorded for the patient, diagnosis data, treatments and procedures performed, and costs associate with the patient's care [43]. However, the medical information systems is rich with data, but it is also exposed to a lot of quality problems [57] and poor information [36] due to lack of efficient analysis tools [58]. Therefore, the diversity of the nature of clinical data from other business data produces several challenges that are being faced relating to the use of DWH technology in medical field. These challenges include the clinical data format, business analysis, data integration, data quality, and powerful ETL technique.

2.2.1 The Clinical Data Format

There is a massive of medical records that collect data about patient during the regular day-to-day events [59] and store in various medical information systems such as HIS (Hospital Information System), RIS (Radiology Information System), PACS (Picture Archiving and Communications System), LIS (Laboratory information system) and etc [60] [59]. However, the medical information systems are not been designed to communicate with one another using a common standard. On the other hand, the clinical data are generated during different patient visit processes at different time. They are stored dispersedly in the above information systems and are usually isolated from one another. This clinical data integrate various medical data systems into CDWH support of new treatment options, medical interventions, drug development, etc [59]. However, the integration process is time consuming and labor

intensive to extract data for quality assessment and research purposes [61]. Furthermore, the medical data must be validate consistency, accurate and timely data [39] [62]. The lack of standardization between hospitals and institutions makes integrating data process is difficult [63]. Furthermore, clinical data can be characterized by many relationships between objects and concepts, which are difficult to identify formally [64]. The clinical data is often presented in an unstructured format, and contain various types of data and the nature of this types of data are complex and poorly characterized mathematically [44]. Additionally, patients have hundreds of different facts describing their current situation [65]. Therefore medical data must be available in a centralized DWH structure equipped with proper tools and mechanisms to integrate data in the process of developing CDWH.

2.2.2 Medical Analysis

Business analysis is identifying business purpose and determining solutions to business problems [66]. One of the most important aspects of developing a CDWH is to define medical purpose. In medical analysis, the medical decision maker aims to retrieve the data, or generate a report that cross-references the cost of delivering a particular service in a particular demographic to a particular patient population. Whatever the medical question, it is essential to realize that the medical institutions are evaluating the quality and effectiveness of their treatment, and support medical research. However, the CDWH does not achieve its objectives without clearly defining the medical purpose [67]. Furthermore, the discussion of the medical analysis phases are significant to study and analyze the existing process from medical perspective as well as to determine project objectives, requirements, constrains and acceptance criteria.

2.2.3 Data Integration

Data Integration is a process of combining data from more than one disparate data sources within one or several institutions into a single physical repository [68] [69], which is optimized for intelligent data analysis purposes [18] [19] [70] [56]. Furthermore, Every unit may use different hardware platforms, different operating systems, different information management systems or different network protocols [59]. However, one of the major CDWH challenges is the integration of several disparate, standalone information repositories into a single logical repository [68] [71] [20].

The data integration becomes a significant issue in situation of developing a CDWH due to the complexity of the hospital environment such as various care practices, and data types and definitions. Furthermore, the clinical data integrate from various medical information systems. These medical systems are different clinical routines, incompatible structures, and incompleteness of clinical information systems [72] [73]. Additionally, the collection of clinical sources has the property that similar data can be contained in several sources but represented in a variety of ways depending on the source. This representational heterogeneity encompasses structural, naming, semantic, and content differences [64]. Therefore, the integration of the medical data is a difficult process and data are not yet turned into useful knowledge; due to the lack of efficient analysis tools, and the lack of international standards in patient care at institutions [58] [63] .

On other hand, the private and security issues are the main problems in CDWH arising from the sensitivity of certain data types. The information content about both physicians and patients should be protected. The protection process considers in the ETL tool processes and data analysis applications [74].

Therefore, the data integration is an important issue in developing CDWH to support decision making and medical research.

2.2.4 Data Quality

Data quality is an essential characteristic that determines the reliability of data for analysis, decisions making and planning [5] to enhance usability of the data warehouse [75] [76] [77]. The acceptable data quality in the medical field is a critical issue to the reliability of medical decision making and research environment. Quality of data is achieved when the required (useful) data that exactly meet the specific needs stored in common format required by CDWH without data quality problems.

The quality of the information depends on 3 things[78]: (1) the quality of the data itself, (2) the quality of the application programs and (3) the quality of the database schema ETL. Therefore, data quality problems produce at various stages of CDWH development; data integration & data profiling, Data staging and ETL, and DWH modeling & schema design[79] [80]. However, the information produced in the healthcare industry is excessive, disjointed, incomplete, inaccurate, in the wrong place, or difficult to make sense [81]. Therefore, poor data quality can occur along several dimensions [5] [79] [80] [82]. These dimensions of data quality attribute include; intrinsic (accuracy, consistency, reliability, integrity, and redundancy dimension), accessibility (availability dimension), contextual (relevancy, freshness, validity, completeness, and Scalability dimension), and representational (medical purpose understandability and data sources understandability dimension) as shown in figure 2.1.

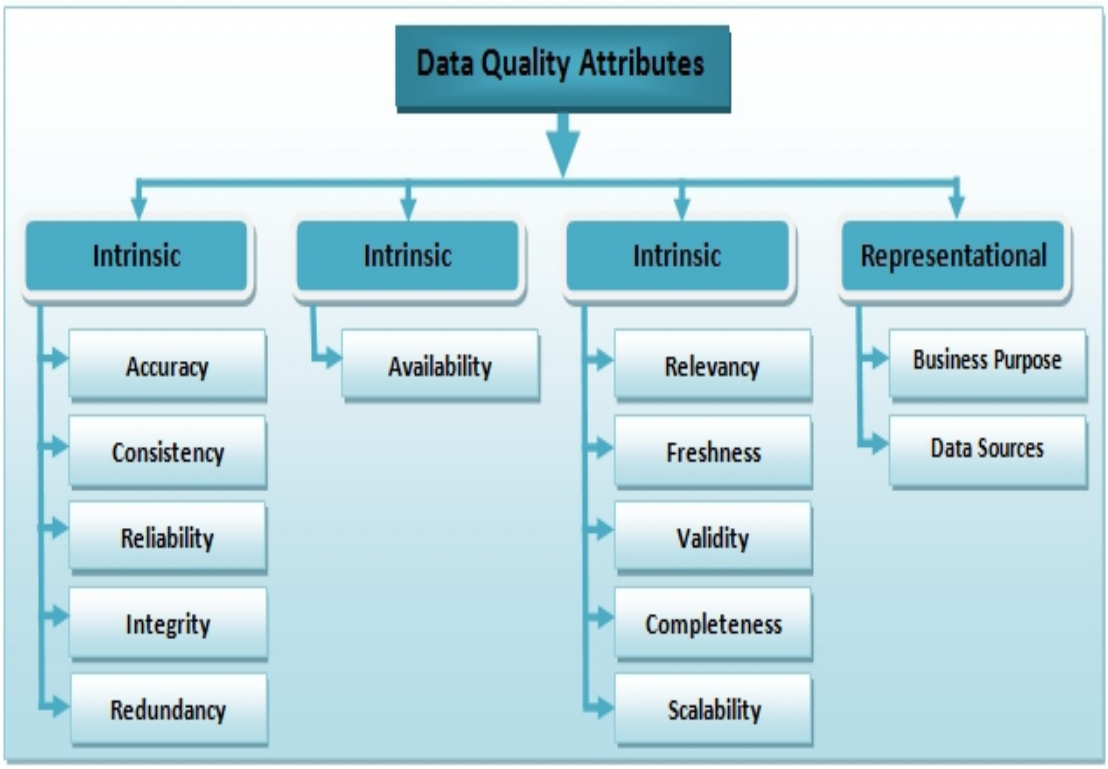


Figure (2.1): Categories and Dimensions of information Data Quality

Furthermore, the data quality problems must be determined and solved to enhance the quality of data. Defining the level of data quality that is appropriate to the organization is required [83]. The data quality process requirements involve understanding the data quality problems from medical perspective. Additionally, understanding the format of data stored by each source is a very important issue to ensure quality of data, there are wide varieties of structured and unstructured types of data.

2.2.5 The Extract, Transform, Cleanse and Load (ETL) Process

ETL plays a vital role in a DWH solutions [32] [33] [74]. Appropriate design of ETL processes are considered as the core component of a successful DWH system [74] [84] [33] [85], which shown in Figure 2.2. ETL is data transformation, data validation, and data cleansing focused. ETL processes responsible for the extract data from heterogeneous data sources, converting extracted data into a common format suitable for analyzing and mining, identifying and data

quality problems, cleansed data to eliminate undesired data, and finally loading these data into the DWH [86] [87].

In medical field ETL process activities are highly sensitive to quality of data and data integration, poor quality of data will affect the revenue of an organization and causes low quality decision making [5]. Building the ETL process is potentially one of the major challenges in CDWH. This is because of its difficulty and lack of formal model for representing ETL activities that map the incoming data from different data sources to be in a suitable format for loading to the targeted DWH [88] [89] [90] [91]. Furthermore, it is a complex, time consuming, and consume most of DWH project's implementation efforts, costs, and resources [87] [92] [93].

Due to the complication of medical data structure and clinical operations in real-world clinical environment, it is important to develop a powerful ETL tool to integrate, transform, and cleanse medical data before loading these data into CDWH as reported in [74] [94]. Furthermore, due to the heterogeneous data sources and the related rich data formats (e.g. Excel, Oracle, and SQL Server), ETL functionality needs both flexibility and expansibility.

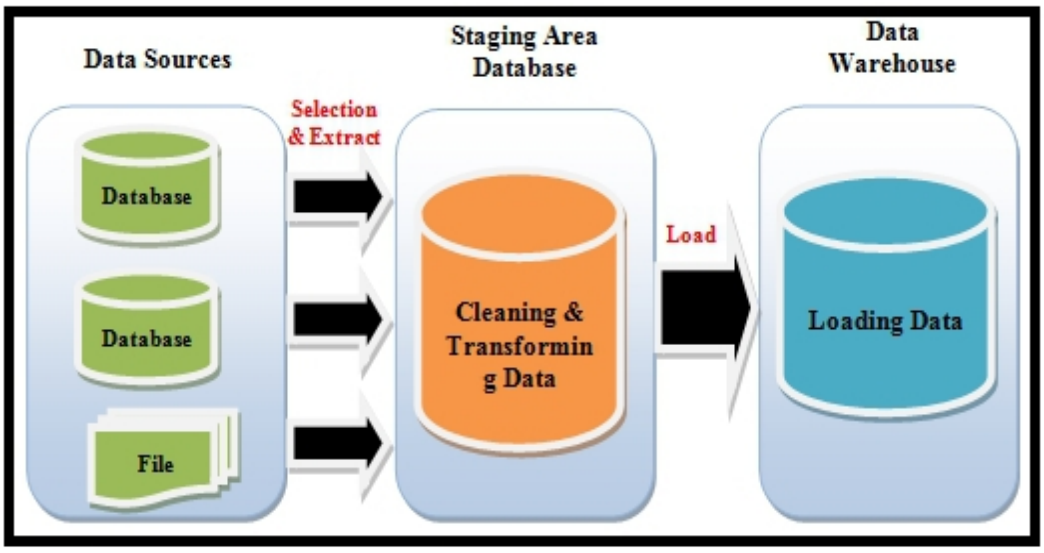


Figure (2.2): Typical Data Warehouse process

An ETL system consists of three consecutive functional steps: extraction, transformation, and loading process:

2.2.5.1. Extraction Process

Extraction process is responsible for extracting data from various heterogeneous data sources. The ETL process requires connecting to the source systems, and selecting the relevant data needed for analytical processing and research within the CDWH [12].

The data extract from numerous disparate source systems and each of these data sources has its distinct set of characteristics that need to be managed in order to effectively extract data for the ETL process as explained in [12]. Furthermore, the process needs to effectively integrate systems that have different platforms, such as different database management systems, different operating systems, and different communications protocols [95]. Additionally, the complexity of the extraction process depends on the data characteristics and attributes, amount of source data and processing time. Therefore, the ETL process needs to effectively integrate technology to extract these data.

2.2.5.2. Transformation Process

Transformation process is to transforms the extracted data into a common format by applying a set of conditions, rules or functions to derive the data to be loaded to the end targeted system. The transformation phase in medical field tends to make multiple data manipulations on the incoming data according to medical needs [3], to transfer data into common standard structures that can be accepted by the healthcare and medical research.

Medical field needs very complex transformations to meet the medical and technical needs of the targeted system. Every patient might have hundreds of different facts describing his current situation [65]. There is a need to aggregate this massive amount of information in a

useful way. Furthermore, the number of dimensions in clinical data is often very large, generating a need for intelligent ways of dimensionally reducing the data into high-level abstractions [96].

The transformation process requires joining the data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules by defining the granularity of fact tables, the dimension tables, DWH schema (star or snowflake), derived facts, slowly changing fact tables and dimension tables.

2.2.5.3. Cleansing Process

Data cleansing is one of the most important issues in ETL process as it ensures the quality of the data in the DWH [79]. Studies conducted by many researchers reported that more than 50 percent of DWH projects will have a limited success or will be outright failures and the result of the lack of attention to data quality issues [97]. The data cleansing deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data [97] [98]. The data cleansing phase involves three steps include: data analysis, data refinement and data verification [3]. The objective of data analysis is to identify problems area and detects the data problem. Data problems include completeness problems, validity problems, accuracy problems, consistency problems, conformity problems and Integrity problem. For each problem area, the data quality issues and acceptance criteria are identified, then, for each data quality issue, the solutions are developed. Furthermore, the data with quality issues will be refined using some of the data cleansing methods to realize their full benefits. Additionally, the cleansed data then will be assessed against the acceptance criteria again to ensure that the data issues have to be resolved the data cleansing process. Finally, after verification, the data will be moved from staging

area into CDWH [98]. Therefore, the objective of data cleansing process is to make cleansing and conforming on the extracted data to gain accurate data of high quality**.**

2.2.5.4. Loading Process

Loading process is the process of loading data from staging area to the CDWH depending on the requirements of the organization. The extracted and transformed data is written into the dimensional structures actually accessed by the end users and applications [97]. A major data loading problem is the ability of ETL process to discriminate between the new and the existing data at loading time; the new rows that need to be appended and rows that already exist need to be updated as discussed in [99] [100]. Furthermore, the loading process should be performed correctly and with as little resources as possible.

Based on the previous discussion, the CDWH is more complicated than the DWH and produces a set of issues and challenges. The clinical data is different from business data where the clinical data produces new issues and requirements if not considered; they affect the quality of data. Furthermore, the complexity of clinical data rises several issues for instance; poorly characterized mathematically, difficult data type for mining, difficult to determine hidden relationships, and required to be (secured and private). In addition, the security and privacy are critical issues in medical institutions; the clinical data contain confidential information, only authorized users logging onto the CDWH.

On the other hand, a clear understanding of the medical purpose represents an important stage in the process of developing CDWH. Moreover, the medical data requirements are collected to understand the problems domain in addition to determine the suitable data model that will be used and derive the architecture of the CDWH. Additionally,

constrains and acceptance criteria determine to evaluate the CDWH in order to ensure that medical objectives is achieved.

CDWH aims to integrate large volumes of data collected from several clinical information systems. The most common type of problems that are reported in literature of data integration are: lack of standardization of data, non standardization of formats, heterogeneity of data sources, Non-Compliance of data in data sources with the standards, missing data, inconsistent data across the sources. Because the complexity of the hospital environment evolves the diagnosis and treatment procedures and their relation with other information such as patient symptoms, disease stage, risk factors, and treatment risks. Moreover, the data collected from different hospitals which use and diverse data format and DBMS. Consequently, the data integration affects with these factors which consider as time consuming and laborious tasks. Therefore, development of effectively integrated systems that have different platforms is required.

Data Quality is one of important issues in CDWH because medical decisions making are based on data stored in CDWH. Determining data quality problems at every stage of data warehousing is required, if not taken care of these problems it will lead to poor data quality. The quality of data depends on three things: the quality of the data itself, the data quality problems, and the quality of the database schema. Additionally, the analysts must be able to identify relationships among various systems and understand the format of data stored in each source, because diversity of structured and unstructured types of data. Furthermore, the understanding of data quality problems from medical perspective is an important issue to build a robust technique that solves the data quality problems. Moreover, the achievement of good CDWH performance and deliver quality information are mainly depends on taking care of some of

the factors related to data quality while designing the CDWH schema. Therefore, implementing of effective data quality technologies provide high-quality data, reduce time and cost, and support clinical decision making.

Appropriate design of the ETL process is considered to be more critical stage of CDWH process where most of the data cleansing and scrubbing of data is done. The medical data consolidated from several source systems and each of these data sources has its distinct set of characteristics. Therefore, the complexity of the medical institution environment issues should be considered during the process of developing of ETL process. These issues involve clear identification of extracting, cleansing, transformation and loading requirements as well as developing and evaluating an ETL mechanism. Additionally, the data in CDWH must be corrected, completed, consistent, and integrated to provide a suitable medical decision making.

2.3. Data Warehouse Architecture

DWH architecture is a way of representing the overall components and services of the warehouse, with details showing how the components will fit together and how the system will grow over time. DWH architectures depend on the specifics of an organization's situation with suitable considerations for DWH requirements, describing structure of data, specifying ETL processes operations, modeling DWH schemas, and management of metadata and technology and bandwidth utilization [30] [101]. However, the architecture focuses on three main components [1] [102]: 1) Data Sources (operational systems and flat files), 2) Warehouse (metadata, summary data, and raw data), and 3) Users (analysis, reporting, and mining).

One of the common architectures is DWH architecture with a staging area [70]; there is a need to cleanse and process the operational

data before putting them into the data warehouse. The staging area is temporary location where data from source systems is copied. Furthermore, the staging area is mainly required in a DWH Architecture for timing and quality purposes [26] [103] [104]. Figure 2.3 show's this typical architecture with a staging area.

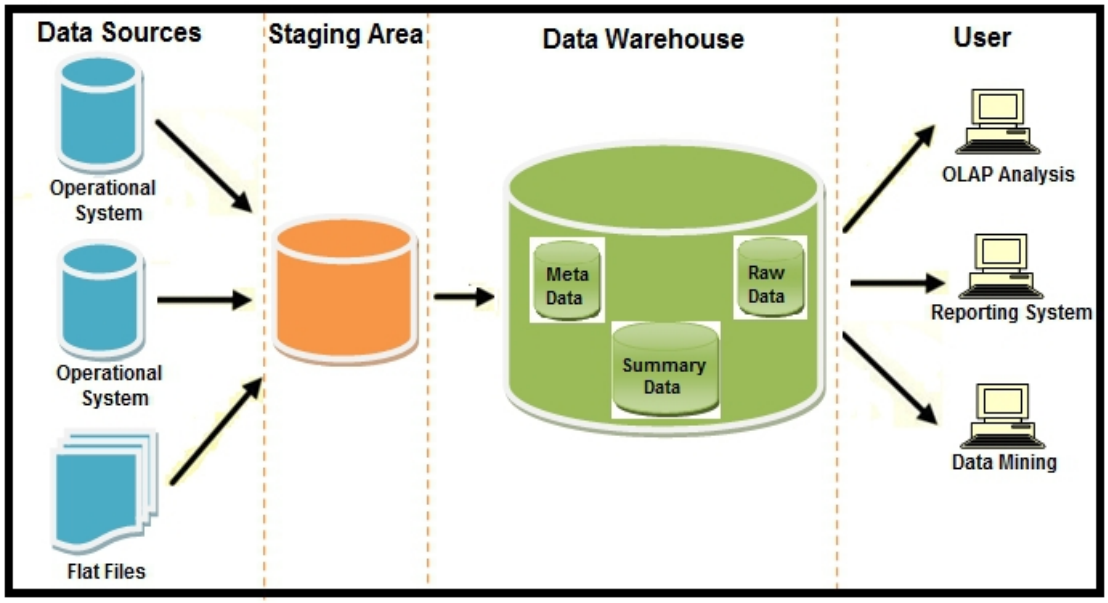


Figure (2.3): Architecture of a Data Warehouse with a Staging Area

On the other hand, The second type of architecture is DWH architecture with a staging area and data marts [70]. This type of architecture is required when there is considerable need to use architectural for different groups within an organization, this is done by adding data marts [26] [105]. Figure 2.4 shows this typical architecture with a staging area and data marts.

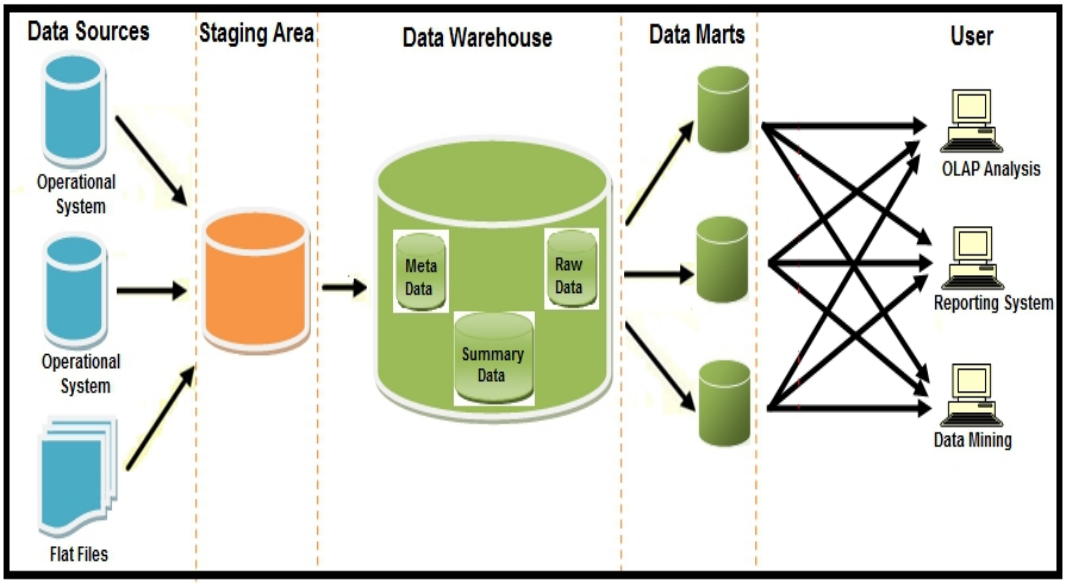


Figure (2.4): Architecture of a Data Warehouse with a Staging Area and Data Marts

The DWH architecture enables to better understanding of the DWH processes, understanding data contents and medical rules, improved timeliness, and flexibility (to store new data for future extension, modification, re-modeling, and reuse) [48]. The proper and clearly defined of the DWH architecture is very important issue for quality of data warehouse. However, the subject areas are not fit together, connections lead to nowhere, and the whole warehouse is difficult to manage and change without clearly defining the DWH architecture. The DWH architecture provides many benefits as follows:

1. Providing an organizing framework: Explaining the individual components, owners, priorities, and how they fit together.
2. Faster development and reuse: Allow understanding the data warehouse process, data base contents, and business rules more quickly.
3. Management and communications: Defining and communicating requirements and scoping to set expectations.

4. Improved flexibility and maintenance: enabling addition of new data sources, interface standards allow plug and play, and the model and Meta data allow impact analysis and single-point changes.
5. Coordinate parallel efforts: Multiple, relatively independent efforts have a chance to converge successfully.

The DWH architecture presents for understanding the functionality required to successfully implementation of DWH processes, independent of the manner in which the DWH is developing, the following section provides brief overview of previous work efforts on the DWH architecture. Laura Hadley in [106] presented the architecture of a DWH as "a description of the elements and services of the warehouse, with details showing how the components will fit together and how the system will grow over time, it is a set of documents, plans, models, drawing and specifications, with separate sections for each key component area and enough details to allow their implementation by skilled professionals". Inmon, W.H and Vassiliadis, in [102] [1] presented that the generic DWH architecture consists of three layers (data sources, DSA, and primary DWH). Gupta et al, in [30] addressed that the data Warehouse Architecture is dependent on an Enterprise's business process with suitable considerations for query requirements, security, data modeling, management of metadata and technology and bandwidth utilization. The study introduces DWH architecture focuses on three main things: 1) Data Sources (operational systems and flat files), 2) Warehouse (metadata, summary data, and raw data), and Users (analysis, reporting, and mining). Rainardi in [107] reported that the DWH two main architectures: the data flow architecture and the system architecture. The data flow architecture is about how the data stored are arranged within a data warehouse and how the data flow from the source systems to the users through these data stores. The

system architecture is about the physical configuration of the servers, network, software, storage, and clients. Labio et al, in [108] discussed four DWH architecture layers. The DWH architecture layers are: 1) Staging layer: The staging layer is where you load, transform, and cleanse data before moving it to the data warehouse, 2) Data warehouse layer: The data warehouse tables are the main component of the database design. They represent the most granular level of data in the data warehouse, 3) Data mart layer: A data mart is a subset of the data warehouse for a specific part of your organization like a department or line of business and 4) Aggregation layer: Aggregating or summarizing data helps to enhance query performance. Devlin et al, in [109] discussed 8-DWH architecture layers and presented the specifics of any one system. The DWH architecture layers are: 1) data source layer; represents the different data sources that feed data into the DWH, 2) data extraction layer; data gets pulled from the data source into the data warehouse system, 3) staging area; where data sits prior to being scrubbed and transformed into a DWH/data mart, 4) ETL layer; where data gains its "intelligence", as logic is applied to transform the data from a transactional nature to an analytical nature, 5) data storage layer; where the transformed and cleansed data sit, 6) data logic layer; where business rules are stored, 6) data presentation layer; refers to the information that reaches the users, 7) metadata layer; where information about the data stored in the DWH system is stored, and 8) system operations layer; includes information on how the data warehouse system operates, such as ETL job status, system performance, and user access history. Therefore, architecture is very critical in the development of DWH, it shows what, how and why a DWH is developed, and it should be driven by the business needs.

2.4. Clinical Data Quality and Clinical Data Integration

Data quality and data integration become significant issues in healthcare institutions especially in the situations of combing data from more than one system within one institution or more, in order to support decision making and medical research. Data quality and data integration solutions should work together seamlessly. Furthermore, data quality is an essential characteristic that determines the reliability of data for decisions making. With data quality technologies, healthcare institutions understand and improve the completeness, accuracy, availability, timely, integrity, and clinical standards-format of data make data appropriate for a specific use. By improving the quality of data, the healthcare institutions can deliver a high quality data, reduce time and cost to implement data warehouse and maximize the return on investments, reduce the time required for data cleansing, improve medical service, provide medical intelligence on individuals and healthcare institutions for medical research. On the other hand, Data integration is the combination of technical and medical processes used to combine data residing in disparate sources into meaningful and valuable information. Additionally, data integration technologies encompasses discovery, monitoring, transforming and delivering of data from a variety of sources into central repository (CDWH), to provide a unified view of the data assets.

Therefore, the utilization of data quality and data integration solutions makes healthcare institutions have more confidence in the information received to decisions making. Thus, data quality and data integration issues have become the focus of researchers, and numerous open problems remain unsolved. The following sections provide review related work in the field of the data quality and data integration in CDWH.

2.4.1. Clinical Data Integration Issues

The issues of accelerating CDWH designing and developing that respect to clinical data integration has discussed by several researchers to observe data integration. W. Reed, et al in [55] developed data warehouse model collaboration between two hospitals. They proposed the establishment of a data warehouse, based on a patient-centered solution, which takes into account the different legal requirements. The model is focus on the importance of making clinical data and biological samples readily available for research; and the need for an efficient and secured way that enabling access and combined data from clinical sources. Furthermore, the model enabled users to explore the warehouse for various analysis and decision support purposes. Souad Demigha [110] presented a methodology of developing a CDWH in Radiologysenology for exploring and analyzing the information collected during the screening operation. This developed CDWH based on case-based reasoning (CBR) and on an ontological representation of radiologicsenologic domain. The aim of CDWH is linking radiology reports to patient demographics, diagnosis, and pathology data, to assist breast cancer screening in diagnosis, education and research. Furthermore, the paper discussed the importance of defining requirements before developing the CDWH. A DWH for combined different sources of diseases registry data into single repository designed by A. K. Hamoud, et al [51]. The proposed CDWH might be implemented by the doctors, clinicians and other healthcare professionals in the Iraqi clinical institutes to support their decisions. The diseases registry warehouse is the first step to build clinical decision support System (CDSS) which can help the clinicians in supporting their decisions. Szirbik, et al in [99] used rational unified process (RUP) framework when designing a medical data warehouse for elderly patient

care systems. Such methodology emphasized current trends, as early identification of critical requirements, data modeling, close and timely interaction with users and stakeholders, ontology building, quality management, and exception handling. This medical data warehouse delivered stakeholders to perform better collaborative negotiations that brought better solutions for the overall systems investigated. As a result, better decision making processes were established that led to a social impact and enhanced global outcomes. X. Zhou, et al in [74] proposed CDWH solution to improve the clinical data organization, management, processing and analysis. The solution based on the structured electronic medical record (EMR). The study aimed to support clinical researches and medical knowledge discovery. The proposed DWH consists of several key components: clinical data model (schema), ETL tool, (OLAP) and integrated data mining functionalities. A solutions and computational tools presented by H. Hackl, et al [111] for aggregation and analysis and integration of bimolecular and clinical data for the identification of cancer markers and targets for therapy. The study described hardware and software requirements and methods and tools for the data analysis. Furthermore, solutions include storage architecture, high-performance computing, and application servers.

2.4.2. Clinical Data Architecture Issues

The issues of accelerating CDWH designing and developing that respect to architectural design has discussed by several researchers to observe clinical data integration and improved clinical data quality. Erhard Rahm, et al [112] presented a CDWH platform for the integrated analysis of clinical information, microarray data and annotations. The study recommended six stages to improve DWH performance: requirements, presented CDWH architecture, develop model to integrate clinical data, develop ETL process, and performs reporting and statistical

analysis. A CDWH platform with online analysis processing (OLAP) developed by Boon Keong Seah [113] proposed cleansing methodology to meet CDWH needs. The developed CDWH based on the following steps: analyzing the business requirement, developing of DWH modeling, developing of ETL process, indexing data model, encrypting the dimensions, and developing of OLAP analysis. Denise C. Ramick [114] presented the techniques of CDWH, and discussed critical issues relating to the preparation, design, and implementation of a successful CDWH. Furthermore, the study proposed the CDWH development six stages that include: (1) Planning process, (2) CDWH design, (3) CDWH Implementation, (4) CDWH Maintenance (5) Data Analysis, and finally (6) Program Enrollment. Furthermore, the study expands the planning process stage to involve consideration of data sources, data cleansing, warehouse growth rates, future expansion, data inconsistencies, data semantics, storage management, and external data sources. A CDWH conjunction with a Clinical Decision Support System (DSS) is developed by Teh Ying Wah et al. [48]. The developed CDWH is specific to the Lymphoma. The aim of this study is to provide better diagnosis and treatment process for various analysis and decision support purposes. The study proposed CDWH development methodology that consisted of five sequential stages: (1) business analysis stage, (2) architecture design stage, (3) physical development stage (4) implementation and development stage and (5) evaluation stage. A CDWH platform developed by Zhu, Yan , et al [49]. The study presented and described the community healthcare service system in china, the technical infrastructure of DWH, and the derision support architecture for such healthcare management system. The developed CDWH aims to improve the healthcare service quality from the point of view of residents. Furthermore, this paper developed ETL process focusing in users'

requirement analysis, definition of decision subject. Additionally, this study depicts star schema architecture to develop DWH. R. Sahama, et al in [20] presented a clinical data warehouse architecture (a centralized data warehouse structure) that capable to treat integration issues, by describing a common set of task for data warehousing methodologies that including business requirements analysis, data design, architectural design, implementation and deployment. In order to maintain medical data store challenging; complexity and time consuming to review a series of patient records. This approach offer dealing with amount of data, security, and minimized data duplications.

2.4.3: The Extract, Transform, Cleanse and Load (ETL) Issues

The issues of accelerating CDWH development that respect to ETL design has discussed by several researchers. A model proposed by S. H. A. El-Sappagh, et al [87] for conceptual design of ETL processes. The study reported there is a lack of a standard model which can be used to represent the ETL scenarios. The proposed model is built upon the enhancement of the models in the previous models to support some missing mapping features. Furthermore, this study represented needs to find a standard conceptual model for representing in simplified way for the extraction, transformation, and loading (ETL) processes. This study designed a novel conceptual model entity mapping diagram (EMD) as a simplified model for representing extraction, transformation, and loading processes. A semantic framework DWH system development is presented by A. Ta'a, et al [101]. The framework focused on the requirement analysis method for designing the ETL processes. The method - RAMEPs (Requirement Analysis Method for ETL Processes) was developed to support the design of ETL processes by analyzing and producing the DWH requirements as requested by the organization, decision-makers, and developers. Furthermore, the ETL processes are

modeled and designed by capturing two important facts: i) DWH schemas, and ii) data sources integration and transformation. Additionally, the validation process emphasized on the correctness of the goal-oriented and ontology requirement model, and validated by using compliant tools that can build both models. A method for modeling and organizing ETL processes presented by A. Kabiri, et al [115]. This study showed the functional modules that should be distinguished in each ETL (the modularization of ETL process). The proposed approach takes four inputs (namely mapping rules, conforming rules, cleansing rules and specific rules) and produces a conceptual model of an ETL processes using a graphical notation of the framework KANTARA. M. Blechner, et al in [116] proposed a star schema and associate extraction process. This study aims to enhance the collection process of contextual and semantic relationships between the data. Furthermore, the integrated and cleansed data sets generated via a health information exchange (HIE) combined to centralize and automate the development and maintenance of a clinical research data warehouse.

A framework for the conceptual, the logical, and the physical design of ETL processes, presented by A. Simitsis, et al [117]. This Study proposed a novel conceptual and a novel logical model for the representation of ETL processes with two main characteristics: generosity and customization. The proposed framework focus on the optimization of the ETL processes, in order to minimize the execution time of an ETL process. A conceptual model for ETL processes proposed by S. Dupor, et al [118] based on the visualization of data flow showing transformations of records accompanied by attribute transformations. The proposed model contributes to simplification of complex processes by showing a simple visual overview of the process. This proposal arose from recognition of diverse needs in planning,

development and maintenance of ETL processes. Furthermore, this proposal meets the requirements of being able to easily present the process so it can easily be read, quickly developed, efficiently maintained and optimized according to the requirements of business users. A CDWH developed by X. Pan, et al [77] as a fundamental data infrastructure for large scale clinical data management and decision support services. This study introduced an enhanced ETL technique framework, which includes operational data store (ODS) model and two step data preprocessing subcomponents to perform the ETL tasks. Furthermore, ETL task has been separated into two core steps in enhanced ETL component: (1) dynamic filter and copy of the original operational data sources to ODS; (2) specialized transforming the ODS data to detailed CDWH. M. M. Awad, , et al in [119] introduced an enhanced ETL technique framework, they explained the important of data extracting, transforming and loading (ETL) in the process of developing clinical data warehouse to ensure high data quality. The ETL solution add new component to ETL to solve special problems in their own business, they aims to integrate various operational data sources into a clinical data warehouse. This technique improves the ETL performance to be used in clinical data center since they would have various kinds of operational data sources that need to be integrated in this data environment.

2.4.4. Summary

Clinical data quality and data integration become significant issues in healthcare institutions, in order to support decision making and medical research. authors in [55] [110] [51] [99] [74] [111] have proposed CDWH development approaches. The approaches aim to allow appropriate integration of medical information between different healthcare institutions. There are number of data integration technologies

and methods proposed to improve the quality of data that stored in the DWH to enhance the process of decision making in medical domain. These technologies and methods discussed clinical data integration issues; it attempts to answer the question how the clinical data integrated from a various clinical operational systems in a proper way, in order to support medical decision making. But these methods do not describe how to efficiently dealing with data integration issues (such data extraction, data cleansing, and data transformations, data loading issues). Furthermore these proposed structures do not discuss how to efficiently dealing with data quality issues.

Furthermore, other authors have proposed CDWH technologies in [112] [113] [114] [48] [49] [20] aims to treat clinical data quality and integration issues, by describing a common set of task for Data warehousing methodologies that include business requirements analysis, data design, architectural design, implementation and deployment. These approaches offer dealing with amount of data, security, and minimized data duplications. Although these technologies are not directly describe efficient ways to deal with data integration issues (such data extraction, data cleansing, and data transformations, data loading issues). Furthermore these proposed structures may not provide information to support the conduct of medical research; it has been constructed as an extension to the normal medical databases.

ETL processes play a critical role in CDWH development, which is responsible for integrating data from a various sources into CDWH. Some technologies and methods discussed the most important issues in the ETL process that involved conceptual model, data quality issues and enhanced ETL. The authors proposed a conceptual model of ETL models for the extraction, transformation and loading data from a various sources into staging area, such as [117] [118]. However, the proposed

model is informal and mainly deal with functional requirements, however the non-functional requirements (reliability, performance, security) is a very important issue in the data warehouse integration. Additionally, these methods are lacking the presentation of specific extraction, transformation and loading algorithms along with their consequent OLTP and OLAP performance issues. Other authors focused on the data quality issues, they aims to improve the quality of data in staging area, such as [87] [101] [115] [116] . However, these proposed technologies and method do not describe how to efficiently dealing with clinical data quality and data integration issues. On the other hand, some authors proposed enhanced ETL model [77] [119] , concerning on adding new component to ETL to solve special problems. However, the limitation of these methods is that the physical implementation of the ETL operations is not taken into consideration.

The clinical data quality and clinical data integration are summarizing in the following table 2.2.

Table (2-2): Clinical Data Quality and Clinical Data Integration Issues

No	***Clinical Data Quality and Integration Issues***		***Description***
1-	Clinical data integration issues [55] [110] [51] [99] [74] [111]		Where the authors are proposed CDWH development approaches concerned in designing appropriate integrates approaches to integrated medical information from different healthcare institutions.
2-	Clinical data quality and data integration [112] [113] [114] [48] [49] [20]		Where the authors are proposed CDWH development approaches concerned in clinical data quality and integration issues in the evolution of the workflow from a certain plan to another.
3-	ETL processes	ETL model [117] [118]	Where the authors are primarily concerned in providing a conceptual model for workflows.
		Data quality issues [87] [101] [115] [116]	Where criteria are established to determine whether a workflow is well formed to improve the quality of data in staging area.
		Enhanced ETL model [77] [119]	Where the authors are concerned in solving problems by adding new component to ETL process technique.

CHAPTER THREE

RESEARCH METHODOLOGY

This chapter discusses the methodology used in this research work to achieve the objectives stated in section 1.4 of chapter 1. The techniques that used to integrate the medical data and improve the quality of data stored in CDWH, and the evaluation explained. The methodology consists of four stages. The following Figure (5.1) shows the four stages of the research methodology.

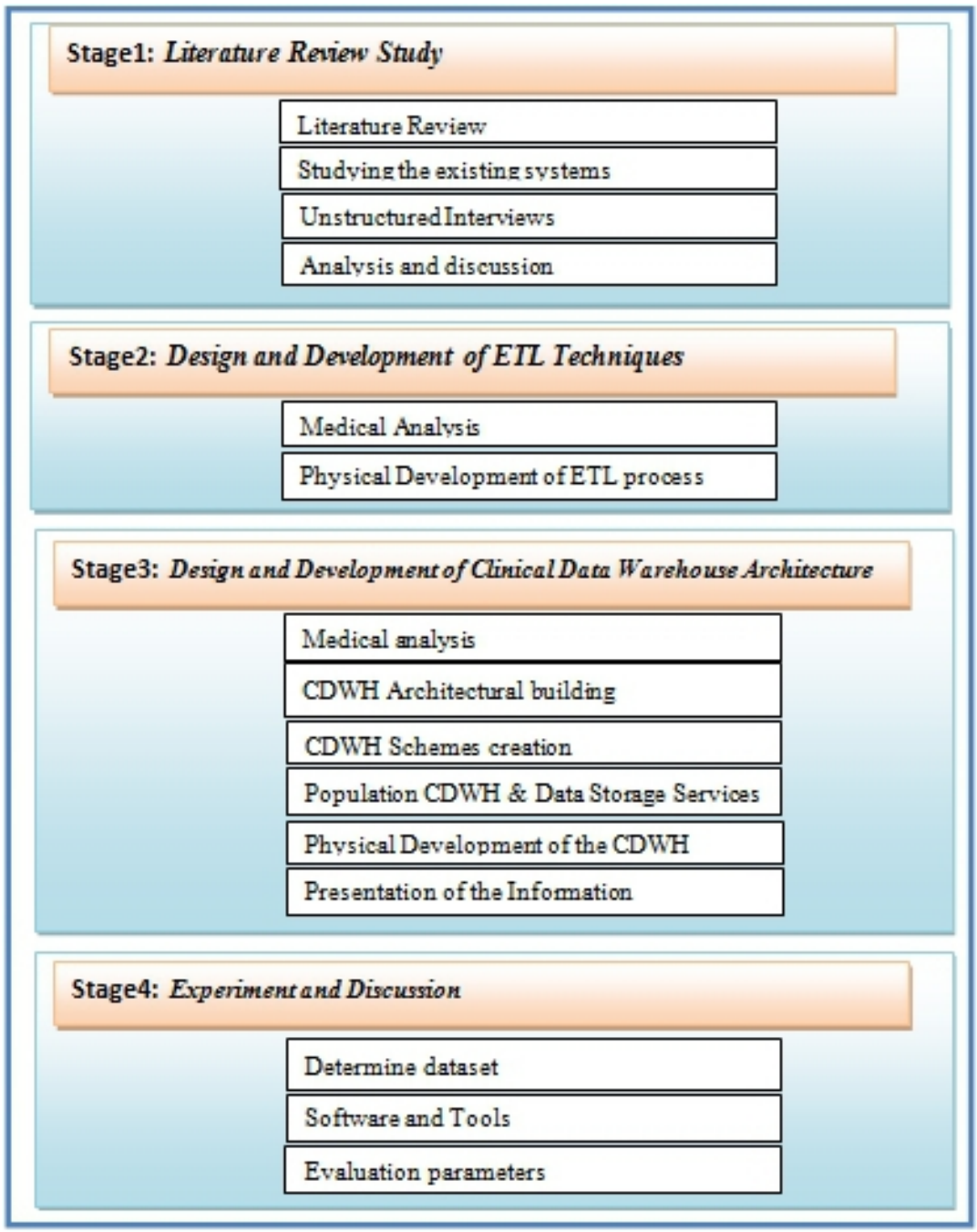


Figure (3.1): Research Methodology Stages

3.1. Literature Review Study

The reviewing of the existing work efforts relevant to the study was completed according to what the integration and data quality are needs, to ensure high quality data in CDWH in order to support medical decision making and enhance medical research environment.

Areas of concentration includes: DWH, using the technologies of DWH in medical fields, CDWH issues and challenges, data warehouse architecture, ETL processes and clinical data quality and clinical data integration. The goal of this literature review is to ensure that the most recent contributions to the relevant literature have been used in the preparation of this study. As a result, this literature review and case studies are important to:

1) Developing background knowledge about DWH and its architecture, CDWH and how it different from DWH and operation systems.
2) Studying and identifying how the DWH technologies are applied in the medical field.
3) Studying data quality issues and identifying the data problems that affect the data stored into CDWH.
4) Studying clinical data integration to determine data integration, data cleansing and ETL issues.
5) Finally Studying ETL processes and analyzing what factors lead to selection of the optimal model to perform these processes.

Furthermore, unstructured interviews and study existing systems have been used to gather more detailed information and identify barriers to the development of the optimal CDWH design.

Upon completion of this stage, we will have comprehensive identified issues in current CDWH, data integration problems that impact

clinical data quality; and to identify issues for designing and developing of ETL technique and designing and developing of CDWH.

3.2. Designing and Developing of ETL Techniques

The ETL processes are present a major part in the process of designing and developing of CDWH, which the efficiency of CDWH is depending mainly on ETL component and its architecture. The proposed ETL techniques designing and developing through three sequential phases to facilitate, manage, and optimize the designing and developing of the ETL processes and introduce techniques for verification and evaluation of ETL Processes. These phases include: medical analysis, and physical development as shown in Figure 5.2.

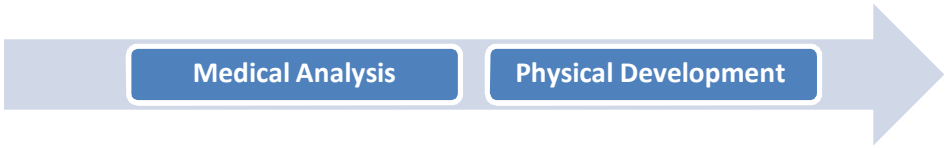


Figure (3.2): Research Designing and Developing of ETL Techniques

3.2.1. Medical Analysis

In the medical analysis the existing processes are studied and analyzed from medical perspective as well as to determine the requirements and the requirements. In medical analysis two aspects are studied in detail to determine the data integration and quality data as well as to determine the medical goals and acceptance criteria.

1) Requirement gathering: To collect the medical requirements. These requirements include: clinical data requirements, clinical data integration requirements, clinical data quality requirements, and ETL technique development requirements for each ETL process.
2) Requirement analysis: The requirements are analyzed to determine the problems domain and to identify the suitable data

model that will be used. Furthermore, the data sources are analyzed in order to comprehend their structure and contents. The ETL technique designed base on these requirements, in order to maintain data integration and improve data quality for CDWH.

3.2.2. Physical Development of the ETL Technique

The physical designing of ETL processes include developed the technologies used to extract, cleanse, transform and load data into the CDWH. The following processes are developed in order to observe the data integration and quality for CDWH.

1) Data Extraction process: This process is responsible for extracting data from the various data sources. The data extraction process involves two steps:
   i. Analysis of Extraction Process Requirements: To determined and defined the data extraction problems that affect the quality of data and integration process, examined how to select the relevant data, and how to transferred these data to staging area.
   ii. Implementation of Extraction Process: Algorithm is developed to handle these data problems in order to extract and integrate the relevant data from various data sources to staging area.
2) Data cleansing process: After data stored in staging area the ETL cleansing process is responsible for cleansing data from any data quality or integration problems.
   i. Analysis of cleansing Process Requirements: To determined and defined data cleansing problems that affect the quality of data and integration process, identified data quality problems and acceptance criteria for each problem area, and how to cleansed these data according to requirements.

ii. Implementation of Cleansing Process: Algorithm is developed to handle these data problems in order to provide a high quality of data in CDWH.

3) Data Transformation Process: This process is responsible for transformed the cleanse data at staging area to the required format for CDWH. Also it responsible for mapping data from staging area fields to CDWH fields.
   i. Analysis of Transformation Process Requirements: To determined and defined data transformation problems that affect the quality of data and integration process, examined how to transform data, and how to map data from staging area to CDWH.
   ii. Implementation of transformation Process: Algorithm is developed to handle these data problems in order to maintain the mapping and transformation process.
4) Data Loading Process: This process is responsible for loading cleansed and transformed data from staging area to the CDWH where, these data accessing by the end users and application systems for reporting and analysis processes.
   i. Analysis of Loading Process Requirements: To determined and defined data loading problems that affect the quality of data and integration process, examined how to select loading strategy options (batch load or simple load), and how to load data to target database.
   ii. Implementation of Loading Process: Algorithm is developed to handle these data problems in order to perform the loading data process correctly and with as little resources as possible.

Upon completion of Designing and Developing of ETL Techniques Stage we will have:

1. *Determined medical requirements which include: clinical data requirements and clinical data integration requirements, clinical data quality requirements, and ETL technique development requirements for each ETL process.*
2. *Identified data quality and integration problems that include: extraction, cleansing, transformation, and loading problems.*
3. *Developed methods to handle data quality problems, and built system involves data extraction, data cleansing, data transformation, and data loading techniques.*

3.3. Designing and Developing of Clinical Data Warehouse Architecture

In this research work, the proposed CDWH is accomplished through six phases, involved: medical analysis, CDWH architectural building, creation CDWH logical schemes, population CDWH and data storage services, physical development of the CDWH, and presentation of the information as shown in Figure 3.3.

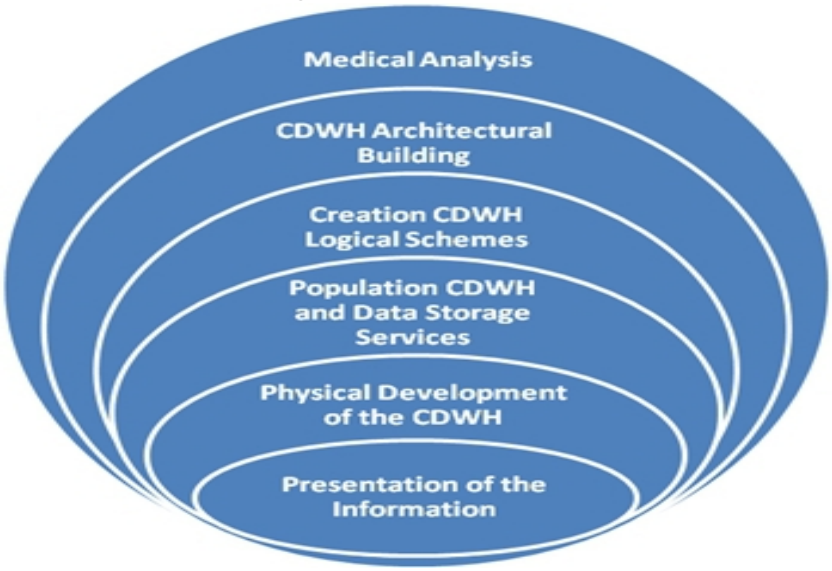


Figure (3.3): Research Designing and Developing of CDWH Architecture

3.3.1. The Medical Analysis

In the medical analysis the existing processes are studied and analyzed from medical perspective as well as to determine project objectives, requirements, constrains and acceptance criteria. In medical analysis four aspects are studied in detail as the following.

1) Requirement gathering: To collect the requirements which, state the medical value of CDWH and drive the architecture of the DWH. The data analysis requirements are collected for the purpose of integrating data before stored into the CDWH to provide more effective analysis environment. These requirements include: dataset, data analysis requirements and system development analysis requirements.
2) Requirement analysis: The requirements are analyzed to understand the purpose of the CDWH. The initial dimensional model of the proposed CDWH is described. A dimensional model consists of two types of tables having different characteristics; dimension and fact tables.
3) Model validation: Is responsible for identifying the data sources of the required data.
4) Requirements modeling: The design of CDWH conceptual data model based on medical data fields is accomplish. The conceptual data model is enabling to understand at high level what the different entities in data are? And how they relate to each other?

3.3.2. Clinical Data Warehouse Architectural Building

The DWH architecture presents for understanding the functionality required to successfully implementation of DWH processes. In this phase the three DWH architecture layers are determined.

1) Data layer: To addresses the configuration of data stored within a CDWH system, it considerate on how the data stores are arranged within a CDWH and how the data flow from the source systems to the users through these data stores.
2) System layer: To addresses all tasks that must be completed to develop and maintain the CDWH. The system layer includes information on how the data warehouse system operates.
3) Infrastructure layer: Is addressing the physical arrangement and connections between the servers, network, software, storage system, and clients.

3.3.3. Creation Clinical Data Warehouse logical Schemes

The design of the logical data model is done in this phase. A logical data model is enabling to understand the details of data without worrying about how they will actually be implemented. The proposed logical data model consists of three fact table and twenty dimension tables.

3.3.4. Population Clinical Data Warehouse and Data Storage Services

This phase involved the design of staging area, system process and infrastructure that include of all necessary hardware, and software that are used for developing the CDWH.

3.3.5. Physical Development of the Clinical Data Warehouse

The objective of this phase is to design and develop the physical data model of the CDWH base on the developing of the CDWH logical data model. The physical development includes all the database processes required to create all table structures, including column name, column data type, column constraints, primary key, foreign key, and relationships between tables. Furthermore, the implementation of a

physical data model includes the creation dimension tables, fact tables, table constraints, and indexing data model.

3.3.6. Presentation of the Information

The objective of this phase is to discover the required knowledge from CDWH and provide many benefits to healthcare institutions. The data mining techniques are integrated with OLAP services in the CDWH system to support the data analysis.

Upon completion of Design and Development clinical data warehouse architecture stage, we will have:

1. *Determined the requirements which include: functional requirements and nonfunctional requirements*
2. *Developed the CDWH Architecture.*
3. *Creation CDWH Schemes logical Mode.*
4. *Accomplish the various processes to develop designing and developing the CDWH.*
5. *Designing and developing of the physical data model of the CDWH*
6. *Developing of analysis application to present the information.*

3.4. Experiment and Discussion

It is the final stage, where the experiment are describing and discussing.

3.4.1. Determine Dataset:

The medical data collected during the regular day-to-day events are recorded and stored in many diverse medical operational systems at the Radiation and Isotopes Centre Khartoum (RICK) – Sudan and Radiation and Isotopes Hospital –Shendi (RICSH) – Sudan. These databases contain medical and clinical data, which divided into categories according to subject domain. These dataset includes patient

demographic information, medical information, diagnostic clinical information, treatment clinical information and, laboratory clinical information. Furthermore, the selection of the relevant data from sources and integrate them into CDWH is represented as an important issues. The selected databases contain comprehensive data on over 26,000 unique patients collected over nearly 5 years from 2009 to 2013. The dataset contains all required data that cover the clinical medical processes.

3.4.2. Software and Tools:

In this research work there are many software and tools are used to developing and implementing ETL and CDWH techniques. These software and tools include the following:

1) PHP Programming language: it used to implement the ETL techniques.
2) SQL Server 2012: It used to create stage area and clinical data warehouse tables.
3) OLAP: it used to discover the required knowledge from CDWH and provide many benefits to medical institutions with quality improvement, data access performance improvement, improves information visualization of the data analysis results, and more informed decision support.
4) Pivot Table: It used to browse OLAP cube, the pivot table is excellent tool to browse OLAP cube.

3.4.3. Evaluation Method for the ETL Techniques

The ETL techniques evaluated using four data warehouse operational perspective, which include: subject-oriented, integrated, non-volatile and time-variant.

1) Subject-oriented: All relevant data according medical purpose are extracted and stored in a CDWH.

2) Integrated: Clinical data that is extracted from a several data sources and loaded into the CDWH must be consistent in format and other aspects.
3) Time-variant: Clinical data in a CDWH support both current and historical perspective measurement. Clinical data are often loaded from the operational databases on a periodic basis, and stored in the data warehouse for a long period of time.
4) Non-volatile: Clinical data in a CDWH always stay stable to enable a highly consistent dimensional view of data. There is no modification or deletion performed against the data after it has been loaded into the CDWH.

CHAPTER FOUR
DESIGN AND DEVELOPMENT OF ETL TECHNIQUE

The objective of this chapter is to design and develop ETL techniques focuses mainly on extraction, transformation, cleansing, and loading processes and its requirements in medical field. The main contribution of this research is designing and developing to enhance ETL techniques which integrate clinical data and improve quality data in CDWH. The remaining of this chapter is organized as follows. Section 4.1 presents the important of the extraction, cleansing, transformation, and loading processes and introduce the proposed ETL model. Section 4.2 discusses the medical analysis process. In section 4.3 the physical ETL technique is developed, while section 4.4 discusses the implementation and evolution of ETL. The summery of this chapter concludes in section 4.5.

4.1 The Extraction, Cleansing, Transformation, and Loading processes (ETL)

ETL process represents a major part in the process of designing and developing of CDWH where the efficiency of DWH is mainly depending on ETL component and its architecture. The general framework for ETL processes are shown in Figure 4.1.

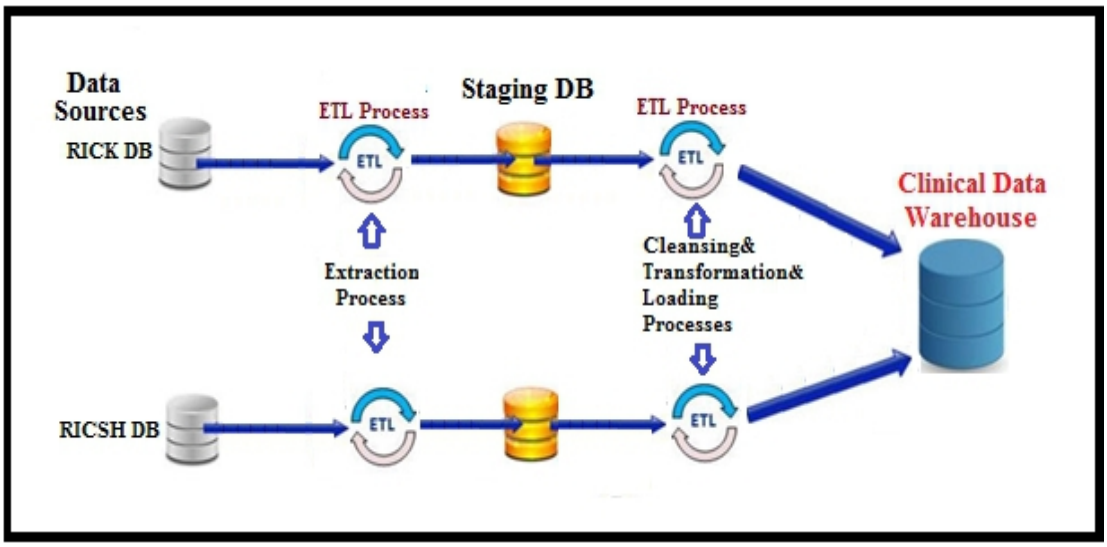


Figure (4.1): Processes of the ETL Model

The data is extracted from various data sources, and then transferred to the staging area where it is cleansed and transformed before being loaded to the target database. However, data sources may have many different data structure formats as flat files, operational systems, and etc. Therefore, the most prominent tasks of the functionality of ETL workflows include: (1) the identification of relevant data at the various data sources, (2) the extraction of this data, (3) the transportation of this data to the stage area, (4) the transformation of the information into a common format, (5) the cleansing of the resulting dataset, on the basis of target database and medical rules, and (6) the propagation of the data to the target database. Furthermore, the metadata plays a decisive role in guaranteeing data quality, it can guide the ETL process more effectively, through many functions. The functions of metadata include; defining the location and properties of data source, determining the corresponding rules of transforming data source and target data, defining the relevant medical logic, and defining other necessary preparation before data loading.

ETL process is critical for the success of a CDWH, and their design and development has been characterized as a labor-intensive and time consuming. Furthermore, these processes are important for improvement of data quality of CDWH contents. Moreover, these data intensive workflows are quite complex in nature, involve integrating various sources, cleansing and transforming activities, and loading facilities. Addition to, the ETL processes must be designed to facilitate modification, because the CDWH is updated periodically due to data sources changing. Therefore, CDWH technique needs to change in order to maintain its value.

Additionally, the ETL process is often a complex combination of process and technology that consumes a significant portion of the DWH

development efforts. However, the ETL process is responsible for the extract data from heterogeneous data sources, converting extracted data into a common format suitable for analyzing and mining, identifying and data quality problems, cleansed data to eliminate undesired data, and finally loading these data into the DWH (Extract-Transform -Cleanse – Load). Due to the complication of medical data structure and clinical operations in real-world clinical environment, it is important to develop powerful ETL techniques. Moreover, in medical field the ETL process activities are highly sensitive to the quality of data and data integration, whereas the poor quality of data will affect the revenue of an organization and causes low quality decision making.

The proposed ETL techniques consists four sequential stages to facilitate, manage, and optimize the designing and developing of the ETL processes and introduced techniques to verification and evaluation of ETL Processes. Where, the design and developing of ETL technique begins with medical analysis (requirements gathering and analysis) stage, where the data sources are analyzed in order to comprehend their structure and contents to observe the integration and quality of data. The deliverable of this stage is a conceptual model for the data store and the processes. In a second stage, the physical design of ETL processes, include design technologies that used to extract, cleanse, transform and load data into the CDWH. Finally, ETL evaluation process stage, where the ETL technique is evaluated.

4.2. Medical analysis

One of the most important aspects of developing ETL techniques is to define medical purpose. The discussion of the medical analysis stage is important to study and analyze the existing process from medical perspective as well as to determine requirements, and the requirements

are further analysis and investigate to determine the data integration and quality data problems.

4.2.1. Requirement Gathering

The requirement gathering is the first phase in medical analysis stage, where the data sources are analyzed in order to comprehend their structure and contents. The requirements are gathered in order to understand the purpose of the ETL process, problems domain and to identify the suitable data model that will be used. Determining and gathering the requirements must be done in proper ways. These requirements include: clinical data requirements and clinical data integration requirements, clinical data quality requirements, and ETL technique development requirements. The ETL technique designing base on these requirements, in order to maintains data integration and improve data quality.

4.2.1.1. Clinical Data Requirements

The data identify and gathering processes must meet specific requirements to improve data quality. The clinical data recorded and stored in various medical systems during different patient visit process at different time. These clinical data, including demographic information, diagnostics and treatment procedures, laboratory test, exam reports, and whether death occurs during hospital stay, and etc. The nature of medical data produces new issues and challenges to DWH technologies. Handling the issues and challenges needs to provide the following requirements [120]:

1. The medical data contain personal information that requires ethical and legal constrains.
2. Medical information systems are of sensitive nature, diverse storage formats, and inherent privacy issues.

3. Medical information systems contain accumulated substantial amounts of data about patients with associated clinical conditions and treatment details. The hidden relationships and patterns within medical information are used to monitor the impact of specific disease, effect of medical processes and their efficiencies and deficiencies.
4. The clinical data contain various types of data such as: text and qualitative format, numeric and quantitative format, Image (such as MRI and Radiology), Ultrasound (such as Echo), Sequential or time series data, Signal data (such as EEG and ECG), and Genetic, microarray and protein data. Consequently, mining these types of data requires transformation mechanism that developed specifically to deal with particular characteristics of medical data.
5. The clinical data require specific mechanism to aggregate data, where the nature of clinical data is complex and poorly characterized mathematically.

4.2.1.2. Clinical Data Integration Requirements

Data integration is a process of combining data from more than one disparate data sources within one or several institutions into target database. This large volume of data is integrated, rearranged and consolidated to provide a unified view to analyze the data. These integrated data are not yet turned into useful knowledge due to the lack of efficient analysis tools, also the lack of standardization between institutions which makes data gathering difficult. Therefore, the data integration is an important issue in developing DWH. On other hand the data integration becomes significant issue in situation of developing a CDWH due to the complexity of the hospital environment such as various care practices, data types and definitions. Additionally, the clinical data integrate from various medical information systems which,

are different clinical routines, incompatible structures, and incompleteness of clinical information systems. Handling medical data integration issues and challenges need to provide the following requirements [120]:

1. Developing enhanced integration technique to combine heterogeneous medical data sources to CDWH.
2. Providing a technique to integrate medical data from heterogeneous clinical information systems and hence needs to be integrated for consistency and analysis.
3. Reducing the dimensions of medical facts describing a current situation of a patient.
4. Minimizing the time requires for extracting, transforming and storing the data in the CDWH.

4.2.1.3. Clinical Data Quality Requirements

Data quality is an essential characteristic that determines the reliability of data for analysis, making decisions and planning. Furthermore, the acceptable data quality in the medical field is a critical issue to the reliability of medical decision making and research environment. Quality of data is achieved when require (useful) data which exactly meet the specific needs stored in common format required by CDWH without data quality problems. However, data quality problems produce at various stages of CDWH development; data integration & data profiling, data staging and ETL, and DWH modeling & schema design. Additionally, these data quality problems must be determined and handled to enhance the quality of data. Handling medical data quality issues and challenges need to provide the following requirements [120]:

1. Needs to lay down strong techniques to manage medical data quality.

2. Defining levels of data quality which are appropriate to the organization.
3. Understanding the data quality problems from medical perspective, there are a wide variety of dimensions on which data quality can be affected.
4. Understanding the format of data stored by each source, there are wide varieties of structured and unstructured types of data.

4.2.1.4. ETL Technique Development Requirements

ETL plays a vital role in DWH solutions, which responsible for the extracting data from heterogeneous data sources, converting extracted data into a common format suitable for analyzing and mining, identifying and data quality problems, cleansed data to eliminate undesired data, and finally loading these data into the DWH (Extract-Transform -Cleanse -Load). In medical field ETL process activities are highly sensitive to quality of data and data integration. Furthermore, poor quality of data will affect the revenue of an organization and causes low quality decision making. Additionally, due to the complication of medical data structure and clinical operations in real-world clinical environment, it is important to develop a powerful ETL tool to integrate, transform, and cleanse medical data before loading these data into CDWH. In addition, the ETL process is quite complex in medical field which requires extracting data from several sources, cleansing and transforming activities, and loading facilities. Thus, each phase in the ETL process has its issues and challenges:

(I) Extraction process: Extraction process is responsible for extracting relevant data from heterogeneous data sources. Where, the ETL process requires connection to the source systems, and selecting the relevant data needed for analytical processing and research within the CDWH. The data extract

from numerous disparate source systems and each of these data sources has its distinct set of characteristics that need to be managed in order to effectively extract data for the ETL process. Furthermore the complexity of the extraction process depends on the data characteristics and attributes, amount of source data and processing time. Therefore, the ETL process needs to effectively integrate technology to extract these data. Handling extraction process issues and challenges need to provide the following requirements to ensure subject-oriented of the CDWH [120]:

1. Analyzing data sources in order to comprehend their structure and contents to understand the data that exist in the databases to identify the relevant data at these sources that depending on the purpose of CDWH, the selection of these data requires:
    a. Identifying source systems that contain the required data and identifying the quality and scope of each data source.
    b. Understanding the format of data stored by each source to determine whether all the data available to fulfill the requirements or not, and the required data fields populate properly and consistently.
    c. Identifying the attributes contain in each data source.
2. Determining the options of extracting the data from the source systems which include update notification, incremental extracts, and full extracts to capture only changes in source files.
3. Determining the protocols for data transferring.
4. Determining encryption standards need to be set with each of the source systems.

5. Monitoring data transfer failures and errors and making notifications through different methods such as control files, metadata files, email notifications, system log writing and file system log writing.

(II) Cleansing process: Data cleansing is one of the most important issue in ETL process as it ensures the quality of the data in the CDWH. The data cleansing deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. The data cleansing phase involves three steps include: data analysis, data refinement and data verification. The objective of data analysis is to identify problems area and detects the data problem. Data problems include completeness problems, accuracy problems, consistency problems and Integrity problems. For each problem area, the data quality issues and acceptance criteria are identified, then, for each data quality issues, the solutions are developed. Furthermore, the data with quality issues will be refined using some of the data cleansing methods to realize their full benefit. Additionally, the cleansed data then will be assessed against the acceptance criteria again to ensure that the data issues have to be resolved after the data cleansing process. Finally, after verification, the data will be moved from staging area to CDWH. Therefore, the objective of data cleansing process is to make cleansing and conforming on the extracted data to gain accurate data of high quality. Handling cleansing process issues and challenges in CDWH need to provide the following requirements [120]:

1. Understanding the data quality problems from medical perspective.

2. cleansing of the extracted data set, according to the required medical rules,
3. All the requisite information is available, free from errors, in a usable state.
4. The data collected is relevant to the medical purpose.
5. The ability to link relative records together to ensure the data consistency in format.
6. The data satisfy a set of constrains, and maintains in a consistent fashion to ensure the data values consistent across datasets.
7. All patient basic information records must contain a unique patient identification number for each patient.

(III) Transformation process: Transformation process is to transform the extracted data into a common format by applying a set of conditions, rules or functions. The transformation phase tends to make multiple data manipulations on the incoming data according to medical needs, to ensure that the data load into CDWH is integrated and accurate. The transformation process requires joining the data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules by defining the granularity of fact tables, the dimension tables, CDWH schema (snowflake), derived facts, slowly changing fact tables and dimension tables. In medical field very complex transformations provide the following requirements to meet the medical needs of the targeted system [120]:

1. Understanding the format of data stored by each source to determine whether all the data fulfill the requirements or not.

2. Figure out a way of mapping the external data sources and internal data sources fields to the CDWH fields.
3. Transforming and coding the medical data into the required content format for CDWH storage.
4. Providing amount of manipulation needed for transformation process according to the medial needs of the CDWH such as normalization, standardization, aggregation and etc, using different methods according to requirement specifications,
5. Providing suitable data model to allow querying by multiple dimensions.

(IV) Loading process: Loading process is the process of loading data from staging area to the CDWH. The extracted, cleansed and transformed data is written into the dimensional structures actually accessed by the end users and applications. A major data loading problem is the ability of ETL process to discriminate between new and the existing data at loading time; the new rows that need to be appended and rows that already exist need to be updated. Handling loading process issues and challenges need to provide the following requirements to ensure that loading process should perform correctly and with as little resources as possible [120]:

1. The ability of ETL process to provide the desired latency in updating the dataset (batch load or simple load).
2. The ability of ETL process to discriminate between new and the existing data at loading time; the new rows that need to be appended and rows that already exist need to be updated.
3. Data is up to date or is provided at the time specified (data tagged with a time).

4. The ability of ETL process to schedule extracts by time, interval, or event (strategy of periodical refreshing).
5. The ability of ETL process to manipulate loss of data during the ETL loading process.
6. Performing the loading process of data correctly and with as little resources as possible.

4.2.2. Requirement Analysis

In the previous discussion the requirements are gathered. Demonstrating these requirements is important for designing an efficient and robust ETL workflow. Thus, the conceptual model for the propose ETL techniques is describing and presenting the necessary processes of ETL process which clarify how ETL workflows, and how the ETL process performs its intended operation during a specified period of time under given conditions? Clinical data quality and integration issues is discussed in many research works; authors are proposed CDWH development approaches concerned in clinical data quality and designing appropriate integrates approaches to integrated medical information from different healthcare institutions.

The clinical data integration and quality problems will be handled within ETL techniques, where the authors concerned in providing a model for workflows to improve the quality of data in staging area. The classification of these clinical data integration and quality problems are summarized in Table 4.1.

Table (4.1): Clinical Data Quality and Clinical Data Integration Issues Classification

Issue	Problem	Problem Description	Most the possible causes of data quality Problems
Extraction Data	**Medical Purpose Problem**	The medical purpose and requirement is not determined in proper ways.	- Incomplete or wrong requirement analysis of the project leads to poor schema design. - Lack of currency in medical rules cause poor requirement analysis which leads to poor schema design.
	Identify Problem	Inadequate selection of data sources required to achieve medical purpose.	- Sources which do not comply with medical rules. - Different medical rules of various data sources. - Usage of decontrolled applications and databases as data sources for CDWH in the organizations.
	Relevancy problem	The data collected is not relevant to the medical purpose.	- Inadequate selection of relevant data from selected data sources.
	Loss of data Problem	The data is loss during the process of transferring of data form source to staging area	- Loss of data during the extraction process (rejected records).
	Scalability Problem	An ETL process is not able to handle higher volumes of data.	- Multiple data sources generate semantic heterogeneity. - As time and proximity from the source increase, the chances for getting correct data decrease.
	Integrity problem	The inability to link related records together may actually introduce duplication across systems, data the Doesn't consistent in format	- Missing or improper relation tables. - Missing in joining data from several sources, - Violate domain values.
	Completeness problem	All the requisite information is not recorded /available, or in an unusable state (the data isn't thorough in the attributes that require them).	- Fields with null values. - Some fields with false or incomplete values.
	Consistency problem	The data is not satisfies a set of constraints, and not maintained in a consistent fashion. Data values doesn't consistent across data sets.	- Inconsistent use of special characters (for Eg. A date uses hyphens to separate the year, month, and day whereas a numerical value stored as a string uses hyphens to indicate negative numbers).
	DWH required format Problem	Data in inappropriate forms for mining.	- Different data types for similar columns (A patient ID is stored as a number in one table and a string in another). - Fields with Inconsistent/Incorrect data formatting (E.g. specific attribute is stored in one table in the specific format and in another table in different format.
Cleansing data	**Accuracy problem**	Missing value (untimely or not current data can impact operational and analytical applications) or conflicts in data.	- Data in an unusable state. - conflicts in data - Duplicate data
	Availability Problem	Data are available when required	- The resources of the system that needed are not available when needed. - The required data in the data sources is not available.
Transformation Data	**Validity problem**	Validity is refers to the correctness and reasonableness of data Conformity.	- Data in inappropriate forms for mining. - Use of different representation formats in data sources.
	Huge data	Complex data analysis and mining on huge amount of data take a very long time, making such analysis impractical or infeasible.	The concept of data reduction is commonly understood as either reducing the volume or reducing the dimensions (number of attributes).

Loading Data	**Freshness problem**	Inability of the system to provide the desired latency in updating the data set correctly and with as little resources as possible.	- Inappropriate ETL process of update strategy (insert/update/delete), - Inappropriate ETL process of load strategy opted (Bulk, batch load or simple load), - Lack of periodical refreshing of the integrated data storage (Data Staging area),
	Reliability problem	Reliability refer to "The probability that the ETL process will perform its intended operation during a specified time period under given conditions".	- Data warehouse architecture undertaken affects (staging, non staging architecture), - Choice of dimensional modeling (Star, Snowflake) schema. - Type of staging area, relational or non relational, - Incomplete/Wrong identification of facts/dimensions, bridge tables or relationship tables or their individual relationships.
	Timeliness problem	Data is not up to date or is not provided at the time specified.	-The data is not tagged with a time. - The data warehouse is not hosted both historical and current data?

The ETL techniques will design and develop to address the following questions about the problems that affect the data integration and the data quality at extraction, cleansing transformation, and loading:
(I) Data extraction problems: The questions about the problems that affect the data integration and the data quality in data extraction process presents as the following:

i. Is the medical rules determine in proper ways?
ii. What are the data extraction problems that affect the quality and integration of data?
iii. Are you using sources in complying with the medical rules?
iv. Do you have multiple formats to be accessed- relational DBs, flat files, etc.?
v. How data extraction process techniques search the relevant data**?**
vi. What are the transferred options of the extracting data from the systems into staging area (update notification, incremental extracts, or full extracts)?
vii. What are the required/available frequencies of the extracts?

(II) Data cleansing problems: The questions about the problems that affect the data integration and the data quality in data cleansing process presents as the followings:

i. What are the data cleansing problems that affect the quality and the integration of data?
ii. What properties of the data need to be addressed in order to ensure that the data are cleansed?
iii. Are all the required data recorded /available?
iv. Is there redundancy in data?
v. Are the required data correct, free from errors and in comply with medical rules?
vi. Are the required data collected relevant to the medical purpose?
vii. Are the required data fields satisfy a set of constraints and maintained in a consistent fashion?
viii. Are all data cleansed before loading into the CDWH?

(III) Data Transformation problems: The questions about the problems that affect the data integration and the data quality in data transformation process presents as the Followings:

i. What are the data transformation problems that affect the quality and the integration of data?
ii. How an effective converts of the extracted data into the proper format for data analysis process can be done?
iii. Are the required data consistent achieved in CDWH?
iv. How the clinical data can be standardized and mapped to CDWH?
v. All the needed mapping are done in a proper way?

(IV) Data loading problems: The questions about the problems that affect the data integration and the data quality in data loading process presents as the Followings:

i. What are the data loading problems that affect the quality and the integration of data?
ii. What properties of the data need to be addressed in order to ensure that data is loaded correctly to the CDWH?
iii. Are the developed data loading technique effective to load the data correctly with as little resources as possible?

4.3. Physical Development of the ETL Techniques

The design and develop of the physical data model of the ETL base in the previous discussion of medical analysis. The main goal of physical ETL Model design is to observe the data integration and quality for CDWH. The most important tasks of ETL process includes:

(I) Understanding the data that existed in the sources database to identify the relevant data at the sources, this data selection depends on purpose of CDWH, therefore, the selection of these data requires:
   i. Identifying source systems that contained the required data and identifying the quality and the scope of each data source.
   ii. Understanding the format of data stored by each source to determine whether all the data available are fulfilling the requirements or not, and the required data fields populated properly and consistently.
   iii. Determining the size of the source.
   iv. Identifying the attributes contained in each data source.

(II) The extraction of these data.

(III) The transportation of these data to the data stage area.

(IV) Data pre-processing:

i. The cleansing of the resulting data set, according to the required medical rules.

ii. The transformation (i.e., summarization, integration, and aggregation) of the extracted data from various sources format into a CDWH format according to the requirement specifications.

(V) The propagation of the data to the CDWH, and establishing the update policy for each local source.

The physical design of ETL processes include developed four techniques used to extract, cleanse, transform and load data from data sources into the CDWH. These techniques are implementing through two phases involved Extracting/Cleansing and Transformation/Loading. The extracting and cleansing phase where the data are extract from sources, transfer to staging area and the datasets are cleanse. On the other phase, data in stage area are transforming to appropriate format to mining, mapping, and loading into CDWH. The following sections are illustrating the four mention techniques.

4.3.1. Data Extraction Process

The ETL extraction process is responsible for extract and integrates the required data from the various data sources as a first step in ETL process. However, each data source has its distinct set of characteristics that need to manage in order to effectively extract data process. Moreover, the data sources are different platforms, such as database management systems, operating systems, and communication protocols. The ETL extraction process technique is proposed to develop the extraction technique. This extraction technique consists of Analysis of extraction process requirements, and Implementation of extraction process.

4.3.1.1. Analysis of Extraction Process Requirements

Analysis of extraction process requirements aim to determining and defining data extraction problems that may affect the quality of data and integration process, examining how to select the relevant data, and how to transfer these data to staging area. These data extraction problems are determined from medical respective. The data problems include clear understanding of medical purposes, identifying the required data, scalability, consistency, integrity, completeness and format problem**.** However, the required data stored in several data sources which are different structure and technologies. This required clear understanding of these sources structure and technologies to understand the format and attribute of data stored by each source. Whereas, identifying of the data sources depend on the purpose of the developing of CDWH.

The required data categorize in three types of medical data related to the patients. These categories involve cancer disease information, demographic information, and patient clinical records information. The first category, cancer disease information refers to general cancer's information, such as cancer types, cancer stages, diagnosis types, symptoms, treatment types, risk factor types, physicians and etc. The second category, demographic data refer to personal patient's information such as location, occupation, sex, education, parent's relatives, tribe and etc. The third category, patient clinical information refers to transaction data that collected about each patient during having treatment. These data involved patient diagnosis, patient treatment procedure, treatment Result, treatment side effects, laboratory test and etc.

The extraction process consists of two phases, initial extraction, and incremental extraction. In the initial extraction, the relevant clinical

data extract from data sources for the first time. This process is done only one time after developing the CDWH. On the other hand, the incremental extraction refreshes the CDWH with the modified and added data in the data sources since the last extraction process. This process is done periodically according to the medical needs. Once data extracted from source systems according specific rules and conditions, the data are transferred to staging area. Furthermore, the transfer process is monitoring and making notifications when failure and error occur. In CDWH update process ETL technique captures only change in data sources since the last extraction process.

4.3.1.2. Implementation of Extraction Process

Based on the previous discussion, algorithm is proposed to handle data problems that may affect the quality and the integration of data during extraction process. The following section describes the algorithm for data extraction processes technique. The extraction algorithm involves number of methods as explain in figure 4.2.

Algorithm 1: Data Extraction Process

i. **Identifying and analyzing the data sources and creating the data sources list for each data source:**

1. Identifying the list of data sources that contain the required data according to medical needs to observe medical goals,

2. Indentifying the type of the databases (format of data stored by each source),

4. Analyzing each of data source to identifying the structure of data sources (format of data stored by each source).

5- Identifying the attributes contain in each data source,

6. Checking if it is a new source add to the data source list or checking if there any object added to data source.

ii. **Establishing connection and extracting data:**

7. Determining the type of the data source,

8. Using appropriate drivers to establish the connection,

9. Identifying the data source object that contains the required data.

10. Select the relevant data according medical objectives;

11. Checking to select the required extraction option (initial/ or incremental) to perform the required processes.

12. Mapping the data source and data staging area schemas,

13. Identifying the data problems that affect the data quality at extraction process and handle these problems with appropriate method.

14. Indentifying the options of extracting data from the source systems into staging area,

15. Identifying the appropriate extraction strategy.

iii. Loading of extracted data into data staging are:

16. Establishing connection with data staging area using appropriate data connection for transferring data,

17. Transferring the data into data staging area.

iv. Modification /updating of data Staging Area:

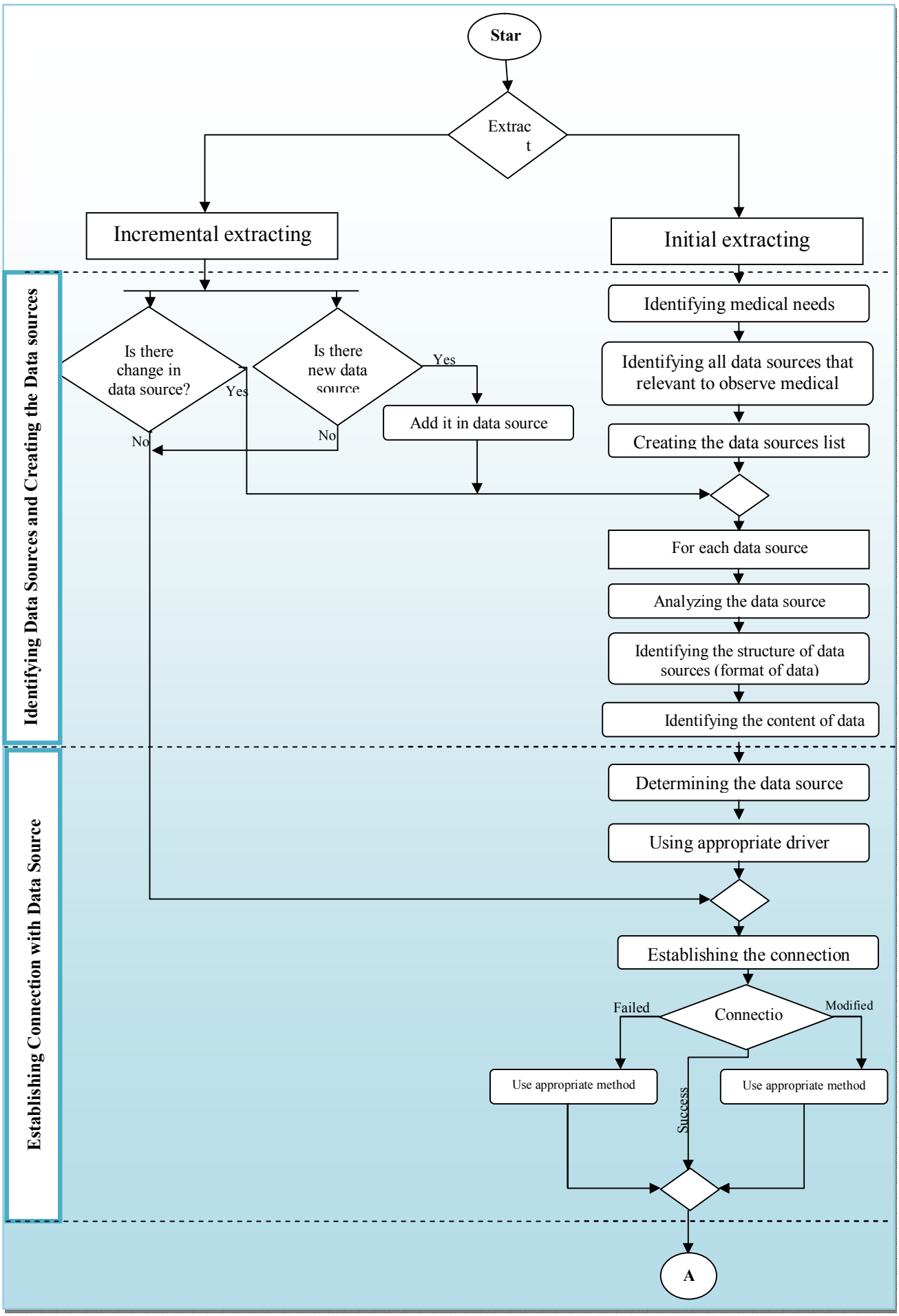
18. Identifying the changes in the data sources,

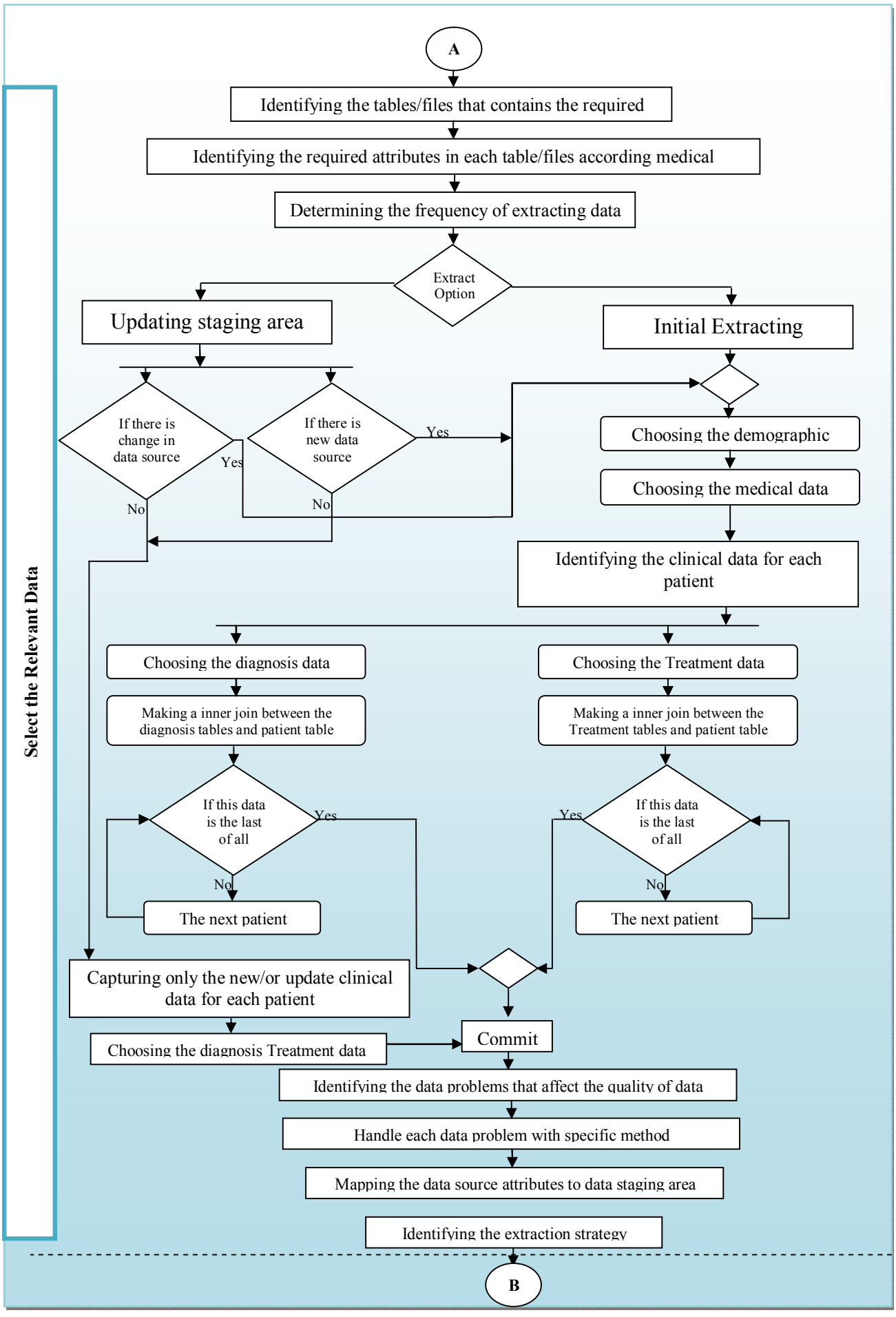
19. Update DSA.

v. Monitoring data transferring:

20. Monitoring data transfer failures and errors,

21. Making notifications.


Star
Extrac t
Incremental extracting
Initial extracting
Identifying Data Sources and Creating the Data sources
Is there change in data source?
Yes
No
Is there new data source
Yes
No
Add it in data source
Identifying medical needs
Identifying all data sources that relevant to observe medical
Creating the data sources list
For each data source
Analyzing the data source
Identifying the structure of data sources (format of data)
Identifying the content of data
Establishing Connection with Data Source
Determining the data source
Using appropriate driver
Establishing the connection
Failed
Connectio
Modified
Use appropriate method
Use appropriate method
Success
A


Select the Relevant Data
A
Identifying the tables/files that contains the required
Identifying the required attributes in each table/files according medical
Determining the frequency of extracting data
Extract Option
Updating staging area
Initial Extracting
If there is change in data source
If there is new data source
Yes
Yes
No
No
Choosing the demographic
Choosing the medical data
Identifying the clinical data for each patient
Choosing the diagnosis data
Choosing the Treatment data
Making a inner join between the diagnosis tables and patient table
Making a inner join between the Treatment tables and patient table
If this data is the last of all
If this data is the last of all
Yes
Yes
No
No
The next patient
The next patient
Capturing only the new/or update clinical data for each patient
Choosing the diagnosis Treatment data
Commit
Identifying the data problems that affect the quality of data
Handle each data problem with specific method
Mapping the data source attributes to data staging area
Identifying the extraction strategy
B

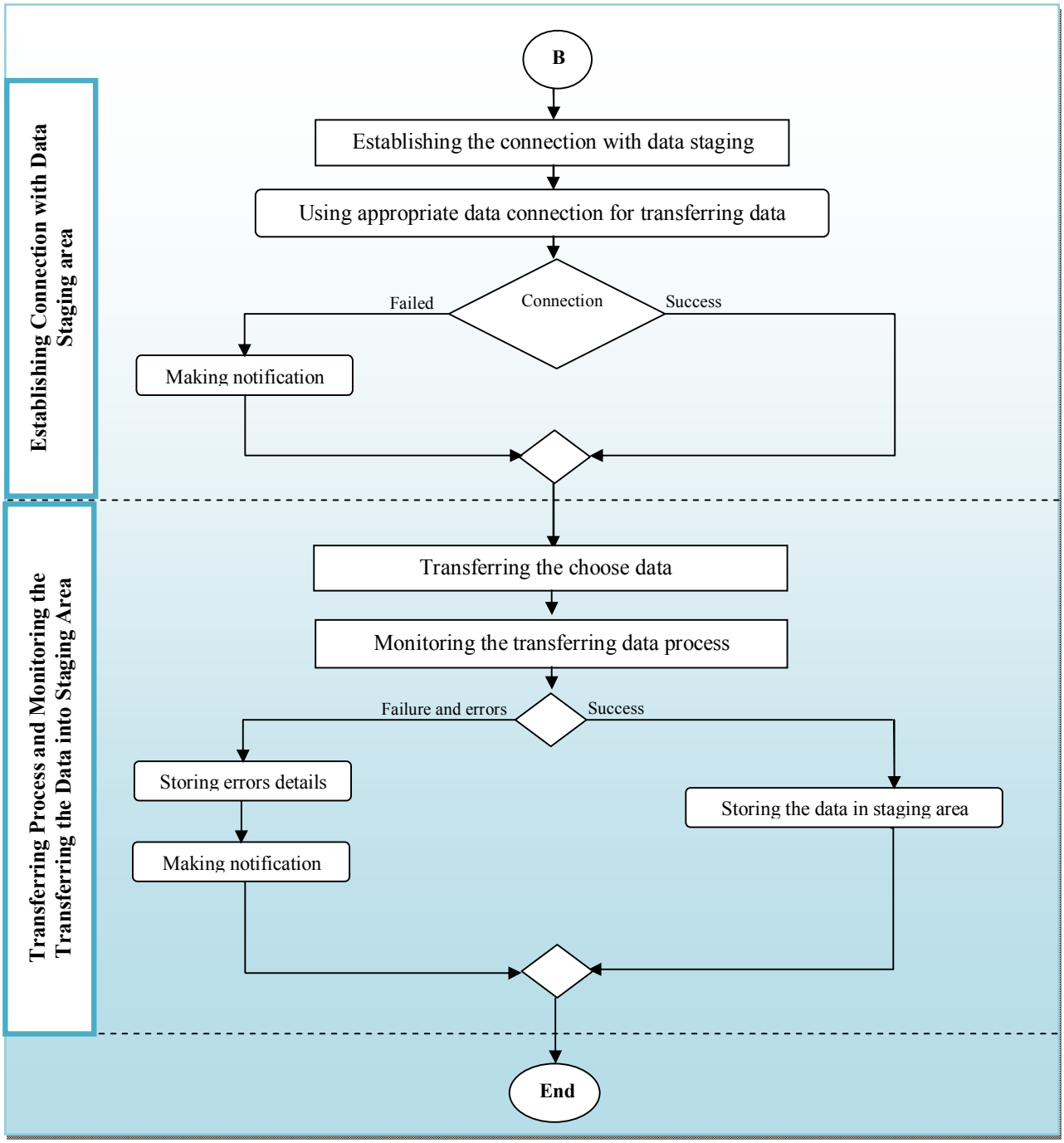


Figure (4.2): Flowchart of Data Extraction Technique

4.3.2. Data Cleansing Process

ETL cleansing process is responsible for cleansing data stored in staging area from any data quality or integration problems. An ETL cleansing technique is proposed to perform data cleansing as a pre-processing step in staging area before storing the resultant data into CDWH. The cleansing technique identifies and handles the data quality problems that may affect data quality from medical respective to obtain integrated and high quality dataset. Furthermore, the cleansing technique consists of analysis of cleansing process requirements, and implementation of cleansing process.

4.3.2.1. Analysis of Cleansing Process Requirements

Analysis of cleansing process requirements aim to determine and define the data cleansing problems that may affect the quality of data and the integration process, identifying data quality problems and acceptance criteria each problem area, and how to cleanse these data according requirements. Data cleansing technique checks whether the data stored in staging area complies with the data quality rules. A data quality rule is the criteria that verify the data from the source systems are within the expected range and in the correct format.

Thus, to obtaining dataset satisfy medical requirements required clear understanding of CDWH needs and goals, clear understanding and determining of the data quality problems from medical perspective. The cleansing process involves a set of cleansing methods perform to handle the various data quality problems. The data problems that affect the quality of data and the integration include: accuracy and availability problem. For each problem area, the data quality issues and acceptance criteria are identified, then, for each data quality issues, the solutions are developed.

4.3.2.2. Implementation of Cleansing Process

The cleansing process is responsible for cleansing the dataset at staging area. Based on the previous discussion, algorithm is proposed to handle these data problems in order to provide a high quality of data in CDWH. Implementations of cleansing process describe the design and develop of cleansing technique where, the data with quality problems are determined and refined. Furthermore, for each problem area, the data quality issues and acceptance criteria are identified. Then, for each data quality issues, the solutions are developed. The following section describes the algorithm for data cleansing process as shown in figure 4.3.

Algorithm 2: Data Cleansing Process

i. Identifying the data Objects Tables/Files:

1. Identifying list of data objects in staging area which contain data that required cleansing according to medical needs,
2. Selecting the needed attributes contain in objects in order to cleanse,
3. Update the data objects list after any extraction process.

ii. Establishing connection and Cleansing data:

4. Using appropriate drivers to establish the connection to staging area according to the type of staging area,
5. indentifying the properties of the data need to be addressed to ensure that the data is cleansed.
6. Searching and identifying the data problems and data problems instances,
7. Identifying manipulation option (methods) needed for cleansing process according to requirement specifications.

iii. Cleansing of data:

8. Establishing connection with staging area,
9. Handling the errors by special method for each instance,
10. Storing the cleansing data in temporary tables,

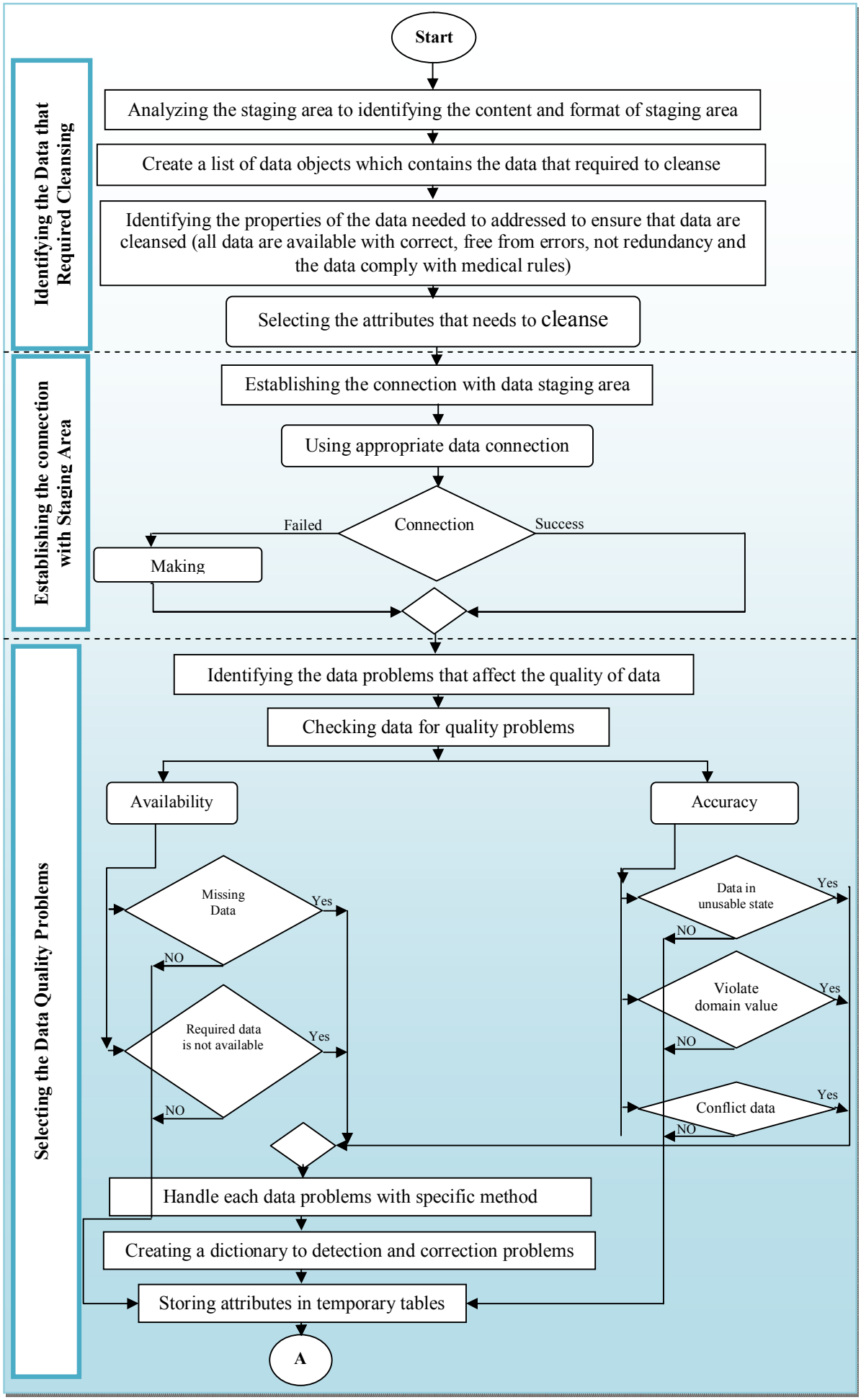
11. Check cleansing process to verify and ensure that the data problems have been handling,

iv. Loading cleanse data into data target:

12. Combining of all attributes in temporary tables,
13. Loading to the target (staging area).

v. Monitoring data transferring:

14. Monitoring data cleansing failures and errors,
15. Making notifications.


Identifying the Data that Required Cleansing
Start
Analyzing the staging area to identifying the content and format of staging area
Create a list of data objects which contains the data that required to cleanse
Identifying the properties of the data needed to addressed to ensure that data are cleansed (all data are available with correct, free from errors, not redundancy and the data comply with medical rules)
Selecting the attributes that needs to cleanse
Establishing the connection with Staging Area
Establishing the connection with data staging area
Using appropriate data connection
Connection
Failed
Success
Making
Selecting the Data Quality Problems
Identifying the data problems that affect the quality of data
Checking data for quality problems
Availability
Accuracy
Missing Data
Yes
NO
Required data is not available
Yes
NO
Data in unusable state
Yes
NO
Violate domain value
Yes
NO
Conflict data
Yes
NO
Handle each data problems with specific method
Creating a dictionary to detection and correction problems
Storing attributes in temporary tables
A

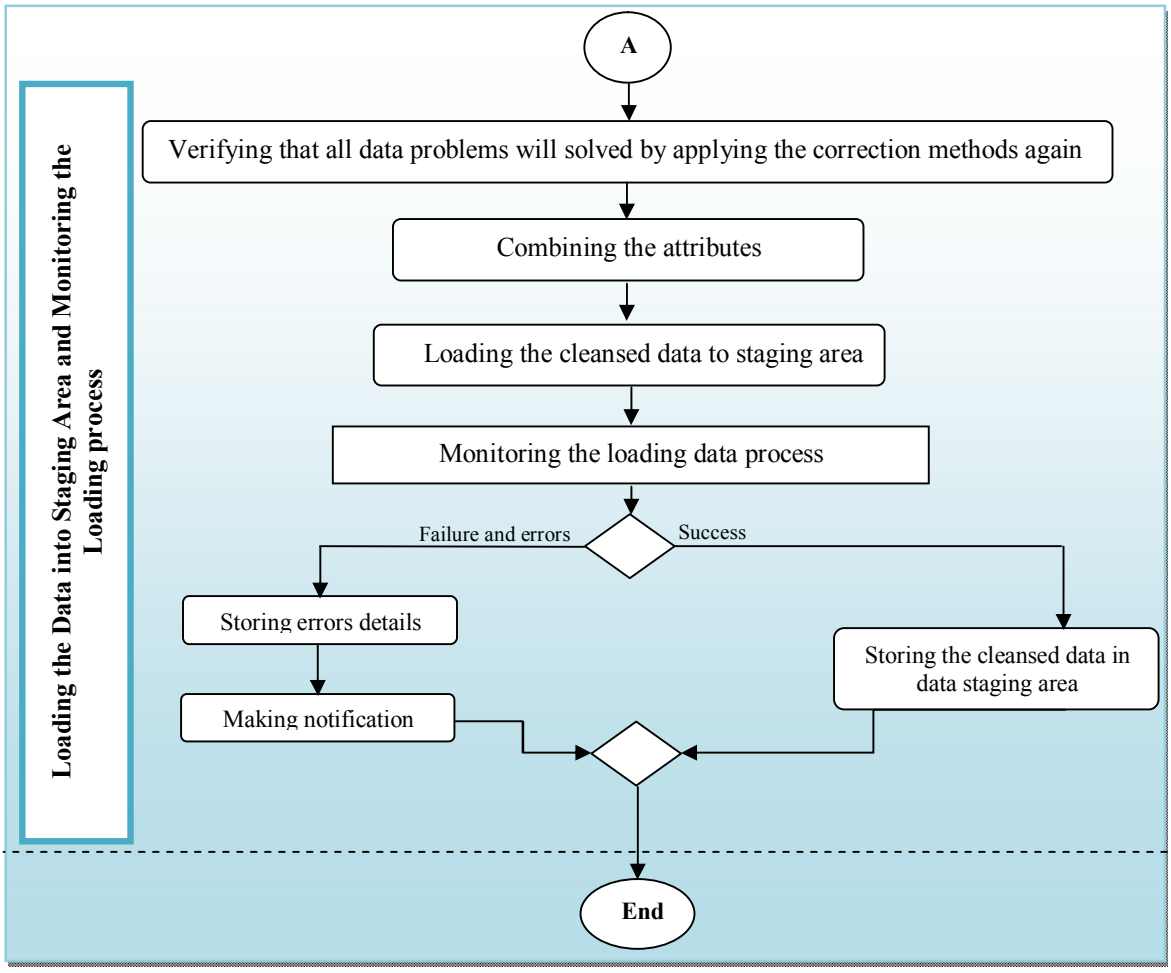


Figure (4.3): Flowchart of Data Cleansing Technique

4.3.3. Data Transformation Process

The transformation process transforms the extracted data from various sources into appropriate format for data mining according to requirement specifications. Furthermore, the transformation process responsible for mapping from staging area fields to CDWH fields. However, each data source has its distinct set of characteristics that need to manage in order to perform effective transformation process.

The transformation technique proposed to transform the clinical data according to requirement specifications and mapping them from the staging area to CDWH. Furthermore this transformation technique identifies and handles the data problems that may affect data quality during transformation process.

The proposed technique consists of Analysis of extraction process requirements, and Implementation of extraction process.

4.3.3.1. Analysis of Transformation Process Requirements

Analysis of transformation process requirements aim to determining and defining data transformation problems that affect the quality of data and integration process, examine how to transform data, and how to map data from staging area to CDWH. The data transformation problems that affect the quality and the integration of data are determined and defined by understanding the mapping and transformation process requirements from medical respective. These data problems include validity, huge data, and data format problems. However, the data stored in various data sources with deferent structure and format of the required clinical data. This required clear understanding of these source structures to understand the format and attribute of data stored by each source. Furthermore, the complex nature of clinical data requires a powerful transformation technique. Moreover, the data sources contain a massive of data about patients with the

associate clinical conditions and treatment details. This requires determining the hidden relationships and patterns within medical information which use to monitor the impact of specific disease, effect of medical processes and their efficiencies and deficiencies. Additionally, the clinical data contain various types of data that required aggregating.

Based on the previous discussion, the transformation involves a set of transformation methods perform to handle the data quality problems that are reflected in the dataset. These methods include:

1. Converting the source-system-specific representation of data into target representation and eliminating technical or semantic differences (conflicts) between multiple sources involved in the ETL process. These rules include:
2. Checking the logic relationship between attributes. This performs to ensure data validity and integrity checks.
3. Aggregation is applied to the same type of data, which can increase sample size and hence improve analysis powerful.
4. Dimension reduction: where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
5. Discretization and concept of hierarchy generation: where raw data for attributes are replaced by range or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and these are powerful tools for data mining.
6. Data standardization: To facilitate matching and integration, attribute values convert to uniform format to the CDWH.
7. Mapping Process: There are a set of mapping rules defining the mapping of attribute available in a specific source to the attributes used in the target. These functional rules include:
8. Simple attribute mapping, where an attribute of the source is assigned to an equivalent one in the target.

9. Constant value assignment, where a constant value is assigned to an attribute of the target.
10. Mapping via formulas, where predefined functions are used to feed an attribute of the target.
11. Complex mapping, to state specifications for complex transformations.

4.3.3.2. Implementation of Transformation Process

Based on the previous discussion, the following algorithm is proposed to handle the data quality problems that may exist during transformation process in order to maintain the transformation and mapping process as shown in figure 4.4.

Algorithm 3: Data Transformation Process

i. **Identifying the data Objects Tables/Files:**
   1. Identifying the list of data object in staging area which contain data that required transformations according to medical needs,
   2. studying the attributes contain in each data object in staging area,
   3. Identifying the format of data stored by data object to determine whether all the data fulfill the requirements or not.
   4. Selecting the needed attributes contain in objects in order to transform,

ii. **Establishing connection and transformation data:**
   5. Using appropriate drivers to establish the connection to staging area,
   6. Selecting all attributes that need to transform, put them in temporary tables,
   7. Identifying the data problems that affect the data quality at transformation process and handle these problems with appropriate method.
   8. Identifying manipulation option needed to determine the appropriate method for transformation process according to requirement specifications such as summarization, integration, and aggregation.

iii. **Transformation of data:**

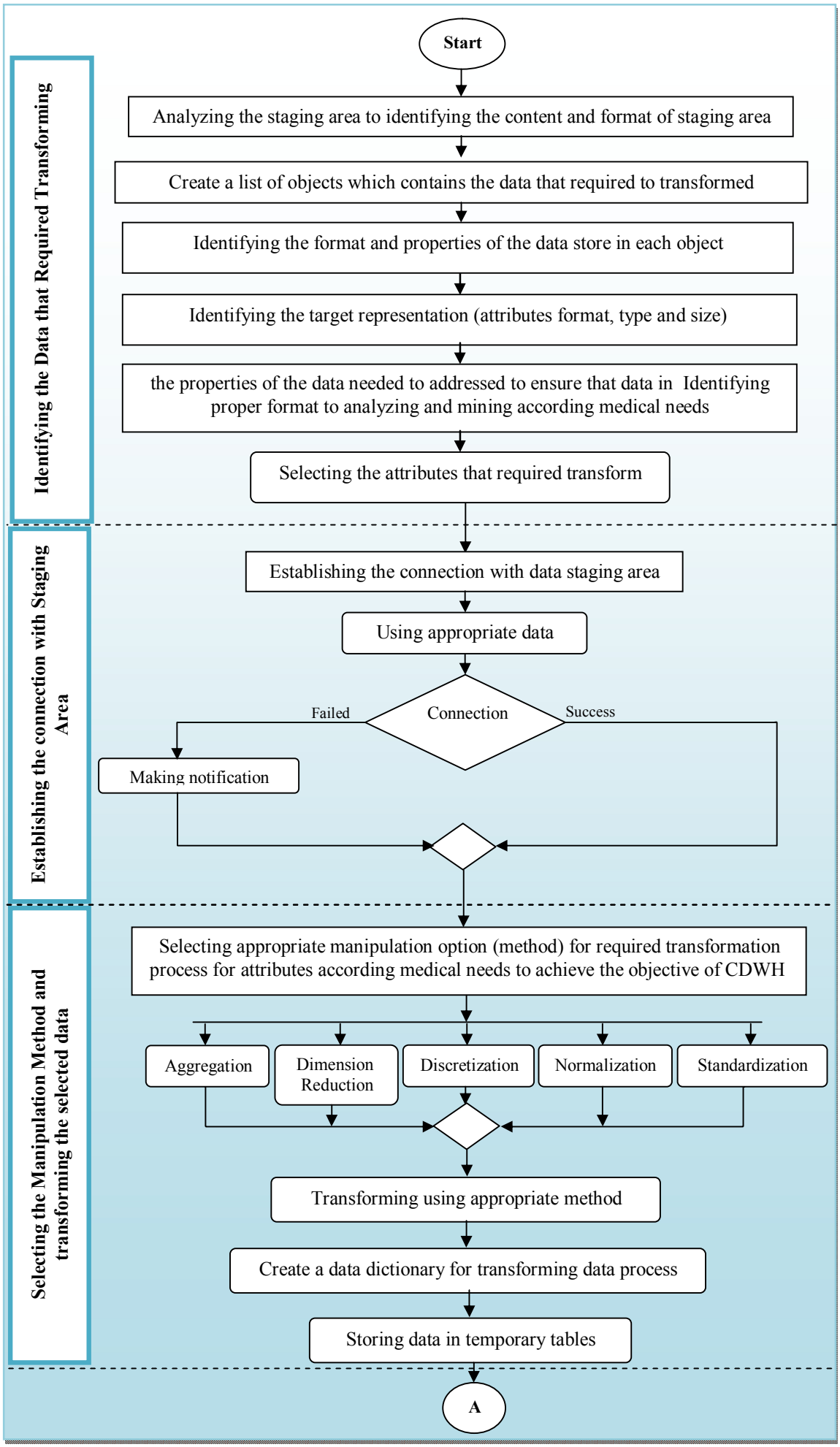
8. Establishing connection with staging area,
9. Transforming all attributes in temporary tables into the required format for target storage.
10. Loading the transformation data to the target data.

iv. **Mapping of data:**

11. Figuring out a way of mapping the data sources fields to the CDWH fields.

v. **Monitoring data transferring:**

12. Monitoring data transformation and mapping failures and errors,
13. Making notifications.


Identifying the Data that Required Transforming
Start
Analyzing the staging area to identifying the content and format of staging area
Create a list of objects which contains the data that required to transformed
Identifying the format and properties of the data store in each object
Identifying the target representation (attributes format, type and size)
the properties of the data needed to addressed to ensure that data in Identifying proper format to analyzing and mining according medical needs
Selecting the attributes that required transform
Establishing the connection with Staging Area
Establishing the connection with data staging area
Using appropriate data
Connection
Failed
Success
Making notification
Selecting the Manipulation Method and transforming the selected data
Selecting appropriate manipulation option (method) for required transformation process for attributes according medical needs to achieve the objective of CDWH
Aggregation
Dimension Reduction
Discretization
Normalization
Standardization
Transforming using appropriate method
Create a data dictionary for transforming data process
Storing data in temporary tables
A

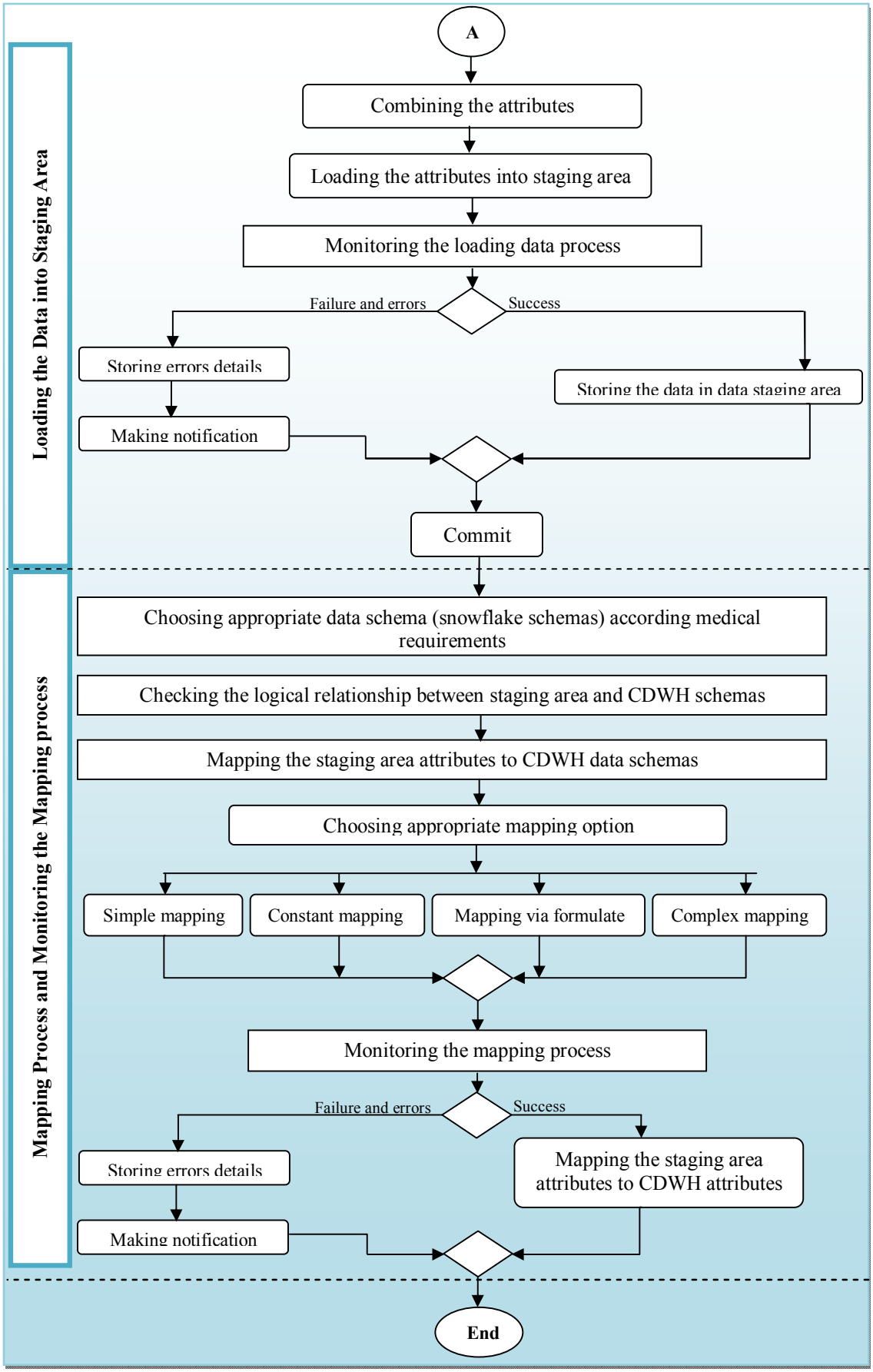


Figure (4.4) Flowchart of Data Transformation Technique

4.3.4. Data Loading Process

The loading process is the final process in the ETL system. The loading technique is proposed to loading cleansed and transformed data from staging area to the CDWH. Furthermore, the cleansing technique identifies and handles the data quality problems that may occur during loading process.

The loading technique consists of Analysis of loading process requirements, and Implementation of loading process.

4.3.4.1. Analysis of Loading Process Requirements

Analysis of loading process requirements aim to determine and define data loading problems that affect the quality of data and integration process, examining how to select loading strategy options (batch load or simple load), and how to load data to the target database. These data problems include freshness problem, and discriminate between new and the existing data at loading time. Furthermore, the loading technique requires mechanism to determine load strategy options, and schedule extracts by time, interval, or event. Additionally, these technique required to determine the strategy of manipulate the loss of data during the loading process and Strategy of periodical refreshing of staging area.

4.3.4.2. Implementation of Loading Process

Based on the previous discussion, algorithm is proposed to handle the data quality problems in order to perform the loading data process correctly and with as little resources as possible. The following section describes the algorithm for data loading process as shown in figure 4.5.

Algorithm 4: Data Loading Process

1-Identifying the data object and creating the Data objects list:

1. Identifying the list of data object in staging area which contain data that loads to update CDWH according medical needs,

2. Identifying frequency of data object loads according medical needs,

3- Identifying the data problems that affect the data quality at loading process and handle these problems with appropriate method.

2- Establishing connection and Loading data:

4. Using appropriate protocols for data transfer.

5. Disable any constraints and indexes to make the loading process efficient.

6. Map the staging area and data CDWH schemas,

7. Identify the desired latency in updating the dataset (batch load or simple load).

8. Indentify the options to schedule extracts by time, interval, or event (strategy of periodical refreshing),

9. Loading data.

3- Loading of data into CDWH:

10. Establishing connection with target database,

11. Loading the data into target database.

4- Modification /updating of CDWH:

12. Identifying the changes in the staging area,

13. Discriminating between new and the existing data at loading time; the new rows that need to be appended and rows that already exist need to be updated.

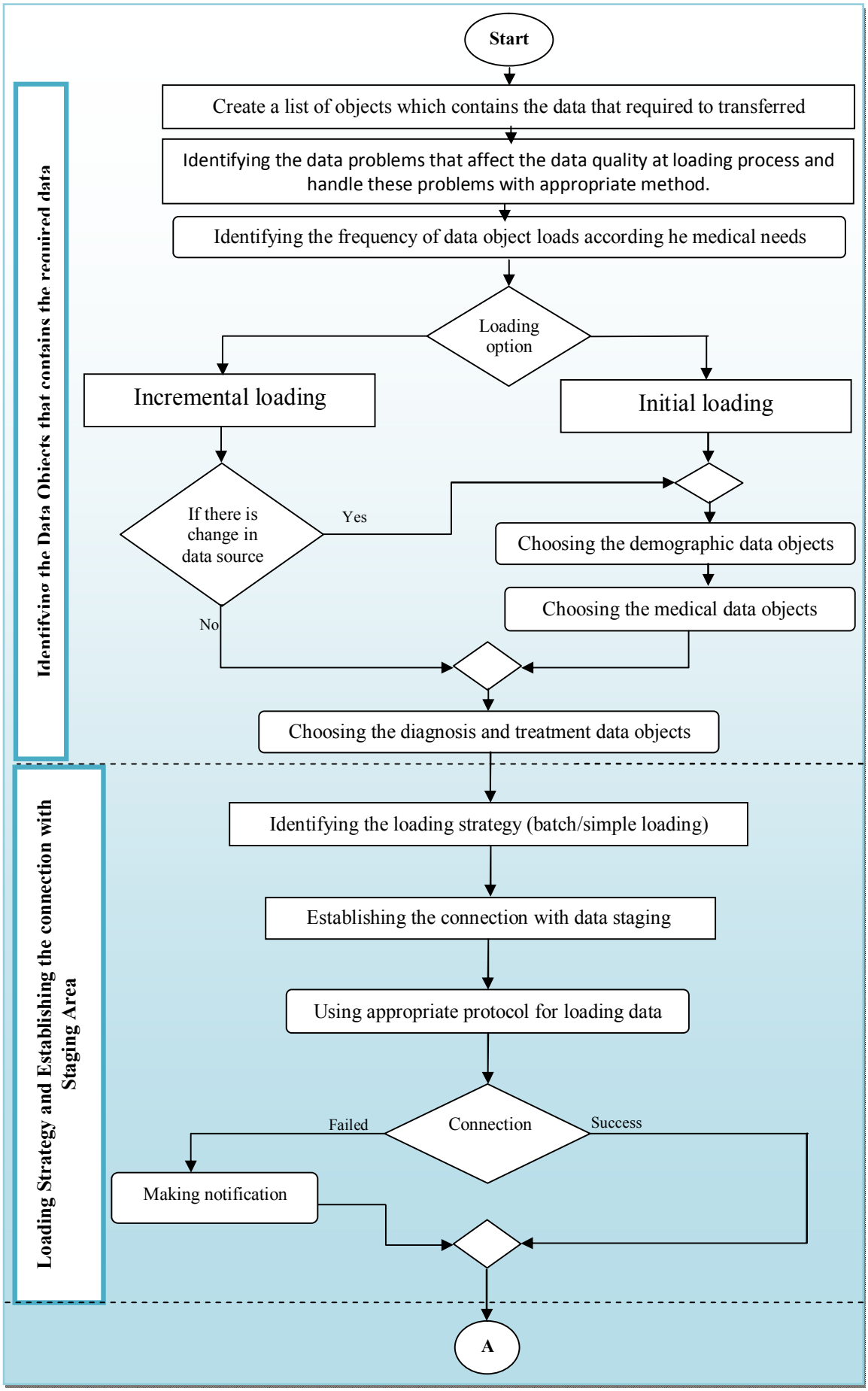
14. Utilizing automate data transfer from staging area into target database,

15. Updating target database.

5- Monitoring data transferring:

16. Monitoring data loading failures and errors,

17. Making notifications.


Identifying the Data Objects that contains the required data
Start
Create a list of objects which contains the data that required to transferred
Identifying the data problems that affect the data quality at loading process and handle these problems with appropriate method.
Identifying the frequency of data object loads according he medical needs
Loading option
Incremental loading
Initial loading
If there is change in data source
Yes
No
Choosing the demographic data objects
Choosing the medical data objects
Choosing the diagnosis and treatment data objects
Loading Strategy and Establishing the connection with Staging Area
Identifying the loading strategy (batch/simple loading)
Establishing the connection with data staging
Using appropriate protocol for loading data
Connection
Failed
Success
Making notification
A

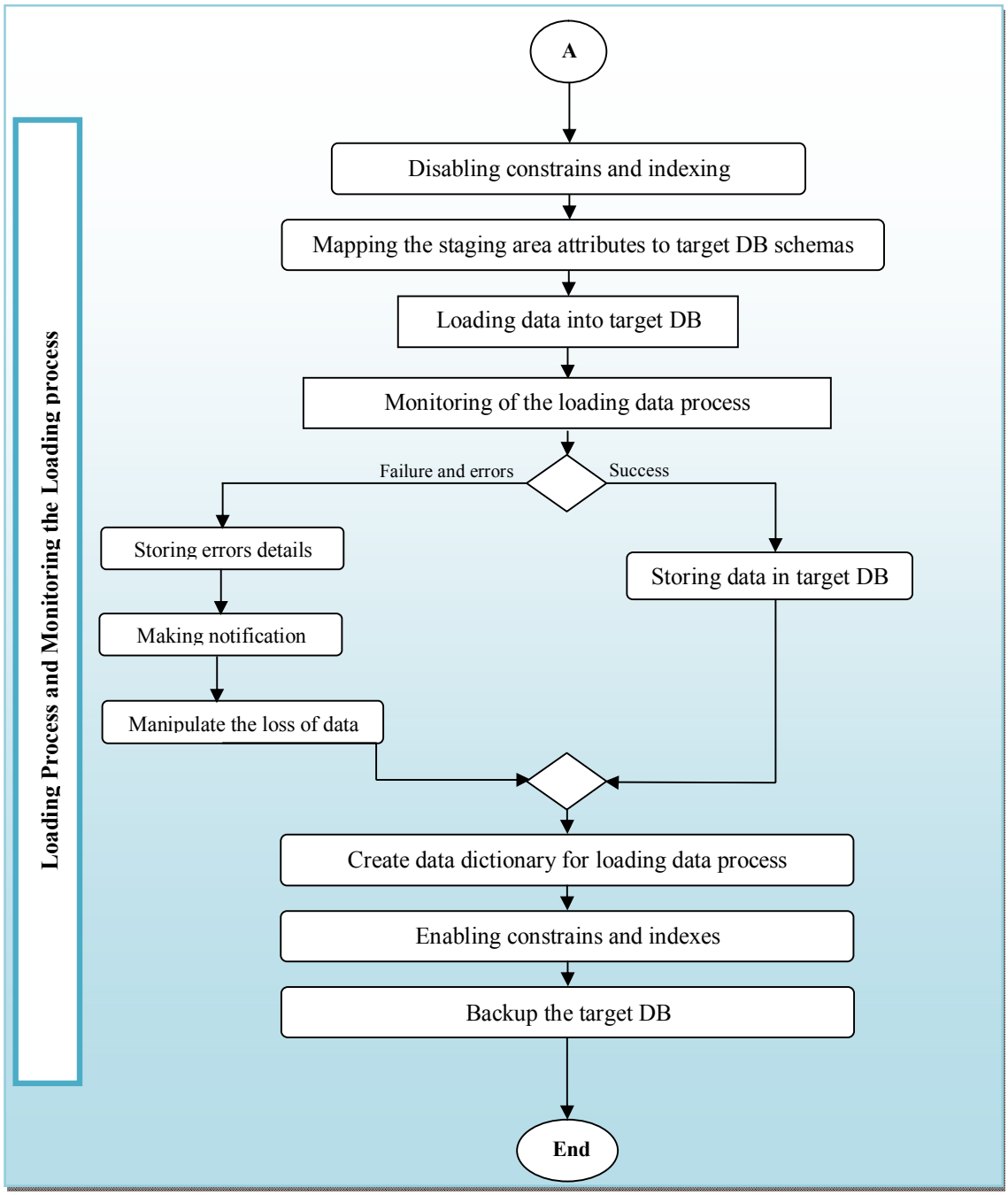


Figure (4.5): Flowchart of Data Loading Technique

4.4. Evaluation of the ETL System

Testing of a data quality at every stage throughout the ETL (extraction, cleansing, transformations, and loading) process is important as more data is being collected and used in medical fields.

The developed ETL techniques are assessed from the ETL process as the following:

1. Data extraction performance (Subject-oriented: Verifying that all the required data have been loaded from data sources to the staging area.
2. Data cleansing and transformation performance (Integrated):
   a) Verifying that the required data are correct, free from errors in staging area and in comply with medical rules.
   b) Ensuring that all data have been transformed and mapped correctly during the ETL process.
   c) Verifying that data values conform with specified formats
   d) Ensuring that the ETL process correctly handle the data quality problems.
3. Data loading performance (Time-variant): Ensuring that all data have been loaded form data sources to stage area and from stage area into CDWH in correct manner and with little resources as possible.
4. ETL performance (Non-volatile): Ensuring the performance of ETL process.

4.5 Summary

Medical field is more complicated than the business area and produces a set of integration issues and challenges. These issues involve clear identification of extracting, cleansing, transformation and loading requirements as well as developing and evaluating the ETL techniques.

These requirements include the clinical data requirements, clinical data integration requirements, clinical data integration requirements, and ETL techniques development requirements. However, the clinical data is different from business data where the clinical data produce new requirements, if not considered the quality of data is affect. The complexity of the hospital environment evolves the diagnosis and treatment procedures and their relation with other information produced distinct set of data characteristics. Moreover, the data collected from different hospitals which use and diverse data format and DBMS. These complexity of clinical data rise several issues such as; poorly characterized mathematically, difficult data type for mining, and difficult to determine hidden relationships.

The functionality of ETL workflows include: the identification of relevant data at the various data sources, the extraction of these data, the transportation of these data to the stage area, the cleansing of the resulting dataset, the transformation data into a common format, and loading of the data to the CDWH. Whereas, clear understanding of the medical purpose represents an important issue in the process of developing ETL techniques.

Consequently, the study discussed how the clinical data extraction, data transformation, data cleansing, data loading processes are performed. The medical data requirements are collected to understand the problems domain in addition to determine the suitable solution to handle these problems. Furthermore, ETL techniques aim to integrate large volumes of data collected from several clinical information systems. Therefore, development of effectively integrated systems that have different platforms is required. The ETL techniques are proposed, these techniques are designed through four sequential phases which include: medical analysis, physical development, and evaluation.

On the other hand, data quality is one of important issues in the process of developing CDWH because medical decisions making made based on data stored in CDWH. The qualities of information depend on three things which includes; the quality of the data itself, the data quality problems, and the quality of the database schema. However, the diversity of structured and unstructured types of data required identifies the relationships among various systems in order to understand the format of data stored in each source. Therefore, the achievement of good performance and deliver quality information required the consideration in the complexity of the medical institution environment issues during the process of developing of ETL process.

As a result, a set of algorithms which used to handle the data quality problems, and ELT techniques, are developed. In addition to, the produced data will be evaluated against acceptance criteria to ensure that the medical objectives are achieved. The data quality assessment process includes:

1. Evaluation of extracted data pattern is performed to identify the truly interesting patterns representing knowledge.
2. Establishing metrics to assess the validity of the data extracted from the various systems.
3. Systematically reviews all the data elements, considering factors such as range of values, number of null records, duplicate records, compliance with medical rules, and inaccurately recorded information.
4. Providing a summary of data problems and a strategy to handle these problems.

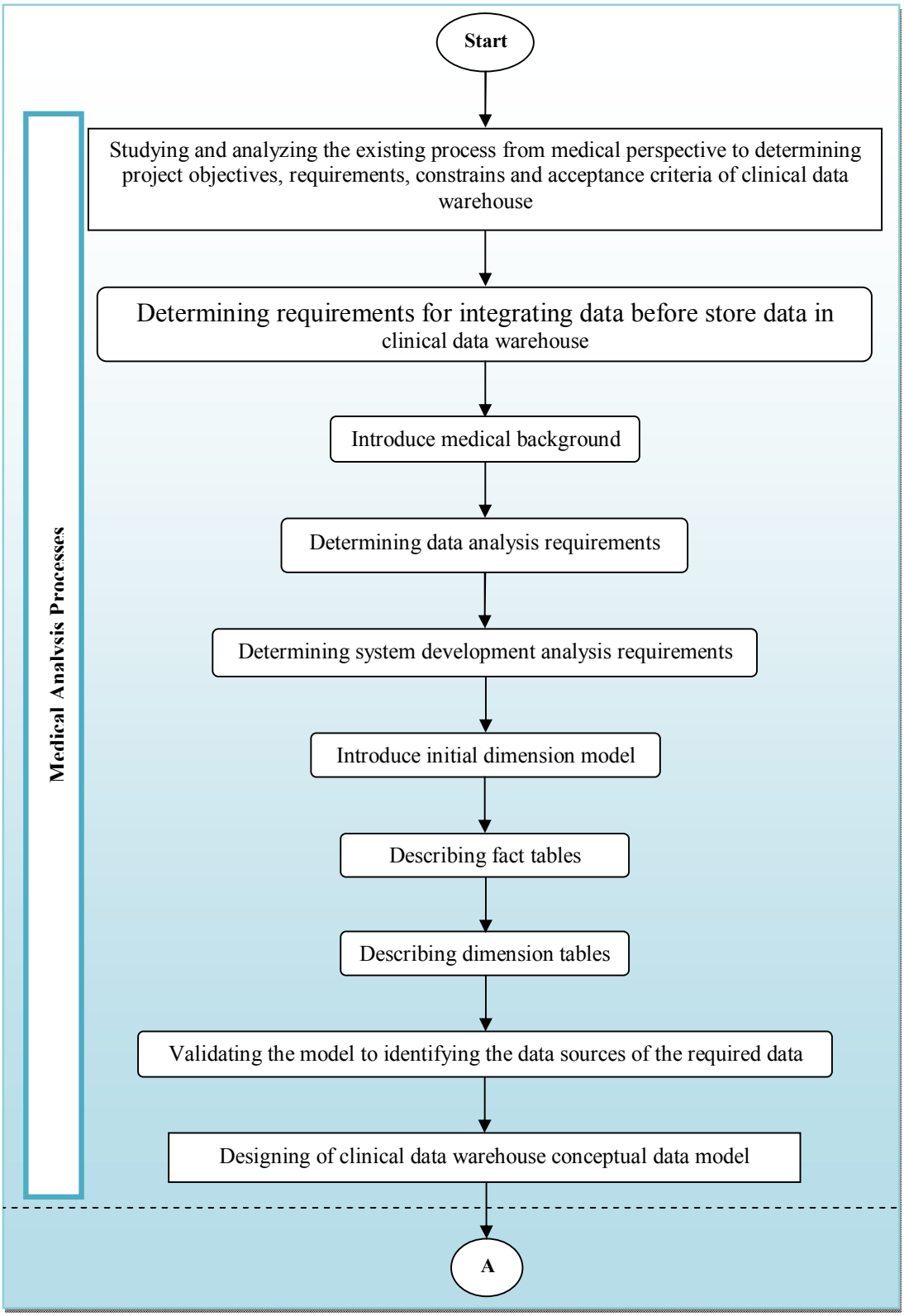
CHAPTER FIVE
DESIGN AND DEVELOPMENT OF CLINICAL DATA WAREHOUSE

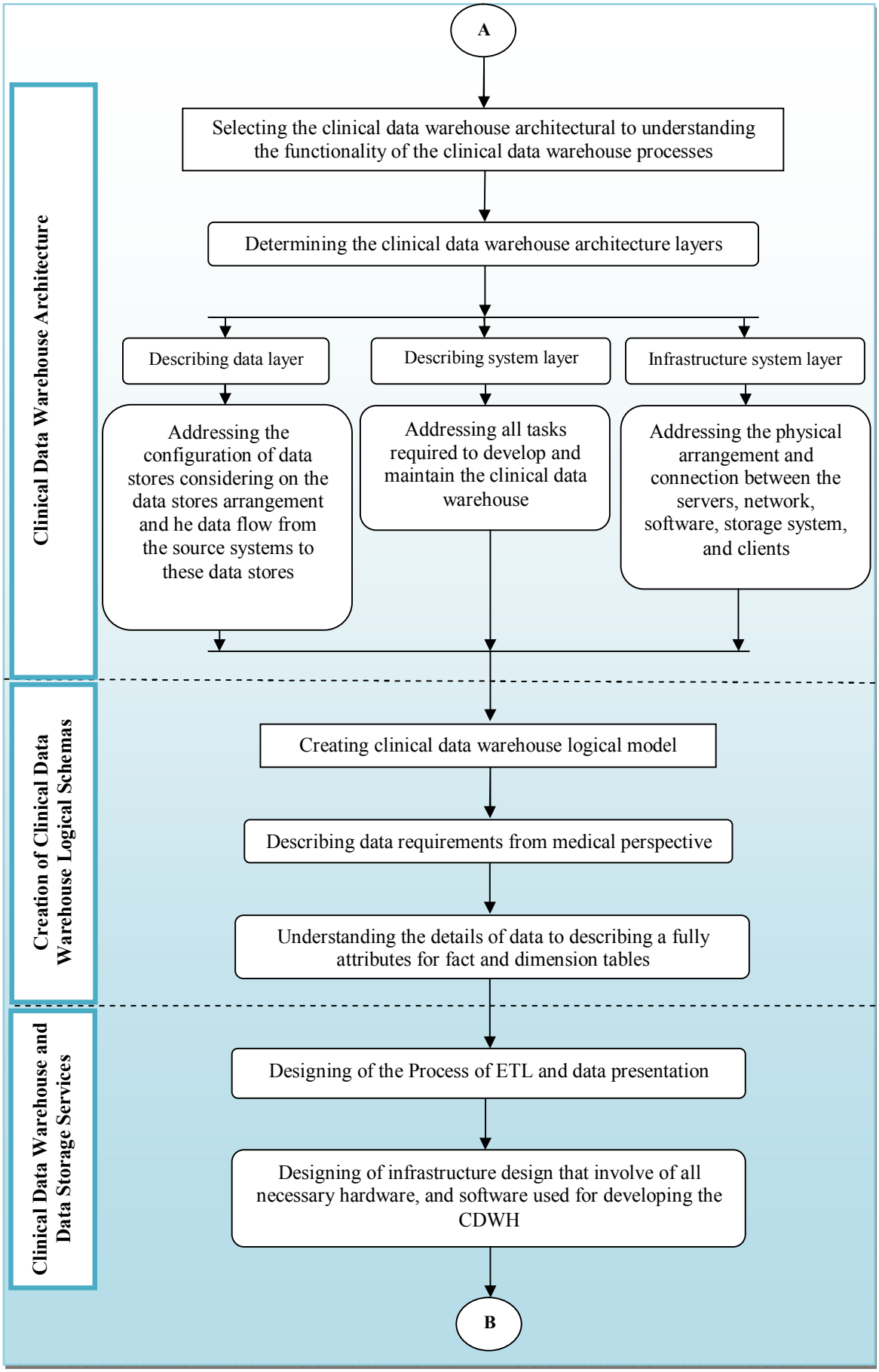
The objective of this chapter is to design and develop CDWH technique to integrate clinical data according to the medical purpose. The remaining of this chapter is organized as follows. Section 5.1 presents the tasks of the proposed CDWH technology. Section 5.2 discusses the medical analysis process. In section 5.3 the CDWH architectural is selected, while section 5.4 performed the processes of creation CDWH logical Schemes. In section 5.5 the processes of population CDWH and data storage services are done. The physical development of CDWH is developed in section 5.6, while section 5.7 and section 5.8 discusses the presentation of the information and evolution of CDWH. The summary of this chapter concludes in section 5.9.

5.1 Clinical Data WareHouse (CDWH)

A massive amount of clinical data and other related data produced by various clinical processes and procedures are generated daily. This clinical data owned by different hospitals and departments are stored in various medical operational systems such as HIS (Hospital Information System), RIS (Radiology Information System), PACS (Picture Archiving and Communications System) and etc. However, these medical data must be integrated to centralized repository in order to provide data analysis and support medical decision making. This chapter presents and discusses the proposed CDWH Design and Development technique base on a life cycle development methodology proposed by Kimball [43] .

The proposed CDWH technology consists of a set of tasks, including the medical analysis, CDWH architectural selection, CDWH schemes creation, population CDWH and data storage services, physical development, and CDWH evaluation, as shown in figure 5.1.


Start
Studying and analyzing the existing process from medical perspective to determining project objectives, requirements, constrains and acceptance criteria of clinical data warehouse
Determining requirements for integrating data before store data in clinical data warehouse
Introduce medical background
Determining data analysis requirements
Determining system development analysis requirements
Introduce initial dimension model
Describing fact tables
Describing dimension tables
Validating the model to identifying the data sources of the required data
Designing of clinical data warehouse conceptual data model
A
Medical Analysis Processes


A
Clinical Data Warehouse Architecture
Selecting the clinical data warehouse architectural to understanding the functionality of the clinical data warehouse processes
Determining the clinical data warehouse architecture layers
Describing data layer
Describing system layer
Infrastructure system layer
Addressing the configuration of data stores considering on the data stores arrangement and he data flow from the source systems to these data stores
Addressing all tasks required to develop and maintain the clinical data warehouse
Addressing the physical arrangement and connection between the servers, network, software, storage system, and clients
Creation of Clinical Data Warehouse Logical Schemas
Creating clinical data warehouse logical model
Describing data requirements from medical perspective
Understanding the details of data to describing a fully attributes for fact and dimension tables
Clinical Data Warehouse and Data Storage Services
Designing of the Process of ETL and data presentation
Designing of infrastructure design that involve of all necessary hardware, and software used for developing the CDWH
B

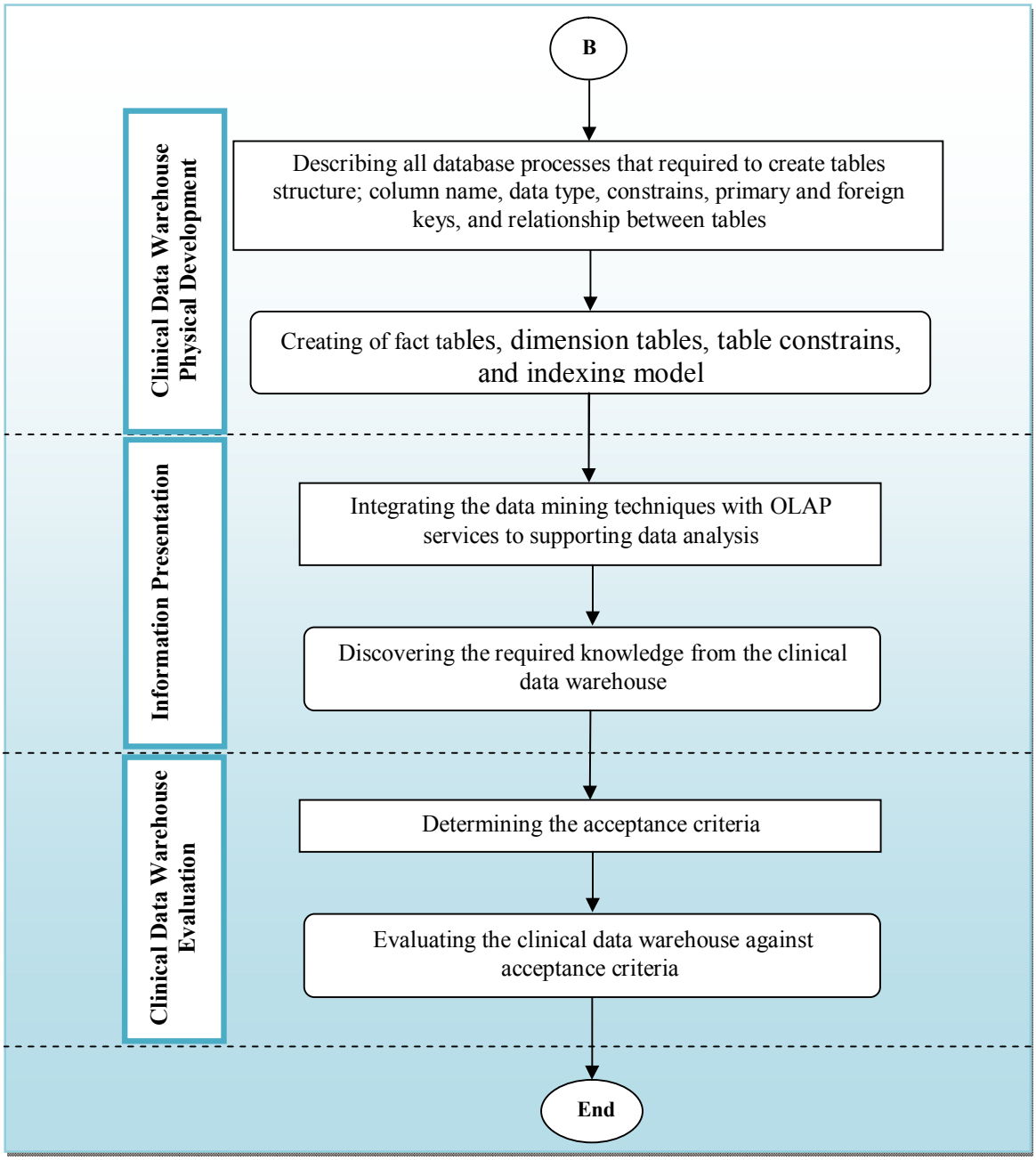


Figure (5.1): Flowchart of CDWH Developing Technique

5.2 Medical Analysis

Medical analysis is identifying medical purpose and determining the solutions to medical problems. Where, the CDWH does not achieve its objectives without clearly defining the medical purpose. Furthermore, the discussion of the medical analysis phases are significant to study and analyze the existing process from medical perspective to determine project objectives, requirements, constrains and acceptance criteria.

The CDWH design and development must meet some functional requirements in order to maintain data integration in CDWH. These requirements include; understanding the medical purpose, requirements and constraints, determining medical objectives, needs and rules, identifying the data sources of the required data, and performing the sizing of the model, and identifying the suitable model that supports data analysis. Therefore, the medical analysis is composed of four phases which, involved requirements gathering, requirement Analysis, validation of model, and requirements modeling

5.2.1 Requirement Gathering

Requirement gathering process is concerned with the understanding of the medical needs and data requirements of the CDWH. During requirements gathering stage, the requirements are collected and documented in order to build a successful CDWH technique. Furthermore, requirements gathering process are very much oriented towards understanding the problem domain for which the modeling will be done. The literature study for the CDWH and analysis of the existing situation in the Radiation and Isotopes Centre-Khartoum and Radiation and Isotopes Hospital –Shendi revealed a set of functional requirements of datasets for the CDWH. Additionally, the system development analysis requirements are determined for the purpose of integrating data

before storage into the CDWH to provide more effective analysis environment.

5.2.1.1 Dataset

Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues. Cancer is not just one disease but many diseases. There are more than 100 different types of cancer; the most cancers are named for the organ or type of cells in which they started [13].

The most common cancer types include: Breast cancer, bladder cancer, Lung cancer, melanoma, colon and rectal cancer, endometrial cancer, pancreatic cancer, kidney cancer, prostate cancer, leukemia and thyroid cancer. Table 5.1 illustrates the important information about the most common cancer types.

Table (5.1): The Common Cancer Types

	Risk Factors	**Diagnosis**	**Treatment types**
Bladder cancer	Smoking, causing chemicals in their workplace, Personal history, Family history, Certain cancer treatments, Arsenic.	Urine tests, Cystoscopy, Biopsy.	Surgery, chemotherapy, biological therapy, and radiation therapy.
Breast cancer	Age, Personal history, Family history, Certain genome changes, Radiation therapy to the chest, Reproductive and menstrual history, Breast density, History of taking DES, Being overweight or obese after menopause, Lack of physical activity, Drinking alcohol.	• Clinical Exam • Mammogram • Biopsy • Lab Tests with Tissue	surgery, radiation therapy, hormone therapy, chemotherapy and targeted therapy
Leukemia	Radiation, Smoking, Benzene, Chemotherapy, Down syndrome and certain other inherited diseases, Myelodysplastic syndrome and certain other blood disorders, Human T-cell leukemia virus type I (HTLV-I), Family history of leukemia.	Physical exam, Blood tests, Biopsy, Cytogenetics, Spinal tap, Chest x-ray.	watchful waiting, chemotherapy, targeted therapy, biological therapy, radiation therapy, and stem cell transplant

Colon Cancer	Age over 50, Colorectal polyps, Family history of colorectal cancer, Genetic alterations, Personal history of cancer, Ulcerative colitis or Crohn disease, Diet, Cigarette smoking.	Urine tests, Blood tests, Ultrasound, CT scan, MRI, IVP, and Biopsy.	Surgery, chemotherapy, biological therapy or radiation therapy.
Endometrial Cancer	Abnormal overgrowth of the endometrium, Obesity, Reproductive and menstrual history, History of taking estrogen alone, History of taking tamoxifen, History of having radiation therapy to the pelvis, Family health history.	Pelvic exam, Ultrasound, Biopsy.	Surgery, radiation therapy, chemotherapy, and hormone therapy.
Kidney Cancer	Smoking, Obesity, High blood pressure, Family history of kidney cancer.	Urine tests, Blood tests, Ultrasound, CT scan, MRI, IVP, and Biopsy.	Surgery, targeted therapy, and biological therapy.
Lung Cancer	Smoke, Radon, Asbestos and other substances, Air pollution, Family history of lung cancer, Personal history of lung cancer, Age over 65.	Physical exam, Chest x-ray, CT scan, Sputum cytology, Thoracentesis, Bronchoscopy, Thoracoscopy, Thoracotomy, Mediastinoscopy.	Surgery, chemotherapy, radiation therapy, targeted therapy.
Melanoma	Sunlight, Lifetime sun exposure, Tanning, Sunlamps and tanning booths, Personal history, Family history, Skin that burns easily, Certain medical conditions or medicines, Dysplastic nevus, A dysplastic nevus is more likely than a common mole to turn into cancer.	Biopsy	chemotherapy, photodynamic therapy, or radiation therapy, biological therapy
Pancreatic Cancer	Smoking, Diabetes, Family history, Inflammation of the pancreas, Obesity.	Physical exam, CT scan, Ultrasound, MRI, PET scan, Needle biopsy.	Surgery, chemotherapy, targeted therapy, and radiation therapy.
Prostate Cancer	Age over 65, Family history, Certain prostate changes, Certain genome changes.	Digital rectal exam, Blood test, Transrectal ultrasound, Transrectal biopsy.	Surveillance, surgery, radiation therapy, hormone therapy, and chemotherapy.
Thyroid Cancer	Radiation, Family history of medullary thyroid cancer, A blood test can detect the changed RET gene, Family history of goiters or colon growths, Personal history, Female, Age over 45, Iodine.	Physical exam, Blood tests, Ultrasound, Thyroid scan, Biopsy, Ultrasound, CT scan, MRI, Chest x-ray, Whole body scan.	Surgery, thyroid hormone treatment, radioactive iodine therapy, external radiation therapy, or chemotherapy.

The CDWH development depends critically on the identifying the required clinical data and identifying the medical purpose. Furthermore the selection of the relevant data from sources and integrate them into CDWH is represented as an important issues. There are several CDWH systems developed to improve the quality of care. Souad Demigha In [110] proposed a DWH system to help assist breast cancer screening in diagnosis, education and research. Where, the dataset collected from multiple and different sources: the BI-RADS (Breast Imaging Reporting and Data System) dictionary, on scientific reports of the EBM (evidence-based in medicine), reports and experience for radiologists-senologists of the Necker Hospital in Paris. The proposed DWH aims to combined breast cancer data into DWH to support decision making. While, Erhard Rahm et al, in [112] introduced the GeWare data warehouse platform for the integrated analysis of clinical information, microarray data and annotations within large biomedical research studies. The dataset is obtained from a commercial study management system while publicly available data is integrated using a mediator approach. Also, Jonathan C. Prather et al, in [121] proposed CDWH at Duke University medical center database, for mining data and discovering knowledge. Where, the dataset collected Duke University medical center systems. However, most of the previous work did use not a real dataset such as from internet, and all the required data was not available.

In this study the medical data collected during the regular day-to-day events are recorded and stored in many medical operational systems at the Radiation and Isotopes Centre Khartoum (RICK) – Sudan and Radiation and Isotopes Hospital –Shendi (RICSH) – Sudan. These medical and clinical data are collected, cleansed, transformed and loaded into CDWH in order to improve the quality of data in the CDWH to

support the decision making processes and provides powerful analysis and researches environment.

The specific RICK and RICSH databases selected for this project include; patient, physician, symptom, clinical diagnosis, treatment, laboratory, location and etc, are represents as subject domain of the require data. Figure 5.2 illustrated the important components of cancer treatment processes.

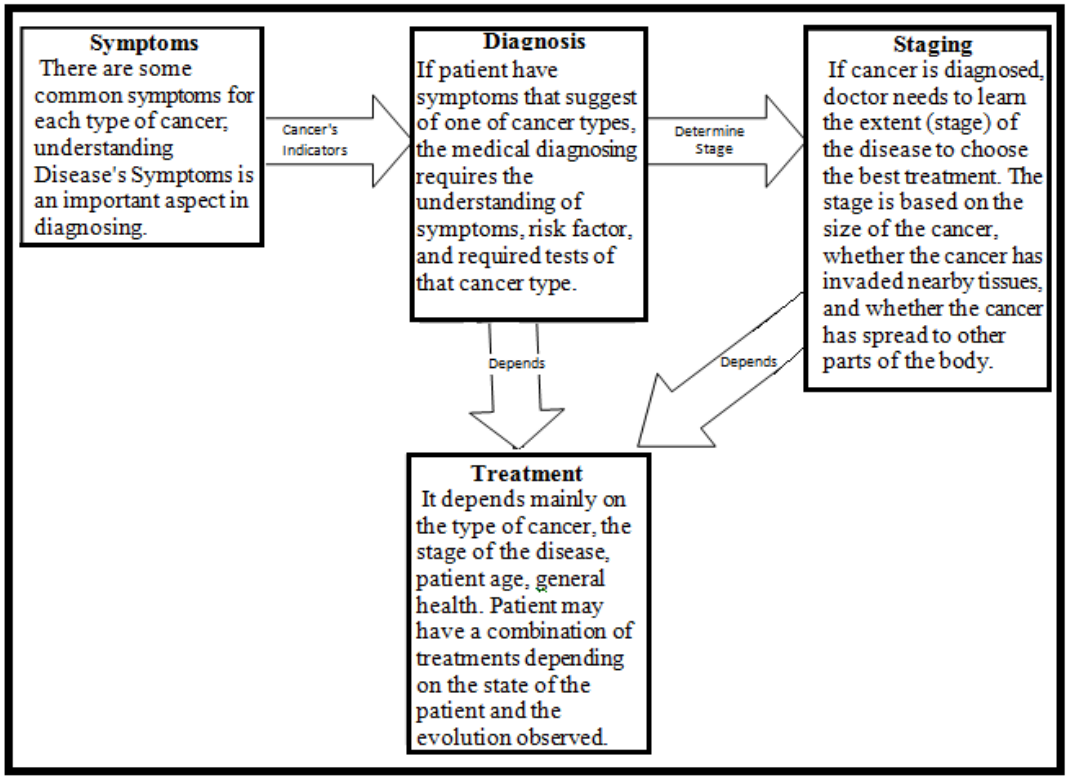


Figure (5.2): Components of Disease Treatment Processes

Additionally, the selected databases contain comprehensive data on over 30,000 unique patients collected over nearly 5 years. The dataset contains all required data that cover the clinical medical processes as shown in figure 5.3.

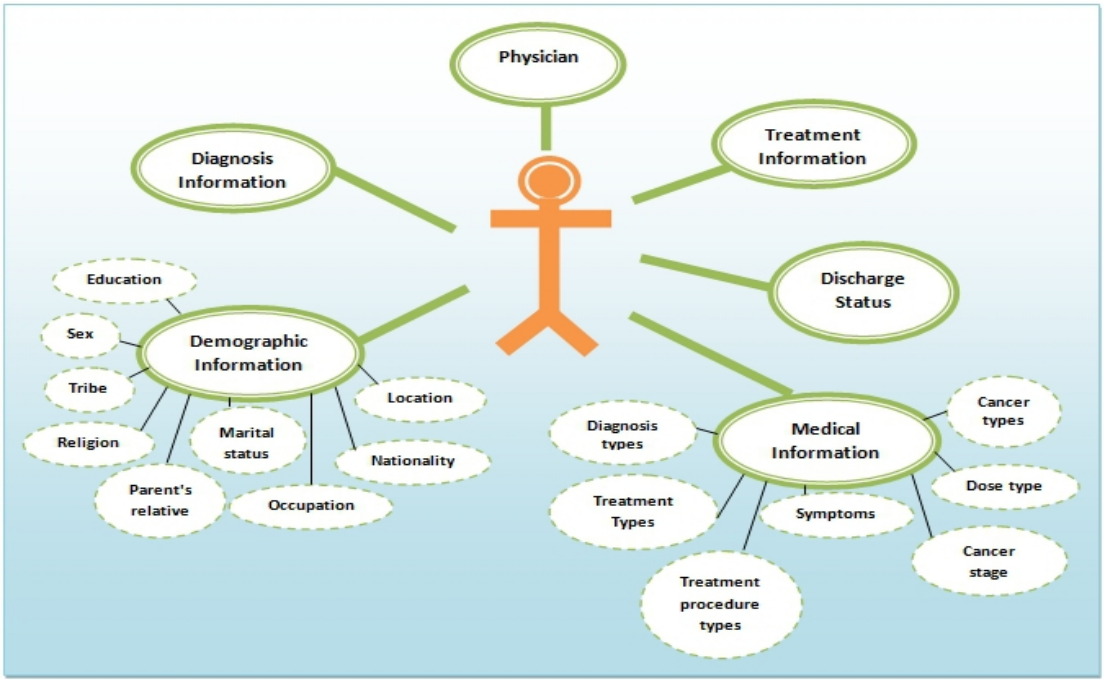


Figure (5.3): Clinical Information System Contents

Furthermore, the medical and clinical data may be divided into categories according to subject domain. These dataset includes patient demographic information, medical information, diagnostic clinical information, treatment clinical information and, laboratory clinical information, as shown in table 5.2.

Table (5.2): Data Requirements of Clinical Data Warehouse

Information Types	Description	Required Data
Demographic Information	It includes the required demographic information to be collected for every patient to provide rich data analysis environment.	Sex, Date of Birth, Tribe, Education, Occupation, Nationality, Resident Address (State/ Province/ Village), Spent Address (State/ Province/ Village), Parents Relatives, Marital Status, income Date, Birth Place (State/ Province/ Village), and Physician Name. Most of this data can be assumed to remain unchanged.
medical information	It includes the required information about patient's diagnosis process.	Cancer type code, Cancer Name, Diagnosis code, Type of diagnosis, risk factor and stage of cancer.
Diagnosis Information	It includes the extraction of the required Clinical details about patient's life habits to enhance data analysis Capabilities.	State of the patient, High risk and Major medical events.
Laboratory information	It includes the required information about tests that patient conducted.	Test name, literal test value, numeric test value, reference range, test ordering time.
Treatments Information	It includes the required information about treatments that patient received.	Cancer type code, Procedure code, type of procedure, Duration, results previous illnesses and treatments and risk factors which is different according the different types of cancer.

5.2.1.2 System Development Analysis Requirements

The existing medical database processes consist of two core entities, involves physician and patient, and five core information elements, involves symptom, clinical diagnosis, treatment, laboratory, location. Each of these entities has multiple attributes. Figure 5.4 shows the relationship between these entities.

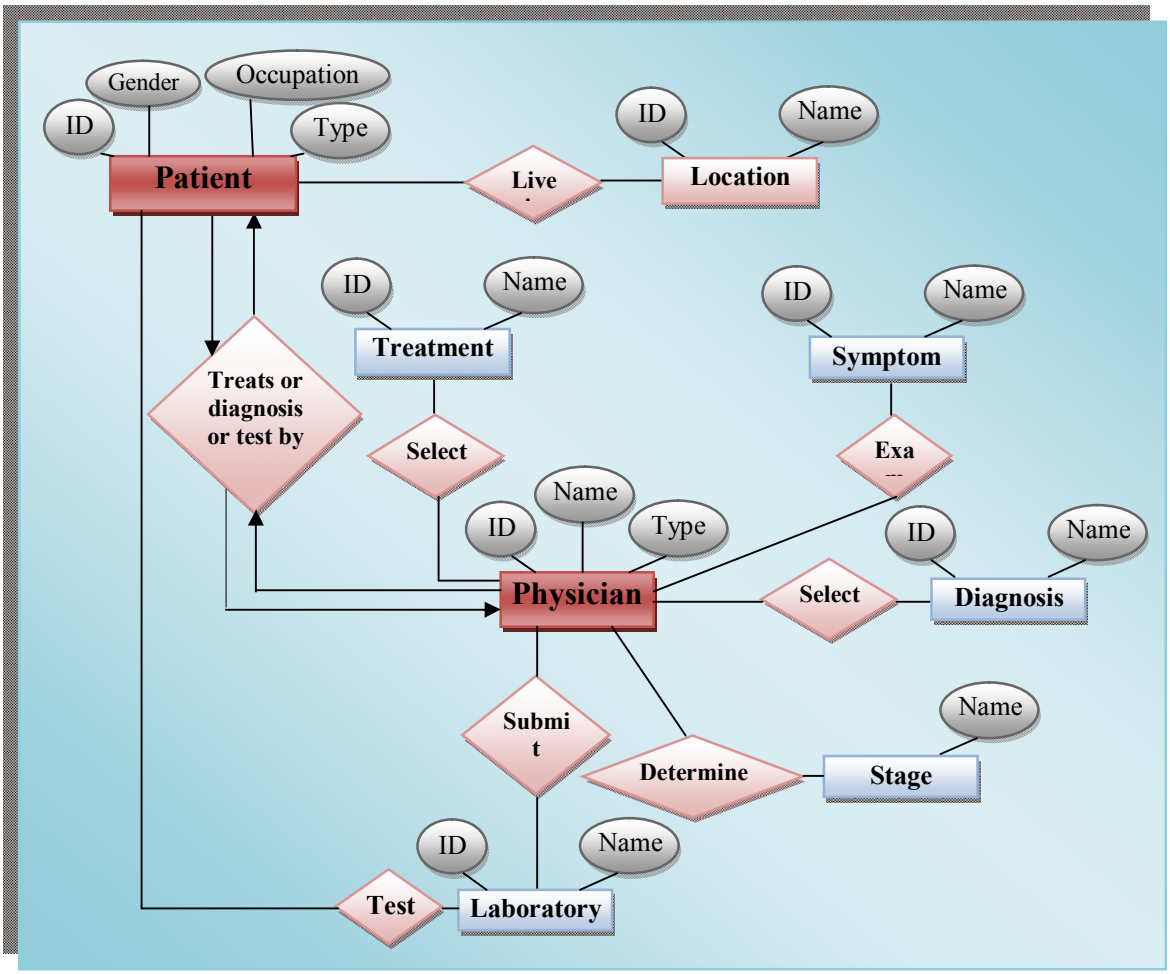


Figure (5.4): Entities of the Medical Database.

The system development must meet some specific functional and nonfunctional requirements in order to maintain the data integration in CDWH. Functional requirements determine what the system does (the features). While, the nonfunctional requirements provide guides and constraints for the system architecture, such as requirements related to

the time, resources, budgets and locations. The functional and nonfunctional requirements are demonstrate as follows:

(I) System Functional Requirements:

1. Understanding the medical purpose; the objectives, and constraints are determined from medical perspective.
2. The CDWH function requires understanding of symptoms, diagnosis, disease staging, risk factors, and treatment procedures.
3. The CDWH will store all recent and historical clinical data about patients required to support data analysis. The clinical data extraction process includes the followings:
   a) Extraction of minimal level of demographic data of the patient from the data sources that include: location types, occupation section, sex, education level, parent relatives, tribe, etc.
   b) Extraction of treatment data details that include: types, procedures, and treatment side effects.
   c) Extraction of clinical data details that include: cancer types, diagnosis, and side effects.
4. Determining the suitable model that support data analysis
5. The ability of clinical data warehouse to display data at both summary and detail levels, and answering the critical questions related to health care from various dimensions.
6. The report and OLAP will have an easy-to-use user interface.
7. The system will be able to display the information in various ways such as figures and charts.

(II) Nonfunctional Requirements:

1. Certain sensitive data are viewable by certain people only.
2. The standard minimum specification for the client is Internet Explorer 6 on Windows XP running on a PC or laptop with

Intel Pentium D 820 with 512MB memory and SVGA resolution (10245768 pixel).

3. The CDWH must be available for 24 hours a day, 7 days a week.
4. The CDWH must be backed up to offline media.
5. The CDWH needs to be flexible so we can enhance it easily and adapt it with changes that happen in the transaction systems.
6. The CDWH front-end applications accessible from anywhere.

5.2.2 Requirement Analysis

The initial dimensional model of the proposed CDWH has to be described after gathering the requirements. The dimensional modeling improves the query performance for reports without affecting data. A dimensional model consists of two types of tables having different characteristics; dimension and fact tables.

5.2.2.1 Dimension Tables

The dimension tables it contains descriptive attributes of each entity in the fact tables to represent various medical processes. The dimension tables store data that extracting from medical operation systems after the process of integrating and cleansing is done. These dimension tables are listed in table (5.3) below.

Table (5.3): CDWH Dimension Tables

No	Dimension Table Name	Dimension Table Description
1.	Patient	The patient dimension represents an important entity in the process of cancer management. This dimension stores demographic information about patient that effects cancer management process, such as Sex, Age, tribe, occupation, location and etc.
2.	Age Range	This dimension stores the age range of each patient.
3.	Sex	This dimension stores patient's gender to studying the effect of this dimension in diagnosis and treatment processes.
4.	Tribe	This dimension stores patient tribe to studying the relationship between tribe and specific type of cancer.
5.	Cancer Type	
6.	Treatment	This dimension stores all treatment options that physician selected to direct the treatment process.
7.	Treatment Procedure	This dimension stores information about the type of treatment procedure that physician selected to direct the treatment process, to studying the effect of this procedure.
8.	State	Geographical location can be able to characterize significant features, to studying the existence of specific type of cancer in specific state or to compare between two states.
9.	Province	Geographical location can be able to characterize significant features, to studying the existence of specific type of cancer in specific province or to compare between two provinces.
10.	Date	A meaningful clinical data cannot use only time points, such as dates when data were collected; it must be able to characterize significant features over periods of time.
11.	Occupation Sector	This dimension stores patient's occupation to studying of the relationship between the types of occupation sector occupation sector environment with a particular type of cancer.
12.	Occupation	This dimension stores patient's occupation to studying of the relationship between the types of occupation with a particular type of cancer.
13.	Discharge Status	This dimension stores the patient's discharge status to studying of the studying the status of patient after treated from specific type of cancer.
14	Education	This dimension stores patient education level to studying the relationship between education and the result of treatment process.
15	Stage	This dimension stores information about patient's first diagnosis stage.

Additionally, dimension hierarchy is the set of members in a dimension and their positions are related to each other. A hierarchy is useful for traversing the multidimensional structures found in Analysis Services. Therefore, designing the dimension tables with proper hierarchies will greatly facilitate the ability to prepare reports that are responsive to user needs and medical research. Figure 5.5 shows the hierarchy of the dimension.

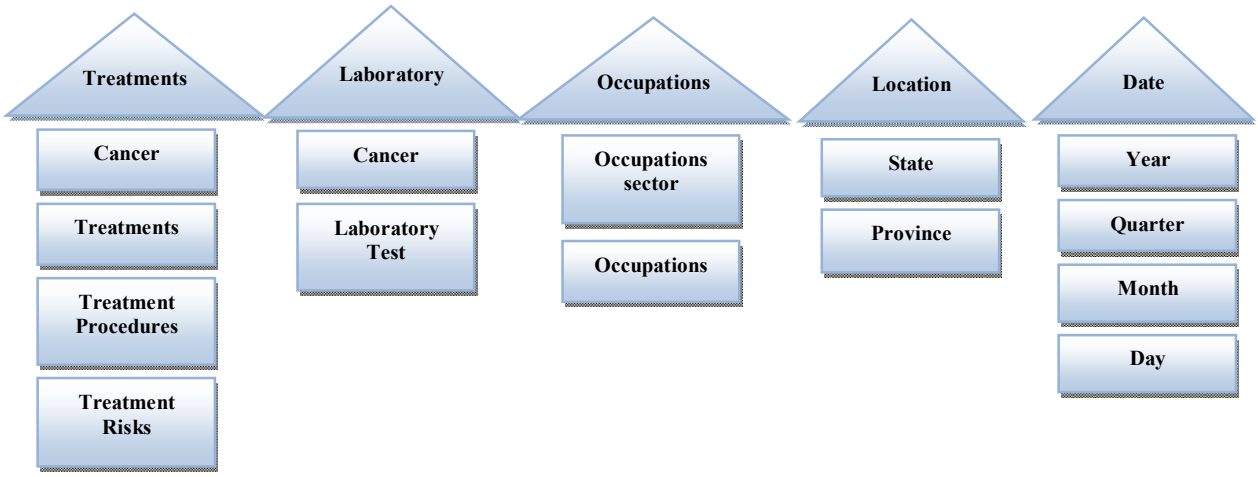


Figure (5.5): CDWH Dimension Hierarchy

5.2.2.2 Fact Table

The fact tables contain the foreign keys to associate dimensional tables to determine the effect of specified dimensions to the measures. These fact tables are listed in table 5.4.

Table (5.4): CDWH Fact Tables

No	Fact Table Name	Description
1.	**Treatment**	The treatment fact table contains related dimension foreign keys, and measurements as key performance indicators for treatment processes.
2.	**Diagnosis**	The diagnosis fact table contains related dimension foreign keys, and measurements as key performance indicators for diagnosis processes.
3.	**Discharge Status**	The Discharge status fact table contains related dimension foreign keys, and measurements as key performance indicators for discharge status.
4.	**Location**	The Location fact table contains related dimension foreign keys.

Additionally, the fact tables contain the measurements as key performance indicators for management decision-making. Measures designed to facilitate drill-down analysis that increases understanding of the operations of the clinical processes. Table 5.5, listed these measures.

Table (5.5): List of Measures

Fact Table	Measures	Medical Rules
Treatment Fact Table	Count total patients treat by special treatment	For all treated patients
	Count total patients treat by special treatment procedure	For all treated patients
Diagnosis Fact Table	Count Total patients diagnosis by special cancer type.	For all diagnosed patients
	Second_Cancer_disease: determined that a specific cancer patient have other cancer diseases;	For all diagnosed patients
	CancerStage: Determine the cancer stage	For all diagnosed patients
	Other _disease: determined that a specific cancer patient have other diseases;	For all diagnosed patients
Discharge Status Fact Table	Status of accident (included improved, not improved, recovery, and died): determined the status of the patient when discharged;	For all discharged patients
	Count total patients treat by other treatment.	For all discharged patients
Location Fact Table	Geographical location of patients (States and provinces): determined the patients' location.	For all patients

5.2.3. Validation

The data identifying and gathering processes must meet specific requirements to improve data analysis processes in CDWH in order to support decision making. The proper data selection required clear understands of the purpose of the CDWH. Validation activity is responsible of identifying the data sources of the required data. The main

activities that are performed as a part of the requirement validation include: analyzed the initial models, and identified the data sources. The data sources are Radiation and Isotopes Centre Khartoum (RICK) and Radiation and Isotopes Hospital -Shendi (RIHSH) databases. Table 5.6, illustrates the sources of required data in (RICK) and (RIHSH).

Table (5.6): Data Sources of the Required Data

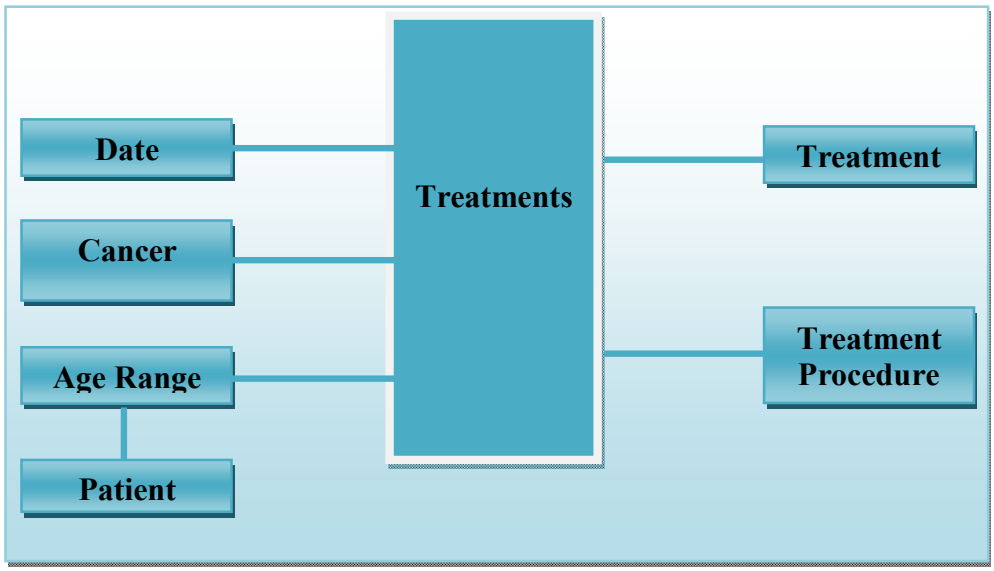
Type of Data	Source Systems of Required Data		Available Dates	
	RICK	RIHSH	RICK	RIHSH
Demographic Information	Statistics System	Statistics System	2009-2012	2010-2013
Clinical Information	Clinical System	Clinical System	2009-2012	2010-2013
Diagnosis information	Clinical System	Clinical System	2009-2012	2010-2013
treatments Information	Clinical System	Clinical System	2009-2012	2010-2013
Laboratory results	Lab System	Lab System	2009-2012	2010-2013

5.2.4. Requirements Modeling

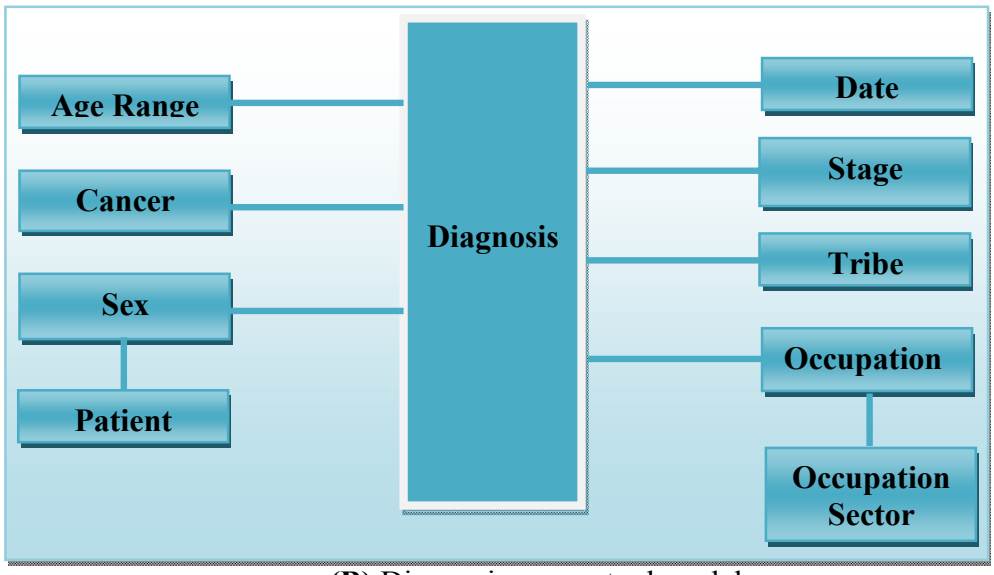
The designing of CDWH conceptual data model based on medical data fields. Whereas, a CDWH organized around a cancer patient's data for the purpose of medical researching, and improving the medical decision support. A conceptual data model describes an entire enterprise, whereas conceptual data model is enabling to understand the medical process at high level by address some questions such as: what the different entities in data are? And how they relate to one another? In this connection, establishment of the data designing such as data modeling, which facilitate measurements of the effectiveness of treatment, and the relationships between clinical processes and conditions. The characteristics of conceptual data model include: (1) design and develop primarily for a medical audience, (2) provide enterprise-wide coverage of the medical concepts, such as physician, patient, symptoms, clinical diagnosis, treatment, laboratory, location and etc, (3) contains relationships between entities, (4) entities will have definitions, and (5)

designed and developed to be independent of DBMS, data storage locations or technologies.

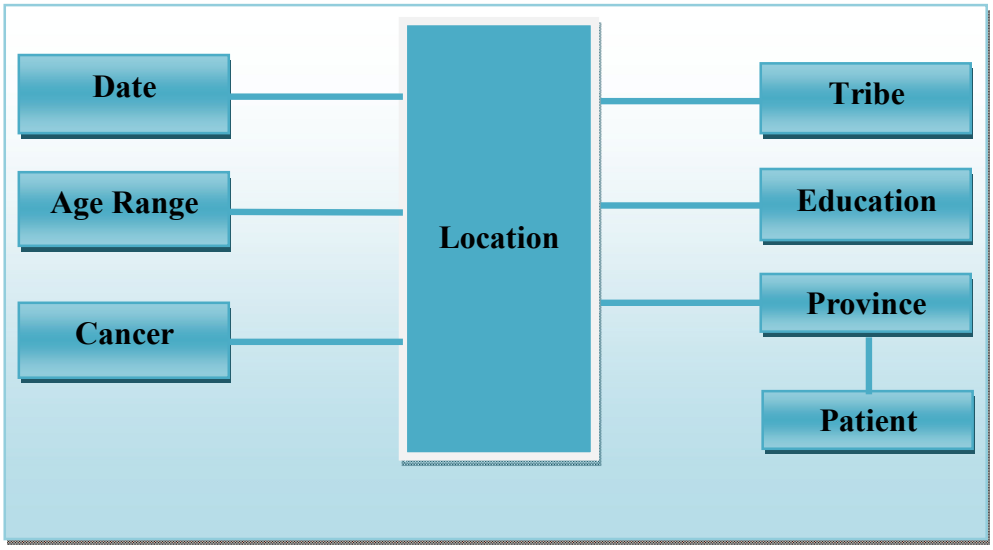
The proposed conceptual model is snowflake schema; it is the most appropriate schema for CDWH technique. Because, the snowflake schema provides many benefits such as: maximizing query performance in the CDWH. Furthermore, the snowflake schema has less data redundancy and clearer relationships between dimension levels than a star schema. Figure 5.6 shows the design of the conceptual model for the proposed CDWH.



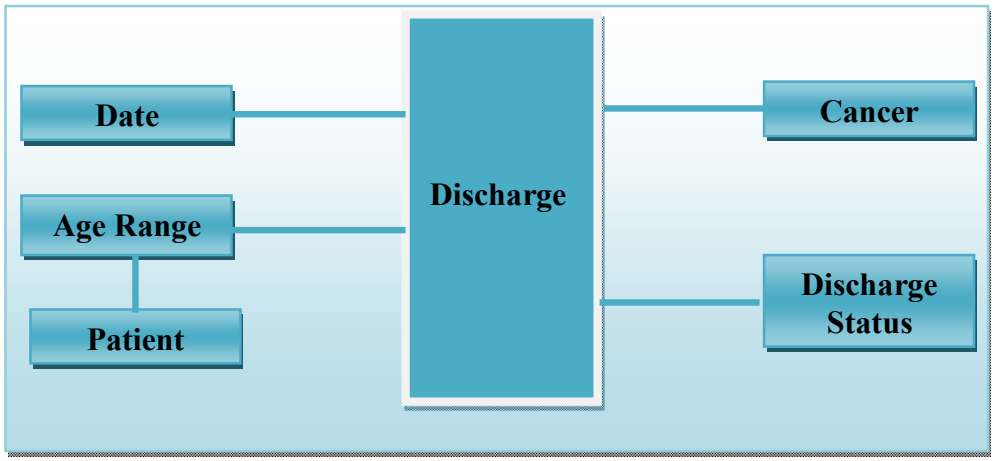
(A) Treatment conceptual model



(B) Diagnosis conceptual model



(C) Location conceptual model



(D) Discharge conceptual model

Figure (5.6): CDWH Conceptual Data Model

5.3. CDWH Architectural Building

The CDWH architecture presents for understanding the functionality required to successfully implementation of CDWH processes, independence of the manner in which the CDWH is developing. Figure 5.7, present the CDWH architecture Diagram.

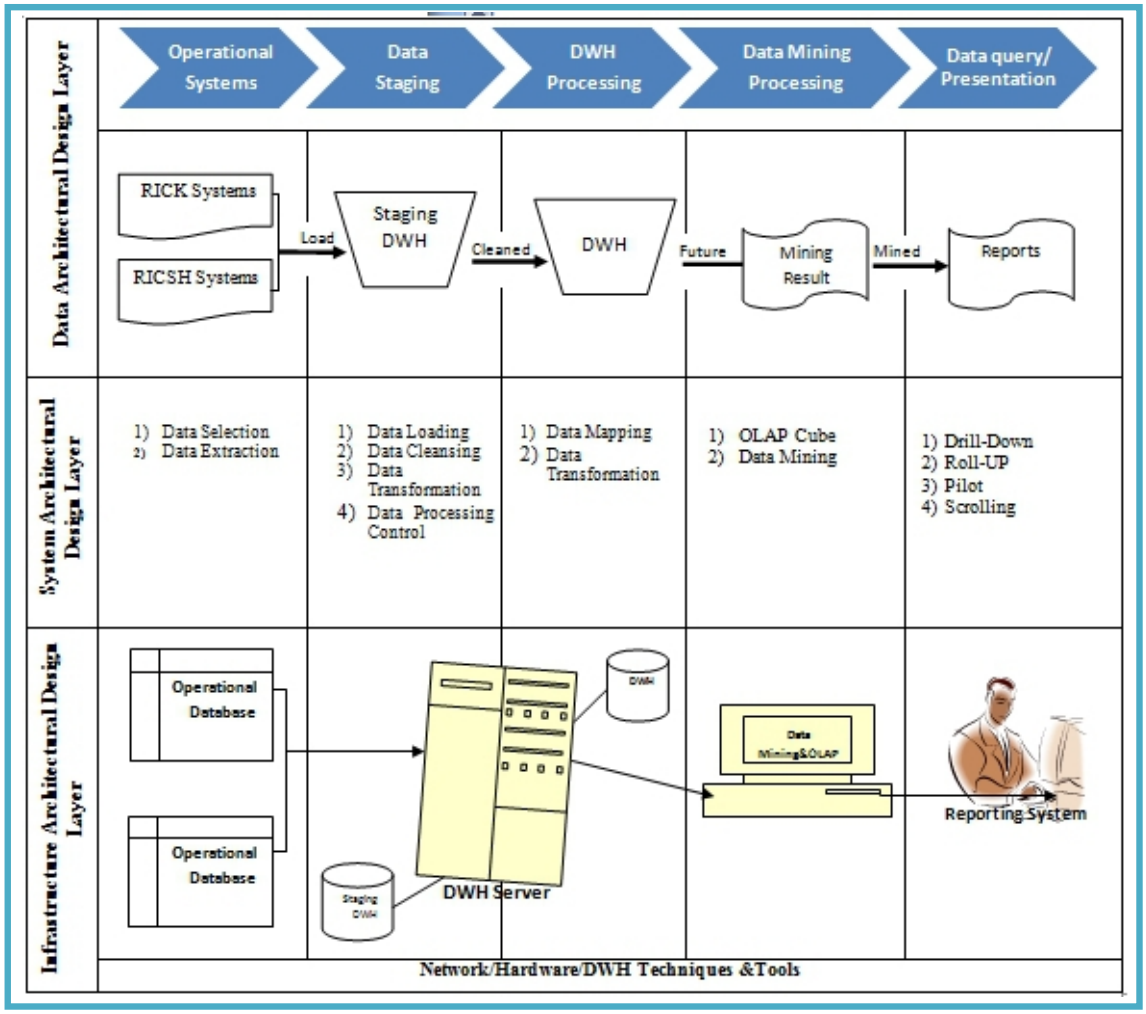


Figure (5.7): CDWH Architecture Diagram

The DWH architecture demonstrates into three layers. The architecture layers include: data architectural layer, system architectural layer, and infrastructure architectural Layer, figure 5.8 illustrates DWH architecture layers.

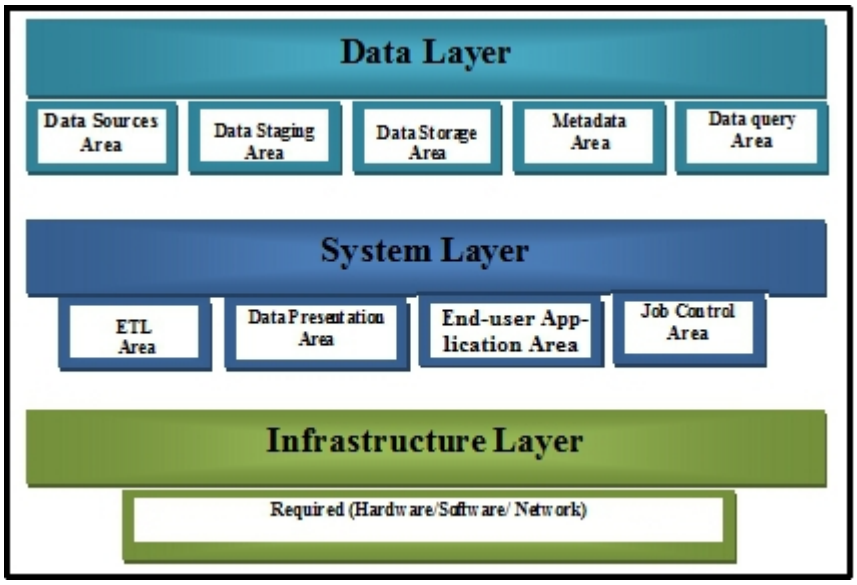


Figure (5.8): Data Warehouse Architecture Layers

5.3.1. Data Layer

The data layer addresses the configuration of data stores within a DWH system. The data layer is discussing the questions about how the data stores are arranged within a CDWH and how the data flow from the source systems to the users through these data stores. The data flow architecture designing based on the data requirements from the medical processes, including the data quality requirements. Therefore, the data layer divide into 5 areas, include: data sources area, data staging area, data storage area, metadata area, and data query area:

(I) Data Sources Systems Area: It represents the different operational source systems that feed relevant data into the DWH. The operational Source systems include legacy systems, online transaction processing and transaction processing systems. The data in data sources can be of any different format such as plain text file, XML file, relational database, other types of database, etc.

(II) Data Staging Area: is an internal data store used for transforming and preparing the data obtained from the source systems, before the data is loaded to target data source (CDWH). Data staging is

considered to be more crucial stage of DWH process where, most of the data cleansing and transformation of data are done. Furthermore, data staging includes all processes necessary to extract, cleanse, and transform before they are stored permanently in the CDWH. As results, the goals of data staging is having one common area makes it easier for subsequent data processing, integration and quality of data to ensure that data loaded into CDWH are of certain qualities needed.

(III) Data Storage Area: The transformed and cleansed data is loaded from staging area in to data storage area. Data Storage stores all data that required for the analysis process. Base on DWH scope and functionality, there are three types of components which can be found in the data storage area: DWH, data mart, and Operational Data Store (ODS). The DWH system may have just one of the three, two of the three, or all the three types.

(IV) Metadata (Directory) Area: Metadata is a data store containing the description of the structure, data, and processes within the DWH. This includes the data definitions and mapping, the data structure of each data store, the data structure of the source systems, the descriptions of each ETL process, the description of data quality rules, and a log of all processes and activities in the data warehouse. Furthermore, metadata supports the information needs of system developers, administrators, users, and applications on the data warehouse. The metadata life cycle activities include: identify and capture metadata in a central repository, establish processes to synchronize metadata with the changing data structure, and provide metadata to users in the right form and with the right tools.

(V) Data Query Area: Data query area is a user-facing data store, where the data are arranged in dimensional format for the purpose of supporting analytical queries.

5.3.2. System Layer

The system layer addresses all tasks that must be accomplished to develop and maintain the DWH. System layer includes information on how the data warehouse system operates, such as the technologies and tools that will be needed to perform the processes for analytical, integrate data, cleanse, and transform data. These system processes include extract, transform and load data area (ETL), data presentation area, and job control:

(I) ETL Area: The ETL area is where data gain its "intelligence", as logic is applied to transform the data from a transactional nature to an analytical nature. Furthermore, ETL techniques are used in this area.

(II) Data Presentation Area: The data presentation area refers to the information that reaches the users. The objectives of data presentation area is to present flexible visualization of the data analysis results and reformat and present data in a form of a tabular or graphical report. Additionally, an emailed report that gets automatically generated and sent or an alert that warns users of exceptions, among others.

(III) Job Control Area: The job control area refers to information about job definition, job scheduling (time and event), monitoring, logging, exception handling, error handling, and notification.

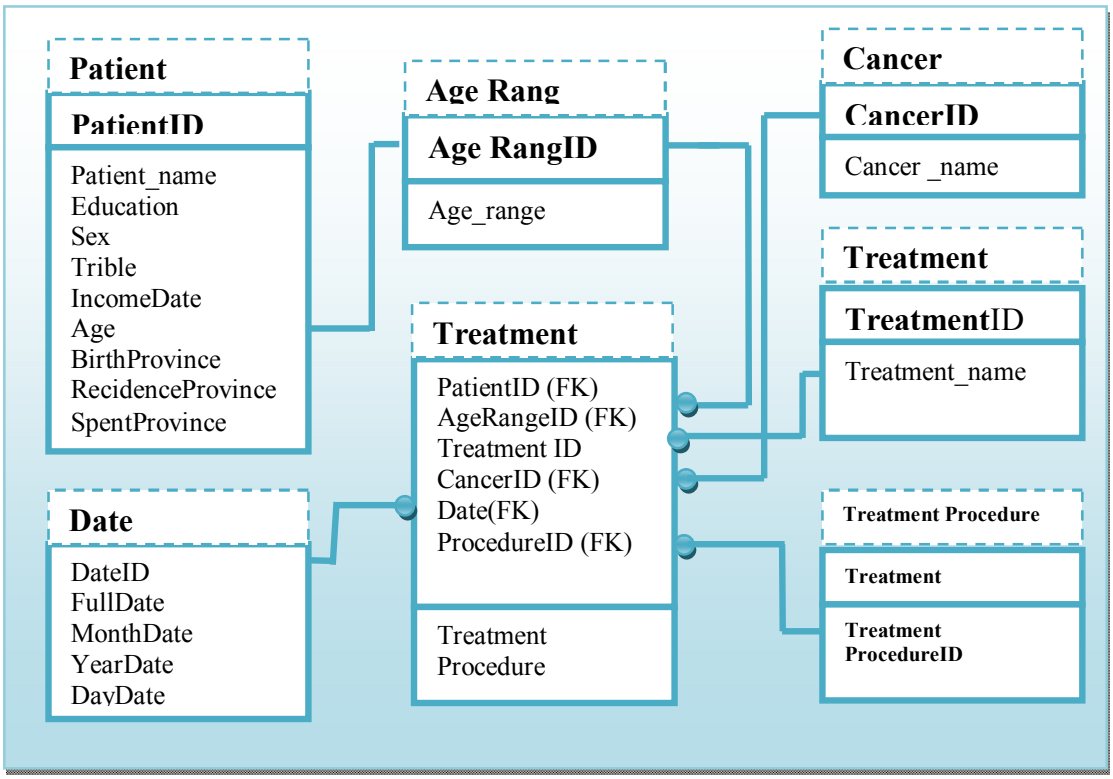
5.3.3. Infrastructure Layer

The infrastructure layer discusses the physical configuration of the servers, network, software, storage, and clients. The design of CDWH infrastructure layer depends on the requirements of the data layer and system layer. Therefore, infrastructure layer is addressing the physical arrangement and connections between the servers, network, software, storage system, and clients. Furthermore, the design of the infrastructure layer required a set of hardware, networking and software components that used to better analyzing the massive amounts of clinical data to make better medical decisions, and medical research. This layer required knowledge about hardware (especially servers), networking (especially checking the data sources, the staging area, and everything in between to ensure that there's enough bandwidth to move data around), and software (especially choice of programming language to building ETL techniques and software to building the DWH such as specific versions of SQL Server services).

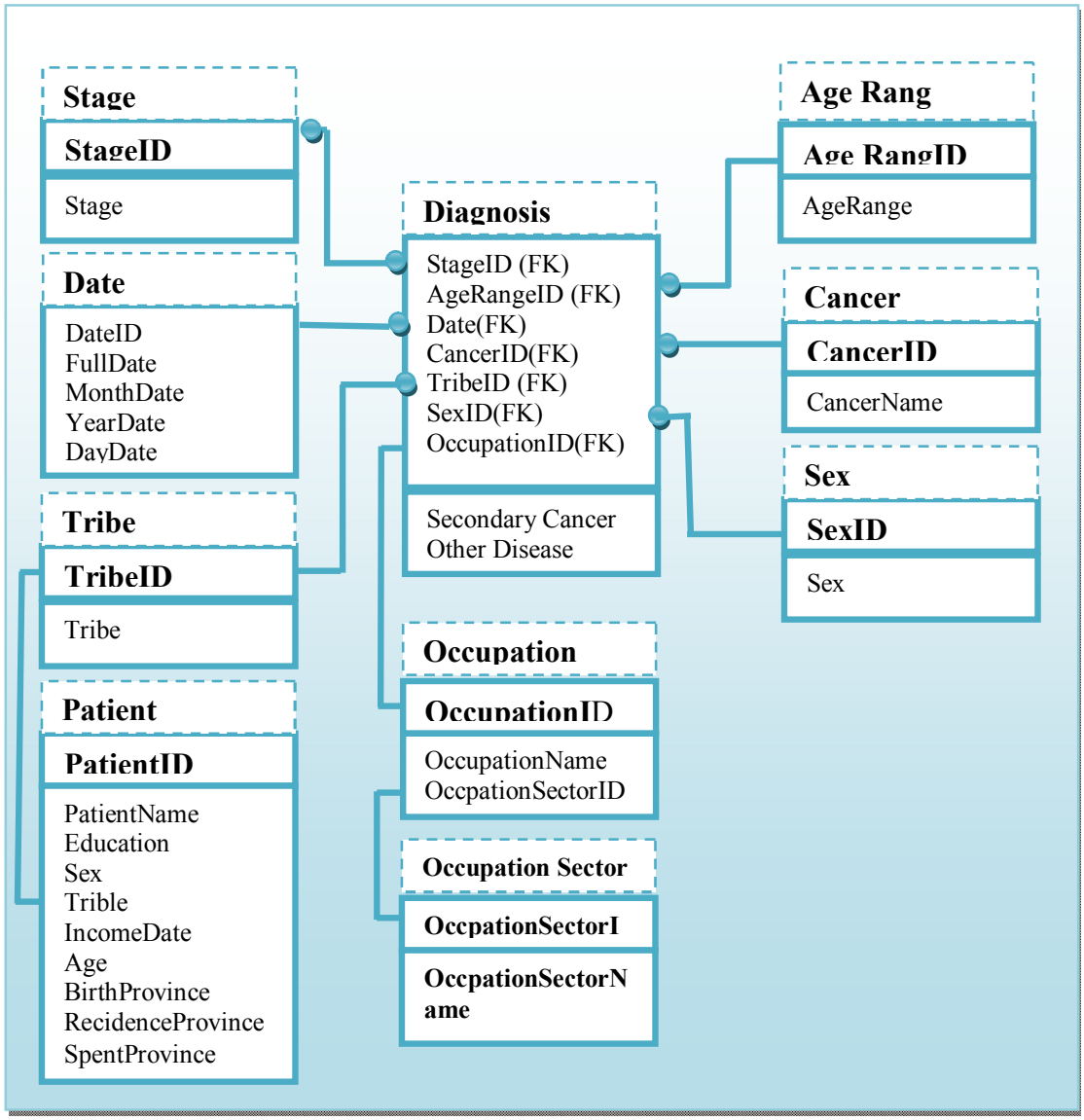
5.4. Creation CDWH Logical Schemes

A logical data model is a fully-attributed data model which is independent from DBMS, technology, data storage or healthcare constraints. The design of logical model based on conceptual model, it describes data requirements from the medical perspective. Furthermore, logical data model enable understanding of the details of data without worrying about how they will actually implement. The characteristics of logical data model include: (1) describing data requirements for a CDWH, (2) containing many entities, although these numbers are highly variable depending on the scope of the data model, (3) containing relationships between entities, (4) designing and developing to be independent of DBMS, data storage locations or technologies, (5)

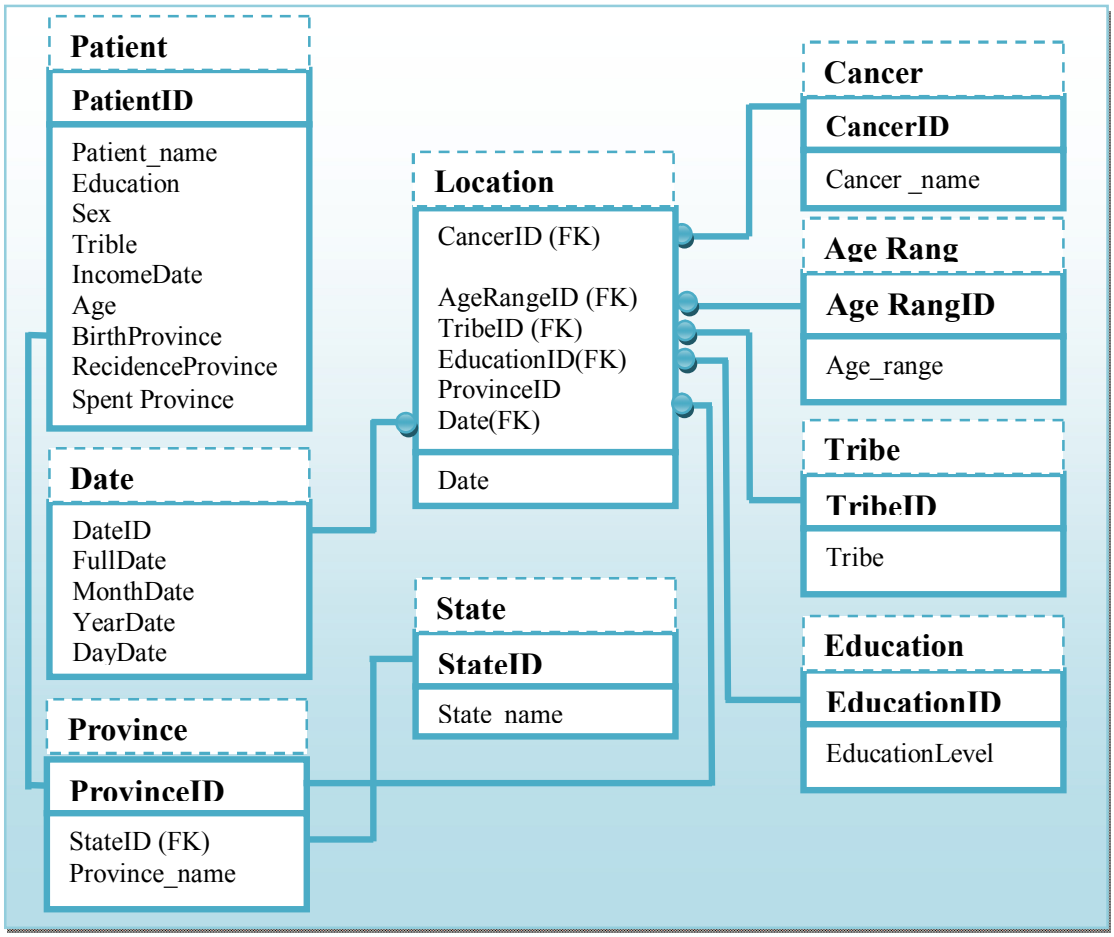
entities and attributes are specified, and (6) the primary key for each entity is specified, (7) foreign keys (keys identifying the relationship between different entities) are specified, and (8) normalization occurs at this level. Figure 5.9, shows the proposed logical data model for the proposed CDWH.



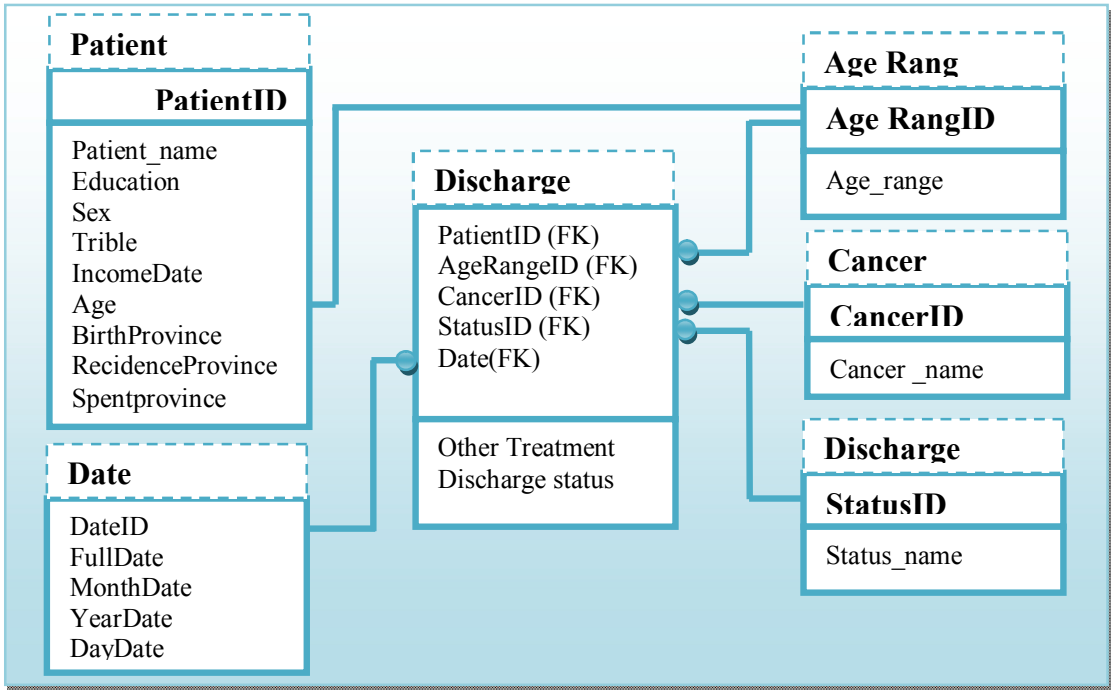
(A) Treatment Logical model



(B) Diagnosis Logical model



(C) Location Logical model



(D) Discharge Logical model

Figure (5.9): CDWH Logical Data Model

The proposed logical data model that consists of fact table is located in the centre of the CDWH and contains foreign keys for all dimension tables. The logical data model of CDWH includes three fact tables and twenty dimension tables. Each fact table contains measures and foreign Key that connected fact table with dimension tables. While, the special set of dimension tables linked to the fact table through a primary key, and describing the context of interpretation of facts.

5.4.1. Dimension Tables

The dimension tables contain descriptive attributes of each entity in the fact tables to represent various medical processes. Each of the dimension tables has a primary key; the primary key contains a uniquely value identified each member of that dimensions to assure uniqueness of these values, and to increase querying efficiency**.** The proposed dimension tables are shown in the following tables.

1. Patient Dimension: Patient dimension represents an important entity in the process of cancer management. This dimension contains description information of the patient as shown in Table 5.7. The patient dimension stores demographic information about patient that affects cancer management process.

Table (5.7): Patient Dimension

Attributes	Description	Data type(length)	Remark
PatientID	Identification number for patient	numeric (10,0)	PK
PhysicianID	Identification number for physician	numeric (10,0)	FK
EducationID	The patient's education level	numeric (10,0)	FK
SexID	The patient's gender	numeric (10,0)	FK
TribeID	The patient's tribe	numeric (10,0)	FK
IncomeDate	The first visit to the hospital	Date	
Age	The patient's age	numeric (10,0)	
BirthProvince	The patient's province where birth	numeric (10,0)	FK
RecidenceProvince	The patient's province where residence	numeric (10,0)	FK
SpentProvince	The patient's province where spent	numeric (10,0)	FK
ReligionID	The patient's religion	numeric (10,0)	FK
OccupationID	The patient's occupation	numeric (10,0)	FK
ParentRelativeID	The patient's parent relative	numeric (10,0)	FK

2. Age range Dimension: The age range dimension stores the age range of each patient. The age range dimension enable to studying the affect of age factor with special caner type. Table 5.8 shows age range dimension attributes.

Table (5.8): Age Range Dimension

Attributes	Description	Data type(length)	Remark
RangeID	A number used to identify the range of age for the patients	numeric (10,0)	PK
AgeRange	The range of age	varchar (50)	
StartRange	The smallest age	numeric (10,0)	
EndRange	The Bigest age	numeric (10,0)	

3. Occupation Sector Dimension: The occupation sector dimension stores patient's occupation field as shown in table 5.9; to studying of the relationship between the types of occupation sector environment with a particular type of cancer.

Table (5.9): Occupation Sector Dimension

Attributes	Description	Data type(length)	Remark
OccupationSectorID	A unique number for each occupation sector	numeric (10,0)	PK
OccupationSector _name	The name of the occupation sector	varchar (50)	

4. Occupation Dimension: The occupation dimension stores patient's occupation to studying of the relationship between the types of occupation with a particular type of cancer. Table 5.10 shows occupation dimension attributes.

Table (5.10): Occupation Dimension

Attributes	Description	Data type(length)	Remark
OccupationSectorID	A unique number for each occupation sector	numeric (10,0)	FK
OccupationID	A unique number for each occupation	numeric (10,0)	PK
OccupationName	The name of the occupation	varchar (50)	

5. State Dimension: Geographical location can be able to characterize significant features, to studying the existence of specific type of cancer in specific state or to compare between two states. Table 5.11 shows state dimension attributes.

Table (5.11): State Dimension

Attributes	Description	Data type(length)	Remark
StateID	A unique identifier for the patient state	numeric (10,0)	PK
StateName	The state of patient	varchar (50)	

6. Province Dimension: The province of patient can be able to characterize significant features, to studying the existence of specific type of cancer in specific province or to compare between two provinces. Table 5.12 shows province dimension attributes.

Table (5.12): Province Dimension

Attributes	Description	Data type(length)	Remark
StateID	A unique identifier for the patient state	numeric (10,0)	FK
ProvinceID	A unique identifier for the patient province	numeric (10,0)	PK
ProvinceName	The province of patient	varchar (50)	

7. Cancer Dimension: The cancer dimension contains description information about the type of cancer, as shown in Table 5.13.

Table (5.13): Cancer Dimension

Attributes	Description	Data type(length)	Remark
CancerID	A unique identifier for the cancer type	numeric (10,0)	PK
CancerName	The name of cancer	varchar (100)	

8. Treatment Dimension: This dimension stores all treatment options that physician selected to direct the treatment process. The selection of suitable treatment is very important because every cancer type requires a specific treatment. Table 5.14 shows treatment dimension attributes.

Table (5.14): Treatment Dimension

Attributes	Description	Data type(length)	Remark
TreatmentID	A unique number for each Treatment	numeric (10,0)	PK
TreatmentName	Notes about the treatment	Varchar(100)	

9. Treatment Procedure Dimension: The treatment procedure dimension stores information about the type of treatment procedure that physician selected to direct the treatment process as shown in Table 5.15, to studying the effect of this procedure.

Table (5.15): Treatment Procedure Dimension

Attributes	Description	Data type(length)	Remark
ProcedureID	A unique number for each procedure treatment	numeric (10,0)	PK
ProcedureName	Notes about the treatment procedure	Varchar(100)	

10. Sex Dimension: The sex dimension stores patient's gender to studying the effect of this dimension in diagnosis and treatment processes; the relationship of specific cancer type with sex. Table 5.16 shows sex dimension attributes.

Table (5.16): Sex Dimension

Attributes	Description	Data type(length)	Remark
SexID	A unique identifier for the each patient gender	numeric (10,0)	FK
Sex	The name of patient gender	varchar (100)	

11. Tribe Dimension: The tribe dimension stores patient tribe as shown in Table 5.17; to studying the relationship between tribe and specific type of cancer.

Table (5.17): Tribe Dimension

Attributes	Description	Data type(length)	Remark
TribeID	A unique identifier for the each patient tribe	numeric (10,0)	FK
Tribe	The tribe name of each patient	varchar (100)	

12. Discharge Status Dimension: The discharge status dimension stores the patient's discharge status to studying of the studying the status of patient after treated from specific type of cancer. Table 5.18 shows discharge status dimension attributes.

Table (5.18): Discharge Status Dimension

Attributes	Description	Data type(length)	Remark
StatusID	A unique identifier for the discharge status each patient.	numeric (10,0)	FK
Status	The status of patient after complete specific treatment	varchar (100)	

13. Stage Dimension: This dimension stores patient cancer status that diagnosed to studying the relationship between the result of treatment process and stage. Table 5.19 shows stage dimension attributes.

Table (5.19): Stage Dimension

Attributes	Description	Data type(length)	Remark
StageID	A unique identifier for the cancer stage each patient.	numeric (10,0)	FK
Stage	The cancer stage of patient after diagnosis	varchar (100)	

14. Education Dimension: This dimension stores patient education level to studying the relationship between education and the result of treatment process. Table 5.20 shows education dimension attributes.

Table (5.20): Education Dimension

Attributes	Description	Data type(length)	Remark
EducationID	A unique identifier for the each patient education	numeric (10,0)	FK
EducationLevel	The education level of each patient	varchar (100)	

15. Date Dimension: A meaningful clinical data cannot use only time points, such as dates when data were collected; it must be able to characterize significant features over periods of time. Table 5.21 shows data dimension attributes.

Table (5.21): Date Dimension

Attributes	Descriptions
DateID	A unique identifier for the Date.
FullDate	Full date
MonthDate	The month patient made the visit
YearDate	The year patient made the visit
DayDate	The day patient made the visit

5.4.2. Fact Tables

A fact is a collection of related data items, consisting of measures and context data. Each fact typically represents a medical item, a medical transaction, or an event that can be used in analyzing the medical process. Additionally, measure is a numeric attribute of a fact, representing the performance or behavior of the medical relative to the dimensions. In fact table the foreign key relationships are established between the dimension tables keys in the corresponding values in the fact table. The proposed fact tables are shown in the following tables:

1. Diagnosis Fact Table: This fact table represents the diagnosis process, containing four measures representing the performance of diagnosis process according seven related dimension tables as shown in Table 5.22.

Table (5.22): Diagnosis Fact Table

Attributes	Description	Data type(length)
RangeID	A number used to identify the range of age for the patients	numeric (10,0)
CancerID	A unique number for each cancer type	
OccupationID	A unique number for occupation of patient	numeric (10,0)
SexID	A unique number for each gender	numeric (10,0)
StageID	A unique number for each cancer stage	numeric (10,0)
TribeID	A unique number for each tribe	numeric (10,0)
DiagnosisDateID	Determine the date of service	Datetime
SecondCancer	determined the number of patients have other type of cancer	numeric (10,0)
OtherDisease	determined the number of patients have other disease	numeric (10,0)
CancerStage Count	Determine the cancer stage count	numeric (10,0)
Count Total patients	Determine the count of patient that diagnosis by special cancer type.	numeric (10,0)

Foreign Key: RangeID references DM_Age range (RngeID)
Foreign Key: CancerID references DM_Cancer (CancerID)
Foreign Key: StageID references DM_stage (stageID)
Foreign Key: OccupationID references DM_Occupation (occupationID)
Foreign Key: SexID references DM_ Sex (SexID)
Foreign Key: TribeID references DM_ Tribe (TribeID)

2. Treatment Fact Table: This fact table represents the treatment process, containing two measures representing the performance of treatment process according five related dimension tables in order to improve the treatment results as shown in Table 5.23.

Table (5.23): Treatment Fact Table

Attributes	Description	Data type(length)
RangeID	A number used to identify the range of age for the patients	numeric (10,0)
CancerID	A unique number for each cancer type	
TreatmentID	A unique number for each treatment	numeric (10,0)
ProcedureID	A unique number for each treatment procedure	numeric (10,0)
TreatmentDateID	A unique number for each treatment risk	numeric (10,0)
Treatment_Patient NO.	Determine the count of patient that treated by special treatment type.	numeric (10,0)
Procedure_Patient NO.	Determine the count of patient that treated by special procedure type.	numeric (10,0)

Foreign Key: RangeID references DM_Age range (RngeID)
Foreign Key: CancerID references DM_Cancer (CancerID)
Foreign Key: TreatmentID references DM_Treatment (TreatmentID)
Foreign Key: ProcedureID references DM_ TreatmentProcedure (ProcedureID)

3. Discharge Status Fact Table: The discharge status fact table represents the status of patients after treated by special procedure. This table contain from two measures representing the performance of treatment process according four related dimension as shown in Table 5.24.

Table (5.24): Discharge Status Fact Table

Attributes	Description	Data type(length)
RangeID	A number used to identify the range of age for the patients	numeric (10,0)
CancerID	A unique number for each cancer type	numeric (10,0)
StatusID	A unique number for discharge status each patient	numeric (10,0)
DischargeDateID	Determine the date of service	Datetime
OtherTreatment	Number of patients have other treatment	numeric (10,0)
Count of Patient status	determined the number of patients status	numeric (10,0)

Foreign Key: RangeID references DM_Age range (RngeID)
Foreign Key: CancerID references DM_Cancer (CancerID)
Foreign Key: PhysicianID references DM_Physician (PhysicianID)
Foreign Key: StatusID references DM_DischargeStatus (statusID)

4. Location Fact Table: The location fact table represents behavior of the medical relative to the dimensions. This table contain from one measures and six related dimension as shown in Table 5.25.

Table (5.25): Location Fact Table

Attributes	Description	Data type(length)
RangeID	A number used to identify the range of age for the patients	numeric (10,0)
ProvinceID	A unique number for the patient province	numeric (10,0)
DateID	Determine the date of service	Datetime
CancerID	A unique number for the cancer that patient's have.	numeric (10,0)
EducationID	A unique number for each education level	numeric (10,0)
TribeID	A unique number for each tribe	numeric (10,0)
PatientCount	A number of Patient	numeric (10,0)

Foreign Key: Range ID references DM_Age range (RngeID)
Foreign Key: Cancer ID references DM_ Province (CancerID)
Foreign Key: EducationID references DM_Education (Education)
Foreign Key: TribeID references DM_ Tribe (TribeID)

5.5. Population CDWH and Data Storage Services

The population CDWH and data storage services stage performs a various processes that required designing and developing the CDWH. This stage includes staging area design, system process design and infrastructure design that involves of all necessary hardware, and software that are used for developing the CDWH.

5.5.1 Staging Area Design

The data staging area is designed to store the extracted data and the processes of cleanse, and transform are performed before the data store into CDWH. The staging area consists of two types of tables setting table and data tables. The first type of these tables, setting tables is the tables that contain data that derives the data flow from sources to staging area. While the second type of these tables contain the data that selected and extracted from sources, as shown in table 5.9.

Table (5.26): Staging Area Tables

Table Types	Tables	Table Description
Setting tables	Sources table	These tables contain data to manage data sources.
	Extraction setting table	These tables contain data to manage the extraction process.
Data tables	Clinical data tables	These tables store clinical data about each patient, patient diagnosis information, Treatment that patient have, and discharge status of patient.
	Medical data tables	These tables store data about risk factor type, treatment type, treatment procedure type, special treatment risk factor, and laboratory test type.
	Demographic tables	These tables store data about sex, religion, education level, parent's relative classification, tribe, occupation sector classification, occupation, state and province.

5.5.2. ELT Processes

ETL is the core component of a successful CDWH technique. Due to the requirement of complex clinical data structure as discussed in chapter two, which requires powerful ETL technique fit well for medical tasks and objectives. These ETL techniques are proposed and discussed in chapter four. The proposed ETL technique consists of data extraction, data cleansing, data transformation, and data loading processes and focused on the particular requirements of these processes. Thus, ETL processes are vital and important to have an effective analysis, mainly concerned with integrate data from various data sources and improve data quality in CDWH.

5.5.3. Infrastructure Design

The technologies and software used in data and system architectural layers determine the infrastructure architectural design layer requirements. The infrastructure design aims to integrate different operational medical sources (e.g. SQL Server, MYSQL and XML format) using a custom-developed ETL technique. Therefore, the infrastructure builds a technological framework for massive clinical data integration, preprocessing, and analysis. Furthermore, the CDWH was created on a centralized server dedicated to data analysis purpose to

evaluate the ETL techniques. The CDWH is a relational database built in the server with:

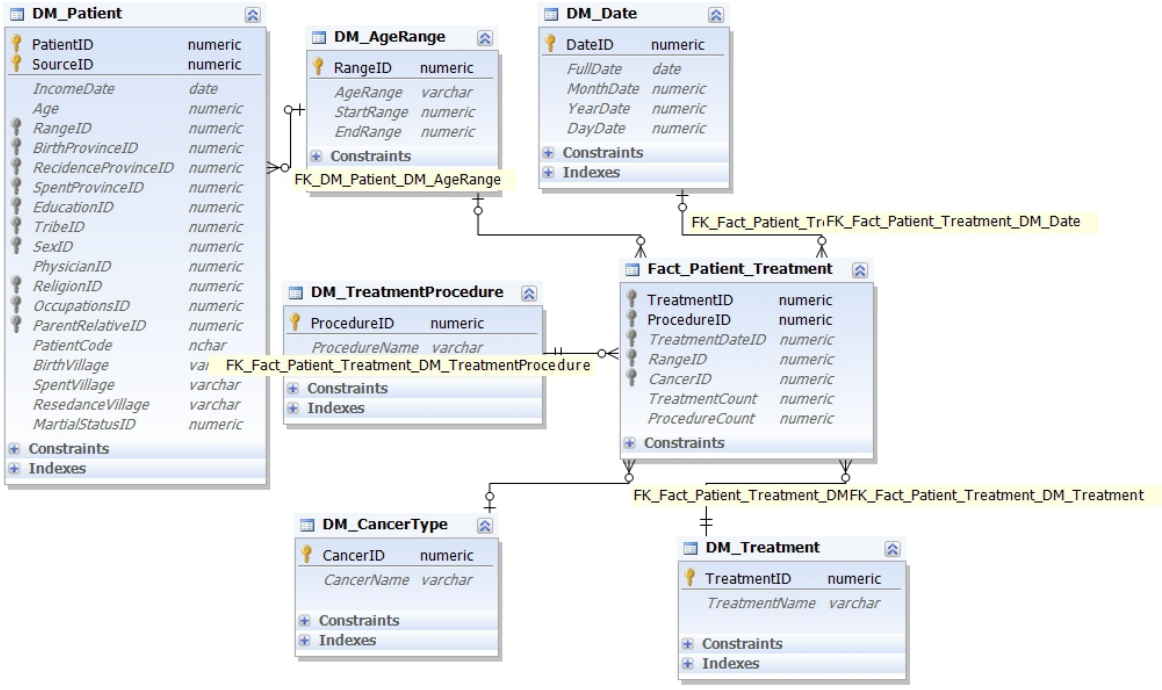
1. Relational database management system: Microsoft SQL 2012 Server is chosen as database engine and installed in the server to create the staging database and CDWH. The SQL 2012 Server includes the following services: integration services, analysis services and reporting services.
2. ETL techniques: custom –developed ETL techniques to extract, cleanse, transform and load integrated data to target database.
3. Custom software applications used to querying, reporting and display information.
4. Hardware: One HP Compaq Elite 8300 CMT, (i i7-3770)3.392 GHz Server; three Hp Intel duo core (i5-2430M) 2.40 GHz II Notebooks.
5. Operating System: Microsoft Windows 2013 Server (64-bit operating System).
6. Hard disk space: 500 Gigabytes.
7. Tape Backup: SAN Disk 3 TB.
8. Power Backup: Smart UPS 1400 (APC).

5.6. Physical Development of the Clinical Data Warehouse

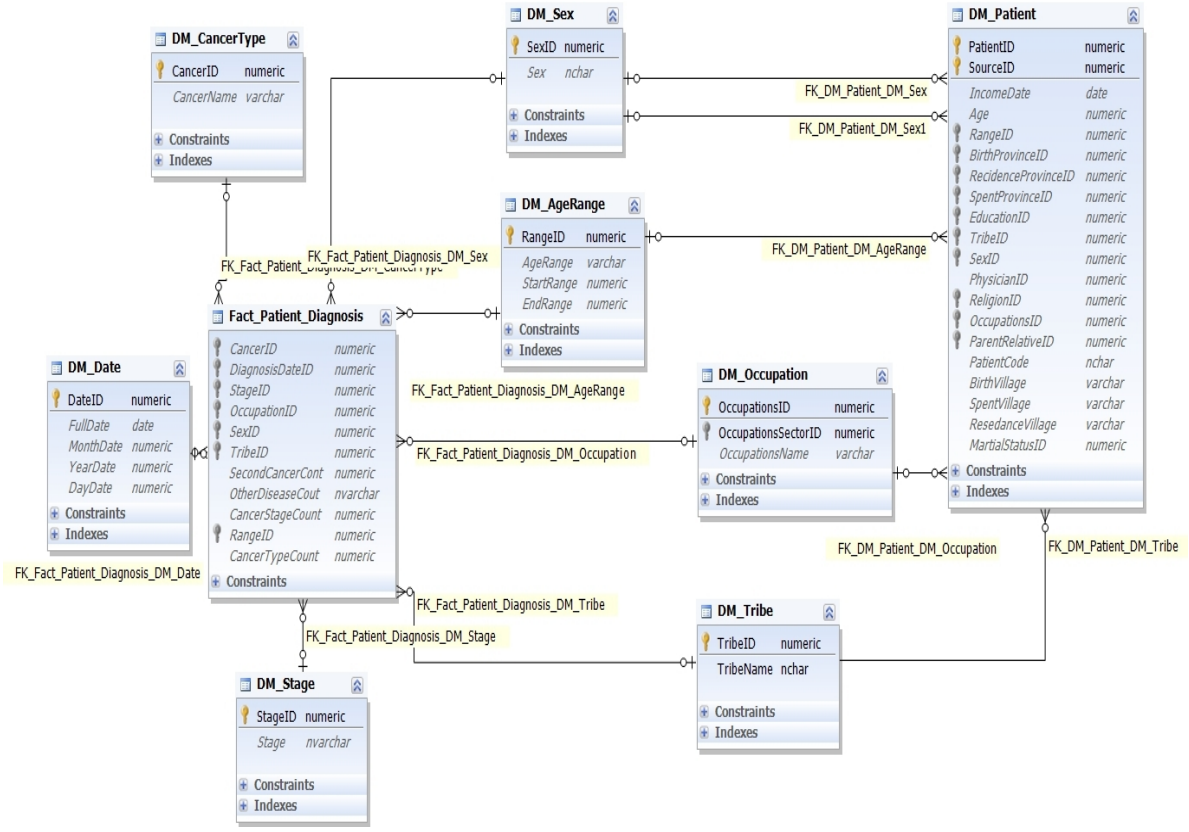
The designing and developing of the physical data model of the CDWH based on the developing of the CDWH logical data model. The main goal of physical CDWH Model design is good query performance. Designing of the physical data model includes all the database processes required to create all table structures, including column name, column data type, column constraints, primary key, foreign key, and relationships between tables or to achieve performance goals. The physical data model illustrated in figure 5.10 shows the tables and

relationships that are used in the CDWH. Furthermore, a physical data model used to calculate storage estimates; it may includes specific storage allocation details for a given database system.

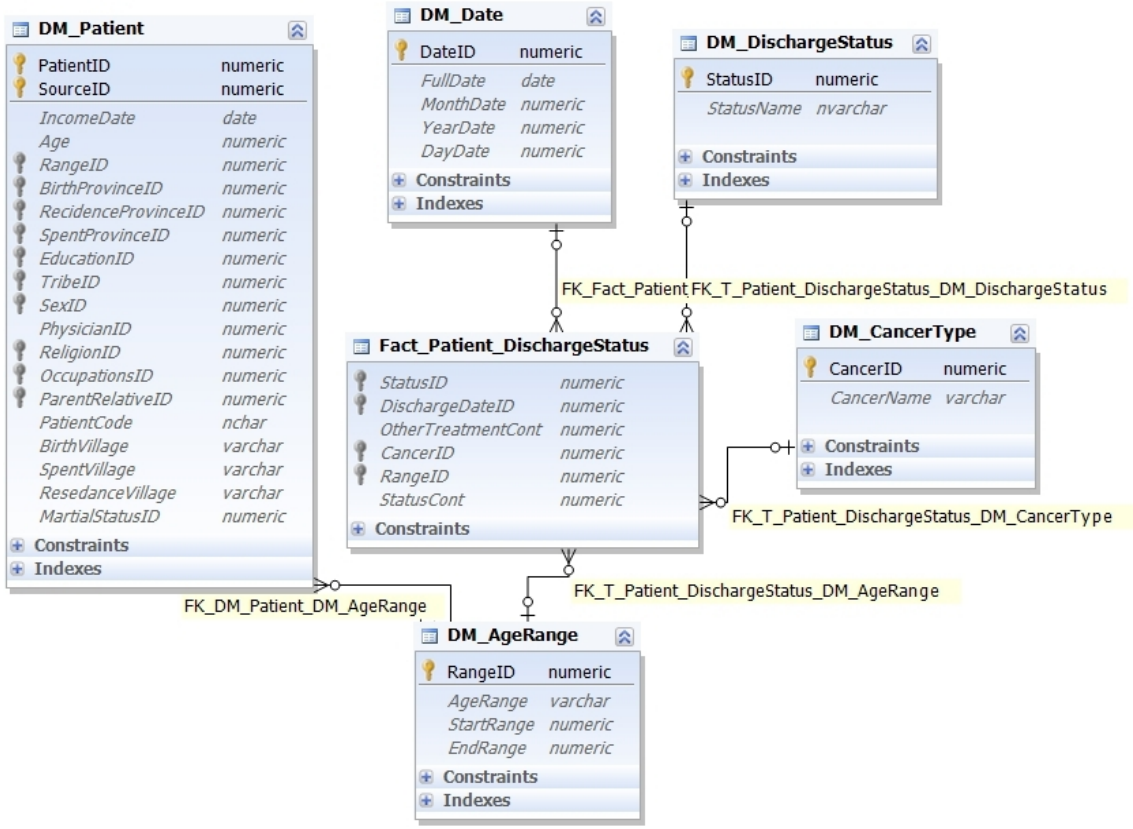
The implementing of a physical data model transforms the physical data model into a physical database by generating the SQL Data Definition Language (DDL) script to create all the objects in the database. This physical data model includes database tables, indexes, and constraints.



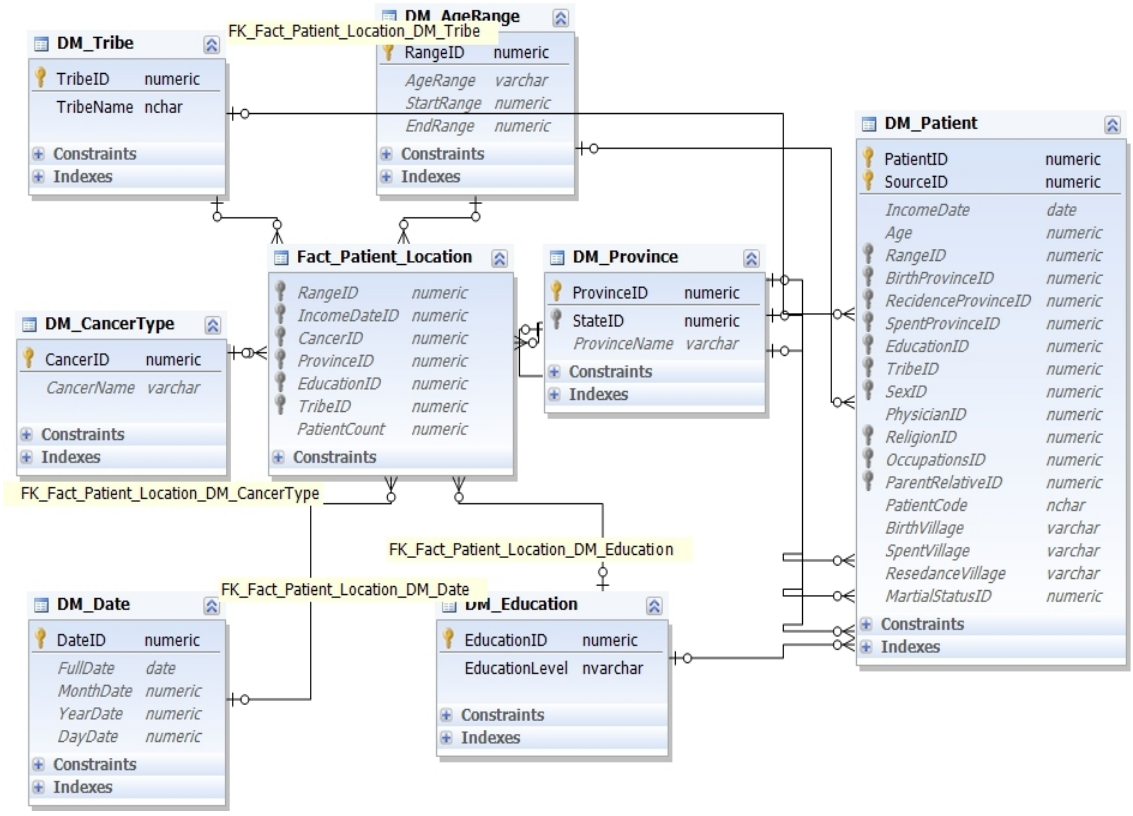
(A) Schema_Treatment:



(B) Schema_Diagnosis



(C) Schema_Discharge



(D) Schema_Locaction

Figure (5.10): Snowflake Schema of CDWH

5.6.1. Dimension Tables

The physical data model for the CDWH consists of multiple dimension tables are created using DDL statements. These dimension tables include: DM_Date, DM_Patient, DM_AgeRange, DM_OccupationSector, DM_Occupation, DM_State, DM_Province, DM_Cancer, DM_Treatment, DM_TreatmentProcedure, DM_Stage, DM_Sex, DM_Trible.

5.6.2. Fact Tables

The first step in designing a fact table is to determine the granularity of the fact table. By granularity, mean the lowest level of information that will be stored in the fact table to achieve the medical goals. This constitutes two steps: (1) determine which dimensions will be included, and (2) determine where along the hierarchy of each dimension the information will be kept. Furthermore, determining the measure

depend on the medical requirements. The relationship between fact table and dimensions is constructing by using surrogate keys. Surrogate key allows to adding more than one record for each single patient. Surrogate key allow using primary key to foreign key in order to connect dimensions with fact table. The physical data model for the CDWH consists of three fact tables which are created using SQL server 2012.These fact table contains the treatments, diagnosis, Discharge and location, which named TB_Treatment_FACT, TB_Dignosis_FACT, TB_Discharge_FACT, and TB_Location_FACT.

5.6.3. Table Constraints

The physical data model implements the table constraints defined for the CDWH, by establishing relationships between the dimension and fact tables using the DDL statements. The physical data model for the CDWH consists of the following table constraints:

(**A**) Diagnosis Fact Table Constraints: This defines the relationship between TB_Diagnosis_FACT table and the dimension tables (DM_Cancer, DM_AgeRange, DM_Stage, DM_Occupation, DM_Sex, DM_Date, and DM_Tribe)

(**B**) Treatment Fact Table Constraint: This defines the relationship between TB_Treatment_FACT table and the dimension tables (DM_Treatment, DM_AgeRange, DM_Cancer, DM_Date and DM_TreatmentProcedure)

(**C**) Location Fact Table Constraints: This defines the relationship between TB_Location_FACT table and the dimension tables (DM_AgeRange, DM_Cancer, DM_Education, DM_Tribe, DM_Province and DM_Date)

(**D**) Discharge Fact Table Constraints: This defines the relationship between TB_Discharge_FACT table and the dimension tables (DM_AgeRange, DM_Cancer, DM_discharge, and DM_Date)

(**E**) Occupation Dimension Table Constraints: This defines the relationship between TB_Occupation_DIM and TB_OccupationSector dimension tables.

(**F**) Treatment Dimension Table Constraints: This defines the relationship between TB_Treatment_DIM and TB_Cancer_DIM

(**G**) Treatment Procedure Dimension Table Constraints: This defines the relationship between TB_TreatmentProcedure_DIM and TB_Treatment_DIM dimension tables.

(**H**) Treatment Risk Dimension Table Constraints: This defines the relationship between TB_TreatmentRisk_DIM and TB_TreatmentProcedure_DIM dimension tables.

(**I**) Province Dimension Table Constraints: This defines the relationship between TB_Province_DIM and TB_State_DIM dimension tables.

(**J**) Laboratory Test Dimension Table Constraints: This defines the relationship between TB_LaboratoryTest_DIM and TB_ CancerType _DIM dimension tables.

(**K**) Risk Factor Dimension Table Constraints: This defines the relationship between TB_ RirkFactor _DIM and TB_ CancerType _DIM dimension tables.

5.6.4. Indexing Data Model

The physical data model includes individual indexes on each of the foreign key columns to promote the use of schemas. In order to improve the speed of query the index on the dimensions and facts table is using.

5.7. Presentation of the Information

A CDWH supports powerful data analysis techniques such as On-Line Analytical Processing (OLAP) and data mining to deliver advanced capabilities. The data mining techniques are integrated with OLAP services in the CDWH system to supporting data analysis. These technologies used to discover the required knowledge from CDWH and provide many benefits to healthcare institutions with quality improvement, data access performance improvement, improves information visualization of the data analysis results, and more informed decision support.

OLAP enables authorized users to access information from multidimensional CDWH, to and present data in a form of a tabular and graphical report. The Common OLAP operations used in OLAP to analyze data include: (1) Slice: is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset, (2) Dice: is a slice on more than two dimensions of a data cube (or more than two consecutive slices), (3) Drill Down/Up: it provide ability to navigate among levels of data ranging from the most summarized (up) to the most detailed (down), (4) A roll-up: it involves computing all of the data relationships for one or more dimensions, (5) Pivot: it rotates the data in order to provide an alternative presentation of data where, the report or page display takes a different dimensional orientation, (6) sort data by ordinal value, and Selection where data is available by value or range.

Additionally, the OLAP cubes are designed based on the dimensions and measures. Where, the output of the cubes use later by the reports to produce the characterized information which give full sight about the relationship between the dimensions. Therefore, cubes

preserve the information and allow browsing at different conceptual levels. It serves as the data source for the data mining task.

5.8. Clinical Data Warehouse Evaluation

The aim of evaluation is to answer the question that is the CDWH achieve its objectives? As presented in Chapter two the DWH is defined as "a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process" [43]. Therefore, the developed CDWH is evaluated against these data CDWH characteristics as the following:

(I) Subject-oriented: The extraction technique integrates all relevant data from three heterogeneous medical data sources according medical requirements and needs of CDWH. These data include medical data, demographic data, and clinical data (patients' data which are related to the diagnosis, treatment procedure, and decision making process to treat the disease).

(II) Integrated: The clinical data are stored in a globally acceptable format with consistent naming conventions, measurements, encoding structures, and physical attributes. The cross reference tables are used to represent multiple instances of the same information with a single instance in the CDWH. For example, instead of storing cancer type in a text format, a cancer dimension is maintained as the reference table.

(III) Time-variant: Clinical data warehouse contains a history of the subject, as well as current information. Data warehouse data represent long-term data from six years. a date_time attribute is used to facilitate both current and historical perspective measurements, This date_time attribute determine the time of service that patient have.

(IV) Non-volatile: The data in the CDWH are stable. These data does not change each time during operational processes are executed. Therefore, the data are consistent regardless to when the clinical data warehouse is accessed.

5.9 Summary

This chapter discussed the major issues in designing and developing of CDWH. However, CDWH is more complicated than the DWH and produced a set of requirements and challenges. The medical analysis is significant to study and analyze the existing process from medical perspective to determine project objectives, requirements, constrains and acceptance criteria. Furthermore, a clear understanding of the medical purpose represents as an important stage in the process of developing CDWH. The medical data requirements are collected and investigate to determine the data integration problems and producing an initial dimensional model. Moreover, the model is assessed to realize medical objectives. Therefore, the CDWH design and development must meet some functional requirements in order to maintain data integration in CDWH. The proposed CDWH technology consists of a set of tasks, including the medical analysis, CDWH architectural selection, CDWH schemes creation, population CDWH and data storage services, physical development, and CDWH evaluation.

This chapter discussed how the development of CDWH is completed? The CDWH life cycle is composed of sex stages; each of these stages has its issues. These stages involved the select suitable CDWH architecture to implement the CDWH processes. Second, design of the logical model to understanding of the details of medical and clinical data. Third, populate CDWH and data storage services to perform a various processes that required to developing the CDWH. This

includes staging area design, system process design and infrastructure design. Fourth, CDWH physical development which includes all the database processes required to create all table structures and relationships between tables or to achieve performance goals. Fifth, present the information where the knowledge is discovering using OLAP and data mining techniques. And the last one, evaluating the CDWH with acceptance criteria, in order to ensure medical objectives is achieved.

Finally, the data in CDWH must be corrected, completed, consistent, and integrated to provide a suitable medical decision making.

Chapter Six

Experiment and Results

The study is conducted to integrate clinical data from different data sources into CDWH with a high data quality. This study, proposed ETL techniques which can efficiently and robustly integrate clinical data and address the data quality problems in medical field. Additionally, the dataset used to test these techniques is real data gathered from the Radiation and Isotopes Centre Khartoum (RICK) – Sudan, and Radiation and Isotopes Hospital –Shendi (RICSH) – Sudan. These dataset stored in different database management systems, which included SQL server, My-SQL and XML format data sources. The dataset contained medical data, demographic data, diagnosis data, treatment data and discharge status data.

In this chapter, ETL techniques pertaining to integrate clinical data of cancer are developed according to the requirements that have been identified in chapter four. Furthermore, the evolution of ETL techniques has been done through a development of CDWH that has been developed according the architecture worked out in chapter five. The experiment involves experiment setup, clinical staging area creation, ETL techniques, analysis and mining the data and results.

6.1. Experiment Setup

The experiment was conducted using network architecture as shown in figure 6.1. The experiment composes four computer conducted in a network. Three nodes (computers) of this network used to represent data sources. The first node used SQL Server 2012 while the second and third node used MY-SQL and XML respectively. On the other hand, the fourth computer used as main node to collect the data from other data sources and analyze the data. This node includes all necessary hardware and software for this research work. Furthermore, SQL server 2012, including integration service and analysis service, are installed onto the

server to analysis the data. Additionally, the ETL techniques are execute in this node to integrate these data before store the data into CDWH.

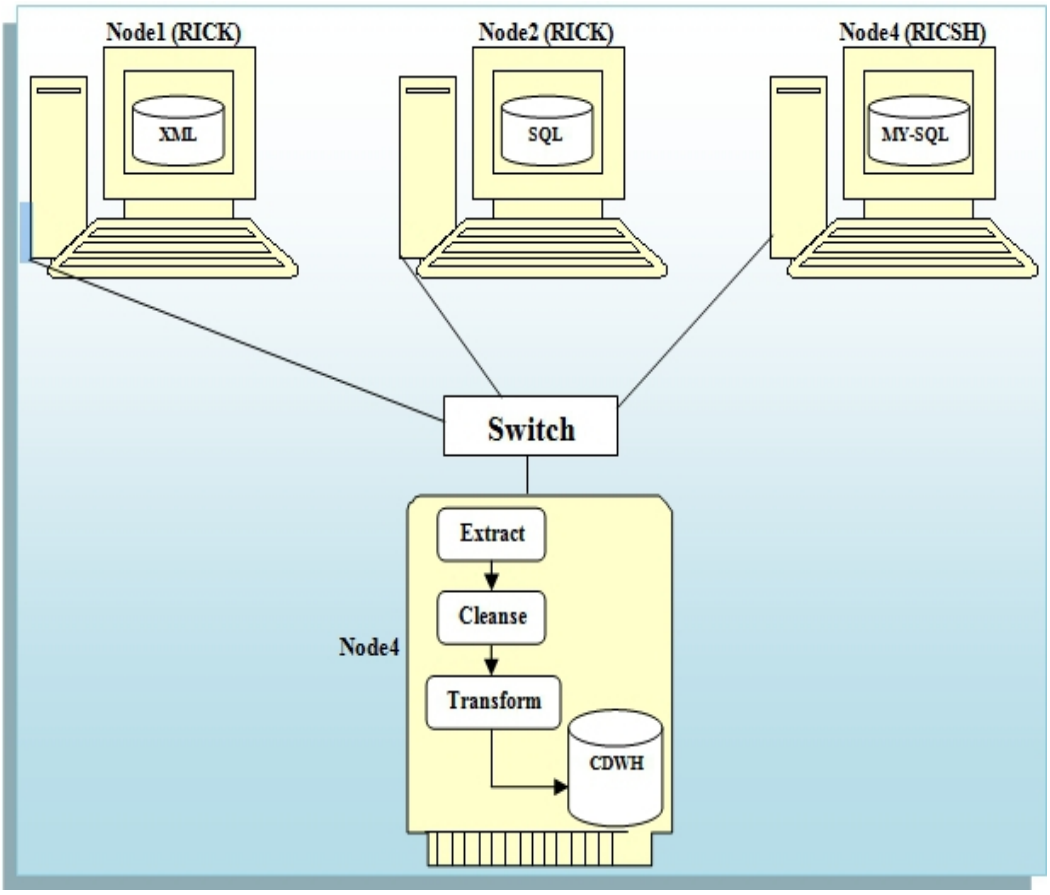


Figure 6.1. Network Architecture

6.2. Staging Database Creation

The clinical staging area is created using the DDL statement to store the extracted data and the processes of cleansing and transforming are performed before the data store into CDWH. The clinical staging area includes all tables that required storing relevant data according to table types that have been identified in chapter five.

6.3. ETL Techniques

The ETL techniques are developed to extract, cleanse, transform and load the required data into CDWH through two phases (extracting/cleansing and transforming/loading) as mentioned in chapter four. The crucial role of the ETL technique is to identify and handle data quality problems that may exist at all the phases of ETL process (extraction, cleansing, transformation, and loading) shown in figure 6.2. Furthermore, the ETL techniques used appropriate methods in order to handle data problems and improve the data quality in the clinical staging database as proposed in chapter four. The extraction time and the scalability are the criteria used in experiment to measure the performance of the ETL techniques.

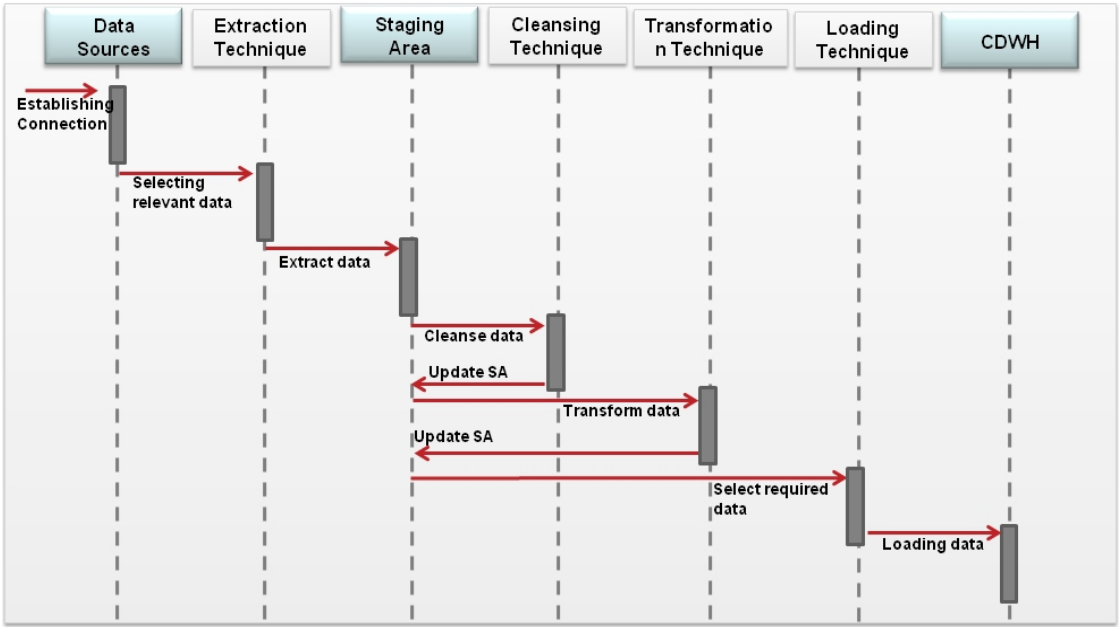


Figure (6.2): ETL Implementation

6.3.1. Extracting/Cleansing phase

In this phase the data are extracting from sources, transfer to staging area and then datasets are cleanse, as the following:

(I) Extracting Technique: The extraction technique is building to extract and integrate the clinical data into stage area. The extraction process technique involved several steps as explain in the following section.

1. Creating the list of data sources that contains the required data according medical needs.
2. For each data source, the data structure and content are identifying. Furthermore, the type of the data source is identifying in order to use appropriate data connection.
3. Establishing the connection with specific data source.
4. Selecting the required medical and demographic data by identifying the tables/files that contains the required data. These data involve two type of data:
    a. Data have not logical relationship between attributes, such as marital status, sex, religion, parent relative, physician, education, tribe, and feeding pattern. These medical and demographic data are select and convert to appropriate format (target stage area representation), and eliminating duplicate on data before store in target table in staging area.
    b. Data have logic relationship between attributes, such as cancer type (symptom, risk factor), treatment (treatment procedure), occupation sector (occupation), and state (province). The relationship between attribute is perform to ensure data validity and integrity. These medical and demographic data are selected and checked whether the attribute fields are complying with or not. Furthermore, the

data convert to appropriate format (target representation), and eliminating duplicate on data before store in target table in staging area.

c. Storing the extracted medical and demographic data into stage area.

5. Selecting the required transaction clinical data. For each data source the required data according medical needs is selected form containing tables/files. These data involve patient's data, patient's diagnosis, patient's treatment, and patient's discharge status data.

   a. The required patient's data is selected according medical needs. The selected data is converts to appropriate format (target (stage area) representation) using appropriate method. Furthermore, there are some constrain applies on attributes such as (0<patient age<120), specific attribute have boundary/ or formula, and there are attribute that have limited value. Additionally, the process of eliminating duplicate on data is performed by using unique identification for each patient (by adding the number of source to patient identification within the source). In addition, mapping from source table to target (stage area table) and using appropriate provider to store the extracted patient's data into stage area table.

   b. The required patient's diagnosis data is selected and converted to appropriate format (target representation) using appropriate method. Furthermore, there are some constrain applies on attributes. In addition, mapping from source table to target (stage area table) and using appropriate provider to store the extracted patient's diagnosis data into stage area table.

c. The required patient's treatment data is selected and converted to appropriate format (target representation) using appropriate method. Furthermore, there are some constrain applies on attributes. In addition, mapping from source table to target (stage area table) and using appropriate provider to store the extracted patient's treatment data into stage area table.
d. The required patient's discharge status data is selected and converted to appropriate format (target representation) using appropriate method. Furthermore, there are some constrain applies on attributes. In addition, mapping from source table to target (stage area table) and using appropriate provider to store the extracted patient's treatment data into stage area table.

Based on the designed algorithm, there are two type of extraction that applying in extraction technique, initial and incremental extraction. The first one, initial extraction performs all mentioned steps above. While the incremental extraction performs a selection of these steps according some conditions such as change in data source list, change in each of data source, or change (insert, update) in each of data source's tables.

As illustrated in previous section, the main task of extraction technique is to handle data quality problems that may exist at extraction process phase. These problem includes medical purpose identify, relevancy, scalability and loss of data during extraction process. The data problems are handled using appropriate methods in order to improve the data quality in the CDWH. Furthermore, the identifying data sources and the relevant data are selected according medical analysis goals. Table 6.1 shows the data sources of the required data, and table 6.2 shows the

required relevant data and number of records that has been loaded into the staging area.

Table 6.1: Relevant Data Sources

No.	Institution Name	Source Type
1.	Radiation and Isotopes Centre Khartoum (RICK)	XML
2.	Radiation and Isotopes Centre Khartoum (RICK)	SQL
3.	Radiation and Isotopes Hospital –Shendi (RICSH)	My-SQL

Table 6.2: Number of Extracted Records into the Stage Area

No.	Medical and Clinical Data	No. of Records			Total No. of records store in Staging Area
		Source XML(1)	Source SQL(2)	Source MySQL(3)	
1.	Patient Data	5615	20029	879	26523
2.	Diagnosis Data	5615	20029	819	26463
3.	Treatment Data	5615	20029	451	26095
4.	Discharge Status Data	5615	20029	599	26243
5.	Cancer Type	139	120	152	152
6.	Treatment Type	2	2	2	2
7.	Treatment Procedure Type	3	3	3	3
8.	Physician	9	12	4	18
9.	Sex Type	2	2	2	2
10.	Religion Type	3	3	3	3
11.	Education Type	4	5	5	5
12.	Parent's Relative Type	4	4	4	4
13.	Tribe	87	118	96	118
14.	State	18	23	17	23
15.	Province	64	97	59	97
16.	Occupation Sector	12	15	14	17
17.	Occupation	52	49	57	69
18.	Discharge Status Type	8	5	6	9
19.	Risk Factor	76	81	75	84
20.	Laboratory Test Type	1	1	1	1
21.	Material status	5	5	5	5
22.	Feeding pattern	7	7	7	7
23.	Symptoms	8	10	10	10

Moreover, the extraction technique treated data problems that affect the quality of data before loading data to staging area. These problems include consistency where, the data is not satisfying a set of constraints, and/or data values are not consistent across data sets. Also, the second problem is completeness where, all the required data are not recorded /available, or fields with null values or fields with default values. The third problem is uniqueness problem the duplicate is exists

in data of patients, diagnosis, treatment, and/or discharge status. The fourth problem is integrity where, the relation between tables is missing, or there is missing in joining data from several sources. The last problem is data formatting where, data in inappropriate forms for mining or using of different representation formats in data sources. Table 6.3 maps these different types of data quality problems and present summary of consistency, completeness, uniqueness, integrity and formatting data problems identified by the extraction technique.

Table 6.3: Summary of Data Extraction Problems

(A) Summary of Data Extraction Problems in Source (1)

No.	Data Quality Problems	No. of Problems	No. of Handled Problems
1.	Completeness	16850	16850
2.	Uniqueness	4	2
3.	Consistency	28080	28080
4.	Integrity	5	5
5.	Formatting data	16845	16845

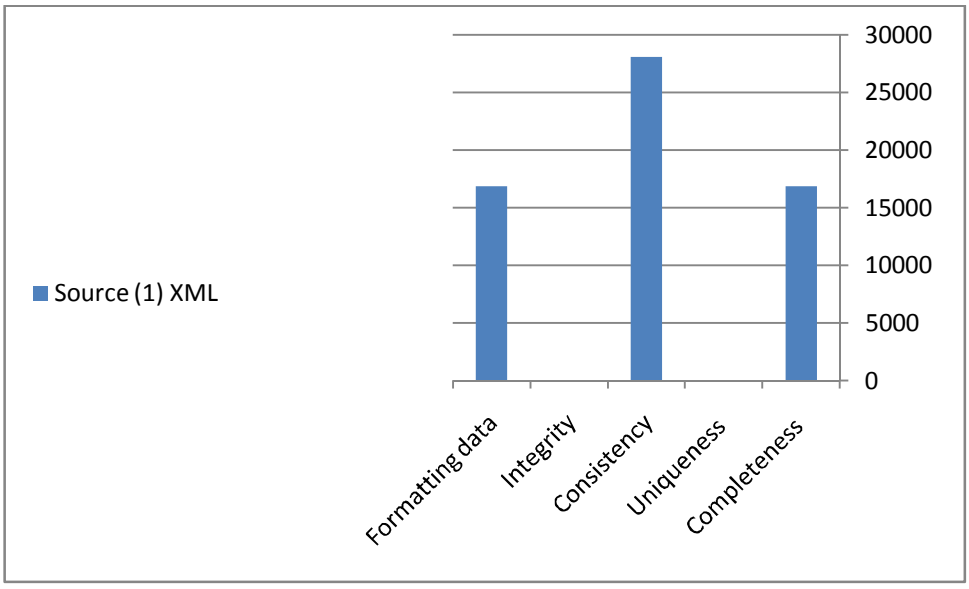
(B) Summary of Data Extraction Problems in Source (2)

No.	Data Quality Problems	No. of Problems	No. of Handled Problems
1.	Completeness	60181	60181
2.	Uniqueness	102	102
3.	Consistency	100150	100150
	Integrity	5	5
4.	Formatting data	6008	60087

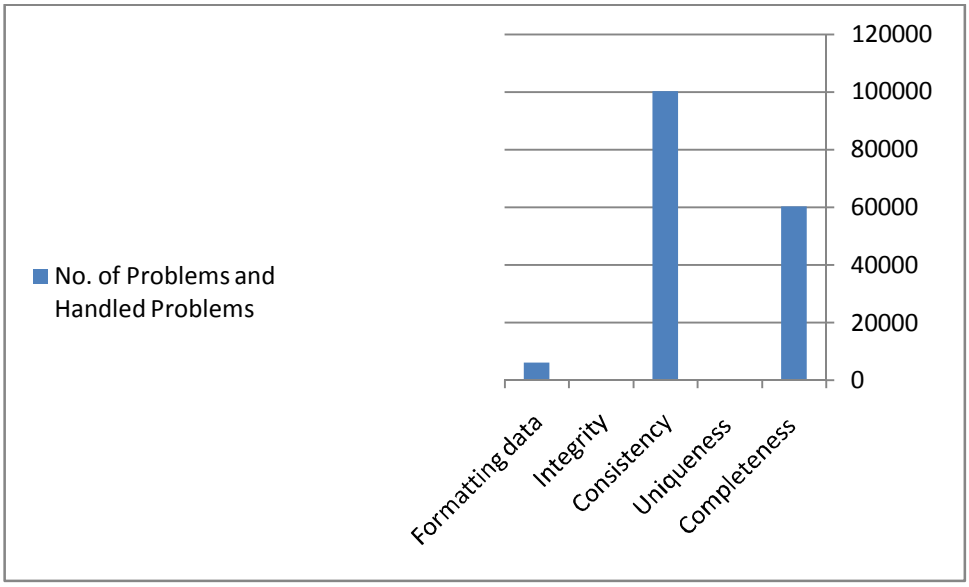
(C) Summary of Data Extraction Problems in Source (3)

No.	Data Quality Problems	No. of Problems	No. of Handled Problems
1.	Completeness	800	800
2.	Uniqueness	0	0
3.	Consistency	4395	4395
	Integrity	0	0
4.	Formatting data	2637	2637

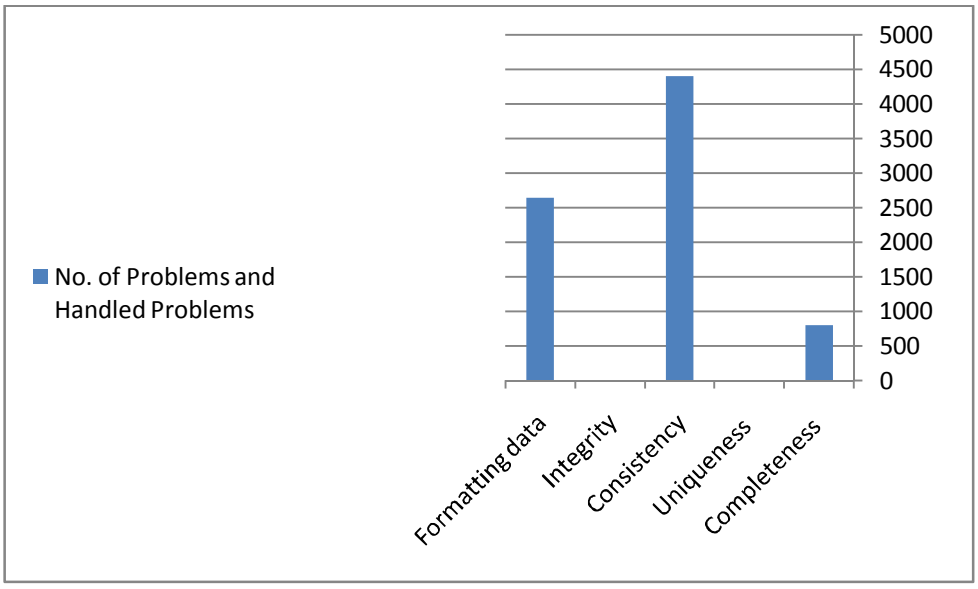
Additionally, the following figures 6.3 below shows the different data quality problems of consistency, completeness, uniqueness, and formatting data problems.



(A): The Number of Handed Data Problems in the Source (1)



(B): The Number of Handed Data Problems in the Source (2)



(C): The Number of Handed Data Problems in the Source (3)

Figure (6.3): The Number Data Problem that Affect Data Quality and Integration Process

On the other hand, the following figure 6.4 below shows the data quality problems in three datasets.

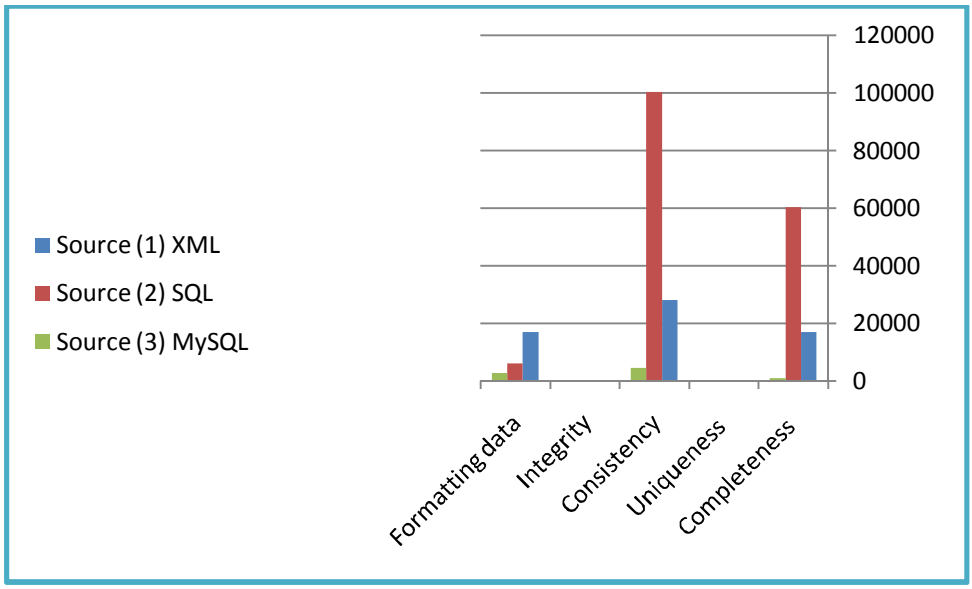


Figure (6.4): The Data Problems No. in Three Datasets

Finally, the data extraction technique performance should be evaluated from Data extraction time. However, there is other factor that affect the extraction time such as memory size, CPU speed and

connection Speed. In order to increase capability and performance of this technique, the extraction technique discriminates between new and existing data at loading time to decrease transferring time by adding new record, and on the other hand, performs the updating process. This method also handles the problems of loss data during transferring process. Also, the extraction technique is capable to handle huge volumes of data, where the technique has capability to check whether the data source is change before the extraction process is begin. The speed of the transferring data from sources to staging area with record number increasing showed the following figure 6.5.

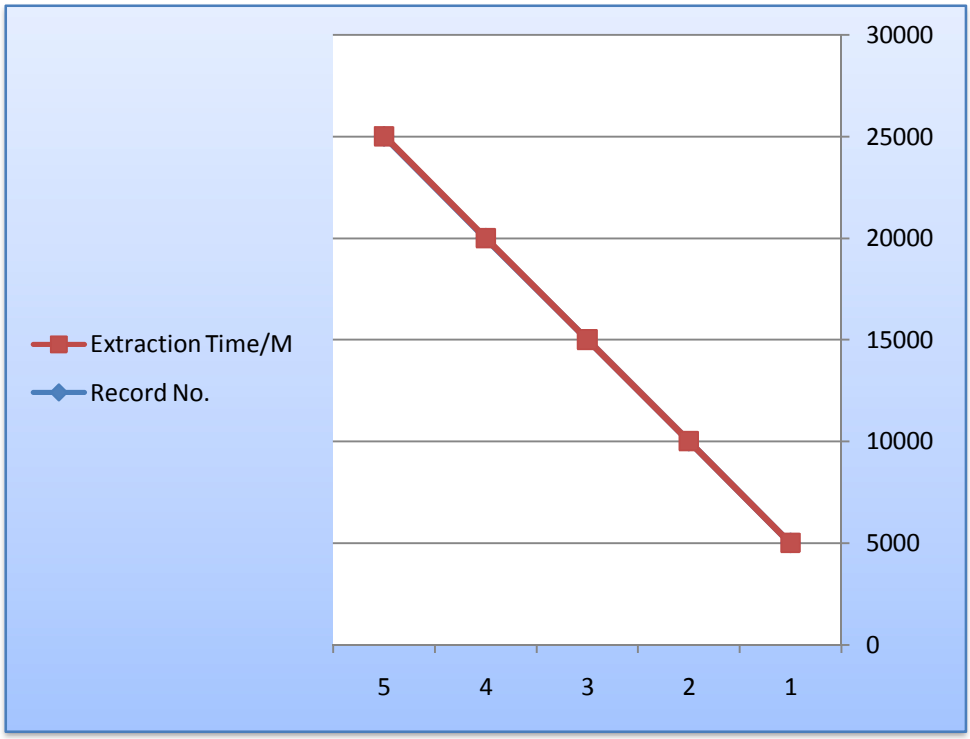


Figure (6.5): The Extraction Time versus Number of Extracted Records

As a result, the extraction technique evaluated on three real data sets and showed that the technique performed well on these sets. Furthermore, the extraction technique scaled efficiently with the number of records, the number of attributes, and the domain size.

(II) Cleansing Technique: The cleansing technique is building to cleanse the dataset in stage area to ensure data set are of certain qualities. However, data loaded to the staging database would still contain certain types of data problems. Therefore, cleansing process technique involved several steps as explain in the following section.

1. Creating the list of data objects that contains the required cleanse according medical needs.
2. For each data object, the data structure and content are identifying, to select the attributes which need to cleanse.
3. Establishing the connection to staging data with appropriate data connection.
4. Identifying the data problems that affect the quality of data in the object of transaction data. These data problems involve accuracy, and availability.
5. Handling each data problems with specific method by using temporary tables.
6. Verifying that all data problems will be handed by applying the correction methods again

The main task of cleansing technique is to identify the problems that may affect the quality of data at staging area. The accuracy and availability are two important data quality problems that are prevalent in real data set. These data problems are handled using appropriate methods in order to improve the data quality in the CDWH. Additionally, the security is important issue arising from the sensitivity of certain types of medical data, to observe this issue the technique hiding the patient names.

As illustrated above, the cleansing technique treat data problems that affects the quality of data at staging area. These problems include accuracy where, the data are not free from errors and the out of the range

of domain values. The second problem is availability where, the required data are not available within specified time and accuracy constraints. Table 6.4 below is present summary of accuracy, integrity, and consistency problems identified by the cleansing technique.

Table 6.4: Summary of Data Cleansing Problems in Saging Area

No.	Data Quality Problems	No. of Problems	No. of handled Problems
1.	Availability	22759	22759
2.	Accuracy	50840	50840
Total of Records			101004

Additionally, the following figures 6.6 below shows the different data quality problems that affect the data in stage area.

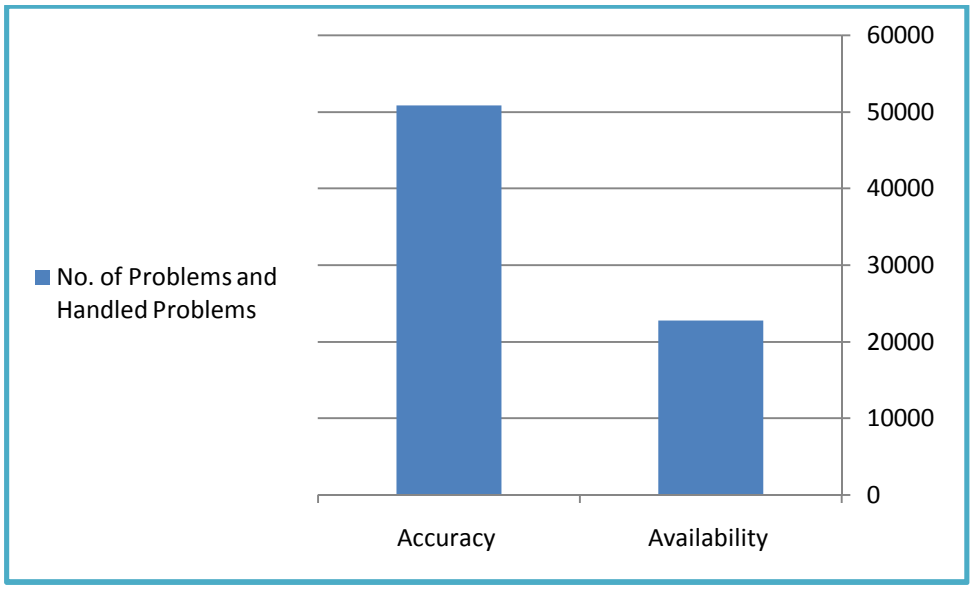


Figure (6.6): The Number Data Problem that Affect Data Quality and Integration Process in Staging Area during Cleansing Process

The time consumed to cleanse the data set at the staging area with record number increasing showed the following figure 6.7.

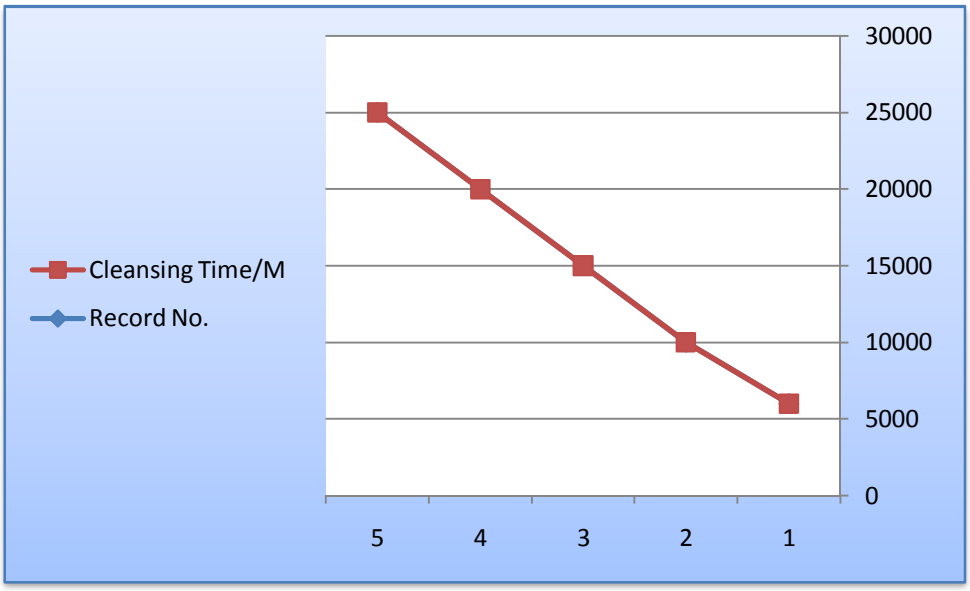


Figure (6.7): The Cleansing Time versus Number of Cleansed Records

6.3.2. Transforming/Loading phase

Transformation/Loading phase: In the transformation and loading phase the data in stage area are transforming to appropriate format to mining, mapping, and loading into CDWH as the following steps:

(I) Transformation Technique: The transformation technique is building to converts the data at staging area to appropriate format according to the data analyzing and mining requirements. The objective of transformation technique is reducing of analysis time at CDWH. The Transformation process technique involved several steps as explain in the following section.

1. Creating the list of data objects that contains the required transformation according medical needs, by analyzing the staging area.

2. For each data object, the format and properties of each object according to the representation of target database is identifying to select the attributes which need to cleanse.
3. Establishing the connection to staging data with appropriate data connection.
4. Identifying the data problems that affect the quality of data in the object of transaction data.
5. Handling each data problems with specific method by selecting appropriate manipulation option (method) the for required transformation process for attributes according medical needs to achieve the objective of CDWH.
6. Performing the mapping processes.
7. Verifying that all data problems will handed.

The main task of transformation technique is to identify the problems that may affect the quality of data at transformation process. The validity and huge data problems are two important data quality problems that are prevalent in real data set. These data problems are solved using appropriate methods in order to improve the data quality in the CDWH. Where, the transformation technique treated data problems which include validity such as, the data are not in appropriate forms for mining or the data in various data source are represented in different format. The second problem is huge data where, the analysis process is taken very long time on huge amount of data. Thus, there is a need of methods to treat these issues. The transformation technique consist multiple tools to observed aggregation, discretization, normalization of data, and dimension reduction to enable better data mining process.

(II) Loading Technique: The loading technique is building to load data from stage area into dimensions and finally load the surrogate keys and measurements into fact table. The objective of loading technique is

reducing of loading time. The loading process technique involved several steps as explain in the following section.

1. Creating the list of data objects that contain the data to transferring.
2. For each data object, the data problems are identified and handled these problems with appropriate method.
3. Establishing the connection to staging area with appropriate data connection.
4. Identifying the type of loading Initial or incremental.
5. Based on the type there are specific step will performed.
6. Performing the mapping processes to CDWH and disable all constrains on CDWH to reduce the loading time.
7. Verifying that all data is stored in CDWH.

The main task of loading technique is to identify the problems that may affect the quality of data at loading process. The freshness, reliability, timeliness and availability problems are four important data quality problems that are prevalent in loading process. These data problems are solved using appropriate methods in order to improve the data quality in the CDWH.

The loading technique treated data problems that affect the quality of data at loading process. These problems include freshness where, there are not of update strategy to updating the data set, the data are not tagged with a time and/or the CDWH is not hosted both historical and current data. The second problem is reliability where, the ETL process will not perform its intended operation during a specified time period under given conditions. The third problem is availability where, the ETL process is not operational during a specific time period because the resources of the system that needed are not available when needed or the

required data sources is not available within specified time. Table 6.9 below shows the number of rows that has been loaded to the CDWH

Table (6.5): Number of Transformed Records Loaded into the CDWH

Table Name		Number of rows loaded
Fact Table	Fact_Patient_Diagnosis	382036800
	Fact_Patient_DischargeStatus	41040
	Fact_Patient_Location	260968800
	Fact_Patient_Treatment	27360
Dimension Table	DM_AgeRange	5
	DM_CancerType	152
	DM_Date	6
	DM_DischargeStatus	9
	DM_Education	5
	DM_Occupation	71
	DM_OccupationSector	17
	DM_Patient	25420
	DM_Province	97
	DM_Sex	2
	DM_Stage	5
	DM_State	23
	DM_Treatment	2
	DM_TreatmentProcedure	3
	DM_Tribe	118

Additionally, the following figures 6.6 below shows the different data quality problems that affect the data in stage area.

Figure (6.8): No. of Records in Stage Area versus No. of Records after Transformation Process

The loading technique solved the problem of losing data during loading process. Also, the loading technique is capable to handle volumes of data.

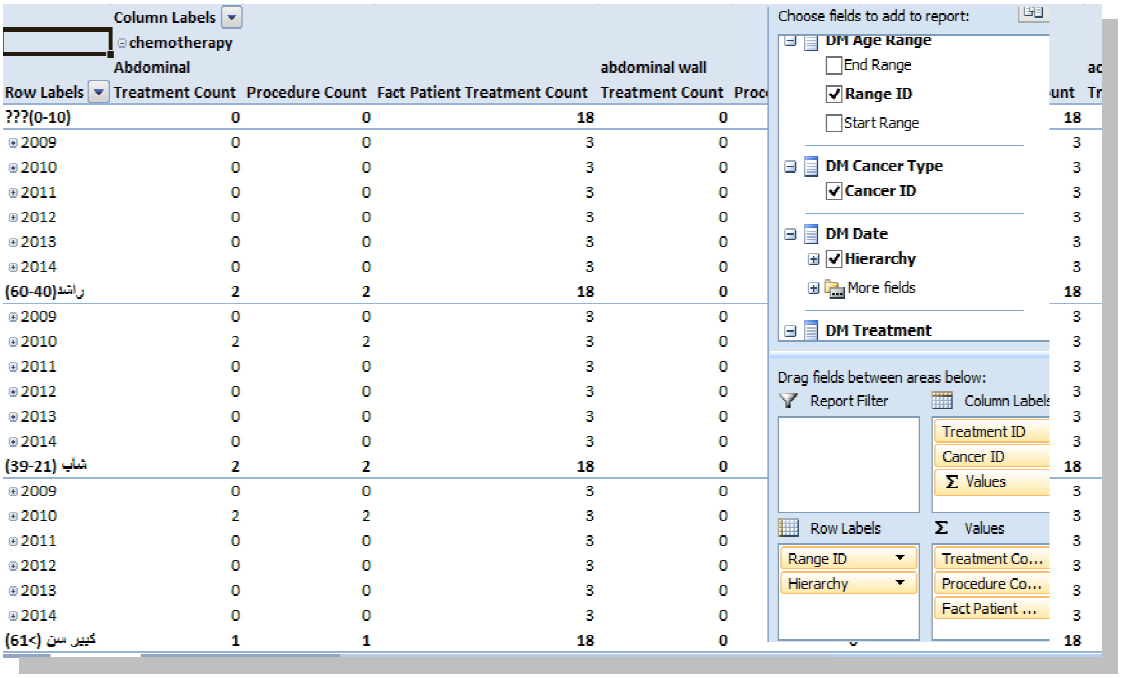
6.4. Analysis and Mining the Data

The CDWH is proposed in this study to show the utility of the ETL techniques. Where, the CDWH is conducted to integrated cancer patients' records from three disciplined sources. This study presented that a disciplined medical data sources could be successfully integrated in order to obtain a high quality data in CDWH to enhance data analysis capabilities for improved medical decision making process. Furthermore, DWH and mining technologies are applicable to medical field; and the conjunction of DWH with data mining has produced promising results.
In this research work the reports are created, the reports are designed based on the cubes and mining tools. The cubes are created using analysis services multidimensional and data mining. Cube design involved the design of data sources, data sources views, dimensions, create cube. Then the dimensions and cubes are process before browsing the data. The reports are created using pivot table according to required information and with desired chart.

This research work use pivot table to browse OLAP cube, the pivot table is excellent tool to browse OLAP cube. The pivot table has capability to configure, filter columns, rows, and measures.

Pivot table about treatment data base on the dimensions cancer type, treatment, treatment procedure, age rage as shown in figure 6.9. Therefore, the designed pivot popup to zoom out the cell pivot table detailed information. For example, the output of cube represents the relationship between count of patients treat by special treatment and count of patients treat by special treatment procedure measures and other

chosen related dimensions. Figure 6.10 shows a pivot table with two dimensions on row (treatment procedure and treatment) and one dimension on column (age range). Figure 6.11 shows pivot table with two dimensions on row (date and age range) and one dimension on column (procedure). And figure 6.12 shows pivot table with one dimension on row (age range) and two dimensions on column (treatment and date).



Column Labels	chemotherapy			abdominal wall	
	Abdominal				
Row Labels	Treatment Count	Procedure Count	Fact Patient Treatment Count	Treatment Count	Proc
???(0-10)	0	0	18	0	
2009	0	0	3	0	
2010	0	0	3	0	
2011	0	0	3	0	
2012	0	0	3	0	
2013	0	0	3	0	
2014	0	0	3	0	
(60-40)راشد	2	2	18	0	
2009	0	0	3	0	
2010	2	2	3	0	
2011	0	0	3	0	
2012	0	0	3	0	
2013	0	0	3	0	
2014	0	0	3	0	
(39-21) شاب	2	2	18	0	
2009	0	0	3	0	
2010	2	2	3	0	
2011	0	0	3	0	
2012	0	0	3	0	
2013	0	0	3	0	
2014	0	0	3	0	
(61<) كبير سن	1	1	18	0	

Figure 6.9: The Effect of the Dimensions with Measures in Treatment Cube

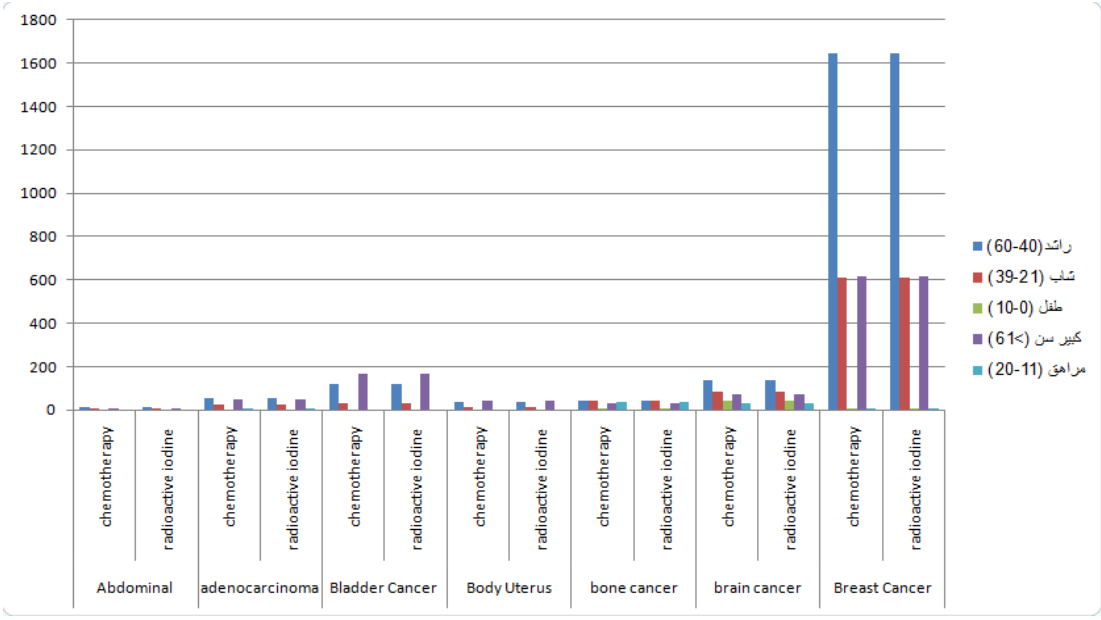


Figure (6.10): Effects of Treatment Procedure versus Age Range and Cancer Type

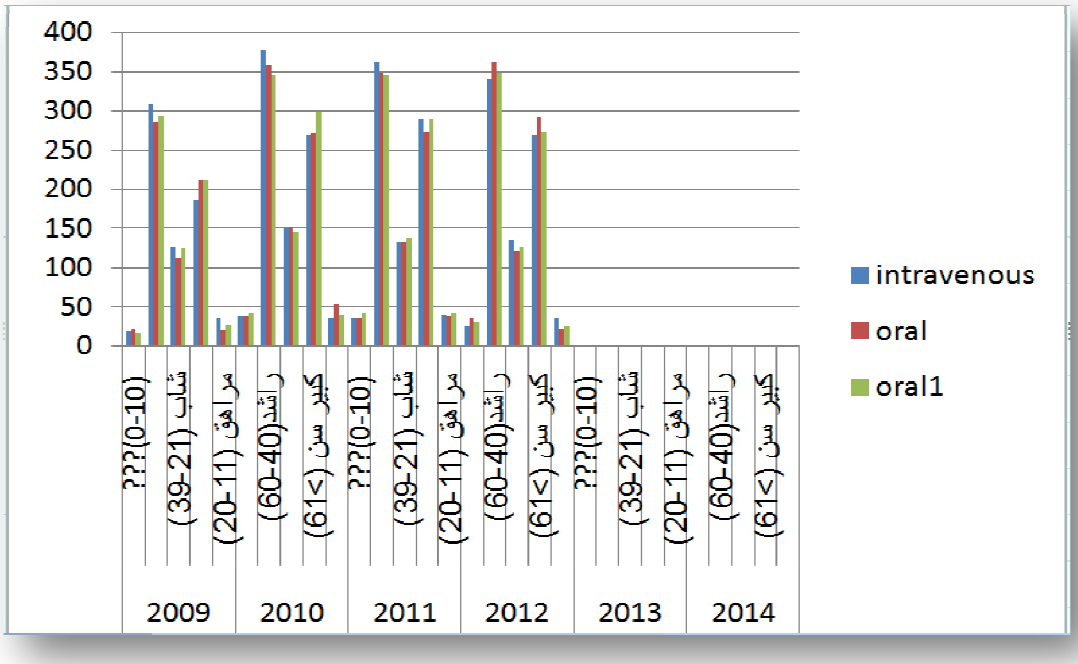


Figure (6.11): Effects of Treatment Procedure versus Age Range and Treatment Date

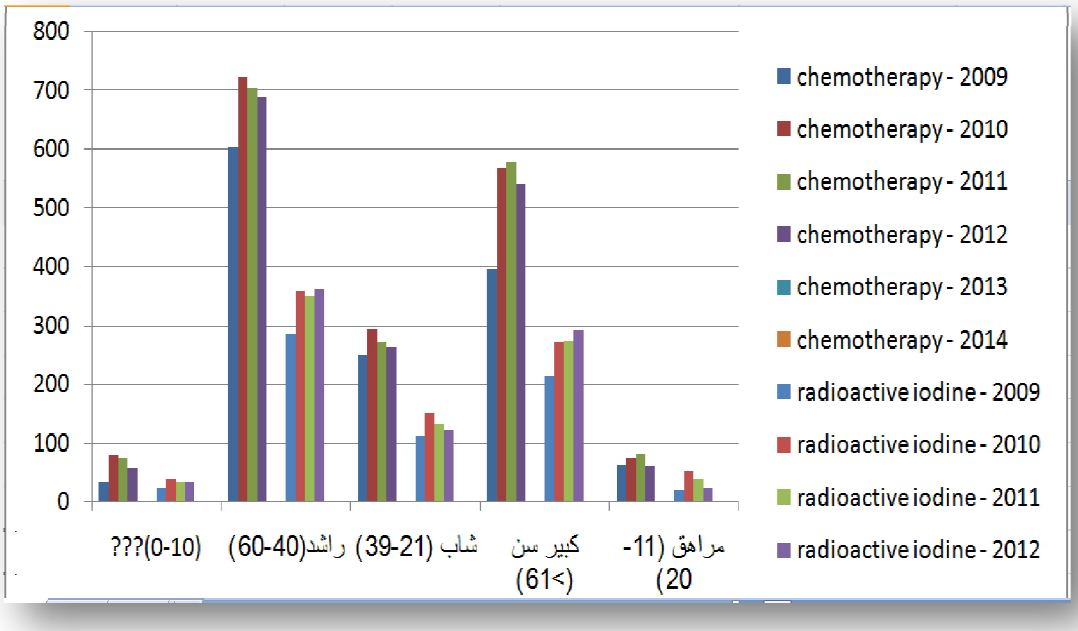


Figure (6.12): The Relationship between Treatment, Date and Age Range,

Furthermore, the diagnosis cube based on the dimensions cancer type, stage, occupation, sex, tribe, age rage, and date as illustrated in figure 6.13. The output of this cube represents the relationship between number of patients who have second cancer, other disease, specific cancer and the number of patient diagnosed in specific stage, and other chosen related dimensions. The analysis shows in figure 6.14, the patient Occupation versus age range of patients their admissions date. This analysis presents the relationship between the patients diagnosed by cancer measure and these mentioned dimensions. In figure 6.15 a pivot chart shows the total number of cancer patients diagnosed by cancer in specific stage and distribute the patients according the age range. Also analyses that could be used the age range and occupation dimensions to determine the relationship with tribe as shown in figure 6.16.

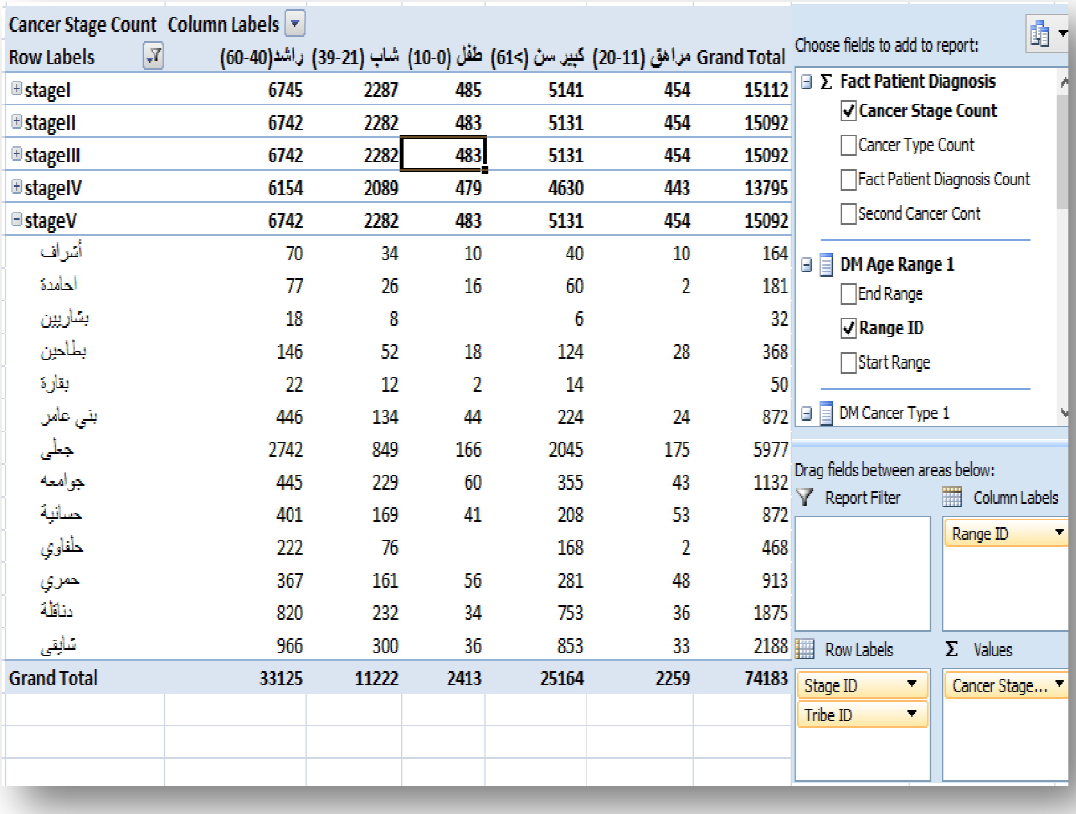


Figure 6.13: The Effect of the Dimensions with Measures in Diagnosis Cube

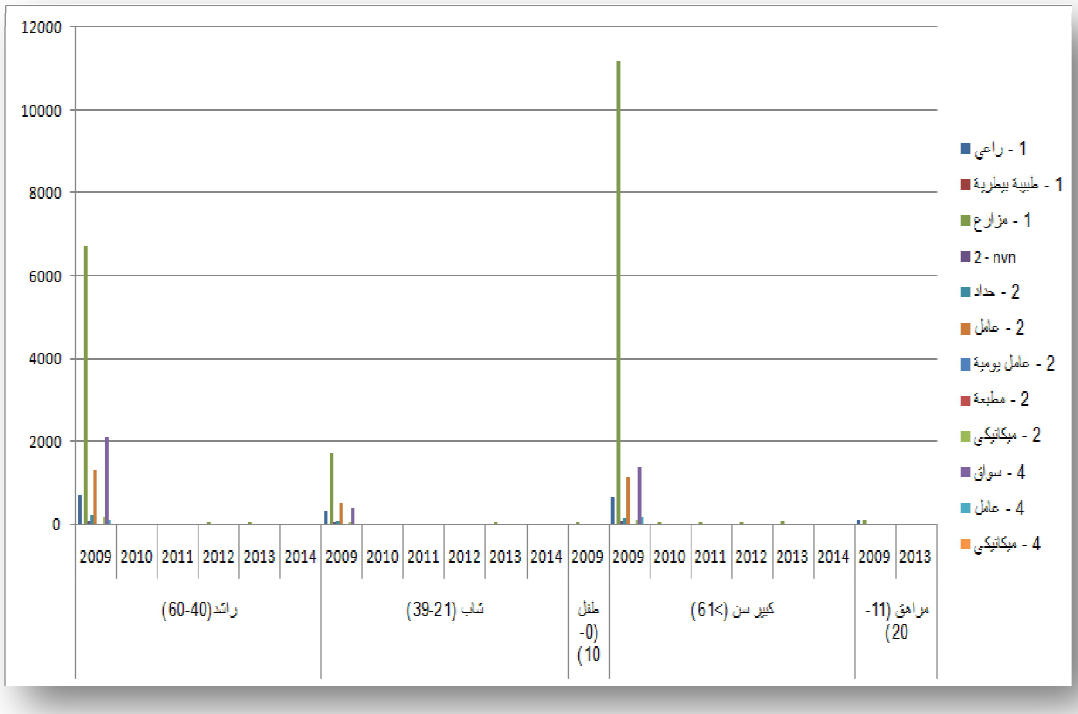


Figure (6.14): Occupation versus Age Range and Admission Date

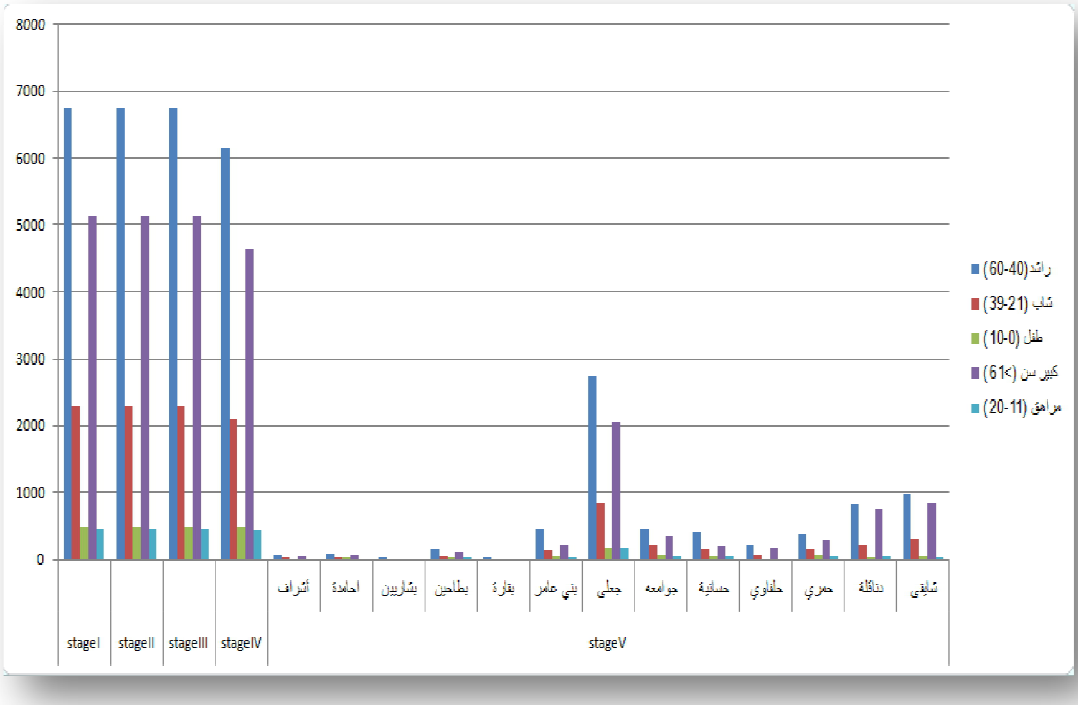


Figure (6.15): Age Range versus Cancer Stage and Tribe

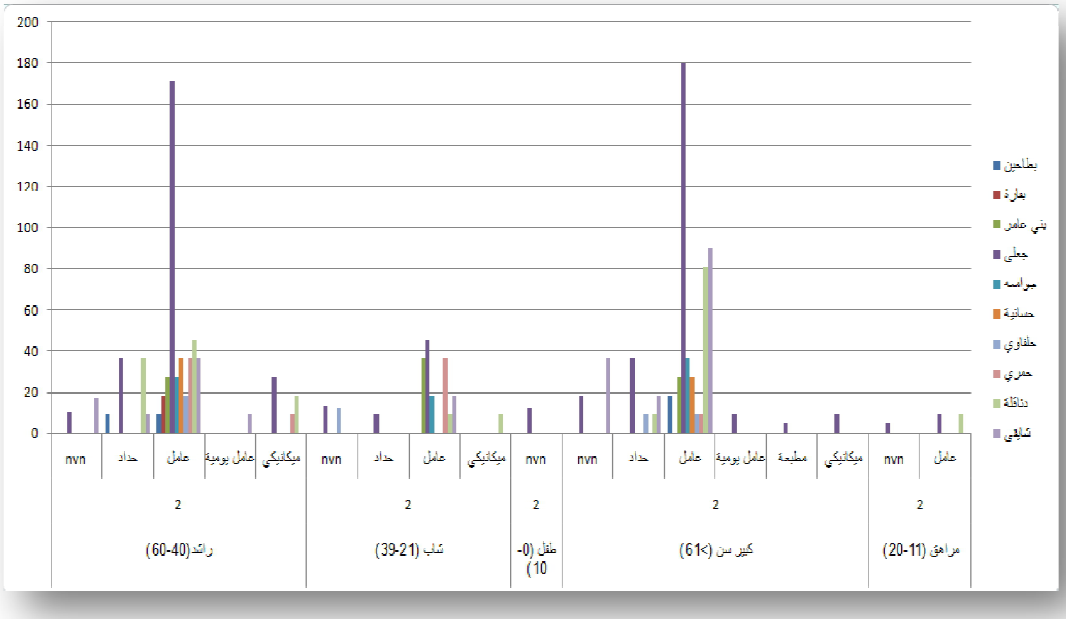
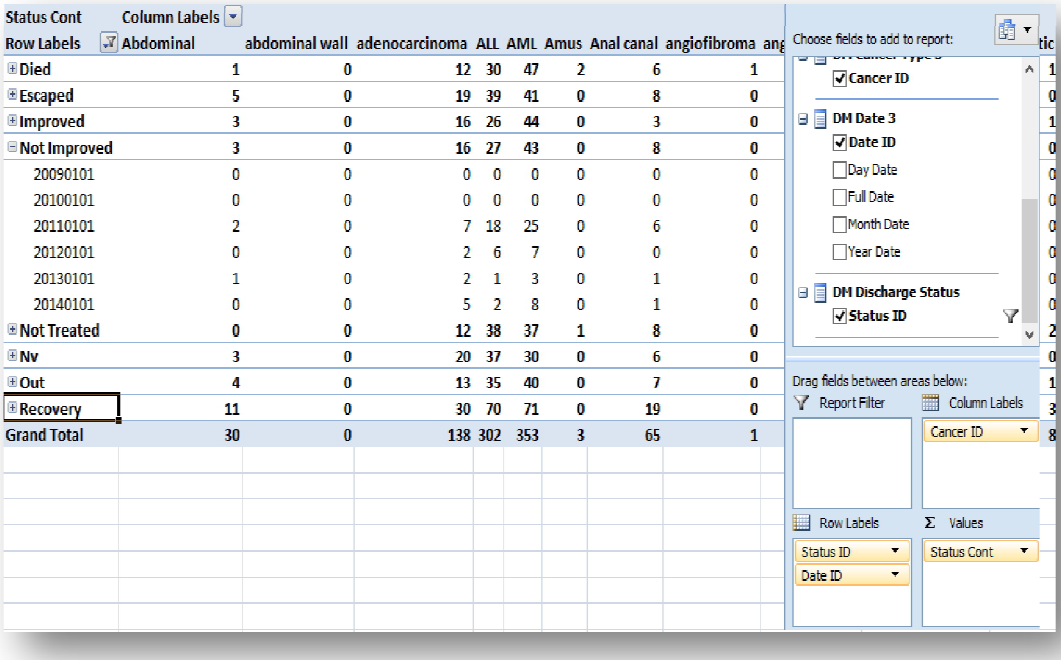


Figure (6.16): Age Range and Occupation versus Tribe

Additionally, the discharge status cube based on cancer type, discharge status, age range and date dimensions are shown in figure 6.16. The output of cube represents a counting number of patients who treated by other treatment after they accomplished the fist or count of patient status when discharged are grouped by state other chosen dimensions. Continuing with the analysis, in figure 6.17 shows the analyst wishes to specify status the patients discharged in specific cancer type distributed by using date of discharge. Also figure 6.18 shows the number of patient's discharge status (improved, not improved, recovery, died and etc) distributed by the age range. Also the number of discharge status of patients whom discharge in special date is identify with the age range as shown in figure 6.19



Status Cont	Column Labels								
Row Labels	Abdominal	abdominal wall	adenocarcinoma	ALL	AML	Amus	Anal canal	angiofibroma	ang
Died	1	0	12	30	47	2	6	1	
Escaped	5	0	19	39	41	0	8	0	
Improved	3	0	16	26	44	0	3	0	
Not Improved	3	0	16	27	43	0	8	0	
20090101	0	0	0	0	0	0	0	0	
20100101	0	0	0	0	0	0	0	0	
20110101	2	0	7	18	25	0	6	0	
20120101	0	0	2	6	7	0	0	0	
20130101	1	0	2	1	3	0	1	0	
20140101	0	0	5	2	8	0	1	0	
Not Treated	0	0	12	38	37	1	8	0	
Nv	3	0	20	37	30	0	6	0	
Out	4	0	13	35	40	0	7	0	
Recovery	11	0	30	70	71	0	19	0	
Grand Total	30	0	138	302	353	3	65	1	

Figure 6.17: The Effect of the Dimensions with Measures in Discharge Status Cube

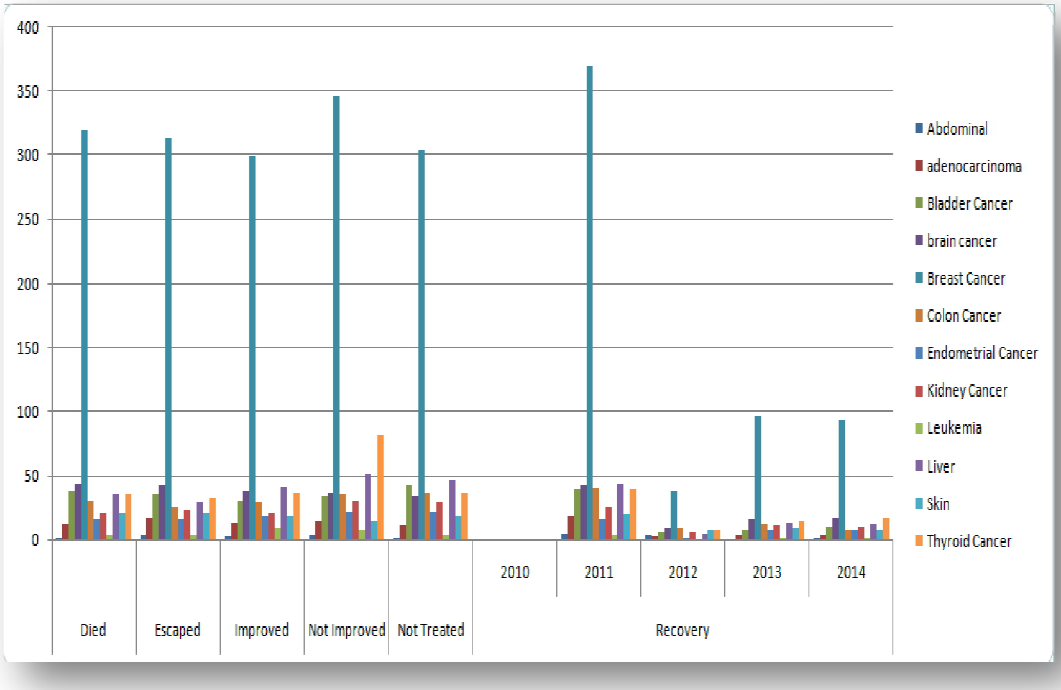


Figure (6.18): Discharge Status of Specify Cancer's Patients According Date Dimension

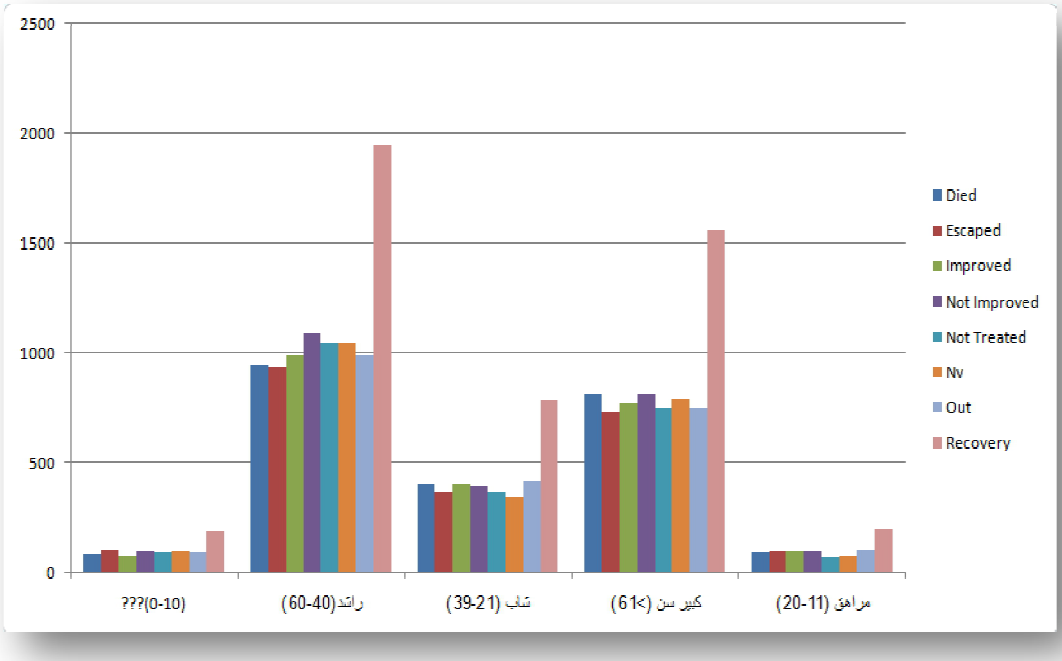


Figure (6.19): The Relationship between Discharge Status and Patients Discharge Status

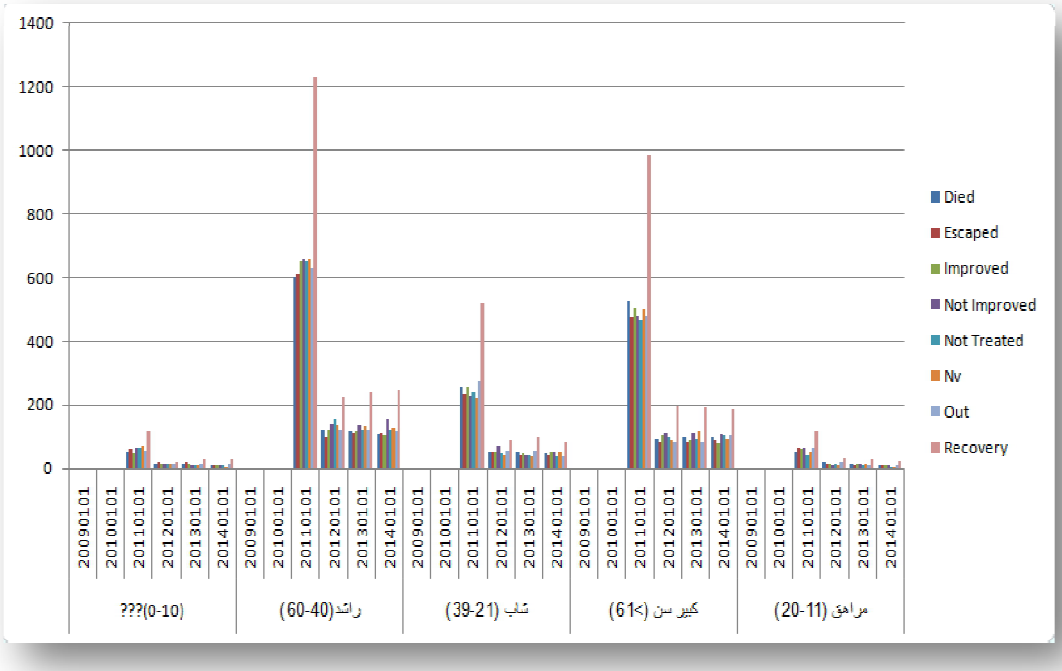
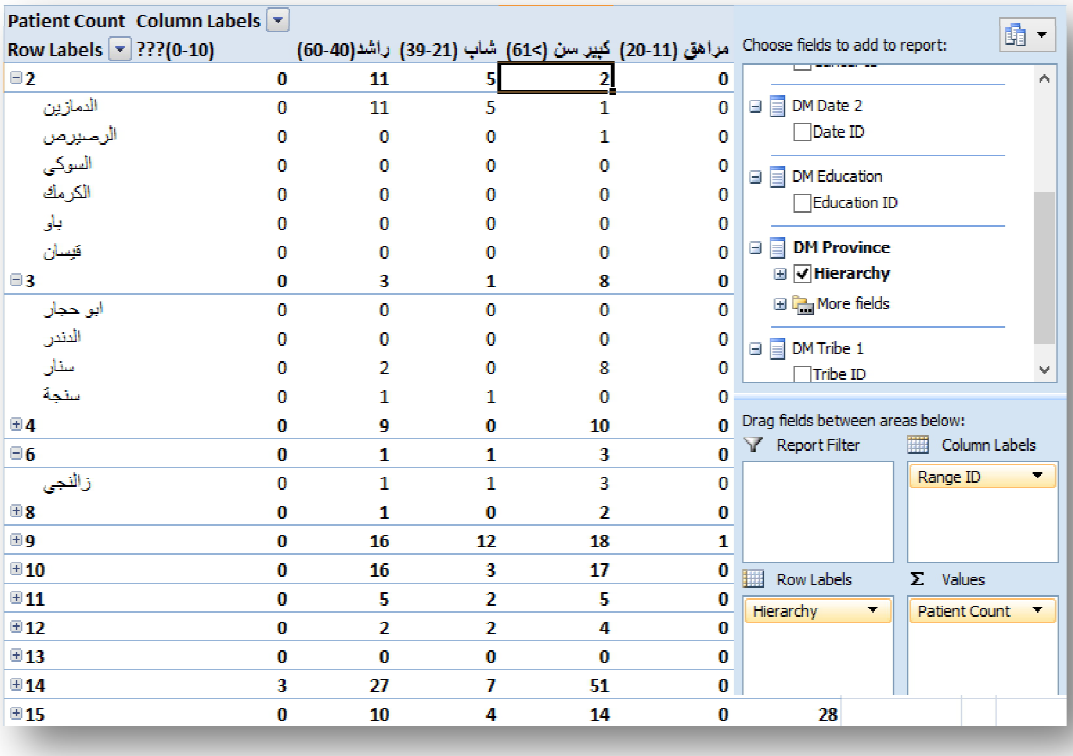


Figure (6.20): Discharge Status of Specify Age Range According Date Dimension

Finally, the location cube based on cancer type, province, tribe, education, age range and date dimensions are shown in figure 6.20. The output of cube represents a geographical location of patients grouped by state other chosen dimension as shown in Figure 6.21, 6.22 and 6.23. Continuing with the same analysis, in figure 6.21 shows the analyst wishes to distribute the patients lived in specific province in special state by using age range. Additional analyses that could be performed with this spreadsheet could include using the education and date dimensions to identify the geographical distribution with the age range as shown in figure 6.22. Also figure 6.23 shows the patients geographical distribution with are range in by specific province in specific state.



Patient Count Column Labels / Row Labels	???(0-10)	راشد(40-60)	شاب (21-39)	كبير سن (61<)	مراهق (11-20)
2	0	11	5	2	0
الدمازين	0	11	5	1	0
الرصيرص	0	0	0	1	0
السوكي	0	0	0	0	0
الكرمك	0	0	0	0	0
باو	0	0	0	0	0
قيسان	0	0	0	0	0
3	0	3	1	8	0
ابو حجار	0	0	0	0	0
الدندر	0	0	0	0	0
سنار	0	2	0	8	0
سنجة	0	1	1	0	0
4	0	9	0	10	0
6	0	1	1	3	0
زالنجي	0	1	1	3	0
8	0	1	0	2	0
9	0	16	12	18	1
10	0	16	3	17	0
11	0	5	2	5	0
12	0	2	2	4	0
13	0	0	0	0	0
14	3	27	7	51	0
15	0	10	4	14	0

Figure 6.21: The Effect of the Dimensions with Measures in Location Cube

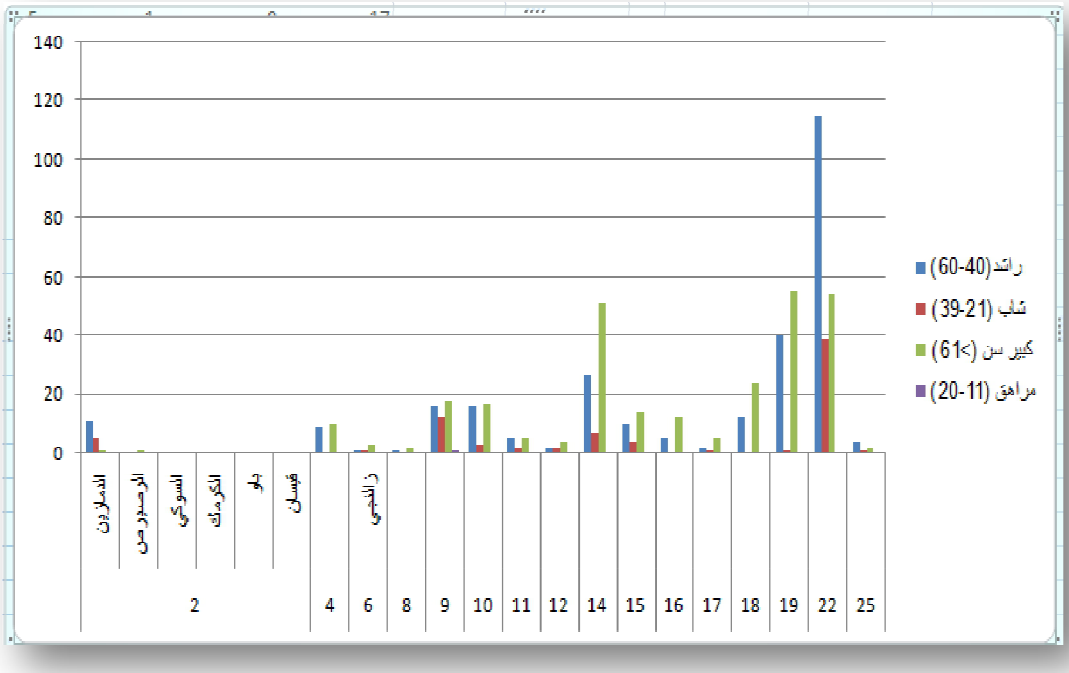


Figure (6.22): Geographical Distribution of Patients VS Age's Range

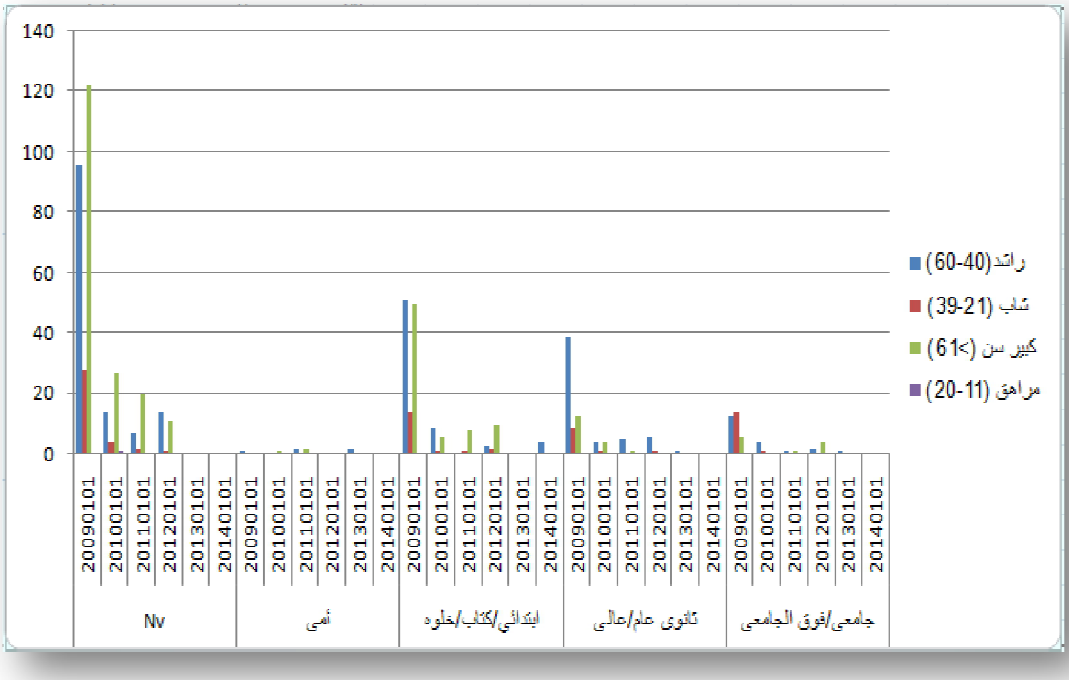


Figure (6.23): Geographical Distribution of Patients Education and Date Admission versus Age's Range

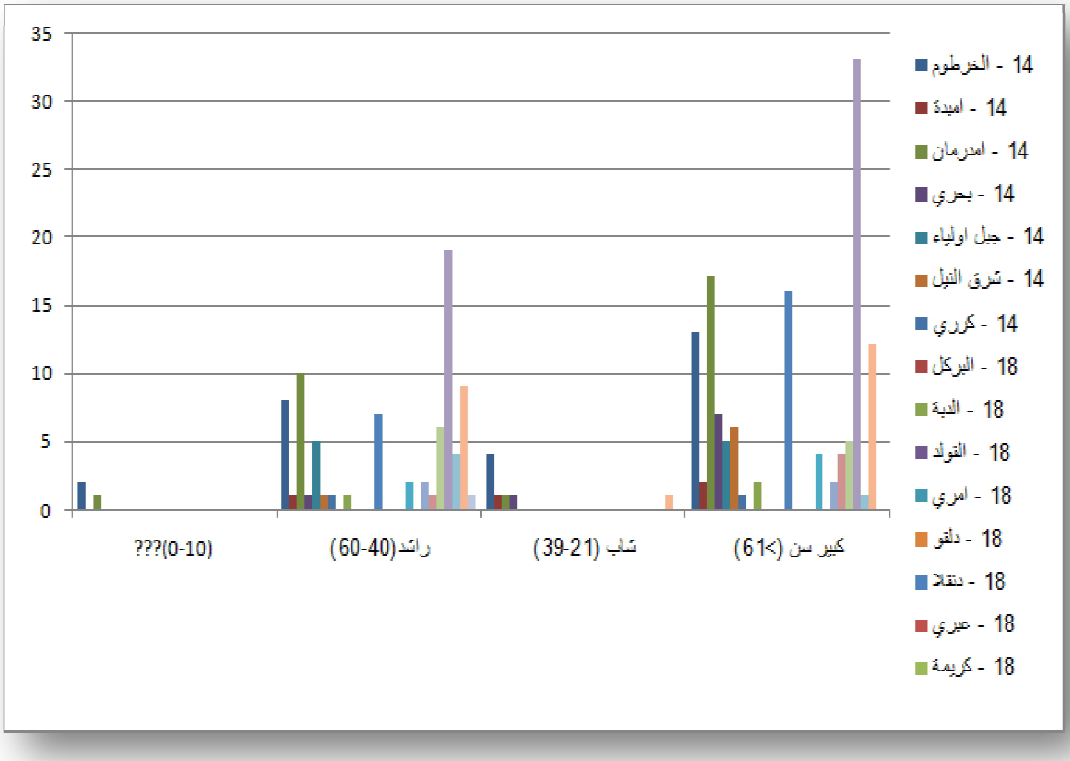


Figure (6.24): Geographical Distribution of Patients (state & province) with Age's Range

6.5 Results

The following sections discuss the results of investigation presented in this chapter. In first the ETL techniques are presented, followed by a presentation of the CDWH technique.

6.5.1 ETL techniques:

The ETL techniques are developed to extract, cleanse, transform and load the data with good quality guarantees into the CDWH. The objectives that conducted are achieved through the developed ETL techniques, which include the following:

1. Designing conceptual model for representing in simplified way the extraction, cleansing, transformation, and loading (ETL) processes, by determining the ETL processes requirements.

2. Improving the quality of data at CDWH by handling the data problems that affect quality of data during the implement of ETL process according to the medical requirements.
3. The flexibility of integrate several types of data sources contain heterogeneous clinical data. The data problems that affect quality at extraction process are identified and handled.
4. Cleanse the data dealing with all data problems that affect the quality of clinical data in staging area to improve the quality of the data before loading them into CDWH.
5. The transformation of the source data format to the target database format makes the data into different storage categories according to the data mining requirements, which handle the mapping relationship between two database tables fields and improves the availability of data.
6. Loading the cleansed data from staging area into CDWH. Each data problems that affect data quality are identified and handled. The main quality issues that were identified are lack of completeness, lack in data accuracy and lack in compatibility.
7. The ETL techniques give a detailed report on the types of errors detected and reported according to the sources of data.

6.5.2 CDWH technique:

The CDWH technique is developed and implemented which, characterized with the following:

1. The structure of the CDWH is designed in a flexible way that it can be extended when needed. Furthermore, the design of CDWH presented functional requirements along with non-functional requirement.
2. The CDWH is built with open source technologies to supports any operating system and hardware platform.

3. The developed CDWH integrate all relevant required medical and clinical data are important to allow follow-up of the treatment of patients, support decision making and improve medical research.
4. The developed CDWH based on OLAP with data mining techniques. The system is powerful because: (a) it discovers hidden patterns in the data, (b) it enhances real-time indicators and discovers bottlenecks and (c) it improves information visualization.
5. Data is presented by pivot table. The pivot table explored and incorporated into the CDWH to enable multidimensional data presentation and querying. Furthermore the CDWH can facilitate analysis by using OLAP tools to answer important physicians and researchers questions such as, retrieved the relationship between patient's occupation and cancer, geographical distribution of cancer, a status of cancer patients discharged, age/sex distribution of cancer.

CHAPTER SEVEN

CONCLUSION, RECOMMENDATIONS AND FUTURE WORKS

7.1 Conclusion

This research work discussed the impact of data integration and data quality for improve decision medical making process in clinical data warehouse (CDWH). CDWH is more complicated than the DWH and produce a set of issues and challenges. However, the integration of a various medical data sources is very complex process, where these data sources are exposed to a lot of quality problems and poor in information. Furthermore, the clinical data is different from business data where the clinical data produce new issues and requirements if not considered, which affect the quality of data.

The research work identified that the usage of DWH technologies in medical field produces new issues and challenges to DWH technologies. These issues and challenges include; clinical data format, business purpose, data integration, data quality, and ETL process issues. Handling these issues and challenges requires determining the medical purpose and requirement are determined in proper ways and integrate the clinical data from various data sources to CDWH. Furthermore the data problems that affect the quality of data are identified and handled through ETL techniques to ensure a high quality of the data in the CDWH.

The first contribution of this research is the investigating of the previous relevant work on data quality and the integration issues. These issues include clinical data integration, clinical data quality, and ETL process. In the clinical data integration issue, the studies concern in designing appropriate integrates approaches to integrate medical information from different healthcare institutions. Furthermore, some studies concern in clinical data quality and integration issues, such as evolution of the workflow from a certain plan to another. In addition to, other studies concern in ETL processes issue such as, ETL model, data

quality, and enhanced ETL model by adding new component to ETL process technique.

The second contribution of this research is designing and developing integration (ETL) techniques which improve quality data in CDWH, by investigating how the data integration techniques efficiency handle data integration and quality problems, showing how the four key activities; extraction, cleansing, transforming, and loading, are performed. However, the medical data consolidated from several source systems and each of these data sources has its distinct set of characteristics. Therefore, the complexity of the medical institution environment issues considered during the process of developing of ETL process. These issues involve clear identification of extracting, cleansing, transformation and loading requirements as well as developing and evaluating ETL techniques. Thus, the design and develop of ETL technique based on the determined requirement of extraction, cleansing, transformation, and loading processes. These developed ETL techniques take four sequential phases to facilitate, manage, and optimize the design and develop of the ETL processes. These phases includes: medical analysis, physical development, implementation and evaluation. Additionally, the ETL techniques are tested on three real data sets and showed that the technique performed well on these sets for metrics.

Finally, the third contribution is designing and developing CDWH technique that integrates medical data from various sources into CDWH Furthermore, CDWH technique provides many benefits such as improved quality, provide rich analysis environment, provide data for clinical research, and improved medical decision making.

7.2 Recommendations and Future Works

The research work may be continued in several areas. Below presents some of the possible areas:

1. The daily growing of the medial dataset size arise a new issues and challenges such as storage and computation. On one hand, these huge data need large storage space and the size of the data grows to exceed the capacity of the storage media. On the other hand, this big data need large computation resources to increase the performance of the medical data processing. The aforementioned issues can be handles by utilizing the distributed computing technologies such as cloud computing to implement the clinical data warehouse.
2. The data integration required transfer the clinical data from different database management systems distributed over several locations. The sizes of the data that transfer between the computers produce a high traffic problem. As the result, the data transferring take long time and will affect the performance of integration process, which need to be consider in future work.
3. This study used heterogeneous datasets belong to different database management systems. However, the current trend of the database is move from relational database to NOSQL. Consequently, NOSQL implement as a part of clinical data warehouse is need further study.
4. The clinical data sources contain sensitive data about the patients. This study handles all the data including the sensitive data and not provides any protection mechanism or technique to keep data secure.

References:

[1] W. H. Inmon, *Building the data warehouse*: John wiley & sons, 2005.

[2] T. Manjunath, R. S. Hegadi, and G. Ravikumar, "Analysis of Data Quality Aspects in DataWarehouse Systems," *IJCSIT) International Journal of Computer Science and Information Technologies,* vol. 2, pp. 477-485, 2010.

[3] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.,* vol. 23, pp. 3-13, 2000.

[4] O. Corporation, "Oracle9i™ SQL Reference. Release 9.2," pp. pp.1777-1780, 2002.

[5] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. of Management Information Systems,* vol. 12, pp. 5-33, 1996.

[6] R. Arora, P. Pahwa, and S. Bansal, "Alliance rules for data warehouse cleansing," in *2009 International Conference on Signal Processing Systems*, 2009, pp. 743-747.

[7] J. Han, M. Kamber, and J. Pei, *Data mining, southeast asia edition: Concepts and techniques*: Morgan kaufmann, 2006.

[8] M. Hua and J. Pei, "DiMaC: a disguised missing data cleaning tool," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 1077-1080.

[9] T. Johnson and T. Dasu, "Data quality and data cleaning: An overview," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003, pp. 681-681.

[10] A. Berson and S. J. Smith, *Data warehousing, data mining, and OLAP*: McGraw-Hill, Inc., 1997.

[11] A. Simitisis, P. Vassiliadis, S. Skiadopoulos, and T. Sellis, "Data warehouse refreshment," 2007.

[12] D. Theodoratos, S. Ligoudistianos, and T. Sellis, "View selection for designing the global data warehouse," *Data & Knowledge Engineering,* vol. 39, pp. 219-240, 2001.

[13] N. C. Institute, "what-is-cancer," *http://www.cancer.gov/cancertopics/cancerlibrary/what-is-cancer,* 12-3-2012.

[14] W. H. Organization. *World Cancer Report 2014*.

[15] F. J. Martel C, Franceschi S, et al "Global burden of cancers attributable to infections in 2008: a review and synthetic analysis," *The Lancet Oncology,* pp. 607-615, 2012.

[16] W. H. Inmon, D. Strauss, and G. Neushloss, *DW 2.0: The Architecture for the Next Generation of Data Warehousing: The Architecture for the Next Generation of Data Warehousing*: Morgan Kaufmann, 2010.

[17] W. H. Inmon, "EIS and the data warehouse: a simple approach to building an effective foundation for EIS," *Database Programming & Design,* vol. 5, pp. 70-73, 1992.

[18] J. Widom, "Research problems in data warehousing," in *Proceedings of the fourth international conference on Information and knowledge management*, 1995, pp. 25-30.

[19] J. Widom, "Special issue on materialized views and data warehousing," *IEEE Bulletin on Data Engineering,* vol. 18, 1995.

[20] T. R. Sahama 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, p. 68.
[21] S. P. Shah, Y. Huang, T. Xu, M. M. Yuen, J. Ling, and B. F. Ouellette, "Atlas–a data warehouse for integrative bioinformatics," *BMC bioinformatics,* vol. 6, p. 34, 2005.
[22] C. Schönbach, P. Kowalski-Saunders, and V. Brusic, "Data warehousing in molecular biology," *Briefings in Bioinformatics,* vol. 1, pp. 190-198, 2000.
[23] O. Ritter, P. Kocab, M. Senger, D. Wolf, and S. Suhai, "Prototype implementation of the integrated genomic database," *Computers and Biomedical Research,* vol. 27, pp. 97-115, 1994.
[24] D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B. B. Briggs*, et al.*, "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles," *Neoplasia,* vol. 9, pp. 166-180, 2007.
[25] H. Hu, H. Brzeski, J. Hutchins, M. Ramaraj, L. Qu, R. Xiong*, et al.*, "Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research," *Pharmacogenomics,* vol. 5, pp. 933-941, 2004.
[26] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp*, et al.*, "EnsMart: a generic system for fast and flexible access to biological data," *Genome research,* vol. 14, pp. 160-169, 2004.
[27] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. W. Stringer-Calvert, J. D. Tenenbaum*, et al.*, "BioWarehouse: a bioinformatics database warehouse toolkit," *BMC bioinformatics,* vol. 7, p. 170, 2006.
[28] S. R. Gardner, "BUILDING the Data Warehouse," *Communications of the ACM,* vol. 41, p. 53, 1998.
[29] F. R. McFadden and J. A. Hoffer, *Modern Database Management: Basic Concepts of Data Warehousing*: Addison-Wesley, 1993.
[30] M. P. Gupta, "Data Warehouse and Data Mining Technology Data Warehouse and Data Mining Technology-A study of its A study of its impact, relevance and need in Enterpr relevance and need in Enterpr relevance and need in Enterprises of Delhi NCR region ises of Delhi NCR region ises of Delhi NCR region."
[31] W. H. Inmon, C. Imhoff, and R. Sousa, *Corporate information factory*: John Wiley & Sons, 2002.
[32] R. Kimball and M. Ross, "The data warehouse toolkit: the complete guide to dimensional modelling," *Nachdr.]. New York [ua]: Wiley,* 2002.
[33] R. Kimball and J. Caserta, "The data warehouse ETL toolkit: practical techniques for extracting," *Cleaning, Conforming, and Delivering Data,* p. 528, 2004.
[34] H. J. Watson, D. L. Goodhue, and B. H. Wixom, "The benefits of data warehousing: why some organizations realize exceptional payoffs," *Information & Management,* vol. 39, pp. 491-502, 2002.
[35] J. Nealon, W. Rahayu, and E. Pardede, "Improving clinical data warehouse performance via a windowing data structure architecture," in *Computational Science and Its Applications, 2009. ICCSA'09. International Conference on*, 2009, pp. 243-253.

[36] J. G. DeWitt and P. M. Hampton, "Development of a data warehouse at an academic health system: knowing a place for the first time," *Academic Medicine,* vol. 80, pp. 1019-1025, 2005.
[37] T. Rajala, S. Savio, J. Penttinen, P. Dastidar, M. Kähönen, H. Eskola*, et al.*, "Development of a Research Dedicated Archival System (TARAS) in a University Hospital," *Journal of digital imaging,* vol. 24, pp. 864-873, 2011.
[38] M. Silver, T. Sakata, H.-C. Su, C. Herman, S. B. Dolins, and M. J. O Shea, "Case study: how to apply data mining techniques in a healthcare data warehouse," *Journal of healthcare information management,* vol. 15, pp. 155-164, 2001.
[39] M. F. Wisniewski, P. Kieszkowski, B. M. Zagorski, W. E. Trick, M. Sommers, R. A. Weinstein*, et al.*, "Development of a clinical data warehouse for hospital infection control," *Journal of the American Medical Informatics Association,* vol. 10, pp. 454-462, 2003.
[40] J. S. Einbinder and K. Scully, "Using a clinical data repository to estimate the frequency and costs of adverse drug events," in *Proceedings of the AMIA Symposium*, 2001, p. 154.
[41] R. D. Aller, "The Clinical Laboratory Data Warehouse An Overlooked Diamond Mine," *American journal of clinical pathology,* vol. 120, pp. 817-819, 2003.
[42] C. G. Chute, S. A. Beck, T. B. Fisk, and D. N. Mohr, "The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data," *Journal of the American Medical Informatics Association,* vol. 17, pp. 131-135, 2010.
[43] R. Kimball, *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*: John Wiley & Sons, 1998.
[44] N. Esfandiary, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge Discovery in Medicine: Current Issue and Future Trend," *Expert Systems with Applications,* vol. 41, pp. 4434-4463, 2014.
[45] M. Banek, A. M. Tjoa, and N. Stolba, "Integrating different grain levels in a medical data warehouse federation," in *Data Warehousing and Knowledge Discovery*, ed: Springer, 2006, pp. 185-194.
[46] T. B. Pedersen and C. S. Jensen, "Research issues in clinical data warehousing," in *Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference on*, 1998, pp. 43-52.
[47] B. Inmon, "Data warehousing in a healthcare environment," *The Data Administration Newsletter-TDAN. com,* 2007.
[48] T. Y. Wah and O. S. Sim, "Development of a data warehouse for lymphoma cancer diagnosis and treatment decision support," *WSEAS Transactions on Information Science and Applications,* vol. 6, pp. 530-543, 2009.
[49] Y. Zhu and J. Guo, "A kind of data warehouse in community healthcare service system," in *Services Systems and Services Management, 2005. Proceedings of ICSSSM'05. 2005 International Conference on*, 2005, pp. 1408-1413.
[50] X. Zhou, S. Chen, B. Liu, R. Zhang, Y. Wang, P. Li*, et al.*, "Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support," *Artificial Intelligence in Medicine,* vol. 48, pp. 139-152, 2010.
[51] A. K. Hamoud and T. A. Obaid, "Building Data Warehouse for Diseases Registry: First step for Clinical Data Warehouse."

[52] J. E. Bekelman, J. A. Deye, B. Vikram, S. M. Bentzen, D. Bruner, W. J. Curran Jr*, et al.*, "Redesigning radiotherapy quality assurance: opportunities to develop an efficient, evidence-based system to support clinical trials—report of the National Cancer Institute Work Group on Radiotherapy Quality Assurance," *International Journal of Radiation Oncology* Biology* Physics,* vol. 83, pp. 782-790, 2012.
[53] J. R. Schubart and J. S. Einbinder, "Evaluation of a data warehouse in an academic health sciences center," *International Journal of Medical Informatics,* vol. 60, pp. 319-333, 2000.
[54] R. L. Leitheiser, "Data quality in health care data warehouse environments," in *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on*, 2001, p. 10 pp.
[55] W. Reed, S. Jor, and R. Bjugn, "How can clinical biobanks and patient information be adapted for research–Establishing a hospital based data warehouse solution," *Norsk epidemiologi,* vol. 21, 2012.
[56] D. J. Berndt, J. W. Fisher, A. R. Hevner, and J. Studnicki, "Healthcare data warehousing and quality assurance," *Computer,* vol. 34, pp. 56-65, 2001.
[57] H. Xueqin, C. Meng, and C. Bing, "Research on Data Quality of Chinese Medicine Scientific Data," *World Science and Technology,* vol. 11, pp. 589-592, 2010.
[58] N. Kerdprasop and K. Kerdprasop, "Higher Order Programming to Mine Knowledge for a Modern Medical Expert System," *International Journal of Computer Science Issues (IJCSI),* vol. 8, 2011.
[59] R. Zheng, H. Jin, Q. Zhang, Y. Liu, and P. Chu, "Heterogeneous medical data share and integration on grid," in *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, 2008, pp. 905-909.
[60] W. J. Lynch and J. E. Ross Jr, "Medical records, documentation, tracking and order entry system," ed: Google Patents, 1998.
[61] S. Nugawela and T. R. Sahama, "Clinical data integration approach using SAS clinical data integration server (CDI) tools," *Health Informatics: Transforming Healthcare with Technology,* pp. 119-123, 2011.
[62] A. El Fadly, B. Rance, N. Lucas, C. Mead, G. Chatellier, P.-Y. Lastic*, et al.*, "Integrating clinical research with the healthcare enterprise: from the RE-USE project to the EHR4CR platform," *Journal of biomedical informatics,* vol. 44, pp. S94-S102, 2011.
[63] A. Baudot, G. Gomez-Lopez, and A. Valencia, "Translational disease interpretation with molecular networks," *Genome biology,* vol. 10, p. 221, 2009.
[64] T. Hernandez and S. Kambhampati, "Integration of biological sources: current systems and challenges ahead," *ACM SIgmod Record,* vol. 33, pp. 51-60, 2004.
[65] P. Chountas, I. Petrounias, K. Atanassov, V. Kodogiannis, and E. El-Darzi, "Representation & querying of temporal conflict," in *Flexible Query Answering Systems*, ed: Springer, 2002, pp. 112-123.
[66] K. B. Hass, R. Vander Horst, K. Ziemski, and L. Lindbergh, *From Analyst to Leader: Elevating the Role of the Business Analyst*: Management Concepts Press, 2007.
[67] G. W. Gray, "Challenges of building clinical data analysis solutions," *Journal of critical care,* vol. 19, pp. 264-270, 2004.
[68] S. Bianchi, A. Burla, C. Conti, A. Farkash, C. Kent, Y. Maman*, et al.*, "Biomedical data integration–capturing similarities while preserving

disparities," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009.* , 2009, pp. 4654-4657.
[69] M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp. 233-246.
[70] P. Lane, V. Schupmann, and I. Stuart, "Oracle database data warehousing guide, 10g release 2 (10.2)," *Oracle Corporation, Redwood City, CA,* 2005.
[71] M. Shepherd, "Challenges in health informatics," in *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, 2007, pp. 135-135.
[72] R. Kimball, L. Reeves, M. Ross, and W. Thornthwaite, "The Data Warehouse Lifecycle Toolkit. 1998," 1998.
[73] R. J. Kachur, *Data Warehouse Managment Handbook*: Prentice Hall PTR, 1999.
[74] X. Zhou, B. Liu, Y. Wang, R. Zhang, P. Li, S. Chen*, et al.*, "Building clinical data warehouse for traditional Chinese medicine knowledge discovery," in *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, 2008, pp. 615-620.
[75] M. Butt and M. Zaman, "Data Quality Tools for Data Warehousing: Enterprise Case Study," *IOSR Journal of Engineering,* vol. 3, pp. 75-76, 2013.
[76] E. Vannan, "Quality Data–An Improbable Dream," *Educause Quarterly,* vol. 1, pp. 56-58, 2001.
[77] X. Pan, X. Zhou, H. Song, R. Zhang, and T. Zhang, "Enhanced data extraction, transforming and loading processing for traditional Chinese medicine clinical data warehouse," in *e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on*, 2012, pp. 57-61.
[78] N. Anand and M. Kumar, "An Overview on Data Quality Issues at Data Staging ETL," in *Int. Conf. on Advances in Signal Processing and Communication*, 2013.
[79] R. Singh and K. Singh, "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing," *International Journal of Computer Science Issues (IJCSI),* vol. 7, 2010.
[80] P. Z. Yeh and C. A. Puri, "An efficient and robust approach for discovering data quality rules," in *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*, 2010, pp. 248-255.
[81] R. S. Craig, J. A. Vivona, and D. Bercovitch, *Microsoft data warehousing: building distributed decision support systems*: John Wiley & Sons, Inc., 1999.
[82] U. Dayal, M. Castellanos, A. Simitsis, and K. Wilkinson, "Data integration flows for business intelligence," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, 2009, pp. 1-11.
[83] G. K. Tayi and D. P. Ballou, "Examining data quality," *Communications of the ACM,* vol. 41, pp. 54-57, 1998.
[84] S. T. March and A. R. Hevner, "Integrated decision support systems: A data warehousing perspective," *Decision Support Systems,* vol. 43, pp. 1031-1043, 2007.
[85] M. D. Solomon, "Ensuring a successful data warehouse initiative," *Information Systems Management,* vol. 22, pp. 26-36, 2005.
[86] F. N. Savitri and H. Laksmiwati, "Study of localized data cleansing process for ETL performance improvement in independent datamart," in *Electrical*

Engineering and Informatics (ICEEI), 2011 International Conference on, 2011, pp. 1-6.
[87] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. El Bastawissy, "A proposed model for data warehouse ETL processes," *Journal of King Saud University-Computer and Information Sciences,* vol. 23, pp. 91-104, 2011.
[88] R. Kimball and J. Caserta, *The data warehouse ETL toolkit*: John Wiley & Sons, 2004.
[89] M. Demarest, "The politics of data warehousing," *Retrieved January,* vol. 2, p. 2010, 1997.
[90] S. Alison, G. Robinson, and P. Terhune, "Oracle9i Warehouse Builder User's Guide," ed: Oracle Corp., 2000.
[91] B. Inmon, "The data warehouse budget," *DM Review Magazine, January,* 1997.
[92] A. Simitsis and P. Vassiliadis, "A method for the mapping of conceptual designs to logical blueprints for ETL processes," *Decision Support Systems,* vol. 45, pp. 22-40, 2008.
[93] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Conceptual modeling for ETL processes," in *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, 2002, pp. 14-21.
[94] B. Scalzo, *Oracle DBA guide to data warehousing and star schemas*: Prentice Hall Professional, 2003.
[95] A. Simitsis, K. Wilkinson, M. Castellanos, and U. Dayal, "QoX-driven ETL design: reducing the cost of ETL consulting engagements," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 2009, pp. 953-960.
[96] P. Chountas and V. Kodogiannis, "Development of a clinical data warehouse," in *Design of Reliable Communication Networks, 2003.(DRCN 2003). Proceedings. Fourth International Workshop on*, 2004, pp. 8-14.
[97] S. Mummana and R. kiran Rompella, "An Empirical Data Cleaning Technique for CFDs," *International Journal of Engineering Trends and Technology (IJETT),* vol. 4, pp. 3730- 3735, 2013.
[98] T. L. de Andrade, R. Gratao de Souza, M. Babini, and C. R. Valêncio, "Optimization of Algorithm to Identification of Duplicate Tuples through Similarity Phonetic Based on Multithreading," in *Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2011 12th International Conference on*, 2011, pp. 299-304.
[99] N. B. Szirbik, C. Pelletier, and T. Chaussalet, "Six methodological steps to build medical data warehouses for research," *International Journal of Medical Informatics,* vol. 75, pp. 683-691, 2006.
[100] D. Lorentz, J. Gregoire, and S. Abraham, *Oracle9i: SQL Reference, Release 2 (9.2)*: Oracle Corporation, 2002.
[101] A. Ta'a and M. S. Abdullah, "Goal-ontology approach for modeling and designing ETL processes," *Procedia Computer Science,* vol. 3, pp. 942-948, 2011.
[102] P. Vassiliadis, "Data warehouse modeling and quality issues," *National Technical University of Athens Zographou, Athens, GREECE,* 2000.
[103] M. Lenzerini, Y. Vassiliou, P. Vassiliadis, and M. Jarke, *Fundamentals of data warehouses*: Springer, 2003.
[104] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM,* vol. 39, pp. 86-95, 1996.

[105] C. Bontempo and G. Zagelow, "The IBM data warehouse architecture," *Communications of the ACM,* vol. 41, pp. 38-48, 1998.
[106] L. Hadley, "Developing a data warehouse architecture," ed: Vienna University of Technology, 2008.
[107] V. Rainardi, *Building a data warehouse: with examples in SQL Server*: John Wiley & Sons, 2008.
[108] W. J. Labio, D. Quass, and B. Adelberg, "Physical database design for data warehouses," in *Data Engineering, 1997. Proceedings. 13th International Conference on*, 1997, pp. 277-288.
[109] B. Devlin and L. D. Cote, *Data warehouse: from architecture to implementation*: Addison-Wesley Longman Publishing Co., Inc., 1996.
[110] S. Demigha, "A Data Warehouse System to Help Assist Breast Cancer Screening in Diagnosis, Education and Research," in *CSA the Second International Conference on Computer Science and its Applications, IEEE, Jeju, Korea (South),(1-6),(Dec 2009)*, 2009.
[111] H. Hackl, G. Stocker, P. Charoentong, B. Mlecnik, G. Bindea, J. Galon*, et al.*, "Information technology solutions for integration of biomolecular and clinical data in the identification of new cancer biomarkers and targets for therapy," *Pharmacology & therapeutics,* vol. 128, pp. 488-498, 2010.
[112] E. Rahm, T. Kirsten, and J. Lange, "The GeWare data warehouse platform for the analysis of molecular-biological and clinical data," *Journal of Integrative Bioinformatics,* vol. 4, p. 47, 2007.
[113] B. K. Seah, "An application of a healthcare data warehouse system," in *Innovative Computing Technology (INTECH), 2013 Third International Conference on*, 2013, pp. 269-273.
[114] D. C. Ramick, "Data warehousing in disease management programs," *Journal of Healthcare Information Management,* vol. 15, pp. 99-106, 2001.
[115] A. Kabiri, F. Wadjinny, and D. Chiadmi, "Towards a Framework for Conceptual Modeling of ETL Processes," in *Innovative Computing Technology*, ed: Springer, 2011, pp. 146-160.
[116] M. Blechner, R. K. Saripalle, and S. A. Demurjian, "A proposed star schema and extraction process to enhance the collection of contextual & semantic information for clinical research data warehouses," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, 2012, pp. 798-805.
[117] A. Simitsis, "Modeling and managing ETL processes," in *VLDB PhD Workshop*, 2003.
[118] S. Dupor and V. Jovanović, "An approach to conceptual modelling of ETL processes."
[119] M. M. Awad, M. S. Abdullah, and A. B. M. Ali, "Extending ETL framework using service oriented architecture," *Procedia Computer Science,* vol. 3, pp. 110-114, 2011.
[120] R. O. Mohammed and S. A. Talab, "Clinical Data Warehouse Issues and Challenges," *International Journal of u-and e-Service, Science and Technology,* vol. 7, pp. 251-262, 2014.
[121] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse," in *Proceedings of the AMIA annual fall symposium*, 1997, p. 101.