International Journal of u-and e-Service, Science and Technology
Vol.7, No.5 (2014), pp.251-262
http://dx.doi.org/10.14257/ijunnesst.2014.7.5.22

Clinical Data Warehouse Issues and Challenges

Razi O. Mohammed[1] and Samani A. Talab[2]

[1]*Department of Computer Science, Shendi University, Sudan*
[2]*Department of Computer Science, Elnileen University, Sudan*
[1]*Razi190@ush.sd,* [2]*Samani_talab@hotmail.com,*

Abstract

The Clinical Data Warehouse is a result of utilizing data warehouse technology in medical field. The clinical data usually collect from various sources (Clinical information system, and Patient data management system), and stored into data warehouse to be analyzed to make better use of their clinical data in order to Support decision making. The diversity of the nature of clinical data from other business data produces several challenges. These challenges include the clinical data format, business analysis, data integration, data quality, and ETL process. The paper is discussing the issues and challenges to address developing of a successful Data Ware Housing for medical organization, which provide a rich knowledge environment to support effective decision making and support research work. The paper concluded significant handling of these issues which affect the development phase of the Clinical Data Warehouse systems.

Keyword: *Data Warehouse (DWH), Clinical Data warehousing (CDWH), Data quality, ETL process*

1. Introduction

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 [1]. The DWH terminology as defined in [2] is "a copy of transaction data specifically structured for query and analysis". While W.H. Inman in [3] defined a DWH as "a subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions". The DWH technology continued providing function and speed for decision making and researches.

The DWH integrate data from two or more operational system, which exist on one or several organizations. The integration process includes three steps [4]: 1) developing a unified model that can accommodate all information from single databases, 2) transferring the data into developed model before loading them into the data warehouse, and 3) extract the data from the source databases and integrate it in one environment and provide access to obtain the required knowledge which the individual sources cannot provide it. However, the integration steps 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 tools to collect, analyze, clean and present information[5]. The DWHs technology provides many benefits that include; enhancing the business intelligence and query performance, improving data quantity and quality, time-saving for users, and support the decision-making[6]. On the other hand, clinical fields are indeed becoming a very attractive research domain for Computer Science in general and DWH in particular. The CDWH (CDWH) is integrating medical data from various operational medical and administrative systems. The CDWH supports research, reporting, and study planning, and improves the

ISSN: 2005-4246 IJUNESST

value of decision making and timely process intervention [7-9]. Furthermore the CDWH facilitates efficient storage, enhances timely analysis and increases the quality of real time decision making processes[10].

The presentation of the study is organized as follows. Section 2 review related work in clinical data warehouse. Then, Section 3 is discussing clinical data warehouse issues and challenges. Section 4 the discussion. Finally Section 5 concludes the study.

2. Related work:

This section provides brief overview of previous work efforts on the CDWH that are relevant to the work in this study. A CDWH platform with OnLine Analysis Processing (OLAP) developed by Boon Keong Seah [11]. The study proposed cleansing methodology to meet CDWH needs. The developed CDWH based on the following steps: analyze the business requirement, developing of DWH modeling, developing of ETL process, indexing data model, encrypting the dimensions, and developing of OLAP analysis. Erhard Rahm, et al [12] presented a CDWH platform for the integrated analysis of clinical information, microarray data and annotations. The proposed approach studies requirements, present CDWH architecture, develop model to integrate clinical data, develop ETL process, and performs reporting and statistical analysis. Denise C. Ramick[13], presents the techniques of CDWH, and discuss 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 expand 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,

These proposed Clinical data ware technology aims to treat 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. But they do not describe how to efficiently dealing with data integration issues (such data extraction, data cleaning, and data transformations, data loading issues). Furthermore these proposed structures do not discuss how to efficiently dealing with data quality issues.

3. Clinical data warehouse (CDWH) issues and challenges:

A CDWH is a DWH tailored for the needs of users in a clinical environment, combing data from various medical database and cleanse medical data to form a centralized data repository to answer the informational needs of all clinical users and support medical decision making[14].The medical data gathered in the healthcare process, this 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 associated with the patient's care[14]. Using CDWH technologies produced new issues and challenges. However, the medical information systems is rich with data, but it also exposed to a lot of quality problems [15] and poor in information [16]. Therefore, several challenges and requirements associated with utilizing of DWH technique in medical domain are produced. The challenges include: the clinical data format, business analysis, data integration, data quality and ETL process technique.

3.1. The Clinical Data Format:

The usage of CDWH technology aims to determining the relationships in clinical data, discovering disease trends, evaluate the performance of different treatments protocols used, support measuring and improving patient outcomes, and provide information to users in areas ranging from research to management. The medical data collect during the regular day-to-day events and store in various systems that include, statistical information system, medical information system, and Laboratory information system and so on[17] , the most important component of medical illustrated in figure (1). However, the clinical data store in various medical systems during the patient visiting times. These types of data include[18]:

- Demographic Information: Information collects once to provide rich data analysis environment.
- Clinical Information: Information about patient's life habits, which use to enhance the data analysis Capabilities.
- Diagnosis information: Describing the diagnosis process.
- Treatments Information: Information about treatment process that involves treatment type, treatment procedure, and treatment risk information.
- Laboratory information: is a laboratory test results.

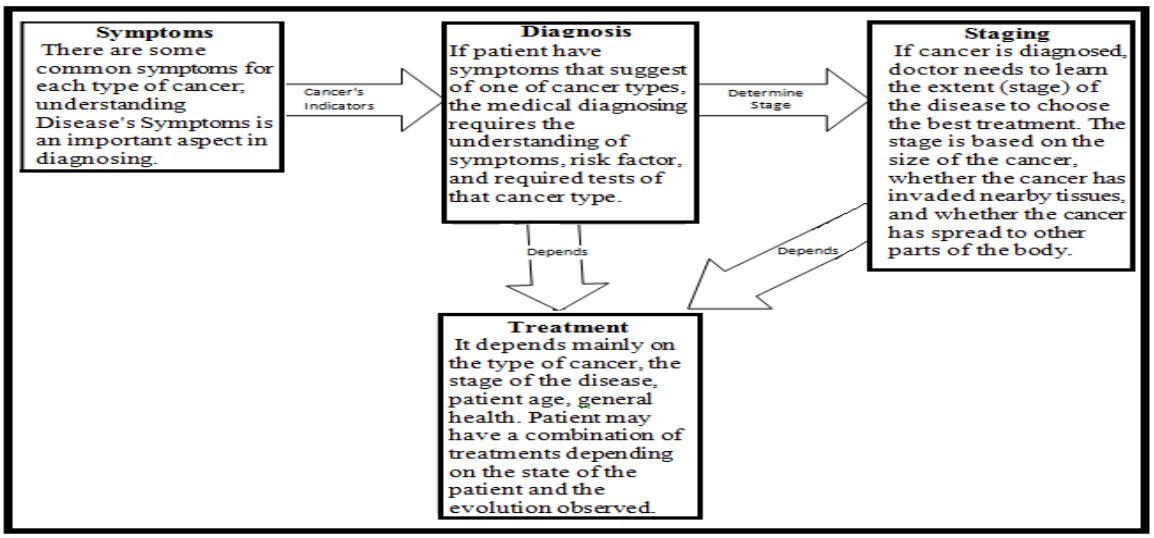

Figure 1. The Most Important Components of Medical Processes

The nature of medical data produces new issues and challenges to DWH technologies. Handling the issues and challenges need to provide the following requirements:

1. The medical data contain personal information that require ethical and legal constrains as explained in[19].
2. Medical information systems are of sensitive nature, diverse storage formats, and inherent privacy issues as reported by berndt in [20].
3. Clinical systems contain accumulate substantial amounts of data about patients with the associated clinical conditions and treatment details as discussed in [10]. 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 as reported in[21]. Consequently, mining these types of data require transformation mechanism that develops 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 are complex and poorly characterized mathematically as explained in [10].

3.2. Business Analysis:

Business analysis is identifying business purpose and determining solutions to business problems[22]. One of the most important aspects of developing a CDWH is to define business purpose. However, the CDWH does not achieve its objectives without clearly defining the business purpose[23]. Furthermore, the discussion of the business 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. The phases are composed of four phases as illustrated in the followings:

- The requirements are gathered in order to understand the purpose of the CDWH, problems domain and to identify the suitable data model that will be used. Determining and gathering the requirements must be done in proper ways, which state the CDWH value and derive the architecture of the CDWH.
- The requirements are further analysis and investigate to determine the data integration problems. This followed by producing an initial dimensional model that showing facts, measures, dimension keys, and dimension hierarchies. Dimension hierarchies can include parallel hierarchical paths.
- Validity of the model is assessed to realize medical objectives and to ensure medical goals and needs which are clearly understood, and the CDWH architecture is designed as per the business requirements.
- Database is planned to be stored on a multidimensional database, showing all elements of the model and their properties. Detailed dimensional models can further be extended and optimized.

The CDWH development must meet some functional requirements in order to maintain data integration in CDWH. These requirements include:

1. Understanding the business purpose, requirements and constraints.
2. Determining medical objectives and needs.
3. Determining the business rules.
4. Determining the suitable model that supports data analysis.
5. Identifying the Data sources of the required data, and performing the sizing of the model.
6. Proving a mechanism that response the queries related to healthcare.

3.3. 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[24]. This large volume of data is integrated, rearranged and consolidated to provide a unified view to analyze the data [25]. These integrated data are not yet turned into useful knowledge due to the lack of efficient analysis tools [26], also the lack of standardization between institutions which makes gathering data difficult[27]. Therefore, the data integration is an important issue in developing CDWH [25]. However, the integration process is time consuming and laborious tasks to separately access and integrates reliably. On other hand the data integration becomes significant issues in situations of developing a CDWH due to the complexity of the hospital environment such as various care practices, and data types and definitions. Additionally, the clinical data integrate from various medical

information systems. These medical systems 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:

1. Developing enhanced integration framework to combine heterogeneous medical data sources to CDWH.
2. Providing a mechanism to integrate medical data from various 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.

3.4. Data Quality:

Data quality is an essential characteristic that determines the reliability of data for analysis, making decisions and planning[28]. The acceptable data quality in the medical field is critical issue to the reliability of medical decision making and research environment.

Quality of data is achieved we when require (useful) data that exactly meet the specific needs stored in common format required by CDWH without data quality problems. However, poor data quality can occur along several dimensions[29]. These dimensions of data quality attribute as shown in figure (2) include; intrinsic (accuracy, consistency, reliability, integrity, and redundancy dimension), accessibility (availability dimension), contextual (relevancy, freshness, validity, completeness, and Scalability dimension), and representational (business purpose understandability and data sources understandability dimension) [28, 30, 31]. Furthermore, data quality problems produce at various stages of CDWH development; data integration & data profiling, Data staging and ETL, and DWH modeling & schema design[29, 30]. Additionally, these data quality problems must be determined and solved to enhance the quality of data. Handling medical data quality issues and challenges need to provide the following requirements:

1. Needs to lay down a strong mechanism to manage medical data quality.
2. Defining levels of data quality that 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.

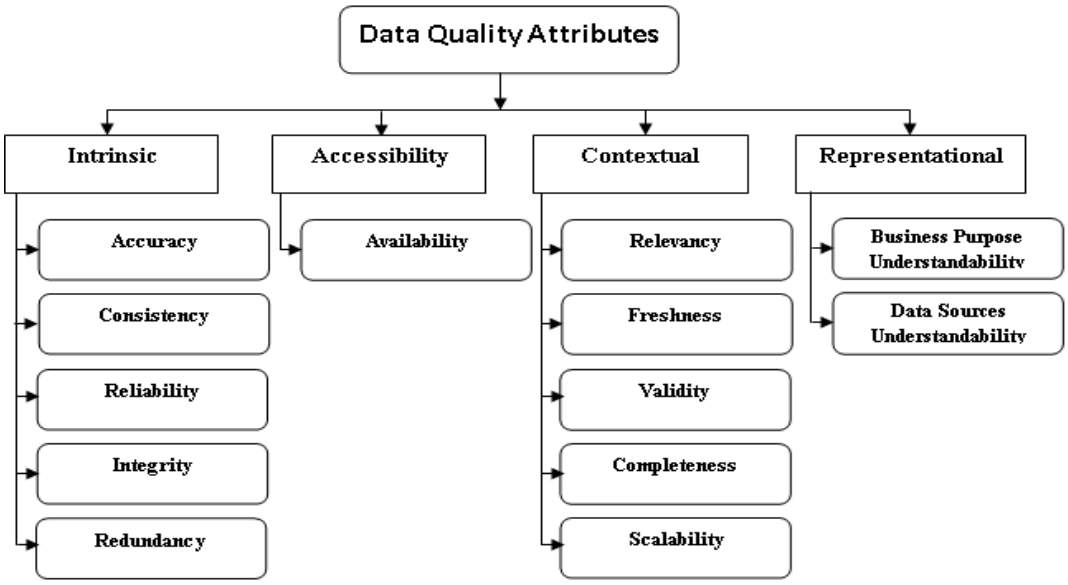

Figure 2. Dimensions of Information Data Quality

3.5. ETL Process:

ETL plays a vital role in DWH solutions [32-34], which shown in Figure (3) and 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) [35]. 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 [28]. 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 clean medical data before loading this data into CDWH. Furthermore, the ETL process is quite complex in medical field which requires extract data from several sources, cleaning and transformation activities, and loading facilities. Additionally, each phase in the ETL process has its issues and challenges:

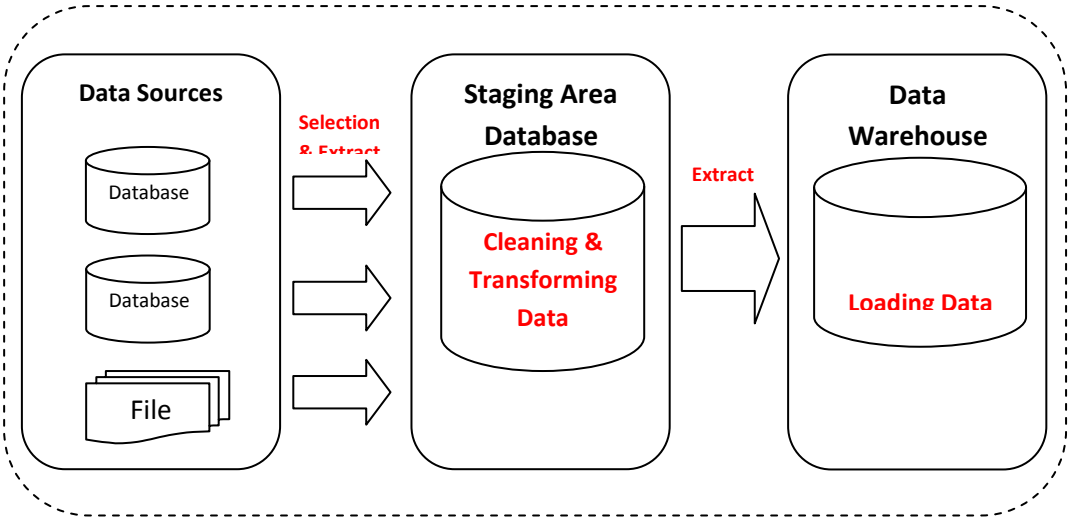

Figure 3. ETL Process

3.5.1. 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[36]. 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 [36]. Furthermore the complexity of the extraction process depends on the data characteristics and attributes, amount of source data and processing time as discussed in [34]. 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:

1. Analyzing data sources in order to comprehend their structure and contents to understand the data that exist in the sources database to identify the relevant data at the sources that needed depending on the purpose of CDWH as discussed in [10, 34, 37], 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 that 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.

3.5.2. Transformation Process: Transformation process is to transform the extracted data into a common format by applying a set of conditions ,rules or functions [34]. The transformation phase tends to make multiple data manipulations on the incoming data according to business needs[38], 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, DW schema (star or snowflake), derived facts, slowly changing fact tables and dimension tables as explained in [34]. In medical field very complex transformations provide the following requirements to meet the medical needs of the targeted system:

a. Understanding the format of data stored by each source to determine whether all the data fulfill the requirements or not as reported in [10, 34, 37].
b. Figure out a way of mapping the external data sources and internal data sources fields to the CDWH fields.
c. Transforming and coding the medical data into the required content format for CDWH storage.
d. Providing amount of manipulation needed for transformation process according to the objective of the CDWH such as summarization, integration, and aggregation using different techniques according to requirement specifications as discussed in [10, 34, 39],
e. Providing suitable data model to allow querying by multiple dimension

3.5.3. 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[30]. The data cleansing deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data [40, 41]. The data cleansing phase involves three steps include: data analysis, data refinement and data verification [38]. 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 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 benefits. Additionally, the cleaned data then will be assessed against the acceptance criteria again to ensure that the data issues have been resolved after the data cleansing process. Finally, after verification, the data will be moved from staging area to CDWH. Therefore, the tends of data cleansing process is to make cleaning 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:

1. Understanding the data quality problems from medical perspective.
2. cleaning of the extracted data set, according to the required medical rules as reported in [10, 34],
3. All the requisite information is available, free from errors, in a usable state.
4. The data collected is relevant to the business purpose.
5. The ability to link relative records together to ensure the data consistency in format.
6. The data satisfies a set of constrains, and maintains in a consistent fashion to ensure the data values consistent across data sets.
7. All patient basic information records must contain a unique patient identification number for each patient.
8. Auto generated primary keys and cross reference tables.

3.5.4. Loading Process: Loading process is the process of loading data from staging area to the CDWH. The extracted and transformed data is written into the dimensional structures actually accessed by the end users and applications[41]. 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 as discussed in [37] [42]. 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:

a. The ability of ETL process to provide the desired latency in updating the dataset.
b. 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.
c. Data is up to date or is provided at the time specified (data tagged with a time).

4. Discussions

CDWH is more complicated than the DWH and produce a set of issues and challenges. The paper discussed the CDWH development issues and challenges such as clinical data format, business purpose, data integration, data quality, and ETL process issues.

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. Furthermore,

the complexity of clinical data rise several issues for instance; poorly characterized mathematically, difficult data type for mining, difficult to determine hidden relationships, and require 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.

A clear understanding of the business 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 medical objectives is achieve.

CDWH aims to integrate large volumes of data collected from several clinical information systems. The complexity of the hospital environment evolve 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 requires.

Data Quality is one of important issues in CDWH because medical decisions making made based on data stored in CDWH. The quality of the information 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 implement quality database 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 as the core component of a successful CDWH development. 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.

5. Conclusion

This paper discussed the CDWH issues and challenges to develop successful CDWH. The usage of DWH technologies in medical field produces new issues and challenges to DWH technologies. Handling these issues and challenges requires determining the clinical data format requirements, business purpose requirements, data integration requirements, data quality requirements, and ETL process requirements. Format of clinical data is differed from business data they are of a complex nature and difficult data type, which require complex transformation methodology. On the other hand, this data require a clear definition of clinical purpose to determine requirements. Whereas, the data integration are important issues that performed the intergradations of large volumes of data from several sources, and build a robust technique to solve all data quality problems at each phase of the ETL process. Furthermore, ETL process must meet special requirements to ensure the quality of the data in the CDWH.

References

[1] Inmon, W. H., 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)**.
[2] Manjunath, T., R. S. Hegadi, and G. Ravikumar, Analysis of Data Quality Aspects in DataWarehouse Systems. IJCSIT) International Journal of Computer Science and Information Technologies. **2**(1). 477-485. **(2010)**.
[3] Inmon, W. H., C. Imhoff, and R. Sousa. Corporate information factory. John Wiley & Sons**(2002)**.
[4] Stein, L. D., Integrating biological databases. Nature Reviews Genetics. **4**(5). 337-345. **(2003)**.
[5] Gardner, S. R., BUILDING the Data Warehouse. Communications of the ACM. **41**(9). 53. **(1998)**.
[6] Watson, H. J., D. L. Goodhue, and B. H. Wixom, The benefits of data warehousing: why some organizations realize exceptional payoffs. Information & Management. **39**(6). 491-502. **(2002)**.
[7] Sahama, T. R. and P. R. Croll. A data warehouse architecture for clinical data warehousing. in Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68. **(2007)**. Australian Computer Society, Inc.
[8] Bernstam, E. V., W. R. Hersh, S. B. Johnson, C. G. Chute, H. Nguyen, I. Sim, M. M. Nahm, M. Weiner, P. Miller, and R. P. DiLaura, Synergies and distinctions between computational disciplines in biomedical research: perspective from the Clinical and Translational Science Award programs. Academic medicine: journal of the Association of American Medical Colleges. **84**(7). 964. **(2009)**.
[9] Chute, C. G., 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. **17**(2). 131-135. **(2010)**.
[10] Esfandiary, N., M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, Knowledge Discovery in Medicine: Current Issue and Future Trend. Expert Systems with Applications. **(2014)**.
[11] Seah, B. K. An application of a healthcare data warehouse system. in Innovative Computing Technology (INTECH), 2013 Third International Conference on. **(2013)**. IEEE.
[12] Rahm, E., T. Kirsten, and J. Lange, The GeWare data warehouse platform for the analysis of molecular-biological and clinical data. Journal of Integrative Bioinformatics. **4**(1). 47. **(2007)**.
[13] Ramick, D. C., Data warehousing in disease management programs. Journal of Healthcare Information Management. **15**(2). 99-106. **(2001)**.
[14] Kimball, R. The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. John Wiley & Sons**(1998)**.
[15] Xueqin, H., C. Meng, and C. Bing, Research on Data Quality of Chinese Medicine Scientific Data. World Science and Technology. **11**(4). 589-592. **(2010)**.
[16] DeWitt, J. G. and P. M. Hampton, Development of a data warehouse at an academic health system: knowing a place for the first time. Academic Medicine. **80**(11). 1019-1025. **(2005)**.
[17] Lynch, W. J. and J. E. Ross Jr, *Medical records, documentation, tracking and order entry system*, 1998, Google Patents.
[18] Lober, W. B., B. T. Karras, M. M. Wagner, J. M. Overhage, A. J. Davidson, H. Fraser, L. J. Trigg, K. D. Mandl, J. U. Espino, and F.-C. Tsui, Roundtable on Bioterrorism Detection Information System–based Surveillance. Journal of the American Medical Informatics Association. **9**(2). 105-115. **(2002)**.
[19] Wylie, J. E. and G. P. Mineau, Biomedical databases: protecting privacy and promoting research. Trends in biotechnology. **21**(3). 113-116. **(2003)**.
[20] Berndt, D. J., J. W. Fisher, A. R. Hevner, and J. Studnicki, Healthcare data warehousing and quality assurance. Computer. **34**(12). 56-65. **(2001)**.
[21] Lavrač, N., Selected techniques for data mining in medicine. Artificial intelligence in medicine. **16**(1). 3-23. **(1999)**.
[22] Hass, K. B., R. Vander Horst, K. Ziemski, and L. Lindbergh. From Analyst to Leader: Elevating the Role of the Business Analyst. Management Concepts Press**(2007)**.
[23] Gray, G. W., Challenges of building clinical data analysis solutions. Journal of critical care. **19**(4). 264-270. **(2004)**.
[24] Lenzerini, M. Data integration: A theoretical perspective. in Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. **(2002)**. ACM.
[25] Lane, P., V. Schupmann, and I. Stuart, Oracle database data warehousing guide, 10g release 2 (10.2). Oracle Corporation, Redwood City, CA. **(2005)**.
[26] Kerdprasop, N. and K. Kerdprasop, Higher Order Programming to Mine Knowledge for a Modern Medical Expert System. International Journal of Computer Science Issues (IJCSI). **8**(3). **(2011)**.
[27] Baudot, A., G. Gomez-Lopez, and A. Valencia, Translational disease interpretation with molecular networks. Genome biology. **10**(6). 221. **(2009)**.
[28] Wang, R. Y. and D. M. Strong, Beyond accuracy: What data quality means to data consumers. J. of Management Information Systems. **12**(4). 5-33. **(1996)**.
[29] Yeh, P. Z. 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)**. IEEE.

[30] Singh, R. and K. Singh, A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. International Journal of Computer Science Issues (IJCSI). **7**(4). **(2010)**.
[31] Dayal, U., 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)**. Acm.
[32] Kimball, R. and J. Caserta. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons**(2004)**.
[33] Kimball, R. and M. Ross, The data warehouse toolkit: the complete guide to dimensional modelling. Nachdr.]. New York [ua]: Wiley. **(2002)**.
[34] Anand, N. and M. Kumar. An Overview on Data Quality Issues at Data Staging ETL. in Int. Conf. on Advances in Signal Processing and Communication. **(2013)**.
[35] Savitri, F. N. 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)**. IEEE.
[36] Theodoratos, D., S. Ligoudistianos, and T. Sellis, View selection for designing the global data warehouse. Data & Knowledge Engineering. **39**(3). 219-240. **(2001)**.
[37] Szirbik, N. B., C. Pelletier, and T. Chaussalet, Six methodological steps to build medical data warehouses for research. International Journal of Medical Informatics. **75**(9). 683-691. **(2006)**.
[38] Rahm, E. and H. H. Do, Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. **23**(4). 3-13. **(2000)**.
[39] Adzic, J., V. Fiore, and S. Spelta, *Data warehouse population platform*, in *Databases in Telecommunications II*. 2001, Springer. p. 9-18.
[40] de Andrade, T. L., 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)**. IEEE.
[41] Mummana, S. and R. kiran Rompella, An Empirical Data Cleaning Technique for CFDs. International Journal of Engineering Trends and Technology (IJETT). **4**(9). 3730- 3735. **(2013)**.
[42] Lorentz, D., J. Gregoire, and S. Abraham. Oracle9i: SQL Reference, Release 2 (9.2). Oracle Corporation**(2002)**.

Authors

Abubaker Elrazi Osman Mohammed, he obtained his BSc degree in computer & Statistical Science from Gazira University-sudan in 2000. He received his MSc in computer science from Gazira University- sudan in 2006. Currently he is PhD student in Shendi University-Sudan. His research interest is data warehouse and data mining. He is working as director of information technology center at Shendi University-Sudan.

Elsamani Abd Eltalab, he received BSc, MSc and PhD degree in computer science from department of computer science, University of Khartoum, Sudan in 1989, 1995, and 2001 respectively. Currently, he is working as dean, associate professor of computer science faculty of computer science and information technology, AL_Neelain University, Khartoum, Sudan. His fields of interest are in data structures, algorithms, teaching and learning, compiler design and numerical computation.