ISSN 2581 – 3781

AJCSE
ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

Available Online at www.ajcse.info
Asian Journal of Computer Science Engineering 2018;3(3):1-13

REVIEW ARTICLE

Framework for Data Warehousing and Mining Clinical Records of Patients: A Review

N. D. Oye*, G. D. Emeje

Department of Computer Science, MAUTECH, Yola, Nigeria

Received on: 25-02-2018, Revised on: 01-04-2018, Accepted on: 05-05-2018

ABSTRACT

The clinical data warehouse (CDW) is a place where health-care providers can gain access to clinical data gathered in the patient care process. It is also anticipated that such data warehouse may provide information to users in areas ranging from research to management. A CDW is a tailored data warehouse for the needs of users in clinical environment. Due to the fast growing data in the health-care sector, there is need for heath industries to open toward adoption of extensive health-care decision support systems. A critical step for achieving precision medicine will be to integrate old and new data into validated information and to convert this information into knowledge directly applicable to diagnosis, prognosis, or treatment. This will entail developing an integrated knowledge environment that continually captures information and grows, accumulates, organizes, and institutionalizes new information, making it accessible to health-care providers. The framework is independent of application domain and would be suitable for users in areas such as data mining and knowledge management. CDWs are complex and time consuming to review a series of patient records. However, it is one of the efficient data repositories existing to deliver quality patient care.

Key words: Clinical data warehousing, clinical record of patients, data mining, hospital information system, knowledge discovery

BACKGROUND OF THE STUDY

Knowledgeable decision-making in health care is vital to provide timely, precise, and appropriate advice to the right patient, to reduce the cost of health care, and to improve the overall quality of health-care services. Since medical decisions are very complex, making choices about medical decision-making processes, procedures, and treatments can be overwhelming.[1] One of the major challenges of information technology (IT) in health-care services is how to integrate several disparate, standalone clinical information repositories into a single logical repository to create a distinct version of fact for all users.[2-5]

A massive amount of health records, related documents, and medical images generated by clinical diagnostic equipment are created daily.[3] Medical records are owned by different hospitals, departments, doctors, technicians, nurses, and patients. These valuable data are stored in various medical information systems such as hospital information system (HIS), radiology information system, picture archiving and communications system in various hospitals, and departments and laboratories being primary locations.[3] These medical information systems are distributed and heterogeneous (utilizing various software and hardware platforms including several configurations). Such processes and data flows have been reported by Zheng *et al*.[3]

All medical records are located in different hospitals or different departments of single hospital. Every unit may use different hardware platforms, different operating systems, different information management systems, or different network protocols. Medical data are also in various formats. There are not only a tremendous volume of imaging files (unstructured data) but also many medical information such as medical records, diagnosis reports, and cases with different definitions and structures in information system (structured data).[3] This causes clinical data stores (CDS) with isolated information across various hospitals, departments, laboratories, and

Address for correspondence:
N. D. Oye
E-mail: oyenath@yahoo.co.uk

related administrative processes, which are time consuming and demanding reliable integration.[6] Data required to make informed medical decisions are trapped within fragmented, disparate, and heterogeneous clinical and administrative systems that are not properly integrated or fully utilized. Ultimately, health care begins to suffer because medical practitioners and health-care providers are unable to access and use this information to perform activities such as diagnostics, prognostics, and treatment optimization to improve patient care.[7]

PROBLEM STATEMENT

Availability of timely and accurate data is vital to make informed medical decisions. Every type of health-care organization faces a common problem with the a considerable amount of data they have in several systems. Such systems are unstructured and unorganized, demanding computational time for data and information integration.[7] Today, patient's data required to make informed medical decisions trapped within fragmented and disparate clinical and administrative systems that are not properly integrated or fully utilized. The process of synthesizing information from these multiple heterogeneous data sources is extremely difficult, time consuming, and in some cases impossible.

Due to the fast growing data in the health-care sector, there is a need for heath industries to open toward adoption of extensive health-care decision support systems (DSS).[8] There is a growing need in the health-care scenario to store and organize sizeable clinical data, analyze the data, assist the health-care professionals in decision-making, and develop data mining methodologies to mine hidden patterns and discover new knowledge. Data warehousing integrates fragmented electronic health records (EHR) from independent and heterogeneous CDS[7] into a single repository. It is based on these concepts that this study plans to design a data warehousing and mining framework that will organize, extract, and integrate medical records.

SIGNIFICANCE OF THE STUDY

- It is clear that advanced clinical data warehousing (CDW) and mining information systems will be a driver for quality improvements of medical care.
- The ability to integrate data to have valuable information will result in a competitive advantage, enabling health-care organizations to operate more efficiently.
- The discovered knowledge in the human leaning technique can be used for community diagnoses or prognosis.
- It will eliminate the use of file system and physical conveyance of files by messengers.
- It will provide a platform for data mining operations on patient's clinical data at the Federal Medical Centre, Yola.
- This study will encourage and challenge many government and non-governmental health-care providers to opt for data warehouse and mining investment to improve information access within their organization, bringing the user of their information in touch with their data, and providing cross-function integration of operation systems within their organization.

LITERATURE REVIEW

Clinical DSS

Clinical decision support (CDS) systems provide clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered, and presented at appropriate times, to enhance health and health care.[9] Clinical informatics is the application of informatics and IT to support health-care delivery services. Its role is rapidly evolving toward providing better clinical decision-making by integrating state-of the-art knowledge with medical record systems.[10] As medicine moves into an era of personalized treatment and precision pharmaceuticals, the application of expertise in EHR/medical record systems and translational research will enhance operating efficiencies for hospitals and reduce costs.[10]

In reality, the populating and analyzing of large amounts of accumulating data in standardized format from EHRs are yet to happen, since protocols and resources have not yet sufficiently matured. Recognition of the importance of applying digitized data and information for patient care have spurred the first class of physicians to become board certified in the newly created subspecialty of clinical informatics.[10] Bioinformatics is the development of storage, analytic, and interpretive

methods to optimize the transformation of increasingly voluminous biomedical and genomic data into proactive, predictive, preventive, and participatory health care.[10]

A critical step for achieving precision medicine will be to integrate old and new data into validated information and to convert this information into knowledge directly applicable to diagnosis, prognosis, or treatment. This will entail developing an integrated knowledge environment that continually captures information and grows, accumulates, organizes, and institutionalizes new information, making it accessible to health-care providers.[10]

DSS

The DSS concept goes back a long time, and the definition varies depending on the evolution of IT and, of course, on the point of view of those who issues such a definition. Looking through several definitions, we can find that Watson, Rainer, and Chang cited by Inuwa and Oye[11] defined DSS as an extensible system, capable of *ad hoc* analysis, and decision modeling, focused on future planning, and used at unplanned and irregular time stamps. Furthermore, Carlson and Sprague cited by Inuwa and Oye[11] defined DSS as being interactive systems that help decision makers to use data and models in resolving unstructured and semi-structured economic problems. Turban cited by Inuwa and Oye[11] defines a DSS as an interactive, flexible, and adaptable system, exclusively designed to offer support in solving unstructured or semi-structured managerial problems, aiming to improve the decisional process. The system uses data (internal and external) and models, providing a simple and easy-to-use interface, thus allowing the decision maker control over the decision process. The DSS offers support in all decision process stages.

DSS exists to help people make decisions. DSS do not make decisions by themselves. They attempt to automate several tasks of the decision-making process, of which the modeling is the core. To comprehend DSS, a person needs to understand the process of making decisions. DSS uses data, provides easy user interface, and can incorporate the decision maker's own insights. The tables in a decision-support database are heavily indexed, and the raw data are often pre-processed and organized to support the various types of queries to be used.[12]

DECISION-MAKING

Decision-making is an indispensable part of everyday life. We make hundreds of decisions each day. To make a good decision, we have to be informed about alternative options. These options can be in different forms such as numbers, graphics, and impressions. Turban classified decisions into three groups as structured, semi-structured, and unstructured. Structured decisions are repetitive and routine decisions. Unstructured decisions are non-routine decisions so that decision maker has to provide judgment. Semi-structured decisions include some characteristics of both structured and unstructured decisions. Decision maker has to provide judgment only for the parts that do not have an accepted procedure.[11] Decision-making is a process of making a choice from a number of alternatives to achieve a desired result (Eisenfuhr, 2011).

DATA WAREHOUSE DEFINITION

Inuwa and Garba (2015)[11] define data warehouse as a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management's decisions. Subject-oriented: Classical operations systems are organized around the applications of the company. Each type of company has its own unique set of subjects.

Integrated

Data are fed from multiple disparate sources into the data warehouse. As the data are fed, it is converted, reformatted, resequenced, summarized, and so forth. The result is that data once it resides in the data warehouse have a single physical corporate image.

Non-volatile

Data warehouse data are loaded and accessed, but it is not updated. Instead, when data in the data warehouse are loaded, it is loaded in a snapshot, static format. When subsequent changes occur, a new snapshot record is written. In doing so, a history of data is kept in the data warehouse.

Time variant

Every unit of data in the data warehouse is accurate as of some one moment in time. In some cases, a record is time stamped. In other cases, a record has a date of transaction. However, in every case, there is some form of time marking to show the moment in time during which the record is accurate.

According to Kimball and Ross[13] as cited by Inuwa and Oye,[11] DW is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model. Kimball and Ross viewed data warehousing as a constituency of data marts. Data Marts are focused on delivering business objectives for departments in the organization, and the DW is a conformed dimension of the data marts. Over the past few years, organizations have increasingly turned to data warehousing to improve information flow and decision support. A DW can be a valuable asset in providing easy access to data for analysis and reporting. Unfortunately, building and maintaining an effective DW have several challenges.[12]

DATA WAREHOUSE MODELING

Ballard cited by Inuwa and Oye[11] gave an assessment of the evolution of the concept of data warehousing, as it relates to data modeling for the DW, and they defined database warehouse modeling as the process of building a model for the data to store in the DW. There are two data modeling techniques that are relevant in a data warehousing environment and they are as follows:

i. Entity relationship (ER) modeling: ER modeling produces a data model of the specific area of interest, using two basic concepts: Entities and the relationships between those entities. Detailed ER models also contain attributes, which can be properties of either the entities or the relationships. The ER model is an abstraction tool because it can be used to understand and simplify the ambiguous data relationships in the business world and complex systems. ER modeling uses the following concepts: Entities, attributes, and the relationships between entities. The ER model can be used to understand and simplify the ambiguous data relationships in the business world and complex systems environments.
ii. Dimensional fact modeling: Dimensional modeling uses three basic concepts: Measures, facts, and dimensions, and dimensional modeling is powerful in representing the requirements of the business user in the context of database tables. Measures are numeric values that can be added and calculated.

DATA WAREHOUSE MODELING TECHNIQUES

Thomas and Carol cited by Inuwa and Oye[11] derived the way a DW or a data mart structure in dimensional modeling into several ways. Flat schema, terraced schema, star schema, fact constellation schema, galaxy schema, snowflake schema, star cluster schema, and star flake schema. However, there are two basic models that are widely used in dimensional modeling: Star and snowflake models.

i. Star schema: The star schema [Figure 1] is a relational database schema used to hold measures and dimensions in a data mart. The measures are stored in a fact table, and the dimensions are stored in dimension tables. For each data mart, there is only one measure surrounded by the dimension tables, hence the name star schema. The center of the star is formed by the fact table. The fact table has a column or the measure and the column for each dimension containing the foreign key for a member of that dimensions. The key for this table is formed by concatenate all of the foreign key fields. The primary key for the fact table is usually referred to as composite key. It contains the measures, hence the name "Fact." The dimensions are stored in dimension tables. The dimension table has a column for the unique identifier of a member of the dimension, usually an integer of a short character value. It has another column for a description.[11]
ii. Snowflake schema: Snowflake schema model is derived from the star schema and, as can be seen, looks like a snow flake. The snowflake model is the result of decomposing one or more of the dimensions, which generally have hierarchies between themselves. Many-to-one relationships among members within a dimension table can be defined as a separate dimension table, forming a hierarchy as can be seen in Figure 2.

DATA MART

A data mart is a small DW built to satisfy the needs of a particular department or business area. The term data mart refers to a subentity of data warehouses containing the data of the DW for a particular sector of the company (department, division, service, product line, etc.,). The data mart is a subset of the DW that is usually oriented to a specific business line or team. Whereas a DW combines databases across an entire enterprise, data marts are usually smaller and focus on a particular subject or department. Some data marts are called dependent data marts and are subsets of larger data warehouses.[15] Also summarized the types of Data Marts as follows.

INDEPENDENT AND DEPENDENT DATA MARTS

An independent data mart is created without the use of a central DW. This could be desirable for smaller groups within an organization. A dependent data mart allows you to unite your organization data in one DW. This gives you the usual advantages of centralization as can be seen in Figure 3.

Architecture of the DW

Data contained in a DW hold five types of data: Data currency, existing data, data summarization (lightly and highly summarized data), and metadata.[11]

This traditional data warehousing architecture in Figure 4 encompasses the following components.[11]

i. Data sources as external systems and tools for extracting data from these sources.
ii. Tools for transforming, which is cleaning and integrating the data.
iii. Tools for loading the data into the DW.
iv. The DW as central, integrated data store.
v. Data marts as extracted data subsets from the DW oriented to specific business lines, departments or analytical applications.
vi. A metadata repository for storing and managing metadata.
vii. Tools to monitor and administer the DW and the extraction, transformation and loading process.
viii. An online analytical processing engine on top of the DW and data marts to present and serve multi-dimensional views of the data to analytical tools.

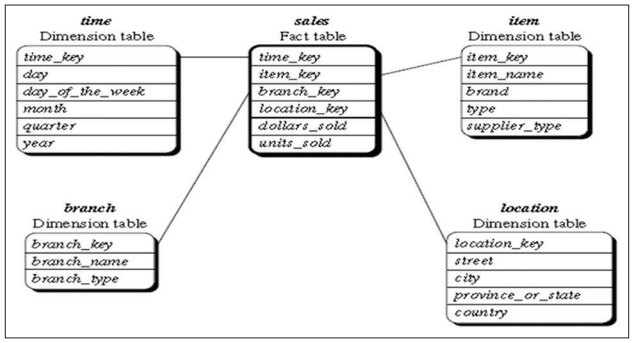

Figure 1: Star schema[14]

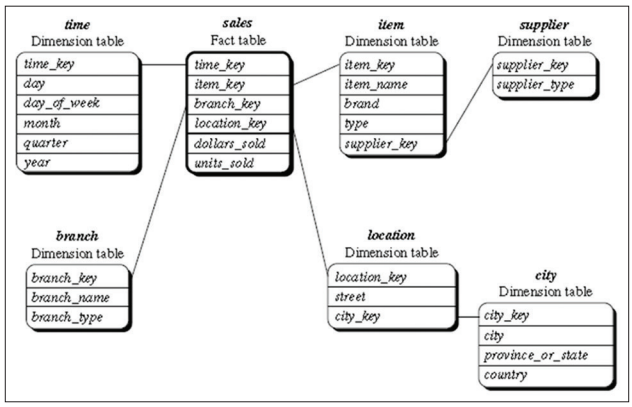

Figure 2: Snowflake schema[14]

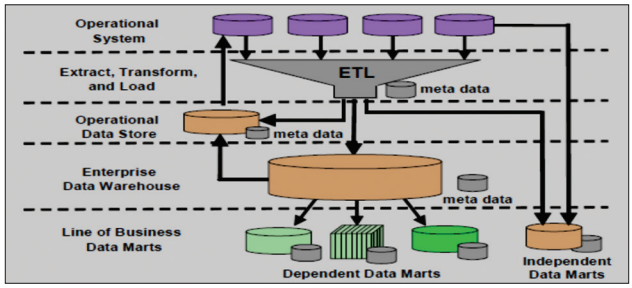

Figure 3: DW architecture[11]

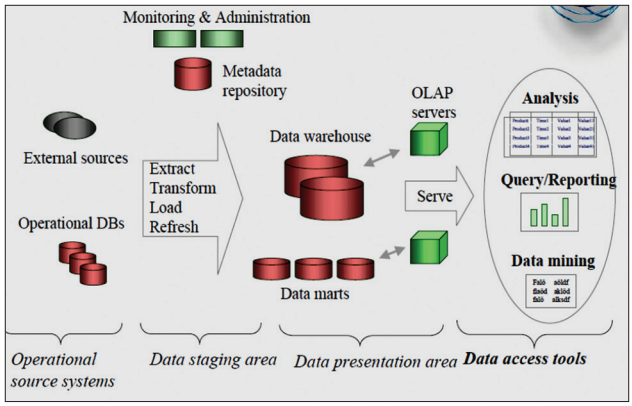

Figure 4: Traditional data warehousing architecture[11]

ix. Tools that use data from the DW for analytical applications and for presenting it to end-users.

This architecture exemplifies the basic idea of physically extracting and integrating mostly transactional data from different sources, storing

it in a central repository while providing access to the data in a multidimensional structure optimized for analytical applications.[16] However, the architecture is rather old, and while this basic idea is still intact, it is rather unclear and inaccurate about several facts.

First, most modern data warehousing architectures use a staging or acquisition area between the data sources and the actual DW. This staging area is part of the extract, transform, and load process (ETL process). It temporarily stores extracted data and allows transformations to be done within the staging area, so source systems are directly decoupled and no longer strained.[16] Second, the interplay between DW and data marts in the storage area is not completely clear.

Actually, in practice, this is one of the biggest discourses about data warehousing architecture with two architectural approaches proposed by Inmon and Kimball.[11] Inmon places his data warehousing architecture in a holistic modeling approach of all operational and analytical databases and information in an organization, the Corporate Information Factory. Atomic DW is a centralized repository with a normalized, still transactional and fine-granular data model containing cleaned and integrated data from several operational sources.[11]

Inmon's approach, also called enterprise DW architecture by Thilini and Hugh,[16] is often considered a top-down approach, as it starts with building the centralized, integrated, enterprise-wide repository and then deriving data marts from it to deliver for departmental analysis requirements. However, it is possible, to build an integrated repository and the derived data marts incrementally in an iterative fashion. Kimball, on the other hand, proposed a bottom-up approach which starts with process and application requirements[13] as cited by Inuwa and Garba.[11] With this approach, first, the data marts are designed based on the organization business processes, where each data mart represents data concerning a specific process. The data marts are constructed and filled directly from the staging area while the transformation takes places between staging area and data marts.

The data marts are analysis oriented and multidimensional. The DW is then just the combination of all data marts, where the single data marts are connected and integrated with each other through the data bus and so-called conformed dimensions that are data mart use and standardized or "conformed" dimension tables.[11] When two data marts use the same dimension, they are connected and can be queried together via an identical dimension table. The data bus is then a net of data marts, which are connected through conformed dimensions. This architecture (also called data mart bus architecture with linked dimensional data marts by Thilini and Hugh)[16] therefore forgoes a normalized, enterprise-wide data model and repository.

In Figure 3, there are a number of options for architecting a data mart. For example:

i. Data can come directly from one or more of the databases in the operational systems, with few or no changes to the data in format or structure. This limits the types and scope of analysis that can be performed. For example, you can see that, in this option, there may be no interaction with the DW Meta Data. This can result in data consistency issues.
ii. Data can be extracted from the operational systems and transformed to provide a cleansed and enhanced set of data to be loaded into the data mart by passing through an ETL process. Although the data is enhanced, it is not consistent with, or in sync with, data from the DW.
iii. Bypassing the DW leads to the creation of an independent data mart. It is not consistent, at any level, with the data in the DW. This is another issue impacting the credibility of reporting.
iv. Cleansed and transformed operational data flow into the DW. From there, dependent data marts can be created or updated. It is a key that updates to the data marts are made during the update cycle of the DW to maintain consistency between them. This is also a major consideration and design point, as you move to a real-time environment. At that time, it is good to revisit the requirements for the data mart, to see if they are still valid.

However, there are also many other data structures that can be part of the data warehousing environment and used for data analysis, and they use differing implementation techniques. Although data marts can be of great value, there are also issues of currency and consistency. This has resulted in recent initiatives designed to minimize the number

of data marts in a company. This is referred to as data mart consolidation (DMC). DMC may sound simple at first, but there are many things to consider. A critical requirement, as with almost any project, is executive sponsorship because you will be changing many existing systems on which people have come to rely, even though the systems may be inadequate or outmoded. To do, this requires serious support from senior management. They will be able to focus on the bigger picture and bottom-line benefits and exercise the authority that will enable making changes.[16]

BENEFITS OF DATA WAREHOUSING

There are several benefits of data warehousing. The most important ones are listed as follows: [13]

i. DW improves access to administrative information for decision makers.
ii. It can get data quickly and easily perform analysis. One can work with better information and make decisions based on data. DW increases the productivity of corporate decision-makers.
iii. Data extraction from its original data sources into the central area resolves the performance problem, which arises from performing complex analyses on operational data.
iv. Data in the warehouse are stored in specialized form, called a multidimensional database. This form makes data querying efficient and fast.
v. A huge amount of data is usually collected in the DW. Compared with relational databases that are still very popular today, data in the warehouse do not need to be in normalized form. In fact, it is usually denormalized to support faster data retrieval.

DW MODELING TOOLS

Building a DW from independent data sources is a difficult process. This process involves extracting, converting, cleaning, integration, and transformation of the data. To do these operations, an ETL tool is required. The key steps that need to be undertaken to transform raw operational data to a form that can be stored in a DW for analysis are as follows:

i. Extraction, the goal of the data extraction step is to bring data from different sources into a database before modification.
ii. Converting the data into a format that is suitable to the DW.
iii. Cleaning of the data, data entry errors, and differences in schema formation can cause, for example, patient dimension table to have several corresponding entries for a single patient.
iv. Integration of the different datasets to suit the data model of the DW.
v. Transformation, of the data through summarization and creation of new attributes, it is a set of rules and scripts that typically handle the transformation of data from an input schema to the destination schema.[17]

DW DATABASE DESIGN MODELING

There are three levels of data modeling. They are conceptual, logical, and physical. Conceptual design manages concepts that are close to the way users perceive data; logical design deals with concepts related to a certain kind of database management system (DBMS); physical design depends on the specific DBMS and describes how data are actually stored. The main goal of conceptual design modeling is developing a formal, complete, abstract design based on the user requirements.[18] DW logical design involves the definition of structures that enable an efficient access to information. The designer builds multidimensional structures considering the conceptual schema representing the information requirements, the source databases, and non-functional (mainly performance) requirements. This phase also includes specifications for data extraction tools, data loading processes, and warehouse access methods. At the end of logical design phase, a working prototype should be created for the end-user.[18]

CDW

The CDW is a place where health-care providers can gain access to clinical data gathered in the patient care process. It is also anticipated that such data warehouse may provide information to users in areas ranging from research to management. A CDW is a tailored data warehouse for the needs of users in clinical environment. The CDW

combines information from a variety of legacy health-care database and extract operational data to form a centralized repository to answer the informational needs of all clinical users. The data in a clinical warehouse are not only used by the patients, nurses and doctors but also used by researchers, scientists, and medical students.[19]

Demands for clinical warehousing are increasing dramatically. From being viewed as a data gathering tool, a CDW is now moving into a new phase of becoming a business-critical platform. Such a platform can support all clinical decisions across the clinical trial portfolio, be central to collaboration, and fundamental to the survival and agility of the business. To fully comprehend the meaning, applicability, and relevance of a CDW, and how it can potentially benefit a company, we must understand how it is different from a typical data warehouse, what has driven its need, and how its adoption can be maximized to drive clinical development.[20]

CHALLENGES IN IMPLEMENTING A CDW

The challenges facing CDW implementation are largely determined by the management's view of the implementing company. "Forward-thinking" senior management views a CDW as being fundamental to progression and essential for collaborative innovation, i.e., the ability to adapt, be agile, acquire organizations, outsource, and generally be more efficient, cost-effective, and competitive.[20]

Demonstrating a clear return on investment is always challenging as CDW is often multiyear programs with abstract, cross-functional concepts. To be successful, a CDW requires continuous senior management commitment and sponsorship. It is important to build a broad, holistic picture for the organization rather than at a departmental level. Buy-in solely from one or two department heads is rarely sufficient.[20]

However, transition from batch-centric data preparation and programming to better solutions, more in keeping with the dynamic nature of the industry, requires a key company visionary, with the gravitas to communicate the overall benefits of a CDW to the wider company.[20] The implementation phase of a CDW can be challenging. This phase requires essential specialist skills in areas such as process design, data standardization, modeling, and system integration. However, once the CDW is in place, fewer specialist skills are required. Key emphasis must also be placed on the management of process design and change, user training, adoption, and cross-departmental coordination to ensuring that investment in the CDW is maximized. A continuous program of monitoring and driving user adoption, streamlining data flow, and extending and enhancing use cases and tools-sets for data exploration, visualization, and analysis are key to ongoing return Figure 5.[20]

NEED FOR CDW

EHRs which describe the diseases and treatments of patients are normally stored in the hospitals or clinics, where they are created. However, patients may be treated in different hospitals and clinics, and therefore, there is a need for integrating health records from different hospitals to enable any hospital to obtain a total overview of a patient's health history. Two different types of heterogeneity problems have to be solved to integrate EHR systems from different hospitals in a consistent way. The first problem is that different hospitals normally do not use a common DBMS, and therefore, the traditional atomicity, consistency, isolation, and durability database properties are missing across the different hospital locations. This may cause performance, autonomy, and consistency problems. The second heterogeneity problem is that there are multiple incompatible standards for how to make EHR entries.[21]

To fulfill the information needs of medical practitioners and patients, it is a necessity to integrate physician records, hospital services, medication histories, and other medical information into a unified digital record that is available to needed persons at home or at the point of care. To realize the benefits of joined-up health care, to provide the right information at the right time and place, health-care organizations must

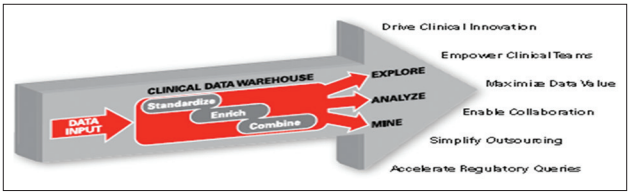

Figure 5: The general capabilities of a clinical data warehouse[20]

deploy and use standards that enable computer systems to exchange information in a way that is safe, secure, and reliable.[4]

An effective health-care establishment requires the availability of efficient software tools that enable easy searching, retrieval, and analysis of all gathered patient data, which are generally stored in distributed and heterogeneous databases. Medical data integration is a significant challenge due to the great variety of data types, formats, and platforms present in the medical scenario.[22] Most of the times, data warehouses fail because they do not meet the needs of the particular application domain or are too difficult/expensive to change with the evolving needs of the domain.[23]

DATA WAREHOUSING IN MEDICAL FIELD

Health-care organizations require data warehousing solutions to integrate the valuable patient and administrative data fragmented across multiple information systems within the organization. As stated by Kerkri cited by Saliya,[7] at a technical level, information sources are heterogeneous and autonomous and have an independent life cycle. Therefore, cooperation between these systems needs specific solutions. These solutions must ensure the confidentiality of patient information. To achieve sufficient medical data share and integration, it is essential for the medical and health enterprises to develop an efficient medical information grid.[3]

A medical data warehouse is a repository where health-care providers can gain access to medical data gathered in the patient care process. Extracting medical domain information to a data warehouse can facilitate efficient storage, enhances timely analysis, and increases the quality of real time decision-making processes. Currently, medical data warehouses need to address the issues of data location, technical platforms, data formats and organizational behaviors on processing those data. Today's health-care organizations require not only the quality and effectiveness of their treatment but also reduction of waste and unnecessary costs. By effectively leveraging enterprise wide data on labor expenditures, supply utilization, procedures, medications prescribed, and other costs associated with patient care, health-care professionals can identify and correct wasteful practices and unnecessary expenditures.[6]

Medical domain has certain unique data requirements such as high volumes of unstructured data (e.g., digital image files, voice clips, and radiology information) and data confidentiality. Data warehousing models should accommodate these unique needs. According to Pedersen and Jensen cited by Saliya,[7] the task of integrating data from several EHR systems is a hard one. This creates the need for a common standard for EHR data. According to Kerkri cited by Saliya,[7] the advantages and disadvantages of data warehousing are given below.

Advantages

i. Ability to allow existing legacy systems to continue in operation without any modification
ii. Consolidating inconsistent data from various legacy systems into one coherent set
iii. Improving quality of data
iv. Allowing users to retrieve necessary data by themselves.

Disadvantages

i. Development cost and time constraints.

CDW architecture

Data warehouse architecture is a description of the components of the warehouse, with details showing how the components will fit together.[8] The data warehouse architecture provides an integrated data warehouse environment while delivering incremental solutions. The architectural design focuses on the application of a centralized data warehouse, data marts, individual marts, metadata repositories, and incremental solution architectures. Different data warehousing systems have different structures. Some may have an ODS (operational data store), while some may have multiple data marts. Some may have a small number of data sources, while some may have dozens of data sources. Since a data warehouse is used for decision-making, it is important that the data extracted from multiple sources should be corrected. It is inevitable that when different data are integrated into the data warehouse, there is a high probability of errors and anomalies. Therefore, tools for data

extraction, data cleaning, data integration, and finally data load are required. Data are stored and managed in the warehouse which presents multidimensional views of data to a variety of front end tools: Query tools, report writers, analysis tools, and data mining tools.[8]

CANCER DATA WAREHOUSE ARCHITECTURE

Cancer, known medically as a malignant neoplasm, is a broad group of various diseases, all involving unregulated cell growth. In cancer, cells divide and grow uncontrollably, forming malignant tumors, and invade nearby parts of the body. The cancer may also spread to more distant parts of the body through the lymphatic system or bloodstream. Not all tumors are cancerous. Benign tumors do not grow uncontrollably, do not invade neighboring tissues, and do not spread throughout the body. Determining what causes cancer is complex. Many things are known to increase the risk of cancer, including tobacco use, certain infections, radiation, lack of physical activity, poor diet and obesity, and environmental pollutants. These can directly damage genes or combine with existing genetic faults within cells to cause the disease. Approximately, five to ten percent of cancers are entirely hereditary. People with suspected cancer are investigated with medical tests. These commonly include blood tests, X-rays, CT scans, and endoscopy.[24]

Figure 6 illustrates the overall architecture of health-care data warehouse specific to cancer proposed by Sheta and Ahmed.[24] Data are imported from several sources and transformed within a staging area before it is integrated and stored in the production data warehouse for further analysis.

INFLUENZA DISEASE DATA WAREHOUSE ARCHITECTURE

Influenza, commonly known as the "flu," is an infectious disease of birds and mammals caused by ribonucleic acid viruses of the family Orthomyxoviridae, the influenza viruses.[25] The most common symptoms are chills, fever, sore throat, muscle pains, headache (often severe), coughing, weakness/fatigue irritated, watering eyes, reddened eyes, skin (especially face), mouth, throat and nose, petechial rash, and general discomfort. Influenza may produce nausea and vomiting, particularly in children. Typically, influenza is transmitted through the air by coughs or sneezes, creating aerosols containing the virus. Influenza can also be transmitted by direct contact with bird droppings or nasal secretions or through contact with contaminated surfaces. Influenza spreads around the world in seasonal epidemics, resulting in about 3 to 5 million yearly cases of severe illness and about 250,000–500,000 yearly deaths.[25] People who suspected influenza are investigated with medical tests. These commonly include blood test (white blood cell differential), chest x-ray, auscultation (to detect abnormal breath sounds), and nasopharyngeal culture.

Figure 7 shows the proposed architecture for the health-care data warehouse specific to influenza disease by Rajib in 2013.[25] Architecture of influenza-specific health care data warehouse system builds with source data components in the left side where multiple data that come from different data source and transform into the data staging area before integrating. The data staging component presents at the next building block. Those two blocks are under data acquisition area. In the middle is the data storage component that manages the data warehouse data. This component also contain Metadata, that keep track of the data and the data marts. Last component of this architecture is information delivery component

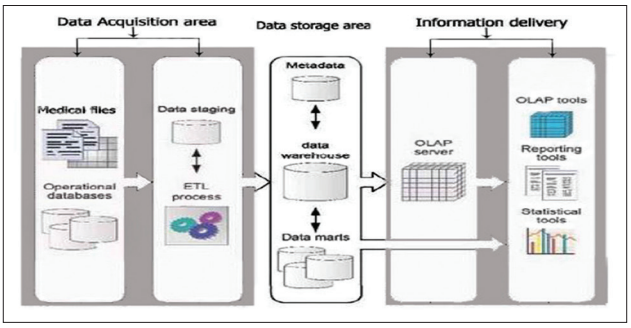

Figure 6: Cancer DW architecture[24]

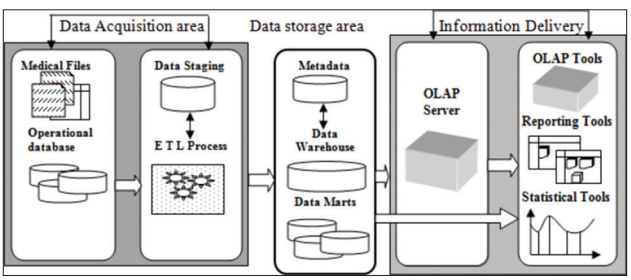

Figure 7: Data warehouse architecture for influenza disease[25]

that shows all the different ways of making the information from the data warehouse available to the user for further analysis.[25]

CARDIAC SURGERY DATA WAREHOUSING MODEL

Cardiac surgery clinical data can be distributed across various disparate and heterogeneous clinical and administrative information systems. This makes accessing data highly time consuming and error prone.[7] A data warehouse can be used to integrate the fragmented data sets. Once the data warehouse is created, it should be populated with data through ETL processes. Figure 8 shows a graphical overview of cardiac surgery data warehousing model.[7]

DIABETIC DATA WAREHOUSING MODEL

Diabetes is a defect in the body's ability to convert glucose (sugar) to energy. Glucose is the main source of fuel for our body. When food is digested, it is changed into fats, protein, or carbohydrates. Foods that affect blood sugars are called carbohydrates. Carbohydrates, when digested, change to glucose. Examples of some carbohydrates are bread, rice, pasta, potatoes, corn, fruit, and milk products. Individuals with diabetes should eat carbohydrates but must do so in moderation. Glucose is then transferred to the blood and is used by the cells for energy. In order for glucose to be transferred from the blood into the cells, the hormone-insulin is needed. Insulin is produced by the beta cells in the pancreas (the organ that produces insulin). In individuals with diabetes, this process is impaired. Diabetes develops when the pancreas fails to produce sufficient quantities of insulin, Type 1 diabetes, or the insulin produced is defective and cannot move glucose into the cells and occurs most frequently in children and young adults, although it can occur at any age. Type 1 diabetes accounts for 5–10% of all diabetes in the United States. There does appear to be a genetic component to Type 1 diabetes, but the cause has yet to be identified. Type 2 diabetes. Either insulin is not produced in sufficient quantities or the insulin produced is defective and cannot move the glucose into the cells, is much more common, and accounts for 90–95% of all diabetes. Type 2 diabetes primarily affects adults; however, recently Type 2 has begun developing in children. There is a strong correlation between Type 2 diabetes, physical inactivity, and obesity.[8] Figure 9 shows the overall propose diabetes data warehouse architecture. The data are stored in the production data warehouse for analysis while imported from several data sources and transformed them within a staging area before it is integrated.

CLINICAL DATA MINING

Data mining refers to a collection of techniques that provide the necessary actions to retrieve and gather knowledge from an exhaustive collection of data and facts. Data are available in enormous magnitude, but the knowledge that can be inferred from the data is still negligible. Data mining concepts are focused on discovering knowledge, predicting trends, and eradicating superfluous data.[26] Discovering knowledge in medical systems and health-care scenarios is a herculean yet critical task.[27] Knowledge discovery describes the process of automatically searching large volumes of data for patterns that can be considered additional knowledge about the data.[26] The knowledge obtained through the process may

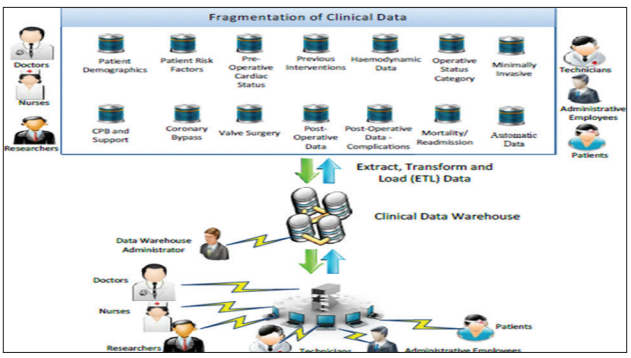

Figure 8: Data warehousing model for integrating fragmented electronic health records from disparate and heterogeneous cardiac surgery clinical data stores[7]

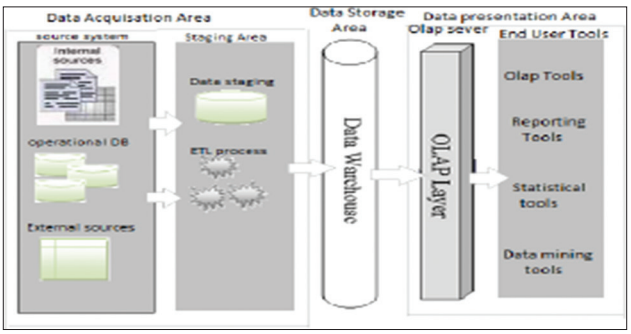

Figure 9: Diabetes data warehouse architecture[8]

become additional data that can be used for further manipulation and discovery.[28] Application of data mining concepts to the medical arena has undeniably made remarkable strides in the sphere of medical research and clinical practice saving time, money, and life.[29] Clinical data mining is the application of data mining techniques using clinical data. Clinical data mining involves the conceptualization, extraction, analysis, and interpretation of available clinical data for practical knowledge-building, clinical decision-making, and practitioner reflection.[30] The main objective of clinical data mining is to haul new and previously unknown clinical solutions and patterns to aid the clinicians in diagnosis, prognosis, and therapy.[31] Moreover, application of software solutions to store patient records in an electronic form is expected to make mining knowledge from clinical data less stressful.

The main activities carried out in the data mart cycle are as follows:

a. Architectural sketch, during which the overall functional and physical architecture of the DW are progressively drawn based on a macroanalysis of user requirements and an exploration of data sources as well as on budget, technological, and organizational constraints.
b. Conformity analysis aimed at determining which dimension of analysis will be conformed across different facts and data marts. Conforming hierarchies in terms of schema and data is a key element to allow cross-fact analysis and obtain consistent results.
c. Data mart prioritization, based on a trade-off between user priorities and technical constraints.
d. Data mart design, which builds and releases the top-priority data mart. After each data mart has been built, the three phases above are iterated to allow the DW plan to be refined and updated.

The activities carried out within a fact cycle are as follows:

i. Source and fact macroanalysis aimed at checking the availability, quality, and completeness of the data sources and determining the main business facts to be analyzed by users.
ii. Fact prioritization that, like for data marts, is the result of a trade-off between user requirements and technical priorities.
iii. Fact design, which develops and releases the top-priority fact. After that, the two phases above are iterated to allow the data mart plan to be refined and updated.

CONCLUSION

The researcher reviewed intensively some literatures related to CDW, clinical decision-making, data warehousing, and clinical data mining. The researcher was able to intensively review some related data warehousing architectures, data warehousing modeling techniques, data warehousing development methodologies, and clinical data mining models relevant to the study. From the various reviews, the researcher found out that the health sector has fast growing amount of patient data that are either stored on files or some HIS databases. The framework is independent of application domain and would be suitable for use in areas such as data mining and knowledge management. CDWs are complex and time consuming to review a series of patient records. However, it is one of the difficult data repositories existing to deliver quality patient care.

REFERENCES

1. Demetriades JE, Kolodner RM, Christopherson GA. Person Centered Health Records. Towards Healthy People. Health Informatics Series. USA: Springer; 2005.
2. Mann L. From “silos” to seamless healthcare: Bringing hospitals and GPs back together again. Med J Aust 2005;1:34-7. Available from: http://www.mja.com.au/ public/issues/182_01_030105/man10274_fm.html. [Last accessed on 2016 Mar 05].
3. Zheng R, Jin H, Zhang H, Liu Y, Chu P. Heterogeneous Medical data Share and Integration on Grid. International Conference on BioMedical Engineering and Informatics; 2008.
4. Goldstein D, Groen PJ, Ponkshe S, Wine M. Medical Informatics 20/20: Quality and Electronic Health Records through Collaboration, Open Solutions, and Innovation. Massachusetts, Sudbury, USA: Jones and Bartlett Publishers, Inc.; 2007.
5. Shepherd M. Challenges in Health Informatics. 40th Hawaii International Conference on System Sciences; 2007.
6. Sahama TR, Croll PR. A Data Warehouse Architecture for Clinical Data Warehousing. First Australasian Workshop on Health Knowledge Management and Discovery; 2007. Available from:http://www.crpit. com/confpapers/CRPITV68Sahama.pdf. [Last accessed on 2017 Apr 24].

7. Saliya N. Data Warehousing Model for Integrating Fragmented Electronic Health Records from Disparate and Heterogeneous Clinical Data Stores. Australia: Queensland University of Technology, School of Electrical Engineering and Computer Science Faculty of Science and Engineering; 2013.
8. Abubakar B. Building a diabetes data warehouse to support decision making in healthcare industry. IOSR J Computer Eng 2014;16:138-43. Available from: http://www.iosrjournal.org. [Last accessed on 2017 Dec 19].
9. Berner E. Clinical Decision Support Systems. University of Alabama at Birmingham, State of the Art Department of Health Services Administration. Birmingham: AHRQ Publication; 2009.
10. Castaneda C, Nalley K, Mannion C, Bhattacharyya P, Blake P, Pecora A, *et al.* Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma 2015;5:4.
11. Inuwa I, Garba E. An improved data warehouse architecture for SPGS, MAUTECH, Yola, Nigeria. West Afr J Indust Acad Res 2015;14:1-2.
12. Güzin T. Developing a Data Warehouse for a University Decision Support System. Atilim University, The Graduate School of Natural and Applied Sciences; 2007.
13. Kimball R. The data warehouse toolkit: The complete guide to dimensional modelling. In: Kimball R, Ross M, editors. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling. New York: John Wiley & Sons, Inc.; 2002.
14. Jiawei H. Data Mining Concepts and Techniques. USA: Morgan Kaufmann Publishers; 2012.
15. Gopinath T, Damodar NM, Lenin Y, Rakesh S, Sandeep M. Scattered across data mart troubles triumph over in the course of data acquisition through the decision support system. Int J Computer Technol Appl 2014;2:411-9.
16. Thilini A, Hugh W. Key organizational factors in data warehouse architecture selection. J Decis Support Syst 2010;49:200-12.
17. Galhardas H. Declarative Data Cleaning Model, Language and Algorithms. Procceeding VLDB Conference. San Francisco: Morgan Kaufmann; 2001. p. 371-80.
18. Başaran, Beril P. A Comparison of Data Warehouse Design Models. Turkey: Atilim University, The Graduate School of Natural and Applied Sciences; 2005.
19. Arunachalam P. Healthcare data warehousing. Imanagers J Computer Sci 2017;4:2-6.
20. Palmer J. The clinical data warehouse, a new mission-critical hub. Int J Palmer Oracle 2013;1:5-12.
21. Frank L, Andersen S. Evaluation of Different Database Designs for Integration of Heterogeneous Distributed Electronic Health Records. The 2010 IEEE/ICME International Conference on Complex Medical Engineering; 2010. p. 204-9.
22. Masseroli M, Bonacina S, Pinciroli F. Java-based browsing, visualization and processing of heterogeneous medical data from remote repositories. Conf Proc IEEE Eng Med Biol Soc 2004;5:3326-9.
23. Sen A, Sinha AP. A comparison of data warehousing methodologies. Commun ACM 2005;3:79-84. Available from: http://www.cacm.acm.org/magazines/2005/3/6272-a-comparison-of-data-warehousing-methodologies/abstractzines/2005/3/6272-a-comparison-of-data-warehousing-methodologies/abstract. [Last accessed on 2017 Jun 08].
24. Sheta O, Eldeen AN. Building a health care data warehouse for cancer diseases. Int J Database Manage Syst 2012;4:39-46.
25. Rajib D. Health Care Data Warehouse System Architecture for Influenza (flu) Diseases. Global Institute of Management and Technology, Krishnanagar. West Bengal, India: Department of Computer Science and Engineering; 2013.
26. Mishra D, Kumar AD, Mausumi D, Mishra S. Predictive data mining: Promising future and applications. Int J Computer Communication Technol 2010;2:20-8.
27. Smith D, Marlow UK. Data Mining in the Clinical Research Environment. PhUSE; 2007.
28. Zhu X, Davidson I. Knowledge Discovery and Data Mining: Challenges and Realities. New York: Hershey; 2007.
29. Prasanna D, Hsu Y, Srivastava C. Data Mining for Health care Management. 2011 SIAM International Conference on Data Mining; 2011.
30. Epstein I. Clinical Data-mining: Integrating Practice and Research. London: Oxford University Press; 2010.
31. Tan PN, Steinbach M, KumarV. Introduction to Data Mining. Boston: Pearson Addison-Wesley; 2005.
32. Barnett V, Lewis T. Outliers in Statistical Data. 3rd ed. New York: John Wiley & Sons; 1994.
33. Beyer M. Agile Techniques Augment but do not Replace Decision Support System and Data Warehouse Best Practice. Gartner Research; 2010.
34. Ibrahim I, Oye N. Design of a data warehouse model for a university decision support system. Inform Knowl Manage 2015;5:89-92.
35. Roddick JF, Fule P, Graco WJ. Exploratory Medical Knowledge Discovery: Experiences and Issues. Annu Rev Inform Sci Technol 2004;38:331À69.
36. Matteo GS. In: Cuzzocrea A, Dayal U, editors. Modern Software Engineering Methodologies Meet DW Design. Bologna: DaWaK. 6862; 2011. p. 66-79.
37. Piatetsky-Shapiro G. Knowledge Discovery in Databases: An Overview. AI Magazine. 1992. Available from: http://www.tesisenred.net/bitstream/handle/10803/9159/Tesi-part3.pdf. [Last retrieved on 2015 Oct 06].
38. Rajendra GS, Dani A. Comparative study of prototype model for software engineering with system development life cycle. IOSR J Eng 2012;7:21-4.
39. Chuck B, Daniel MF, Amit GC, Stanislav V. Dimensional Modeling: In a DSS Environment. New York 10504-1785 USA: IBM Corporation, North Castle Drive Armonk; 2006.