Designing Framework for Data Warehousing of Patient Clinical Records using Data Visualization Technique of Nigeria Medical Records

Dr. Oye, N. D.
Department of Computer Science,
Modibbo Adama University of Technology,
Yola, Adamawa state, Nigeria

Emeje, G.D
Department of Computer Science
Modibbo Adama University of Technology,
Yola, Adamawa state, Nigeria

Abstract:

The availability of timely and accurate data is vital to make informed medical decisions. Today patients data required to make informed medical decisions are trapped within fragmented and disparate clinical and administrative systems that are not properly integrated or fully utilized. Therefore, there is a growing need in the healthcare sector to store and organize size-able clinical record of patients to assist the healthcare professionals in decision making processes. This research is about integrating and organizing disparate clinical record of patients into a Data Warehouse (DW) for data analysis and mining, which will enable evidence-based decision-making processes. This study uses SQL Server Integration Service (SSIS) for the extraction transformation and loading (ETL) of patient clinical records from fragmented administrative systems into the DW, Data visualization technique was used for data presentation while SQL Server Reporting Services and Business Intelligent (BI) tools for designing the output results. This research will assist medical experts and decision makers in the healthcare industry in planning for the future. This research also provides an architecture for designing a clinical DW that is not limited to single disease and will operate as a distributed system, Periodic update on a daily basis is possible because contemporary technologies have narrowed the gap between updates which enable organizations to have a "real time" DW which can be analyzed using an OLAP Server. In conclusion if this research is deployed it will aid medical decision-making process in the Nigerian medical sector.

Keywords: Clinical records; Data marts; Data Warehousing; Framework; Patient record

1. INTRODUCTION

1.1 Background of the study

Knowledgeable decision making in healthcare is vital to provide timely, precise, and appropriate advice to the right patient, to reduce the cost of healthcare and to improve the overall quality of healthcare services. Since medical decisions are very complex, making choices about medical decision-making processes, procedures and treatments can be overwhelming. (Demetriades, Kolodner, & Christopherson, 2005). One of the major challenges of Information Technology (IT) in Healthcare 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 (Mann, 2005; Zheng et al., 2008; Goldstein et al., 2007; Shepherd, 2007).

A massive amount of health records, related documents and medical images generated by clinical diagnostic equipment are created daily (Zheng, Jin, Zhang, Liu, & Chu, 2008). 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 HIS (Hospital Information System), RIS (Radiology Information System), PACS (Picture Archiving and Communications System) in various hospitals, departments and laboratories being primary locations (Zheng, Jin, Zhang, Liu, & Chu, 2008). 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, Jin, Zhang, Liu, & Chu (2008).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 is 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), Zheng et al., (2008).

This causes Clinical Data Stores (CDS) with isolated information across various hospitals, departments, laboratories and related administrative processes, which are time consuming and demanding reliable integration (Sahama, & Croll, 2007). 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, healthcare begins to suffer because medical practitioners and healthcare providers are unable to access and use this information to perform activities such as diagnostics, prognostics and treatment optimization to improve patient care (Saliya, 2013).

1.2 Problem Statement

The availability of timely and accurate data is vital to make informed medical decisions. Every type of healthcare organization faces a common problem with the considerable amount of data they have in several systems. Such systems are unstructured and unorganized, demanding computational time for data and information integration (Saliya, 2013). Today

Patient's data required to make informed medical decisions are 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 healthcare sector there is need for health industries to be open towards adoption of extensive healthcare decision support systems, (Abubakar, Ahmed, Saifullahi, Bello, Abdulra'uf, Sharifai and Abubakar, 2014). There is a growing need in the healthcare scenario to store and organize sizeable clinical data, analyze the data, assist the healthcare professionals in decision making, and develop data mining methodologies to mine hidden patterns and discover new knowledge (Ramani, 2012). Data warehousing integrates fragmented electronic health records from independent and heterogeneous clinical data stores (Saliya, 2013) into a single repository. It is based on these concepts that this study plan to design a Data Warehousing and Mining Framework that will organize, extract, and integrate medical records of patients.

1.3 Aim and Objectives of the Study

The aim of this study is to design a Data Warehousing and Mining Framework for clinical records of patients and the objectives of the study are to:

i. Design a database for the Data Warehouse (DW) Prototype Model using Dimensional Modeling and Techniques.

ii. Simulate the data warehouse database in order to generate reports, uncover hidden patterns, and knowledge from the DW to aid decision making, using the SQL Server 2014, Business Intelligence (BI) Tools and Microsoft Reporting Services.

iii. Develop a web platform that will integrate the front-end, middle-end and the back-end using visual studio 2015 as the development platform. ASP.NET, bootstrap Cascading Style Sheet (CSS), HTML5 and JavaScript will be used for the frond end, C# as the middle end programming language and Microsoft SQL Server Management Studio for the back end development.

1.4 Significance of the Study

It is clear that advanced clinical data warehousing and mining information systems will be a driver for quality improvements of medical care. This ability to integrate data to have valuable information will result in a competitive advantage, enabling healthcare 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. This research will encourage and challenge many government and non-governmental healthcare providers to opt for data warehouse and mining investment in order to improve information access within their organization, bringing the user of their information system in touch with their data, and providing cross-function integration of operation systems within the organizations.

1.5 Definition of Terms

In this section, we have operationally defined some technical terms that were used in this research.

i. Decision Support System: This refers to the system that uses data (internal and external) and models, to provide a simple and easy to use interface, hence, allowing the decision maker to have control over the decision process.

ii. Prototype: Refers to the replica of the complete system ("DW prototype framework") working as if it were real.

iii. Architecture: Is the process that focuses on the formation of data stores within a DW system, along with the procedure of how the data flows from the source systems to the application use by the end users.

iv. Heterogeneous: consisting of or composed of dissimilar elements or ingredients.

v. Data Mining: The process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis.

vi. Data Visualization: Technique used in presenting results obtained in a graphical view for easier understanding and comprehension.

1.5.1 *Data warehouse definition*

Inuwa and Garba (2015) 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 is fed from multiple disparate sources into the data warehouse. As the data is fed it is converted, reformatted, re-sequenced, summarized, and so forth. The result is that data once it resides in the data warehouse has a single physical corporate image.

Non-volatile: Data warehouse data is loaded and accessed, but it is not updated. Instead, when data in the data warehouse is 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 at given moment in time. In some cases, a record is time stamped. In other cases, a record has a date of transaction. But 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 as cited by Inuwa and Oye (2015) DW is the conglomerate of all Data Marts within the enterprise. Information is always stored in the dimensional model. Kimball view 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 last 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 has several challenges (Güzin, 2007).

2. LITERATURE REVIEW

2.1.1 *Data warehouse modelling*

Ballard cited by Inuwa and Oye (2015) gave an assessment of the evolution of the concept of data warehousing, as it relates to data modeling for the DW, they defined database warehouse modeling as the process of building a model for the data in order to store in the DW. There are two data modeling techniques that are relevant in a data warehousing environment and they are:

i. Entity Relationship (ER) Modelling: 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, 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.

2.1.2 *Data warehouse modelling techniques*

Thomas and Carol cited by Inuwa and Oye (2015) derived the way a DW or a Data Mart structure in dimensional modelling can be achieved. 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 (in Figure 2.1) 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 centre 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 contain 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. (Inuwa&Oye, 2015).

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 2.2.

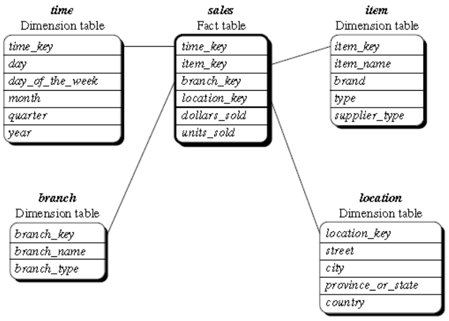


Figure 2. 1: Star Schema (Jiawei, 2012)

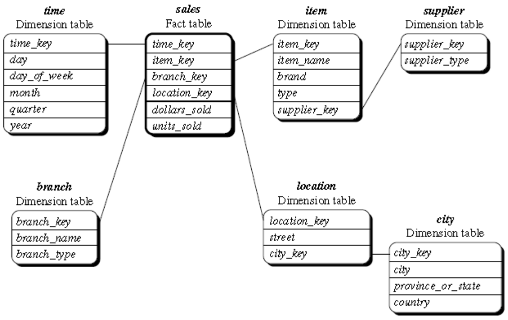


Figure 2. 2: Snowflake Schema (Jiawei, 2012)

2.1.3 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 sub entity 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. Gopinath, Damodar, Lenin, Rakesh& Sandeep, 2014, also summarized the types of Data Marts as follows:

2.1.3.1 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's data in one DW. This gives you the usual advantages of centralization as can be seen in Figure 2.3.

2.1.3.2 Hybrid Data Marts

A hybrid Data Mart allows you to combine input from sources other than a DW. This could be useful for many situations, especially when you need Ad-hoc integration, such as after a new group or product is added to the organization as can be seen in Figure 2.4:

a). A hybrid Data Mart transform data to combine input from sources other than a DW.

b). Extracting the data from hybrid Data Mart based on required conditions.

c). After extracting load into as a departmental Data Marts.

The Data Mart typically contains a subset of corporate data that is valuable to a specific business unit, department, or set of users. This subset consists of historical, summarized, and possibly detailed data captured from transaction processing systems (called independent Data Marts), or from an existing enterprise DW (called dependent Data Marts). It is important to realize that the functional scope of the Data Mart's users defines the Data Mart, not the size of the Data Mart database (Chuck, Daniel, Amit, Carlos and Stanislav, 2006).

2.1.4 Architecture of the DW

Data contained in a DW holds five types of data: data currency, existing data, data summarization (lightly and highly summarized data), and Metadata (Inuwa&Garba 2015). This traditional data warehousing architecture in Figure 2.5 encompasses the following components (Inuwa&Garba 2015):

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.

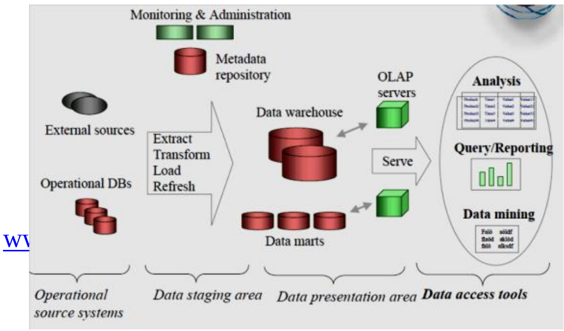
viii. An OLAP (online analytical processing) engine on top of the DW and Data Marts to present and serve multi-dimensional views of the data to analytical tools.

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 multi-dimensional structure optimized for analytical applications. However, the architecture is rather old and, while this basic idea is still intact, it is rather unclear and inaccurate about several facts:

Firstly, 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 (Thilini& Hugh, 2010). Secondly, the interplay between DW and Data Marts in the storage area are not completely clear.

Figure 2. 3: Traditional Data Warehousing Architecture (Inuwa&Garba, 2015).



Actually, in practice this is one of the biggest discourses about data warehousing architecture with two architectural approaches proposed by Bill Inmon and Ralph Kimball (Inuwa&Garba, 2015). Inmon places his data warehousing architecture in a holistic modelling approach of all operational and analytical databases and information in an organization, the Corporate Information Factory (CIF). What he calls the 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 (Inuwa&Garba, 2015).Inmon's approach, also called enterprise DW architecture by Thilini and Hugh (2010) 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 and in an iterative fashion. Kimball on the other hand proposes a bottom-up approach which starts with process and application requirements (Kimball, Reeves & Ross) as cited by Inuwa and Garba (2015). With this approach, first the Data Marts are designed based on the organization's 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 multi-dimensional. The DW is then just the combination of all Data Marts, where the single Data Marts are connected and integrated with each other via the data bus and so-called conformed dimensions that are Data Marts use, standardized or 'conformed' dimension tables (Inuwa&Garba, 2015).

When two Data Marts use the same dimension, they are connected and can be queried together via that identical dimension table. The data bus is then a net of Data Marts, which are connected via conformed dimensions. This architecture (also called Data Mart bus architecture with linked dimensional Data Marts by Thilini and Hugh (2010) therefore forgoes a normalized, enterprise-wide data model and repository.

In Figure 2.4, there are a number of options for architecting a Data Mart. For example:

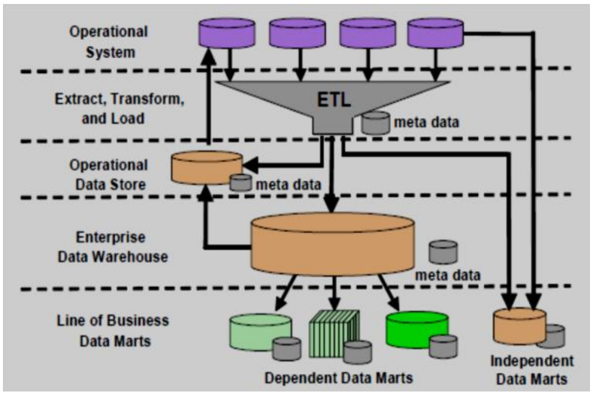


Figure 2. 4: DW Architecture (Inuwa&Garba, 2015)

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 flows 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). Data Mart consolidation 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 (Chuck et al., 2006).

2.1.5 Benefits of data warehousing

There are several benefits of data warehousing. The most important ones are listed as follows (Kimball & Ross, 2002):

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, make decisions based on data. DW increases 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 is 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 does not need to be in normalized form. In fact, it is usually de-normalized to support faster data retrieval.

2.1.6 Data warehousing in medical field

Health care organizations require data warehousing solutions in order to integrate the valuable patient and administrative data fragmented across multiple information systems within the organization. As stated by Kerkri cited by Saliya (2013), at a technical level, information sources are heterogeneous, 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 (Zheng et al., 2008).

A medical data warehouse is a repository where healthcare 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, and data formats; organizational behaviors on processing the data and culture across the data management population. Today's healthcare 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 labour expenditures, supply utilization, procedures, medications prescribed, and other costs associated with patient care, healthcare professionals can identify and correct wasteful practices and unnecessary expenditures (Sahama & Croll, 2007).

Medical domain has certain unique data requirements such as high volumes of unstructured data (e.g. digital image files, voice clips, radiology information, etc.) and data confidentiality. Data warehousing models should accommodate these unique needs. According to Pedersen and Jensen cited by Saliya (2013) 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 (2013), the advantages and disadvantages of data warehousing are given below.

Advantages:

1. Ability to allow existing legacy systems to continue in operation without any modification
2. Consolidating inconsistent data from various legacy systems into one coherent set
3. Improving quality of data
4. Allowing users to retrieve necessary data by themselves

Disadvantages:

1. Development cost and time constraints

2.1.7 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 tumours, 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 tumours are cancerous. Benign tumours do not grow uncontrollably, do not invade neighbouring 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 (Sheta & Ahmed, 2012).

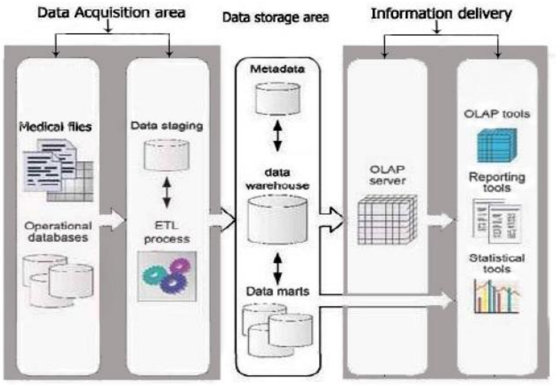


Figure 2. 5: Cancer DW Architecture (Sheta& Ahmed, 2012)

2.1.8 *Influenza disease data warehouse architecture*

Influenza, commonly known as the 'flu', is an infectious disease of birds and mammals caused by ribonucleic acid (RNA) viruses of the family Orthomyxoviridae, the influenza viruses, (Rajib, 2013). 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 three to five million yearly cases of severe illness and about 250,000 to 500,000 yearly deaths (Rajib, 2013). 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), Nasopharyngeal culture.

Figure 2.6 shows the proposed architecture for the health care data warehouse specific to Influenza disease by Rajib in 2013. Architecture of Influenza specific health care data warehouse system builds with Source Data components in the left side where multiple data that comes from different data source and transform into the Data Staging area before integrating. The Data staging component present at the next building block. Those two blocks is under Data Acquisition Area. In the middle Data Storage component that manages the data warehouse data. This component also with Metadata, that also keep track of the data and also with Data Marts. Last component of this architecture is Information Delivery component that shows all the different ways of making the information from the data warehouse available to the user for further analysis (Rajib, 2013).

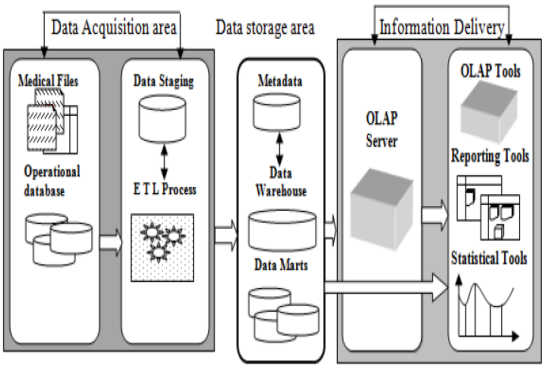


Figure 2. 6: Data Warehouse Architecture for Influenza Disease (Rajib, 2013)

2.1.9 *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 .

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 Extract, Transform and Load processes. Figure 2.12 shows a graphical overview of cardiac surgery data warehousing model (Saliya, 2013).

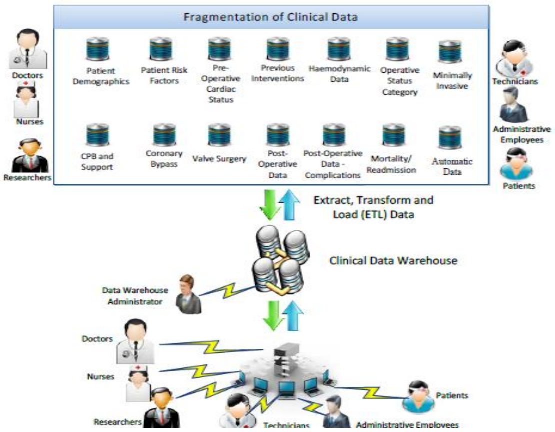


Figure 2. 7: Data Warehousing Model for Integrating Fragmented Electronic Health Records from Disparate and Heterogeneous Cardiac Surgery Clinical Data Stores (Saliya, 2013).

2.1.10 *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, (Abubakar et al., 2014).

Gap Analysis

The research carried out by other researchers, only focus on building a data warehouse for a particular disease. No single study exists which adequately covers the integration of clinical record of patients into a data warehouse for data analysis and mining that is not specific to a single disease. There is also no specific research that covers the integration of medical record of patients using the Nigerian medical records for analysis and mining using data visualization for decision making.Therefore this research is designing a framework for integrating patient clinical records into a single data warehouse for data analysis and mining using data visualization technique that is not specific to single disease using Nigeria medical record.

3. METHODOLOGY

The following research method and tools were used in achieving these research objectives:

System Analysis: UML was used in analysing all the physical and logical models. **OLAP:** Online Analytic processing (Server) was used for processing and analysing data from the server on a web browser. **ETL:** Extraction transformation and loading of data into the data warehouse for mining and visualization purpose **SSIS:** Integrating the operations data with the main data warehouse database.

Visualization Technique: Presentation of mine data from the data warehouse as visual effects rather than row and columns. Data are displayed in form of graphs and shapes for easy interpretation and understanding. Microsoft Visio 2015 as the UML tool, was used in in designing all the logical models of the proposed system. Aspx, HTML 5 and Bootstrap CSS were used in designing the GUI and SQL Server Reporting Services was used in designing the data visualization, clinical decision support and mining outputs of the system. The programming language of choice for the middle end was C# pronounced C-sharp. The database used in designing the physical models is Microsoft SQL Server 2014, while, SQL Server 2012 server tools were used for writing the ETL (Extraction, Transformation and Load) program.

4. RESEARCH RESULTS

System Prototype Development and Validation

The system prototype development is the actual application of the analysis and design that has been carried out. In this phase of the study, we have designed the DW (Fact and dimension tables) prototype, the ETL (Extract, Transform and Load), the front end (GUI) and the Middle end of the application for the purpose of this study. Validation process involves the confirmation by medical system administrators and personnel's that an information system (DW prototype) has been realized appropriately and it is in conformity with the User's needs and intended use. Figures 1-7 are available in the appendix.

FIGURE:1 Shows the architectural design of the Framework for Data Warehouse and Mining Clinical Records. The architecture has the followings abilities:

i. Integration of the independent and dependent Data Marts within the architecture.

ii. It has a multi-tier ETL which are simpler and in different stages.

iii. It has a standard access points for all medical and non-medical experts (Using a three tier server).

iv. It has the ability to Mine and analyse records using Data Visualization

The architecture has a back and frond end system in which so many activities are carried out. The back end systems comprise of the operational data source system, data staging area and the data presentation area. Data are first extracted from different operational data source systems using ETL and then stored at the data staging area where it is being processed as soon as it is captured. The activities of the ETL at the data staging area include data cleansing and validation, data integration, data fixing and data entry errors removal, transforming and refreshing data into a new normalized standard. As soon as data is cleaned, the transformed data are loaded and indexed into the data presentation area where the DW is located. The frontend systems in the other hand comprise of the main servers (OLAP) and data access tools. The OLAP server hold the copy of the cleansed clinical records. The data access tool is the interface where applications are stored, which allows for data Analysis, Reporting/Querying and Data mining activities.

FIGURE: 2, is about ***Creating and loading of the Clinical DW database***

The DW database was created on an MSSQL Server 2014 database, and the data loading was also done through the SSIS package. Fact and Dimension tables were pushed into the DW database, Figure 2 shows Physical database of the DW Model by Star Schema. The DW database is the one that integrates the Fact and Dimension tables of patient's clinical records into the MSSQL Server 2014 database. It was part of the MSSQL Server database that we used as the database repository. The DW database was populated with the correct data of good quality that we can make use of as the data repository. Data visualization is all about presenting data in graphical format rather than rows and columns thereby making data interpretation easy and is best use by decision makers to take decisions for organizations.

FIGURE 3, shows some of the Dimension and Fact tables that were extracted from the source system into the staging database area. The tables were cleansed and refreshed at this point and ready for transfer into their respective Data Marts. Having designed the Fact and Dimension tables and the extraction of data from the source system, then the researcher populated the Data Marts with the data that was extracted earlier from the staging databases. It is now from the Data Marts that another ETL was performed to transport the data into the DW database.

System Verification, Reports and Data Visualization

The best way to verifying the data in the DW is to prepare queries on these data. In this study, certain reports, data analysis and data visualization were presented and the purpose of these reports and analysis is to demonstrate the usefulness of the DW approach to data presentation and decision making. Even though these reports and analysis were based on some random requirements, this set of sample reports and analysis can be used as a basis for generating more comprehensive sets by applying complex queries on the data. We have classified the Analysis, Reports and Data Visualization that we generated from the DW based on all the analysis on patient diagnosis and details. **Figure 4**, illustrates the general system architecture for this study. It shows how the Source System database, Servers and the User/administrator system are connected for possible data visualization and decision making. **FIGURE: 5**, shows data visualization of Malaria patients and their LGA of residence within the year 2015, 2016 and 2017. **Figure 6**, shows data visualization of Tuberculosis patients and their LGA of residence within the year 2015, 2016 and 2017. Decision makers can use the visualized report to determine the trend of infection between the various patients LGA of residence. Proper decision can be taken on how to curb the rate of infection within the LGA's with high rate of illness by setting appropriate infrastructure and medical personnel in those LGA**. Figure 7**, shows a report of patients diagnosed with typhoid-fever based on their gender, month and the year of diagnosis.

5. CONCLUSION

This research has design and implemented a Clinical DW and Mining system using data visualization technique within the context of the healthcare service, to better incorporate patient records into single systems for simpler and improved data mining, analysis, reporting and querying. The Clinical DW we built contains only the data that is required for data mining, reporting and analysis for the purpose of this study and it can be updated periodically, such that all the data can be integrated from different source systems into the central DW. The system is design to operate as a distributed system. Periodic update on a daily basis is possible because contemporary technologies have narrowed the gap between updates. This can enable organizations to have a "real time" DW which can be analyzed using an OLAP Server.

The research focused on developing a Framework for DW and Mining of clinical records of patient, for improve data analysis and mining using data visualization. The study also considered the ideologies of data warehousing in the course of this research and demonstrated how data can be incorporated from diverse desperate heterogeneous clinical data stores into a single DW for mining and analysis purpose, to aid medical practitioners and decision makers in decision making. The researchers have also been able to develop a web based data mining, reporting and analysis tool (GUI) where users can interact with the system to get a speedy and timely information needed for the clinical decision making and community diagnosing.

The Framework for DW and Mining of clinical record of patients was design not to be specific to only a number of disease but to accept as many diseases as possible without any limitation. The ideologies that the researchers followed to develop this system makes it scalable and as such, it can be adopted for any form of disease, infections analysis and mining by medical institutions and healthcare providers in Nigeria. The designed framework can be used by industry professionals and researcher for implementing data warehousing system in the medical field, for long time data analysis and mining. The framework can also be used for community diagnosing in cases of outbreak of certain disease. Developing a Framework for DW and Mining of Clinical records is very essential, particularly for medical decision-makers, academic researchers, IT professionals and non-professional. Clinical data mining must not only support medical professionals and decision makers to understand the past, but also it strive professionals to work towards new prospects.

6. ACKNOWLEDGMENTS

We thank the experts who have helped to review this paper.

7. REFERENCES

[1] Abubakar, etal (2014). Building a diabetes data warehouse to support decision making in healthcare industry. *IOSR Journal of Computer Engineering (IOSR-JCE), 16*(2), e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. IX (Mar-Apr. 2014), PP 138-143 www. Retrieved from iosrjournal.org

[2] Chuck, B., Daniel, M. F., Amit, G. C., & Stanislav, V. (2006). *Dimensional Modeling: In a DSS Environment.* NY 10504-1785 U.S.A.: IBM Corporation, North Castle Drive Armonk

[3] Demetriades, J. E., Kolodner, R. M., & Christopherson, G. A. (2005). Person Centered Health Records. Towards Healthy People. . *Health Informatics Series*. USA: Springer. Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] Goldstein, D., Groen, P. J., Ponkshe, S., & Wine, M. (2007). *Medical informatics 20/20: quality and electronic health records through collaboration, open Solutions, and innovation.* Massachusetts, Sudbury, USA: Jones and Bartlett Publishers, Inc.

[5] Gopinath, T., Damodar, N. M., Lenin, Y., Rakesh, S., & Sandeep, M. (2014). Scattered Across Data Mart Troubles Triumph Over in the course of Data Acquisition through the Decision Support System. *International Journal of Computer Technology and Application, 2*(5), 411-419.

[6] Güzin, T. (2007). *Developing a Data Warehouse for a University Decision Support System.* Atilim University, The Graduate School of Natural and Applied Sciences.

[7] Ibrahim I., & Oye, N. (2015). Design of a Data Warehouse Model for a University Decision Support System. *Information and Knowledge Management, 5*.

[8] Inuwa, I., & Garba, E. (2015). An Improved Data Warehouse Architecture for SPGS, MAUTECH, Yola, Nigeria. *West African Journal of Industrial & Academic Research, 14*(1).

[9] Kimball, R. &. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling. In R. Kimball, & M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling.* New York: John Wiley & Sons, Inc.

[10] Mann, L. (2005). From “silos” to seamless healthcare: bringing hospitals and GPs back together again. *Medical Journal of Australia, 1*(182), 34-37. Retrieved from http://www.mja.com.au/public/issues/182_01_030105/man10274_fm.html

[11] Rajib, D. (2013). *Health care data warehouse system Architecture for influenza (flu) Diseases.* Global Institute of Management & Technology, Krishnanagar., Department of Computer Science & Engineering, West Bengal, India.

[12] Sahama, T. R., & Croll, P. R. (2007). A data warehouse architecture for clinical data warehousing. *First Australasian Workshop on Health Knowledge Management and Discovery.* Retrieved from http://crpit.com/confpapers/CRPITV68Sahama.pdf

[13] Saliya, N. (2013). *Data warehousing model for integrating fragmented electronic health records from disparate and heterogeneous clinical data stores.* Queensland University of Technology, School of Electrical Engineering and Computer Science Faculty of Science and Engineering, Australia.

[14] Shepherd, M. (2007). Challenges in health informatics. *40th Hawaii International Conference on System Sciences.* doi:10.1109/HICSS.2007.123

[15] Sheta, O., &Ahmed, A. N. (2012). Building a Health Care Data Warehouse for Cancer Diseases. *International Journal of Database Management Systems (IJDMS), 4*.

[16] Thilini, A., & Hugh, W. (2010). Key organizational factors in Data Warehouse architecture selection. *Journal of Decision Support Systems, 49*(2), 200–212.

[17] Zheng, R., Jin, H., Zhang, H., Liu, Y., & Chu, P. (2008). Heterogeneous medical data Share and integration on grid. *International Conference on BioMedical Engineering and Informatics.* doi:10.1109/BMEI.2008.185

[18] Zhu, X., & Davidson, I. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities.* New York: Hershey, New York.

APPENDIX

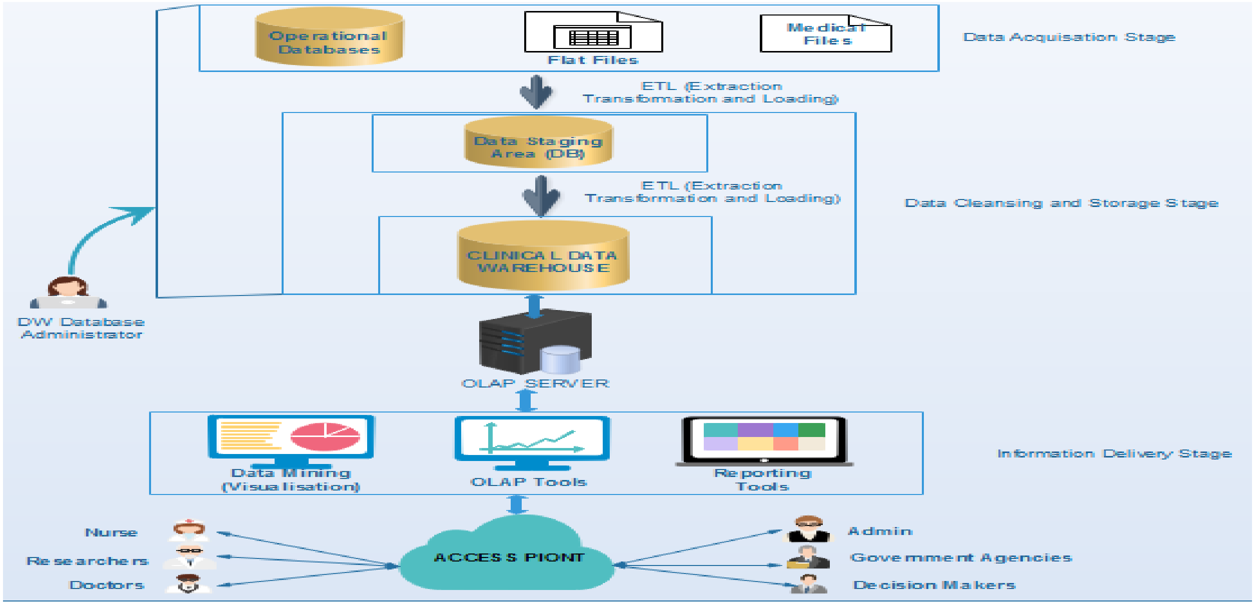


Figure 1: Framework for Data Warehousing and Mining Clinical Records of Patients

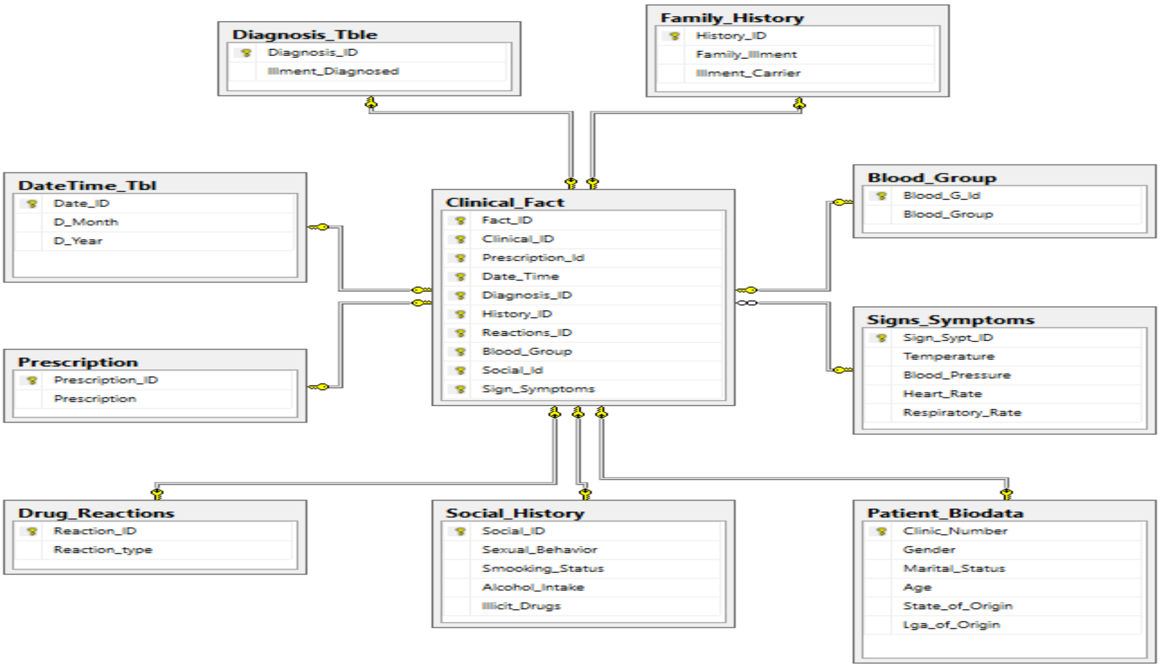


Figure 2: Star schema of the designed Framework for Data Warehousing and Mining Clinical Records of Patients

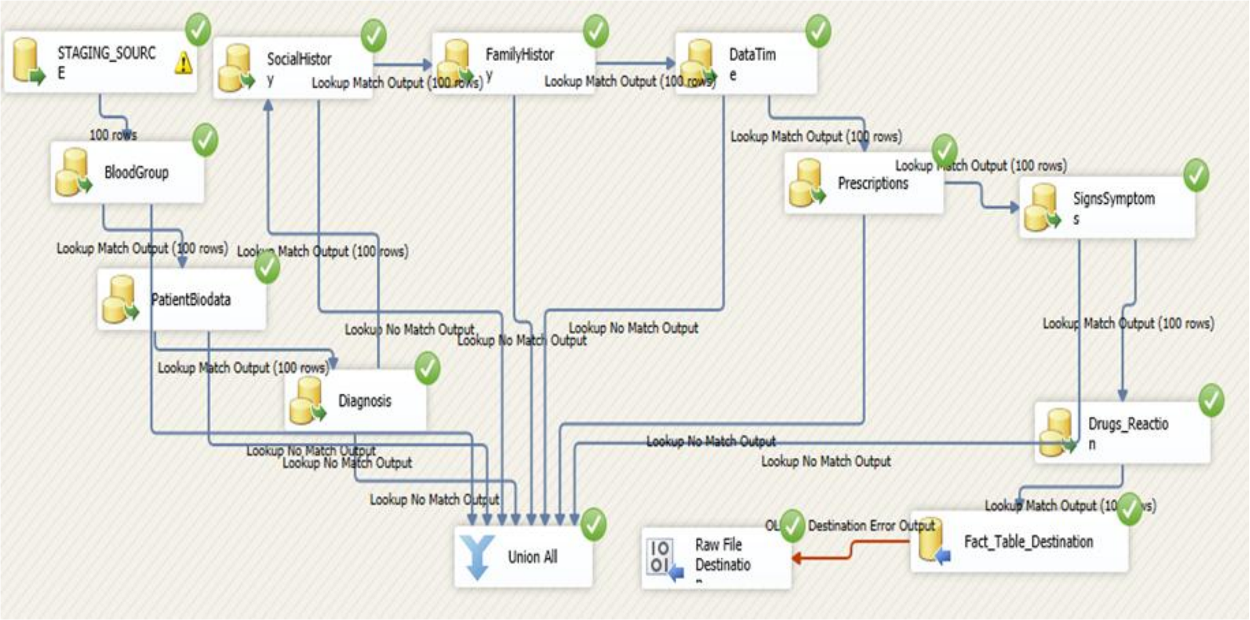


Figure 3: ETL for data cleansing and loading of the DW

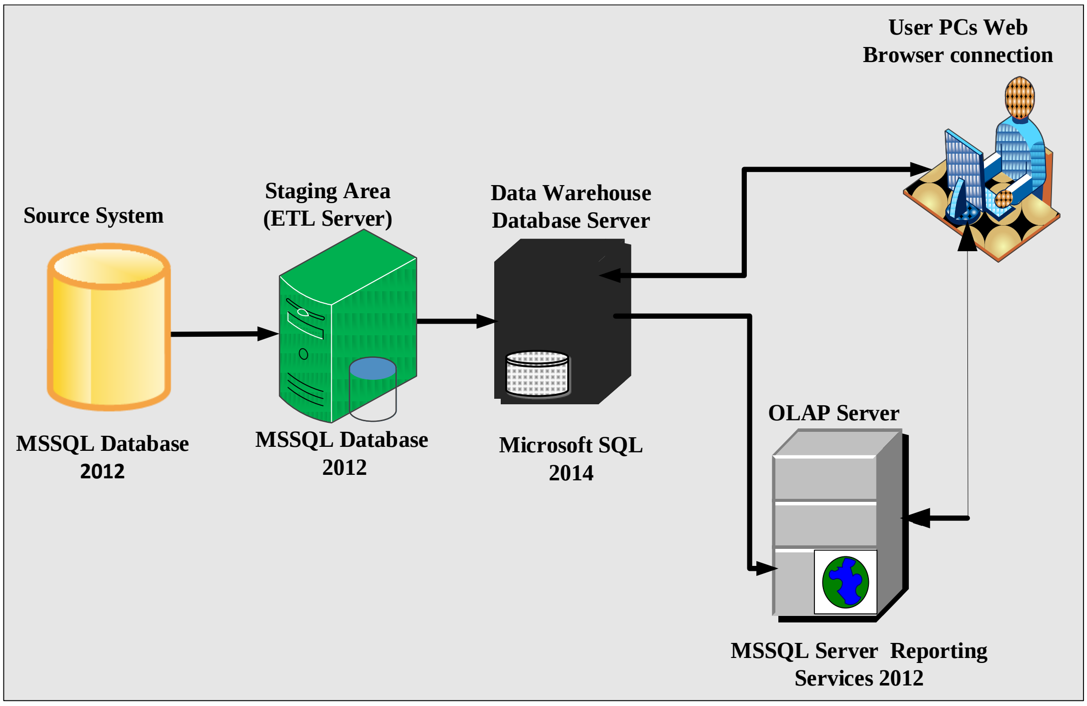


Figure 4: Showing the structural architecture of the system

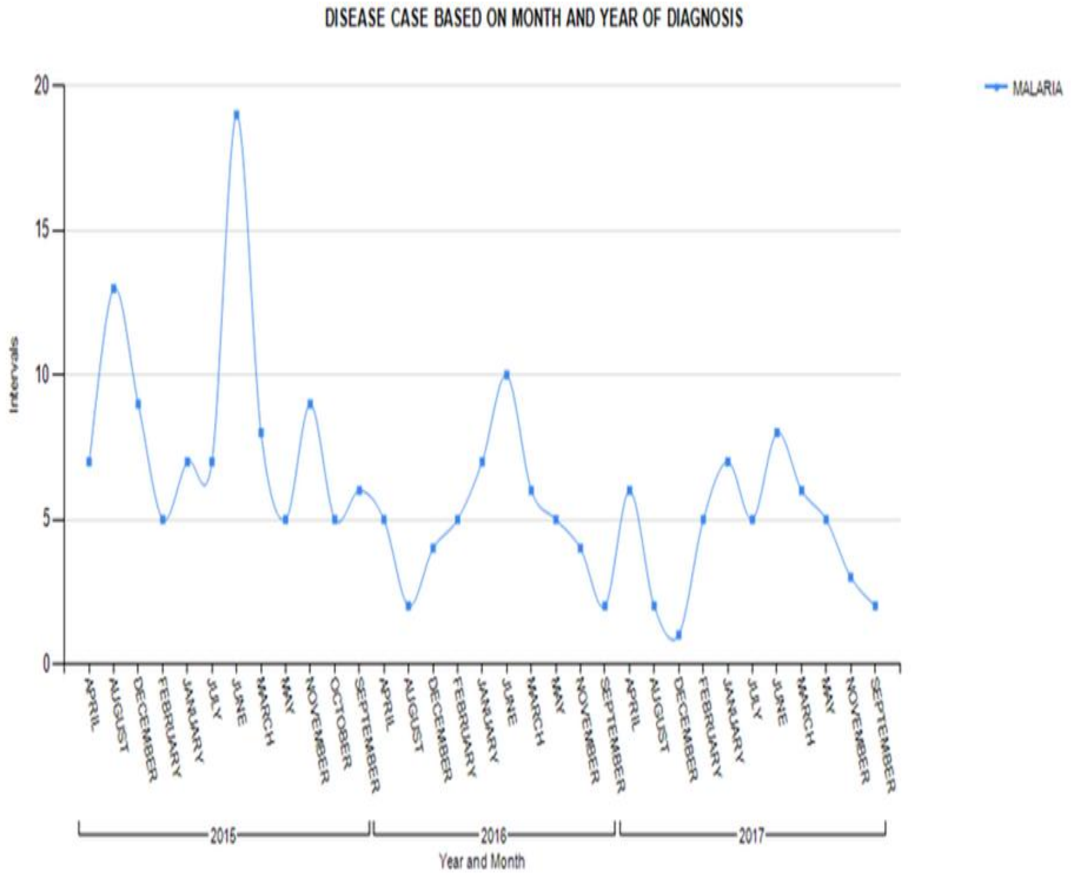


Figure 5: Malaria output result, based on month and year of diagnosis

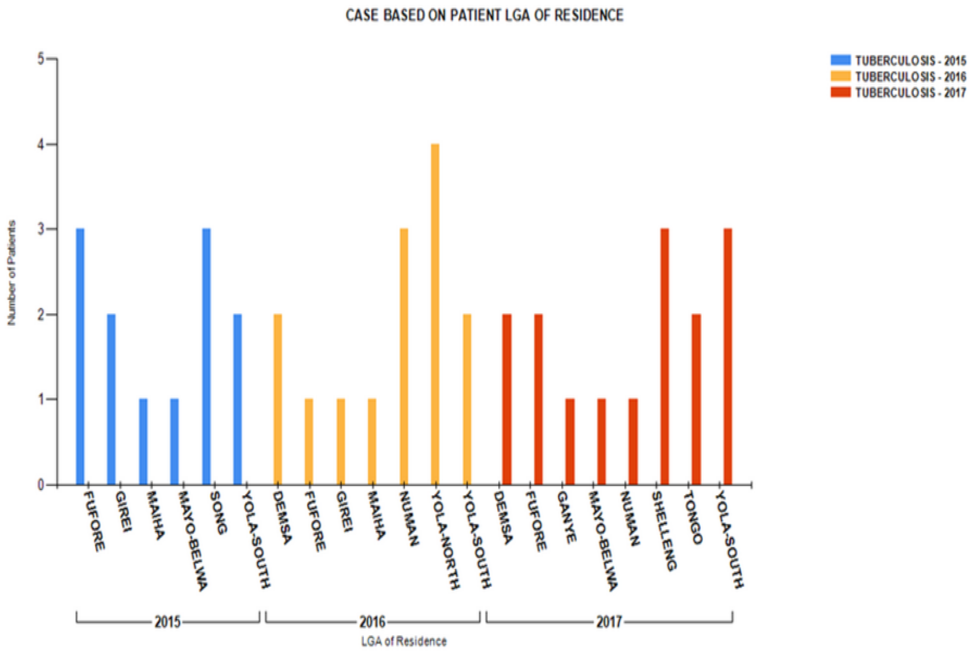


Figure 6: Tuberculosis report based on LGA of residence and year of diagnosis

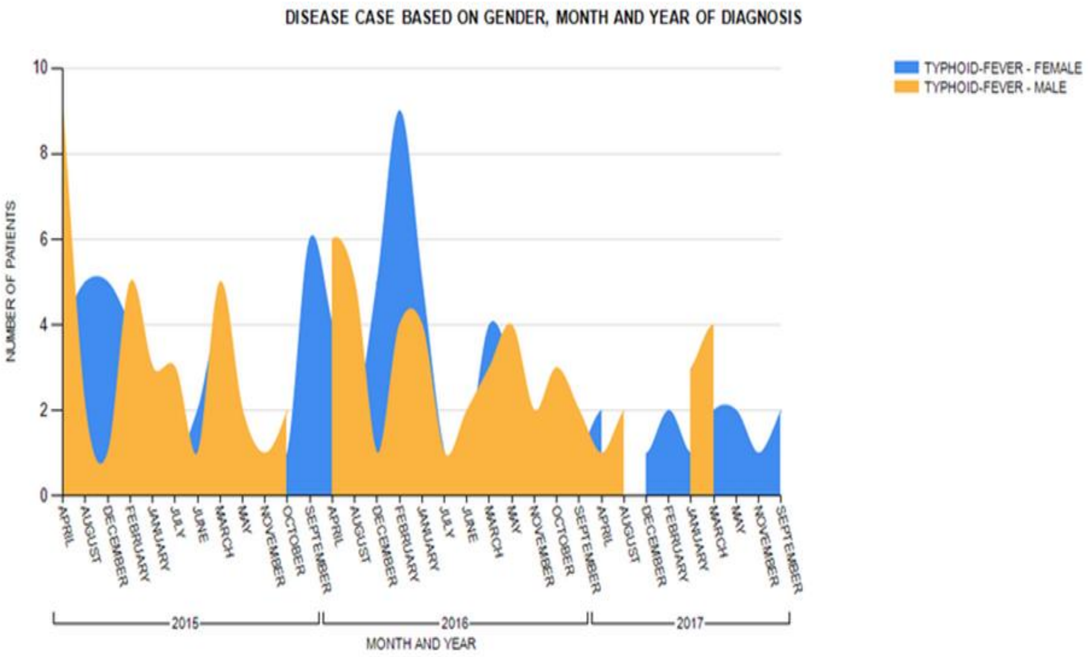


Figure 7: Typhoid-Fever report based on gender, month and year of diagnosis