Developing a Decision Support System Using Cancer Data Warehouse

P.Ramachandran, Dr.N.Girija, Dr.T.Bhuvaneswari

Abstract— A data warehouse is a powerful repository of large, multi dimensional data cubes that answers complex queries and helps in decision making process. Data warehouses often act as a predecessor in the data mining process. Data warehousing has become essential in medical intelligence. This paper reviews the implementation and use of data warehouse in health and service sector specific to cancer disease. Initially A Clinical Data Warehouse is developed which integrates data by automatically performing the ETL procedure i.e. extracting the data from different sources, transforming and gets itself loaded to supports the data mining system which could be used by doctors and medical analysts as a Decision Support System (DSS), to predict cancer in its earlier stages and provide the needed treatment. A sequential methodology to develop the data warehouse architecture using cancer databases is discussed at length in this paper.

Index Terms— Data warehouse, Cancer disease, ETL procedure, Decision Support System, Data Mining

———————————————————— ◆ ————————————————————

1. INTRODUCTION

A Data ware house is a repository for historical, integrated and consistent data. It contains large volume of data which can be stored and processed. Data warehouse is equipped with necessary tools so that relevant results can be obtained that helps in decision making process. The results from data warehouse are obtained using Data mining techniques which involve the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. When beginning to work on a data mining problem, it is first necessary to bring all the data from different data sources together into a set of instances. Integrating data from different sources usually presents many challenges. The data must be assembled, integrated, and cleaned up. The idea of area wide database integration is known as data warehousing and processing the data in the ware house to extract results is known as data mining. Thus data mining and data warehousing are interrelated. Data warehouses provide a single consistent point of access to multiple data sources spread across different locations and can be used to gain knowledge about a particular concept and also to take important decisions. A data warehouse is built on OLAP technology which supports data integration, analyses and decision making, while the OLTP database is used to update and process individual records of a patient on a day to day basis. The results of the knowledge gained are also stored in a Knowledge base. The important concept of data warehouse is that it minimises redundancy by storing a non volatile record

————————————————

• *P.Ramachandran, Ph.D Research Scholar, Department of CS&A,*
SCSVMV University, Kanchipuram, India,
E-Mail: rkmvc.rc@gmail.com
• *Dr.N.Girija, Lecturer, Department of IT, Higher College of Technology,*
Ministry of Manpower, Muscat,
E-Mail: Nbgir2004@gmail.com
• *Dr. T. Bhuvaneswari, Asst.Professor, Department of CS&A,*
L.N. Government of Arts and Science College, Chennai, India,
E-Mail: t_Bhuvaneswari@yahoo.com

in a particular place instead of storing the same in multiple database. To devise a Data warehouse data assembly, integration, cleaning, aggregating, and general preparations are made. This is achieved by ETL process. The data is first extracted from several databases extending across a large geographical area and converting it into a suitable format such as a multi dimensional data cube. While preparing the data for this step lot of errors arise out of inconsistency, redundancy, missing data and anomalies in the records. To reduce this data preprocessing is done. Finally the data is loaded and integrated into data warehouse and is ready for the knowledge to be mined. The movement toward data warehousing is recognition of the fact that the fragmented information that medical institutions and research centers uses to support day-to-day operations at a management level can have immense strategic value when brought together. Clearly, the presence of a data warehouse is a very useful precursor to data mining.

Data warehouse systems are currently well established in many business organizations to improve their marketing strategies and to help in decision making process. There is a need to implement a data warehouse system in health care sectors. Medical data bases are scattered across a large area, when brought together and integrated they can be used to obtain results and make decisions which ranges from earlier detection of the disease to predicting the exact treatment, its results and the curability of the disease. Cancer database is one such instance which is spread across wide geographical area where the pieces of much needed information are scattered.

Cancer is one of the most leading causes of death worldwide. Every year equal number of men and women are diagnosed with cancer. Uncontrolled growth of cells in any of the tissues or parts of the body causes Cancer. Cancer may occur in any part of the body and may spread to several other parts. Once cancer starts spreading to other parts then it is known as

malignant tumour which is very difficult to cure and may lead to death. Early diagnosis of cancer in its beginning stage is very helpful in curing the disease. Cancer may occur due to genetical, biological and environmental factors. Several attributes such as age, gender, habits, family history of cancer etc., plays an important role in causing cancer. A data warehouse built using cancer database will provide clear information to a medical analyst to make decisions on crucial subjects such as the type of service to be provided, the geographical area in which the awareness program should be concentrated, the resources to be purchased etc,. This data warehouse is built in a way that it aids doctors to make more efficient decisions and provides a greater insight into the disease.

2. REVIEW OF LITERATURE

Dr. Mohamed F. AlAjmi, Shakir Khan et al [1] explained analysis of the content state of e-Learning standard also, presented a functional model of dissimilar learning objects using Web usage mining technique of taking out helpful and previously unknown blueprints from the use of Web.

Souâd Demigha [2] proposes to design and develop a data warehouse system in radiology-senology (DWRS) to assist breast cancer screening in diagnosis, education and research.

Teh Ying Wah et al [3] proposed a 5-stage sequential methodology for the clinical data warehouse development and have developed a methodology and architecture for a specific disease clinical data warehouse to improve the quality of Lymphoma diagnosis and treatment decision support.

 Eric Zapletal et al [4] have presented the methodology for integrating a Clinical Data Warehouse based on the I2B2 framework into the clinical information system. The CIS is the main source of data for the CDW and the CDW produces data (indicators, reports) that enhance the overall healthcare activity, which is in turn processed by the CIS. An important issue is the usefulness of the CDW for routine selection of candidate patients for clinical research studies.

Susan Maskery et al [5] has build an Aggregated Biomedical – Information Browser Interface of the DW4TR to perform manual data mining model based on their own clinical observation on African American women and Caucasian American women. The ABB has cohort retrieval functionality. The Aggregated Biomedical –Information Browser empowers physicians and scientists to directly manually mine the data in a clinical data warehouse, thus offering a highly desired service to this group of users.

Rajni Jindal et al [7] surveyed and evaluated the literature related to the various data warehouse design approaches on the basis of design criteria and proposed a generalized object oriented framework for data warehouse conceptual design based on UML.

Shaker H. Ali El-Sappagh et al [9] have classified the approaches for ETL processes into three categories, modelling based on mapping expressions and guidelines, modelling based on conceptual constructs, and modelling based on UML environment. They have also proposed a novel conceptual model entity mapping diagram (EMD) as a simplified model for representing extraction, transformation, and loading processes.

Joaquin Perez-Ortega et al [10] developed a population-based data warehouse on cancer and a variant of the K-means clustering algorithm to show the centroids and the districts of groups on a Map, This tool proved to be particularly useful for assessing and communicating the results because of its visual expressiveness.

Neera Bhansali [13] integrated both historical and current patient data along with imaging and genotyping data into the data warehouse to improve healthcare delivery.

Abubakar Ado et al [14] proposed a architecture for Health care data warehouse for Diabetes diseases which could be use to monitor Diabetes disease, measure cost of infections and to detect prescription errors using ETL and OLAP technologies.

Joyce C. Niland et al [16] defined the various types of Electronic Medical Records and clinical research data systems in use and described the goals and rationale for integrating these two types of systems to enhance research as well as quality of care.

M.Thenmozhi et al [17] presents an ontology driven tool which helps to automatically derive the conceptual model and logical model for the data warehouse from data source and business requirements by making use of ontology and its reasoning capabilities at various stages of the tool to facilitate the design task.

Young Choi, et al [22] Implemented clinical data warehouse technology to test direct EMR link method with St. Mary's Hospital system. This system includes three different ways of clinical data collection to produce a comprehensive data base, direct data extraction from electronic medical record (EMR) system and manual data entry after linking EMR documents like magnetic resonance imaging findings.

Alaa Khalaf Hamoud et al [23] Converts vast amount of patient's records in Electronic Health Record (EHR) to Electronic Medical Records (EMR) by developing a CDW which combines different sources of clinical data into single repository and use it to produce analytical information to support decisions.

Dr. Osama E.Sheta et al [24] presents the evaluation of the architecture of healthcare data warehouse specific to cancer diseases. The evaluation model is based on Bill Inmon's

definition of data warehouse is proposed to evaluate the Cancer data warehouse.

3. MATERIALS AND METHODS

The data warehouse architecture is developed using MySQL server for storing and retrieving the data and JAVA platform for user interface.

3.1. Data Sources

The Data that is needed for this study is collected from cancer hospital records of three districts in Tamilnadu, India. It consists of cancer and non cancer patient's data. The results were configured by considering levels of aggregation such as Districts (Living area), age, gender and symptoms. The information in the databases is then pre-processed to clean the records and minimize the error. This consumes a lot of time where a missing field is substituted manually by considering its relevance to other fields. The data bases are reduced into multiple tables to minimise redundancies and a very poor record with a lot of missing fields and error is omitted. The Pre-processed data is then integrated into a data warehouse.

3.2 Building Architectural Data Warehouse for Cancer Disease

To develop a cancer data warehouse it is first necessary to integrate the different data sources into a central data repository. An ETL model is developed to extract, transform and load data automatically from different databases with OLAP and OLTP techniques.

3.2.1 ETL Process for cancer data warehouse

- **Data Extraction**

In this preliminary stage data from the chosen three different databases and a user interface is extracted so that they can be transformed and loaded into the data warehouse.

- **Data Transformation**

In the next step the extracted data are transformed into the form that is suitable to load the data warehouse by normalizing them and eliminating redundant and noisy data. Transformation means to map the extracted data to the same format as the cancer data warehouse repository. In this step the fields such as district, state, living area from the three databases are converted into a single field Living Area in the data warehouse. Likewise many alterations are made to transform the data from the databases to data warehouse. The built-in transformation code acts a data conversion transformation tool that enables the data warehouse to convert columns from one data type to another.

- **Data Loading**

Finally the transformed data is loaded and merged into the different dimensions of the data warehouse and this act as a training source for further tasks. The transformation code performs this process automatically for every new user entry

in any of the three electronic data bases which are located in different geographical locations. Every day patients across different geographical area can insert their data in these databases and the data warehouse is updated in planned periodic cycles and not frequently as it contains non-volatile data.
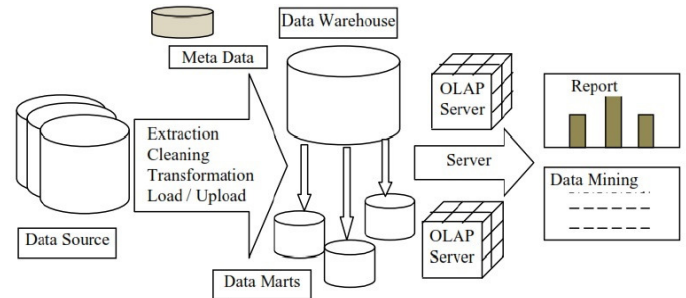

Figure 1: Data Warehouse Architecture

3.2.2 Data Storage

This stage covers the process of loading and storing the transformed data from the ETL staging area into the data warehouse repository. The data is stored in a star schema with each table representing a fact that is directly associated with a patient and a type of cancer.

The ETL code performs incremental loads at regular prescribed intervals and stores the data in this data storage area which support loading integrated data into multiple tables.

3.2.3 The OLTP and OLAP techniques

- **On-Line Transaction Processing**

A Transactional database which is used for On-Line Transaction Processing covers a data that is much shorter in time period and includes only few databases and tables. Also it is very volatile in nature as the updates occur very frequently and the data security is very less. It is not intended to support analytical processing such as generating reports using queries. Thus a Data Warehouse is built here to overcome these flaws of a transactional database by developing an OLTP process which supports OLAP process. This Data warehouse covers a much longer time horizon, includes multiple databases that have been cleaned and processed so that the warehouse's data are only cancer disease oriented and defined uniformly. It also contains non-volatile data which are controlled by planned periodic update and not frequently by the administrator and helps OLAP technique by answering queries from direct users and applications.

- **On-Line Analytical Processing**

The data warehouse is generally an On-Line Analytical Processing based idea which analyse the data that enable users to access information quickly and accurately. OLAP is a tool

that helps to analyse, aggregate, store historical and current data, has a variety of front end tools such as query tools, report writers, analysis tools, and data mining tools used by analyst for planning and decision making. Data mining models and tools can be integrated inside OLAP which can help us to analyze the current situation and predict the future trend by reporting various forms of data report and graphical display of the analysis result. To enable On-Line Analytical Processing a multidimensional schema such as a star schema is adopted here. The developed OLAP based Cancer data warehouse contains a Patient general information table, habits table, risk factor table and cancer table. This OLAP has a integrated data mining technique with which the medical experts can analyse whether the patient has cancer or not,

what are the most common symptoms that a specific type of cancer has, the risk factors associated with type of cancer etc. This OLAP supports operations like Selection, Roll-up, Drill-Down and Slice, through which a person can view data from a patient or a doctor's perspective simultaneously. Thus by integrating both OLTP and OLAP techniques a more secure and powerful data warehouse id built here. The queries and results will be displayed on OLAP built in interface. They are also displayed in several formats like built in report generating tool and charts. The data warehouse becomes intelligent with time as the results from this OLAP technique are stored in the knowledge base for future analyses.
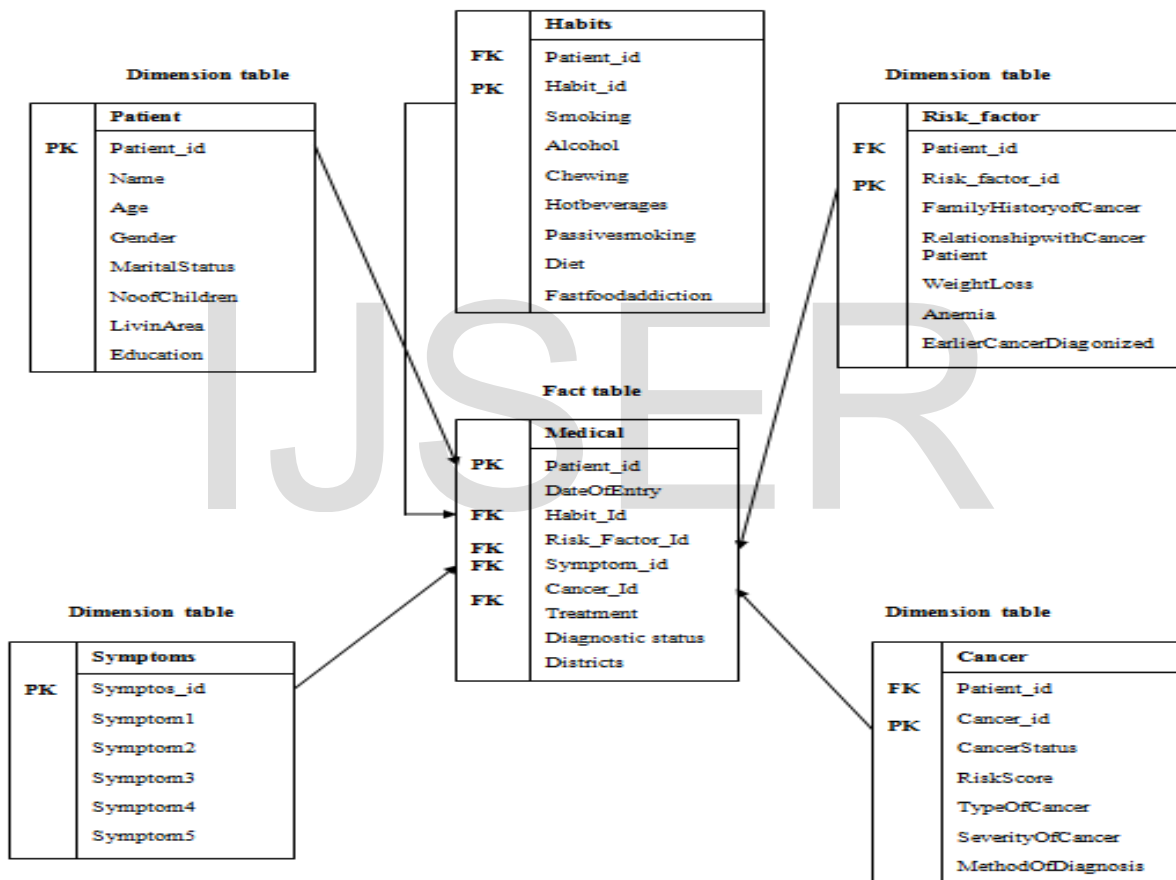

Figure 2: Star Schema representing Cancer Data Warehouse

3.2.4 Multidimensional Data Model

The multidimensional data model is based on the key concepts cube, dimension and hierarchy. This model allows its users to view and extract results from the cancer data warehouse in several different ways. The OLAP structure comprises of cubes, dimensions, measures, hierarchies and levels. A data cube is a multi dimensional data storage unit. Habits, risk factors and symptoms are few dimensions of cancer data warehouse. Measures are the facts that are to be analyzed. This Cancer data warehouse contains a Fact table named

Medical that has fields for Patient_id as primary key, Habit_Id, Risk_Factor_Id, Symptom_id, Cancer_Id as foreign keys representing other dimensions and Treatment, diagnostic status, Districts and DateOfEntry as measures. If a medical analyst generally analyse the quantity of drugs required for a month, the diagnostic status of patients in a month will be the measure and the cancer cube will be the dimension.

4. EVALUATION AND EXPERIMENTAL RESULT

Evaluation of the cancer data warehouse is the final stage where techniques involved are evaluated against some acceptance criteria. The final data warehouse contains all information relevant to the patient and cancer disease. The data warehouse provides the source for current and historical data that helps medical analysts and doctors to improve the decision making process and predict the future trend of cancer. The following are some criteria for evaluation of the cancer data warehouse.

• Reliability: The data in this data warehouse is collected and stored from cancer hospitals in Tamilnadu and the three electronic databases which provide data to the data warehouse server are located in cancer hospitals. Also the developed ETL model ensures that the information that is not relevant to cancer is eliminated as noise.

• Integrity: The developed transformation code ensures the integrity of the data warehouse by converting the columns in the databases from one data type to the type that is suitable for the data warehouse.

• Capability: The data warehouse is built using SQL Server which has a maximum limit for storing data and can accommodate both historical and current data. The data in the warehouse is updated in planned periodic cycles and tagged with date which makes it highly non volatile.

• Manageability: The data in the warehouse is stored in a Star schema in multi dimensional data cubes which makes the data warehouse highly manageable.

• Data querying: The data warehouse is build to answer pre defined complex queries with reports and charts which helps the medical analysts to make useful decisions

The Cancer Data warehouse is useful to record and save large volumes of sensitive information which can be used to gain knowledge about the disease and its treatment. The results from the cancer data warehouse will provide clear information to a medical analyst to make decisions on crucial subjects such as the type of service to be provided, the geographical area in which the awareness program should be concentrated, the resources to be purchased etc. It can also be used to predict the future trend of cancer as the data warehouse stores historical and current records. The results are from the Cancer Data Warehouse are presented below.
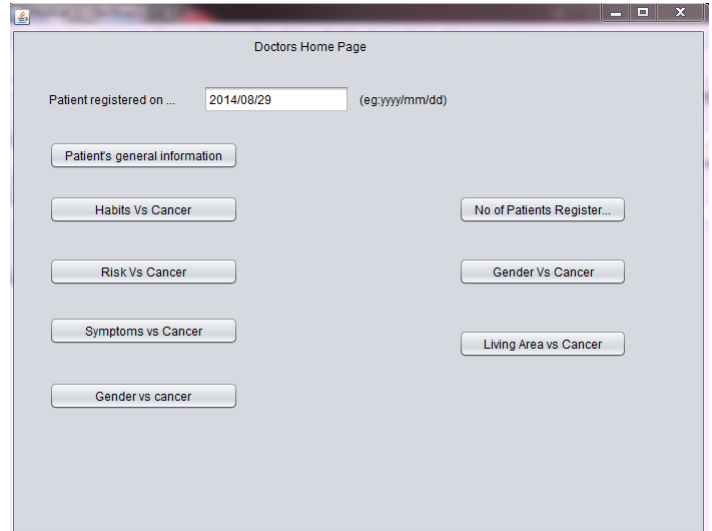

Figure 3. User Interface showing a Medical Experts home page containing pre defined queries for reports and charts generation
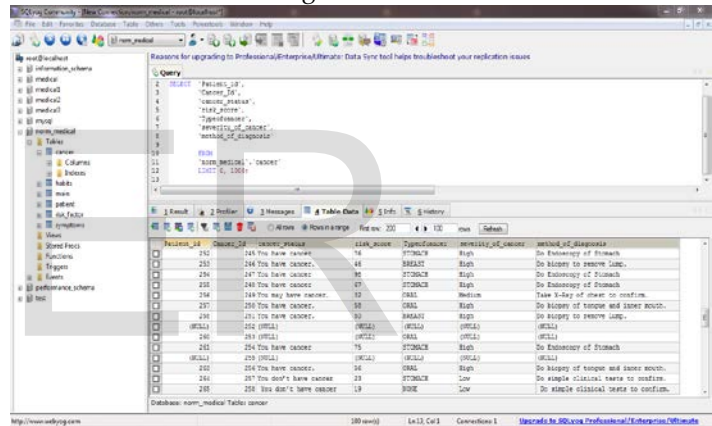

Figure 4. SQL query space to extract reports from Data Warehouse

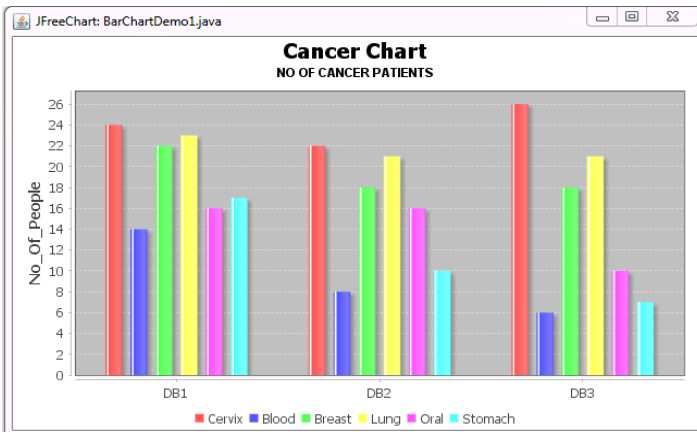Figure 5. Several Reports extracted from the Cancer Data Warehouse

Figure 6. Bar Chart representing the number of cancer patients in different districts

5. CONCLUSIONS

Developing a Data Warehouse specifically for cancer disease is the immediate need in India as a large amount of valuable knowledge can be gained from the vast data stored in rural and urban hospitals in physical form. The Cancer Data Warehouse is developed in a novel method by integrating On-Line Transaction Processing and a Data Mining model with On-Line Analytical Processing. This developed Cancer Data warehouse helps several medical professionals like medical analysts, diagnostic centers, drug administrators, physicians and patients alike who are separated by geographical locations.

In future this cancer data warehouse would be modified to include several dimensions such that it could automatically predict the treatment for an individual patient with a specific type of cancer, its feasibility, survival time, the death rate due to cancer etc.

ACKNOWLEDGEMENT

REFERENCES

[1] Dr. Mohamed F. AlAjmi, Shakir Khan "Studying Data Mining and Data Warehousing with Different E-Learning System" *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 4, No.1, 2013

[2] Souâd Demigha "A Data Warehouse System to Help Assist Breast Cancer Screening in Diagnosis, Education and Research" *World Academy of Science, Engineering and Technology,* Vol: 4 2010-08-29

[3] Teh Ying Wah "Development of a Data Warehouse for Lymphoma Cancer Diagnosis and Treatment Decision Support" *Wseas Transactions on Information Science And Applications,* Issue 3, Volume 6, March 2009, ISSN: 1790-0832

[4] Eric Zapletal "Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case" *MEDINFO 2010 IMIA and SAHIA,* doi: 10.3233/978-1-60750-588-4-193

[5] Susan Maskery "Aggregated Biomedical-Information Browser (ABB): A Graphical User Interface for Clinicians and Scientists to Access a Clinical Data Warehouse" *Journal of Computer Science System Biology (JCSB),* Volume 7(1)020-027 (2014) – 020 ISSN: 0974-7230

[6] Shakir Khan "Studying Data Mining and Data Warehousing with Different E-Learning System" *International Journal of Advanced Computer Science and Applications (IJCSA),* Vol. 4, No. 1, 2013

[7] Rajni Jindal "Comparative Study of Data Warehouse Design Approaches: A Survey" *International Journal of Database Management Systems (IJDMS),* Vol.4, No.1, February 2012

[8] Manya Sethi "Data Warehousing and OLAP Technology" *International Journal of Engineering Research and Applications (IJERA),* ISSN: 2248-9622 www.ijera.com, Vol.2, Issue 2, Mar-Apr2013, pp.955-960

[9] Shaker H. Ali El-Sappagh " A Proposed model for data warehouse ETL Processes" *Journal of King Saud University – Computer and Sciences (2011),* 23, 91-104

[10] Joaquin Perez-Ortega "Spatial Data Mining of a Population –Based Data Warehouse of Cancer in Mexico" *International Journal of Combinational Optimization Problems and Informatics,* Vol. 1, No. 1, May-Aug 2010, pp. 61-67, ISSN: 2007-1558

[11] Monica Chiarini Tremblay "Design of an information volatility measure for health care decision making" *Decision Support Systems,* 52(2012), 331-341

[12] Muhammad Saqib, Muhammad Arshad "Improve Data Warehouse Performance by Preprocessing and Avoidance of Complex Resource Intensive Calculations" *International Journal of Computer Science Issues (IJCSI),* Vol. 9, Issue 1, No 2, January 2012 ISSN (Online): 1694-0814

[13] Neera Bhansali "Challenges of Data Quality in Medical Informatics Data warehouses" *MIT Information Quality Industry Symposium*, July 15-17, 2009

[14] Abubakar Ado1, Ahmed Aliyu "Building a Diabetes Data Warehouse to Support Decision making in healthcare industry" *IOSR Journal of Computer Engineering (IOSR-JCE)* e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. IX (Mar-Apr. 2014), PP 138-143 www.iosrjournals.org

[15] Yang Kehua, Abdoullahi Diasse "A dynamic materialized view Selection in a Cloud-based Data Warehouse" *IJCSI International Journal of Computer Science Issues,* Vol. 11, Issue 2, No 1, March 2014, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784

[16] Joyce C. Niland and Layla Rouse "Clinical Research Systems and Integration with Medical Systems" *Biomedical Informatics for Cancer Research,* DOI 10.1007/978-1-4419-5714-6_2, © Springer Science +Business Media, LLC 2010

[17] M.Thenmozhi and K.Vivekanandan "A Tool for Data Warehouse Multidimensional Schema Design using Ontology" *IJCSI International Journal of Computer Science Issues,* Vol. 10, Issue 2, No 3, March 2013 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784

[18] Polly Hitchcock Noel "VHA Corporate Data Warehouse height and weight data: Opportunities and challenges for health services

research" *JRRD Journal of Rehabilitation Research & Development*, Volume 47, Number 8, 2010 Pages 739–750

[19] Larissa Nekhlyudov and S. M. Greene "Cancer research network: using integrated healthcare delivery systems as platforms for cancer survivorship research" *Journal of Cancer Survivorship Research and Practice,* Springer Science+Business Media New York, Volume 6, Number 4, December 2012 ISSN 1932-2259 DOI 10.1007/s11764-012-0244-8

[20] Antonio Augusto Gonçalves "The Implementation of Clinical Cancer Data Warehouse at the Brazilian National Cancer Institute" Viii Congresso Nacional De Excelência Em Gestão 8 E 9 De Junho De 2012

[21]Vineetha Appidi, Dr Syed Umar, Sushma Vallamkonda "Development of a Data Warehouse for Cancer Diagnosis andTreatment Decision Support" *International Journal Engg Techsci Vol 5(3) 2014, 22 – 26*

[22] In Young Choi, Seungho Park "Development of prostate cancer research database with the clinical data warehouse technology for direct linkage with electronic medical record system" Prostate *International* 2013; 1(2):59-64 http://dx.doi.org/10.12954/PI.12015 http://p-international.org/ pISSN: 2287-8882 eISSN: 2287-903X

[23] Alaa Khalaf Hamoud, Dr Talib A.S. Obaid "Building Data Warehouse for Diseases Registry: First step for Clinical Data Warehouse" International Journal of Scientific & Engineering Research, Volume 4, Issue 11, November-2013 ISSN 2229-5518

[24] Dr. Osama E.Sheta "Evaluating a Healthcare Data Warehouse for Cancer Diseases" IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 3, No.3, June 2013

[25]Paulraj Ponniah "Data Warehousing Fundamentals A Comprehensive Guide for IT Professionals",Wiley India Pvt.Ltd, ISBN: 978-81-265-0919-5

IJSER