# Design and Implementing Cancer Data Warehouse to Support Clinical Decisions

Alaa Khalaf Hamoud, Dr. Hasan Adday, Dr.Taleb A. S. Obaid,
Mrs. Rabab Abdllmajed Hameed

**Abstract:** In the last decade, data warehousing became the most important technology in the field of decision support especially related to critical situations. Most of Decision Support Systems (DSS) related to clinical decisions depend on Data Warehouse (DW) and On-line Analytical Processing (OLAP) in analyzing the related information in order to assist clinicians in making right decisions. In this work, cancer data warehouse (CDW) is designed and implemented based on medical paper records of cancer infections. At first, thousands of medical cancer records were turned from paper based into electronic based by registering them in excel files and access database. These files became the source of the proposed cancer DW. Secondly, the proposed cancer DW is depended as a platform to perform different (OLAP) operations to get analytical results that support decisions maker related to cancer infections.

**Keywords:** Cancer Data Warehouse, Extract Transform and Load (ETL), On-Line Analytical Processing (OLAP), Star Schema, Decision Support System (DSS).

## 1. Introduction

Data warehouse is an informational environment that provides an integrated and total view of the enterprise to make the enterprise's current and historical information easily available in decision making. DW makes decision support transactions possible without hindering operational systems, renders the organization's information consistent. It presents a flexible and interactive source of strategic information [1].

DW can be considered as the base of any successful decision support system. Users involving in DSS are differing from the normal users of simple database in tools that they using and decisions which they making. Users (analysts involving in DSS) are making strategically decisions based on facts provided by data warehousing. In our work, the analysts dealing with CDW need answers to their questions related to these infections.

It is very important to the analysts to understand the behavior of each infection. Many excel files which contain cancer infections are collected from different medical centers in order to consider them as a source for the proposed CDW. In the preprocessing stage, these excel files are combined in one single file so it can be processed easily. Some operations of transforming like cleaning, aggregation and grouping are made as a preprocessing operation on the staging table. These operations make loading dimensions and fact table easy.

The second stage is constructing star schema and building ETL package to load dimensions and fact table of CDW. After these stages, cancer cube is built based of some chosen dimensions in order to perform

OLAP and build predefined reports to view the general behavior of each infection and help the analysts in support their decisions. Many tools are used such as SQL Server Management service tools (SQL Server Management Service (SSMS), SQL Server Integration Service (SSIS), SQL Server Analytical Service (SSAS) and SQL Server Reporting Service (SSRS)) beside Microsoft excel pivot table to view the cube.

## 2. Literature Review

P. Ramachandran et al [2] reviewed 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.

Souad Demigha [3] proposed method to design and develop a data warehouse system in radiology-senology (DWRS) to assist breast cancer screening in diagnosis, education and research.

Ortega J. P. et al [4] 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.

Abubakar Adol et al [5] proposed an architecture for health care data warehouse for diabetes diseases which could be used to monitor diabetes disease. They measure cost of infections and to detect prescription errors in addition it can also be used by healthcare executive managers, doctors, physicians and other health professionals to support the capture, healthcare process and analysis of data, and offer the potential of radically altering the practice and delivery of healthcare and medical research.

Hamoud A. K. and Dr. Obaid T. A. S. [6] implemented healthcare data warehouse depending on Electronic Health Records (EHR) by processing them to Electronic Medical Records (EMR) and made them as source of the proposed DW. OLAP operations are applied to view more analytical results to support clinical decisions.

Dr. Sheta O. E. et al [7] presented 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.

Wah T. Y. and Sim O. S.[8] reviewed the development and use of a clinical data warehouse specific to the Lymphoma or Lymph Node cancer, which could be used by doctors, physicians and other health professionals, in conjunction with a clinical DSS, to support the clinical process as well as to formulate the appropriate model to improve the quality of diagnosis and treatment recommendation decision making. This paper proposes a 5-stage sequential methodology for the clinical data warehouse development.

Choi Y. et al [9] developed prostate cancer research database system which incorporates information about a prostate cancer research including demographics, medical history, operation information, laboratory, and quality of life surveys. Their system includes three different ways of clinical data collection to produce a comprehensive data base; direct data extraction from electronic medical record (EMR) system, manual data entry after linking EMR documents like magnetic resonance imaging findings and paper-based data collection for survey from patients.

## 3. Data Warehouse

According to W.H.Inmon, data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management decision making process. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining, as shown in figure(1) [10].

A data warehouse is not a single software or hardware product you purchase to provide strategic information. It is, rather, a computing environment where users can find strategic information, an environment where users are put directly in touch with the data they need to make better decisions. Strategic information cannot be gained from day to day database transactions but it depended on historical data. Data warehouse can be considered as a huge repository which contains the full historical information related to target company or enterprise. DW turns the ancient data to valuable information in support decisions [1][11].

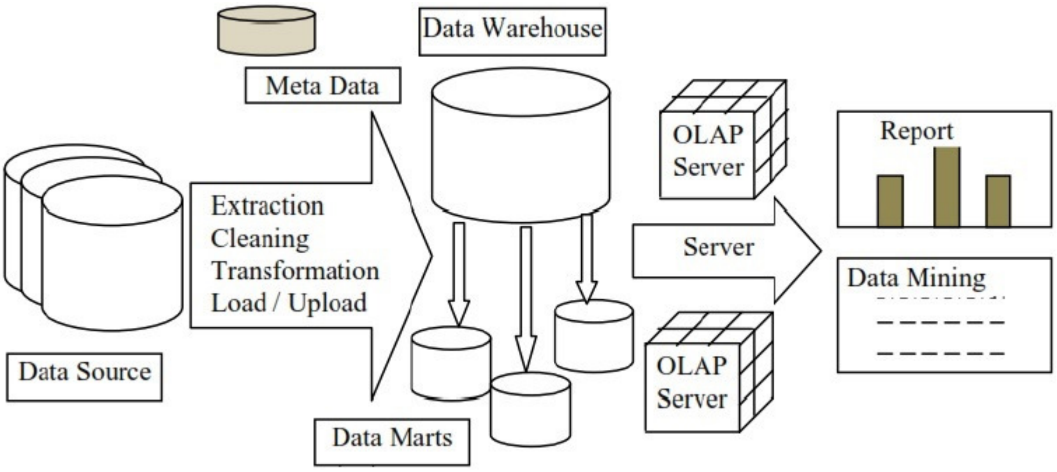


**Figure (1): Data Warehouse**

Data warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis. However, a data warehouse brings high performance to the integrated heterogeneous database system since data are copied, preprocessed, integrated, annotated, summarized, and restructured into one semantic data store. Furthermore, query processing in data warehouses does not interfere with the processing at local sources [12].

## 4. Star Schema

The architecture of data warehouse can be simply viewed by its schema. Start and snow-flack schema are the base to build data warehouses. Star schema is a modeling paradigm in which the data warehouse contains a large central table called fact table, and a set of smaller attendant tables called dimension tables, one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table [12].

Star schema is simple and effective schema to build data warehouse since it is easy to design and understand by the analysts. The logical data model is usually structured as a star schema, with each dimension collapsed into a single table. We generally prefer this flattened structure to a more normalized set of tables because:

- Users who query the relational database directly are better able to navigate the model.
- Relational queries usually perform better against this structure.
- The flattened table is often easier to manage in the ETL process with fewer tables and associated keys [13], as in figure (2).

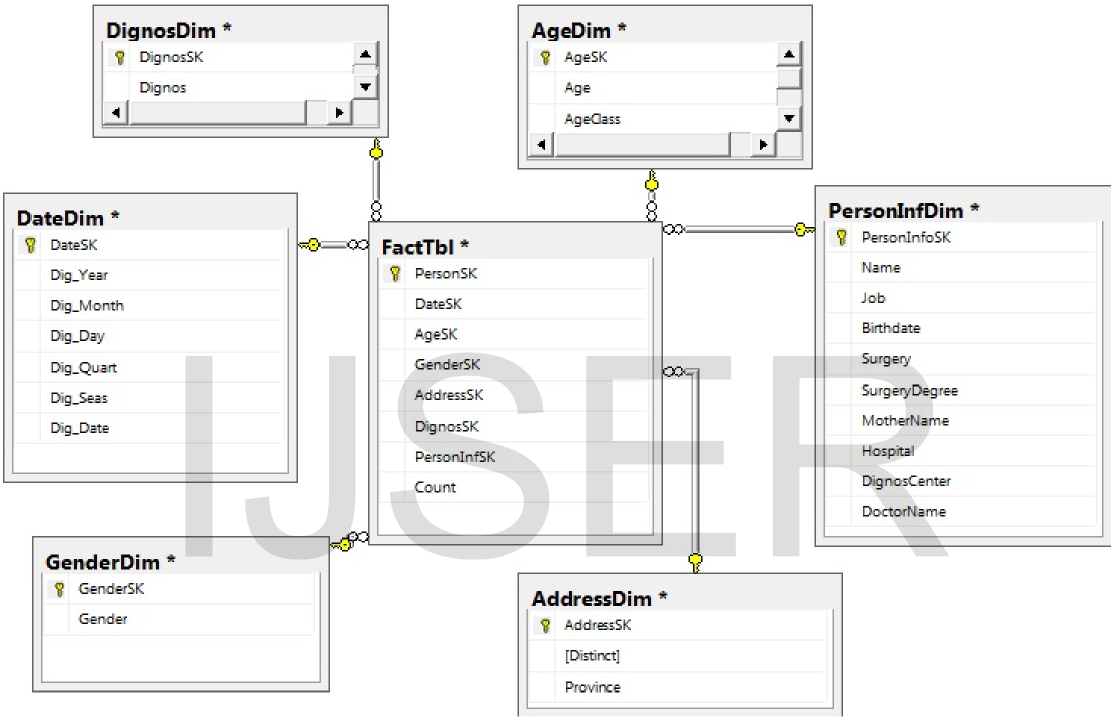


**Figure (2): CDW Star Schema**

## 5. Extract, Transform and Load (ETL)

The Extract, Transform and Load (ETL) system is the foundation of the data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conform data so that separate sources can be used together. Finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions. The ETL system makes or breaks the data warehouse. The ETL system is easily consumes 70 percent of the resources needed for implementation and maintenance of a typical data warehouse [14].

Staging table is the intermediate place between data warehouse and data source. Most ETL systems need a set of staging tables to support the

ETL process. Different ETL systems may use more or fewer staging tables. Create a separate database or schema to hold the staging tables rather than mixing them in with the query able data warehouse database. Segregating staging tables keeps the data model tidier, and more importantly provides flexibility for moving the staging data to a different server to support a very high ETL load [13].

## 6. On-Line Analytical Processing (OLAP)

OLAP Council defines On-Line Analytical Processing (OLAP) by "OLAP is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user" [1].

OLAP offers analytical modeling capabilities, including a calculation engine for deriving ratios, variance, etc., and for computing measures across multiple dimensions. It can generate summarizations, aggregations, and hierarchies at each granularity level and at every dimension intersection. OLAP also supports functional models for forecasting, trend analysis, and statistical analysis. In this context, an OLAP engine is a powerful data analysis tool [12].

Server-based OLAP products are an increasingly popular component of the data warehouse infrastructure. OLAP servers deliver two primary functions:

- *Query performance:* using aggregates and specialized indexing and storage structures. The OLAP servers automatically manage aggregates and indexes, a benefit whose value may become clear by reviewing the previous section that discusses how to manage aggregate tables.
- *Analytic richness*: using languages that, unlike SQL, were designed for complex analytics. OLAP servers also have mechanisms for storing complex calculations and security settings on the server, and some are integrated with data-mining technologies [14].

## 7. Proposed Model

The CDW proposal, designed and implemented based on many steps, as illustrated in figure (3). As mentioned before, the excel files are aggregated from different medical centers and consolidated in single excel file so it can be used as a source for staging table. Staging table holds all medical records of cancer infections in order to perform the related preprocessing ETL operations like filling missing values, converting data types and deriving new columns. Star schema is chosen as CDW schema because it is easy to design and implement. Then, it is flexible to future changing such as adding new dimensions and facts. The proposed model designed and implemented by using SQL Server Management Service (SSMS), SQL Server Analysis Service (SSAS), SQL Server Integration Service (SSIS) 2012 and Microsoft Excel (Pivot Table) 2010.

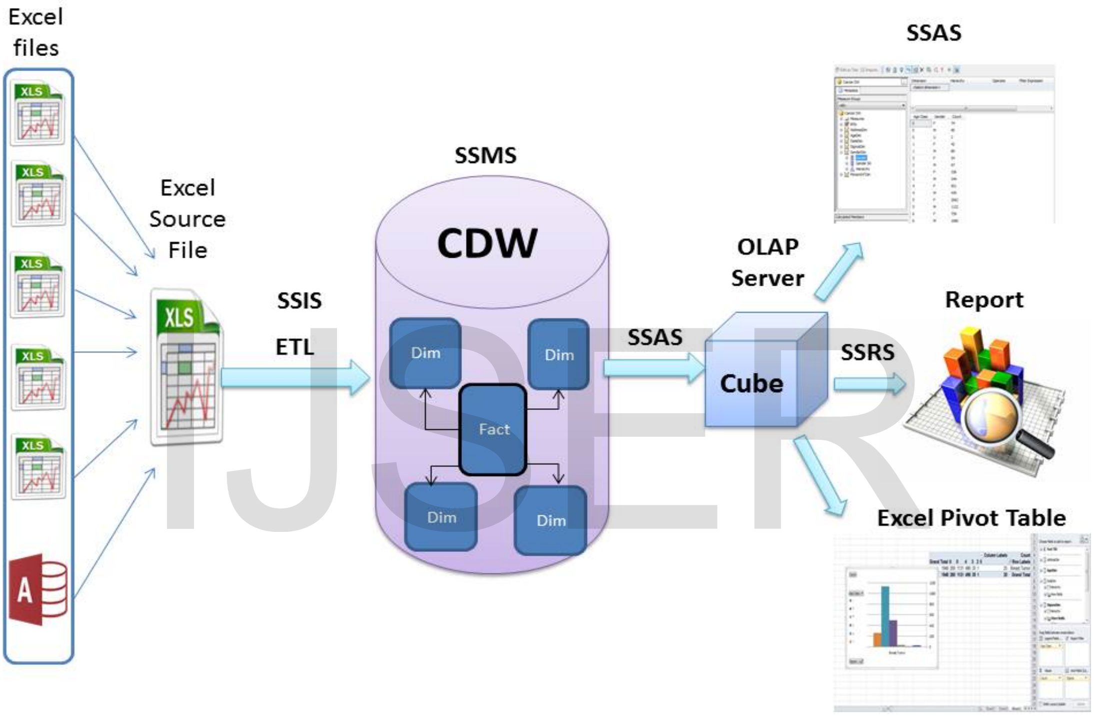


**Figure (3): CDW Model**

### 7.1 Preprocessing Operations

1. *Collecting Medical Records*

Thousands of paper based medical records are collected from different medical centers and hospitals in Basrah province which deal with cancer infections treatment. Most of these medical records contained important attributes while some of them contained patient's name and the infection name only. The operations of registering these records in excel files and access database takes too long time and resources. Since paper medical records were missing some values and not clear, so the work of data entries was very difficult.

2. *Consolidating Data Sources*

The output data sources (excel files and access database) are unified in single excel file to hold all medical records in single place. The processes of emerging excel files takes too long time since the attributes columns in excel files are not the same. The output excel file also emerged with access database which held more than one thousands of medical record. Many of derived attributes are added based on the original attributes such as (Age Class from Age), (Day, Month, Year, Quarter and Season from date of diagnosis). Some transformation processes are applied such as removing spaces inside attribute's values and converting gender into single character (Male into M, Female into F and other values into U).

## 7.2 Data Warehousing

### 1. Design Star Schema

Based on the attributes in data source excel file, star schema dimensions and fact tables designed using SSMS. The schema consists of six dimensions (AgeDim, GenderDim, DisnosDim, PersonInfDim, DateDim and AddressDim) with Fact Table (FactTbl) with one measurement (Count) to calculate number of infections along dimensions, see figure (2). Each dimension contains surrogate key to connect to fact table.

### 2. ETL System

ETL system consists of all processes of Extracting data from data source from staging table, transforming data types and filling missing values. It also includes operations of deriving new attributes by grouping or classification. Tools like Slowly Changing Dimension (SCD), lookup, derived column, aggregate and data conversion are used to process and load dimensions and fact table. Dimensions are load first and fact table loaded based on the data in dimensions by using lookup tool. Lookup tool takes the key (surrogate key) from each dimension and loads it in fact table to make it as reference to dimension table.

### 3. Constructing Cube

A cube is constructed based on dimensions of CDW in order to perform OLAP operations. It designed and implemented by using SQL Server Analytical Service. The dimensions and hierarchies are implemented to allow applying OLAP operations (slice, dice, drill through, drill up and drill down). Two effective hierarchies are implemented which are (date hierarchy and address hierarchy). All dimensions are chosen to build the cube. After cube implementing completed, it can be viewed directly by dropping dimensions members and measurement. Special cubes can be constructed to analyze data based on specific dimensions and measurement.

## 7.3 Results

### 7.3.1 Reports

SSRS provides ability to design and deploy reports based on predefined query or by using parameters. The analysts can view the reports on server by using any web browser, see figure (4). This report shows the number of infections of all cancer classified by age classes and gender. It shows that gender (Female with character F in the figure) has maximum number of infections in the second quarters of all years.

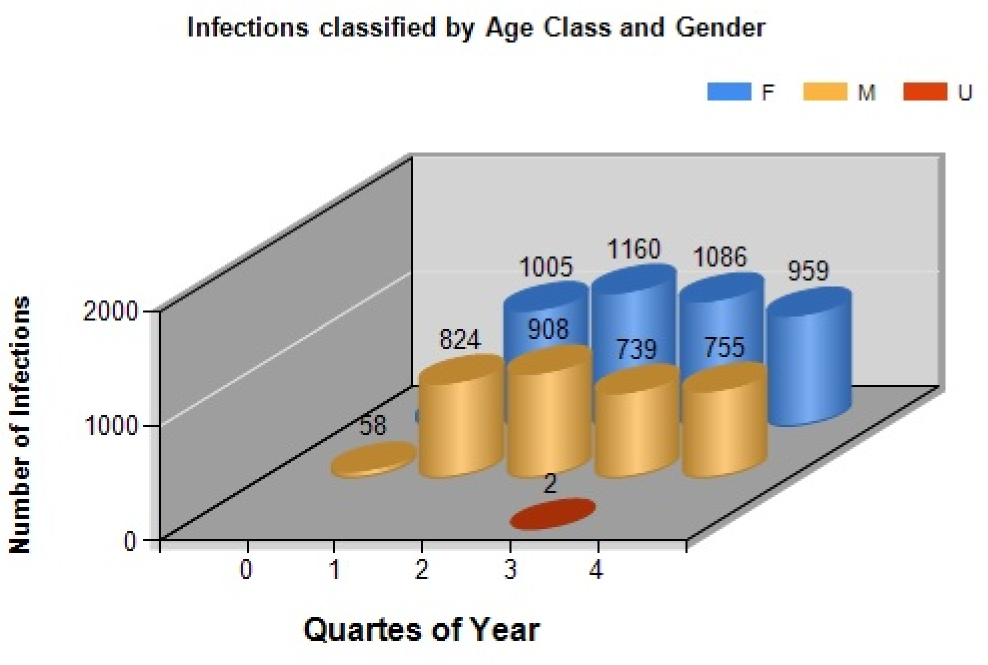
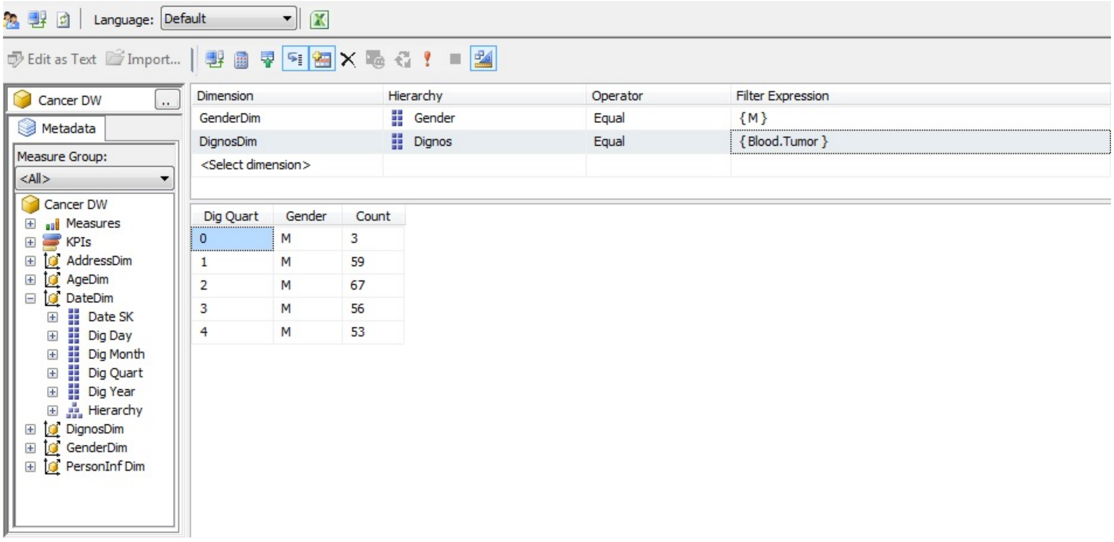


**Figure (4): All Infections classified by Quarter of Year and Gender**

### 7.3.2 SSAS Cube View

After implementing SSAS package, the next step is view the cube using SSMS. SSMS allows the analysts to view the resulting cube by dropping each dimension member easily. It also provides filtering conditions to exclude some values or add more values to get accurate results. Figure (5) shows number of specific infection (blood tumor) classified by quarter of year for specific gender (Male). The analysts can also grouping many infections for many locations and genders.



| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| GenderDim | Gender | Equal | { M } |
| DignosDim | Dignos | Equal | { Blood.Tumor } |
| <Select dimension> | | | |

| Dig Quart | Gender | Count |
|---|---|---|
| 0 | M | 3 |
| 1 | M | 59 |
| 2 | M | 67 |
| 3 | M | 56 |
| 4 | M | 53 |

**Figure (5): SSAS Cube View**

### 7.3.3 Excel Pivot Table Reports

Excel Pivot Table provides tool to connect to SQL Server Analysis Server and view cube in flexible way. As soon as the connection established to Cube, the analyst can chose the dimension members and measurements easily. Figure (6) shows the Breast Tumor infection classified by Gender and Age Class. Each color represents age class and shows the accurate number of infection. It is obvious that age class (5) (41 to 60 years) are infected with high rate number among the other age class.

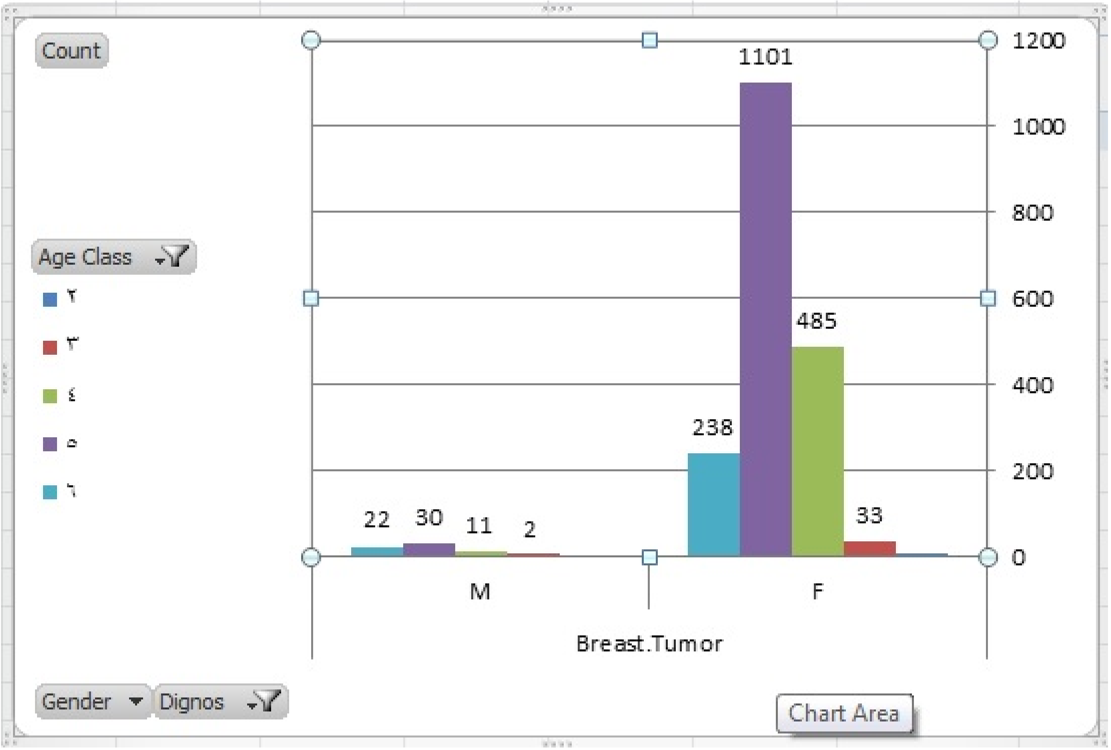


**Figure (6): Breast Tumor with Age Class and Gender**

In the second report, see Figure (7), date dimension member (quarter) is added to show more accurate result. The quarters (1 and 4) are selected to show the number of infections in these quarters. The report shows that both genders in age class (5) had the maximum number of infections in these two quarters of all years. The analyst can select either all or specific members of dimensions to show more results which can be depend in support decisions.

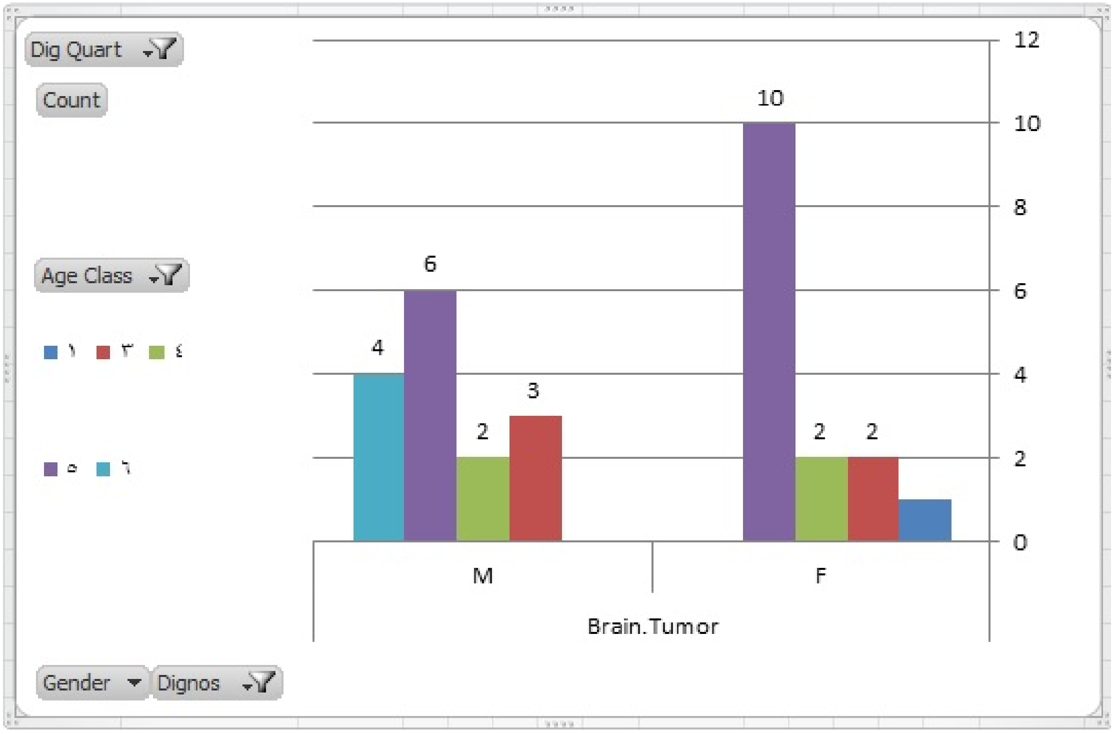


**Figure (7): Brain Tumor with Age Class and Gender Classified By Quarter of Year**

## 8. Conclusion and Discussion

Data warehousing is essential tool in support critical decisions rather than understanding the behavior of cancer infection. By applying OLAP operations, the analysts can get valuable information related to each cancer case. The goal of this work is to help clinicians involved in studying critical medical cases to understand deeply the relationship among the factors of medical records (members on dimensions). It also can help the professionals whom study the behavior of each infection related with geographical location since it is available to view number of infections in each quarter of the year in specific distinct.

## References

[1] Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals, Paulraj Ponniah Copyright © 2001 John Wiley & Sons, Inc. ISBNs: 0-471-41254-6 (Hardback); 0-471-22162-7 (Electronic)

**[2]** P.Ramachandran, Dr.N.Girija, Dr.T.Bhuvaneswari " Developing a Decision Support System Using Cancer Data Warehouse ", International Journal of Scientific & Engineering Research, Volume 5, Issue 9, September-2014 238 ISSN 2229-5518.

**[3]** 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

[4] 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.

**[5]** 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.

[6] A. K. 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.

**[7]** Dr. O. 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

[8] T. Y. Wah, O. S. Sim " 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.

[9] Choi, In Young et al. "Development of prostate cancer research database with the clinical data warehouse technology for direct linkage with electronic medical record system." Prostate international 1.2 (2013): 59-64.

[10] Reddy, G. Satyanarayana, et al. "Data Warehousing, Data Mining, OLAP and OLTP Technologies are essential elements to support decision-making process in industries." International Journal on Computer Science and Engineering 2.9 (2010): 2865-2873.

[11] Inmon WH "Building the Data Warehouse", John Wiley & Sons Inc., ISBN 0-471-08130-2, USA, 2002.

[12] Han, Jiawei, Micheline Kamber "Data mining: concepts and techniques (the Morgan Kaufmann Series in data management systems)." (2000).

[13] Kimball, Ralph "The data warehouse lifecycle toolkit", John Wiley & Sons, 2008.

[14] Kimball, Ralph, and J. Caserta. "The data warehouse ETL toolkit: practical techniques for extracting." Cleaning, Conforming, and Delivering Data, 2004.