ISSN 1870-4069

Design and Implementation of a Data Warehouse to Support Decision-Making in a Health Environment

Simon G. Cornejo[1], Karina Caro[2], Luis-Felipe Rodriguez[1], Roberto Aguilar A.[1], Cynthia B. Perez[1], Luis A. Castro[1]

[1] Sonora Institute of Technology (ITSON), Ciudad Obregon, Sonora, Mexico

[2] Drexel University, Philadelphia, Pennsylvania, USA

{scornejo175319 raguilar175318}@alumno.itson.edu.mx, karinacaro@drexel.edu, {luis.rodriguez, cynthia.perez}@itson.edu.mx, @itson.edu.mx, luis.castro@acm.org

Abstract. One of the most common challenges in the management of electronic medical records (EMR) is to extract critical knowledge that could serve to enhance collaboration among medical doctors to support decision-making. In this paper, we present the design and implementation of a data warehouse designed to strengthen decision-making related to epidemiological patterns, trends, areas of influence and other pathologies. The data warehouse is based on a relational model that contains information about EMR of different states of Mexico. We used the Business Event Analysis and Modeling (BEAM) methodology for the design and implementation of the data warehouse, a novel methodology to design agile data warehouses. Using BEAM, we created a star dimensional model and defined the Extract, Transform and Load (ETL) processes to transfer the data to the new model. In order to show the potential of our data warehouse, an interactive dashboard with different indicators was built. We close discussing how the medical doctors could use our data warehouse to support the decision-making process.

Keywords: decision support systems, data warehouse, health environment, BEAM, ETL.

1 Introduction

In recent years, the use of electronic medical records (EMR) has increased considerably, changing the way traditional records are stored and managed [1]. In EMR, the data is managed and stored digitally, enabling health professionals to maintain information in one place. The use of these systems has enabled the accumulation of large amounts of data, opening up opportunities for analyzing and obtaining relevant information from them [2].

However, in most cases, besides data queries, these data are not analyzed, leaving aside all the knowledge that can be obtained from them. These large amounts of data and the knowledge that can be obtained from them open the possibility of supporting decision-making in the health environment, as well as promoting communication and collaboration among health professionals. Having a tool to support decision-making that could provide knowledge based on patients' historical data stored in EMR, might help to a great extent to improve health services.

An alternative to support this problem is the use of a data warehouse, a repository that preserves the historical context to accurately assess an organization's performance over time [3]. It is optimized for high-throughput queries, since user queries often require that hundreds or thousands of transactions are searched for and compressed into a set of responses. Data warehouses are the basis of the processing of Decision Support Systems (DSS). They facilitate the analysis since all data are concentrated in a single data source, which can integrate both structured and unstructured data, with different granularity. A data warehouse based on the data of an EMR could support health professionals to obtain relevant knowledge to enhance the decision-making process. However, the design and construction of a data warehouse is not an easy task. First, the preparation and cleaning of large amounts of data present some challenges such as the concentration of heterogeneous data, incomplete records, and integrity errors, among others. On the other hand, the design of the data warehouse has to be done in such way that the response time of the queries is fast and the results are correct.

An optimal design of the data warehouse is required to facilitate the extraction and analysis of the stored data. The use of agile methodologies (e.g., SCRUM [4]) have presented multiple advantages in software development such as customer satisfaction by the rapid and continuous delivery of useful software. Agile methodologies are characterized by emphasis on stakeholders and interactions rather than processes and tools [4]. On the side of the design of data warehouses, the Business Event Analysis and Modeling (BEAM) methodology [5] offers several advantages for designing data warehouses. For example, individuals and interactions over processes and tools, working software over comprehensive documentation, and customer collaboration over contract negotiation. BEAM upholds these values and the agile principle of data warehouse practitioners to work directly with stakeholders to produce data models rather than requirements documents, and working Business Intelligence (BI) prototypes of reports / dashboards rather than mockups.

This paper presents the design and implementation of a data warehouse using BEAM methodology, which information was obtained from EMR. This data warehouse provides information in the form of indicators that summarize what is happening in a health environment and support appropriate and timely decision making through the identification of diseases by geographic zones, diseases by stage of life and diseases by season of the year.

The illustration of the usefulness of the designed data warehouse, following, we present a usage scenario where health professionals use the information they visualize in an interactive dashboard to support decision making:

A health professional is concerned that many patients are presenting with a rare disease in their geographical area and he is not sure that the medications he is prescribing to his patients are the most appropriate. The health professional reviews the information provided by an interactive dashboard and he realizes that the disease which

he is dealing with is very common in another geographic area of Mexico. Thus, the health professional starts to get in touch with other health professionals in that geographical area to ask for opinions and share experiences in the treatment and intervention of that disease.

2 Related Work

Research contributions have been made in the areas of data warehouse design, data staging for ETL processing, data quality assurance, and healthcare data warehouse applications, mainly in developed countries, such as the United States. Existing EMR [6] data are made available in a standardized and interoperable format, thus opening up a world of possibilities for semantic or concept-based reuse, consultation and communication of clinical data. The Community Health Applied Research Network (CHARN) [7], in the United States, represents more than 500,000 patients from diverse safety nets in 11 states, aims to create a national and centralized data warehouse with multiple partners from the Center for Community Health using different EMR systems.

The work [8] describes a virtual data warehouse (VDW) of the Health Maintenance Organization Research Network (HMORN), a public, research-centric data model implemented in 17 health care systems across the United States. At the Catholic Health Initiatives research institute, data consultation tools [11] are implemented to enable end users to access the VDW for simple consultation and research readiness activities, capture for collection of study-specific data and results reported by the patient. On the other hand, the decision support system [9] based on multi-criteria data analysis – Annalisa, is an online decision support tool for individuals and clinicians interested in making a shared decision.

In Mexico, there have been initiatives and programs of innovation and technological development of the public and private sectors that have begun to get involved in the subject of e-health. In the early 1990s, the State Basic Information System (SEIB) was centralized and encompassed all 32 states by The Ministry of Health (SSA, for its acronym in Spanish). In 1995, the National Epidemiological Surveillance System (SINAVE) was created. Its coordination is carried out by SSA and it is supported by the Single Information System for Epidemiological Surveillance (SUIVE).

In 2007, SSA initiated the development of the Mexican electronic clinical record standard. In 2011, a study was conducted in Mexican Institute of Social Security (IMSS, for its acronym in Spanish) [10] to develop Quality of Care Indicators (QCI) for Type 2 Diabetes Mellitus (T2DM). The goal was to determine the feasibility of constructing QCI using IMSS's EMR data and assessing the Quality of Care (QC) provided to IMSS patients with T2DM. As a result of this study [10], 18 QCIs were developed, of which 14 were possible to construct using available EMR data. ICQs comprised both the care process and health outcomes.

The related work shows that there is a potential of using EMR to enhance health care services. However, there are few studies that use the EMR data to obtain knowledge that could support the decision-making process in the healthcare environment. In this work, we propose to use EMR data to design and implement a data warehouse aimed at supporting the decision-making in the healthcare environment. In the next section,

we describe the BEAM methodology, following by the results of the design of the data warehouse and its implementation.

3 BEAM Methodology

We used the Business Event Analysis and Modeling (BEAM) methodology to design the proposed data warehouse. To the best of our knowledge, BEAM is one of the most recent and the first agile methodology in the area of Data Warehouse and Business Intelligence (DW / BI) [5]. The BEAM methodology comprises a set of collaborative techniques for modeling BI data requirements and translating them into dimensional models on an agile time scale. Among the techniques used by the BEAM methodology are the 7W's Framework, BEAM*tables, Event Matrix and Enhanced Star Schema.

3.1 7W's Framework and BEAM*tables

The 7W's framework uses questions about who, what, where, when, how many, why, and how[5], data modelers design the model by asking BI stakeholders to tell data stories using these questions. BEAM uses tabular notation and data stories to define business events in a format that is easily recognizable and understandable to BI stakeholders. It uses spreadsheets that enable an easy translation into detailed star schemas.

BEAM*tables help engage BI stakeholders to define reports that answer their specific business questions. They are used to define fact and dimension tables, and they use natural language enable BI stakeholders easily imagine, sort, and filter the low-level detail columns of a business event using the top-level dimensional attributes. BEAM*tables can describe facts, events in terms of measures, and dimensions, descriptions of the facts, which can be used to filter, group and aggregate measurements.

3.2 Event Matrix

The Event matrix documents the relationships between all events and dimensions within a model. Event matrices record events in value chain sequences and promote the definition and reuse of conformed dimensions through dimensional models.

3.3 Enhanced Star Schema

A star schema consists of a central fact table surrounded by a series of dimension tables. The fact table contains facts: the numerical (quantitative) measures of a business event. Dimension tables contain mainly textual (qualitative) descriptions of the event and provide the context for the measurements. Enhanced star schemas are standard star schemas that use BEAM short codes to record dimensional properties and design techniques that are not directly supported by generic data modeling tools.

In the following section, we describe the obtained results of applying the above BEAM techniques to design a data warehouse, as well as, the data warehouse' implementation is described.

DiseasesxVectorxMunicipality [RE]

Patient	Disease	Municipality	State	Date	Amount
[who]	[what]	[where]	[where]	[when]	[how many]
Jaime Rodriguez	Peste bubónica	Cajeme	Sonora	29/01/2011	1
María Luisa Aceves	Tifus	Salina Cruz	Oaxaca	22/06/2013	3
Fernando López	Tifus	El Fuerte	Sinaloa	20/01/2017	3
José García	Tifus	Ahome	Sinaloa	28/03/2017	3
Luis Soto	Peste neumónica	Huetamo	Michoacan	12/03/2017	1

Fig. 1. Event to show diseases by municipalities and states of Mexico.

TREATMENT [TF]

PRESCRIPTION_ID	PATIENT	DISEASE	DOCTOR	MEDICAL
	[who]	[what]	[who]	[what]
74525	Paciente 49697	Meningitis viral, sin otra especificacion	Medico 21263	AUTRIN 600
74525	Paciente 49697	Meningitis viral, sin otra especificacion	Medico 21263	BENAXIMA
70091	Paciente 105454	Fiebre exantematica enteroviral [exantema de boston]	Medico 21115	BLEMIL PLUS ARAC
73050	Paciente 218889	Vertigo epidemico	Medico 21354	DIODOQUIN
74404	Paciente 105507	Otras infecciones virales del sistema nervioso central	Medico 13388	INVIRASE
317922	Paciente 105546	Otras infecciones virales del sistema nervioso central	Medico 21374	INVIRASE
70909	Paciente 105569	Otras fiebres virales transmitidas por mosquitos	Medico 20873	
361269	Paciente 218936	Fiebre del valle del rift	Medico 21181	PAMIGEN
74729	Paciente 218956	Fiebre del valle del rift	Medico 22178	PAMIGEN

Fig. 2. Fact table that contains the IDs that relate the dimensions.

	Dimensions	Importance	Estimate	Doctor	Patient	Medical	Disease	Colony	Municipalit	State	Treatment	Prescription	Doctor	Patient
EVENT (who does what)				who		what		where			why & how		Stakeholder group	
Importance														
Estimate														
Evaluate the patient	5	300		✓	✓			✓	✓	✓			*	✓
Diagnose the patient	7	200		✓	✓		✓	✓	✓	✓	✓		*	✓
Give medical treatment	6	100		✓		✓		✓	✓	✓		✓	*	✓

Fig. 3. Event matrix.

4 Results

The data repository used to design and implement the data warehouse is a relational database derived from EMR system, containing 316,295 records of medical consultations from different patients living in different cities and states of Mexico. The structure of this relational database was analyzed and only the required tables to create the data warehouse with the dimensional model obtained using BEAM were extracted.

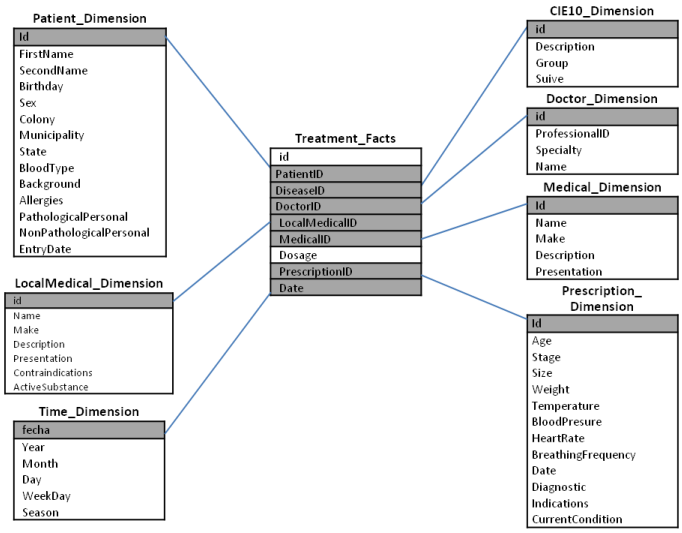

Fig. 4. Star Schema showing the dimensional model of the data warehouse.

Fig. 5. Interactive dashboard connected to the dimensional model of the data warehouse.

The tables were identified according to the indicators that the stakeholders (doctors from Obregon city identified and prioritized the indicators) defined in the business events of the BEAM methodology, which would be useful for decision making. The indicators are disease by stage of life, diseases by season of the year, epidemiology surveillance diseases, all can be filtered by municipality and state. The relational database consists of 172 tables. The tables that contain the required data to create the

business events are only 7. These tables record data about patients, doctors, medications, diseases, treatments of the patients and the prescriptions that the doctors issue at the end of each medical appointment.

Once the relational database was analyzed, following the BEAM methodology, the 7W's framework technique was applied, several BEAM*tables and the event matrix were created and finally the star dimensional model was created.

7Ws Framework and BEAM*Tables. Events were defined with the help of stakeholders. In our case, the stakeholders are healthcare professionals. Examples of the defined events are Disease by stage of life, Disease by season of the year, Epidemiology surveillance diseases by municipality and state. The event shown in Fig. 1 is an example of a query to display records of patients' diseases, with granularity by municipalities and states of Mexico.

Fig. 2 shows an example of the fact table of the model that contains the IDs of the prescriptions, patients, diseases, doctors, medicines. All the descriptions corresponding to the IDs of this table are found in the derived dimension tables. This can be observed in the star dimensional model (Fig. 4), where the relationship between the IDs of the fact table and the dimension tables is defined.

Event matrix. Fig. 3 shows the event matrix with the events that happen during a medical consultation and relates them to the data of the dimensions that are involved in the development of the medical consultation.

Star dimensional model. As a result of using the BEAM techniques, we obtained the star dimensional model (Fig. 4). The dimensional model has seven dimension tables and a fact table, which contains the IDs that relate the facts to each of the dimensions. Patient dimension contains the records of the registered patients. Medical dimension contains the names, and specialty of health professionals. Medication and Medication Local dimensions are the Descriptions and brands of prescription drugs. CIE10 dimension includes the diseases and their classification SUIVE, in case of being considered epidemiological. Prescription dimension contains the symptoms of the patients at the time of the medical appointment. Time dimension is tied to the dates of the facts recorded and it enables to achieve the required granularity for the events. The relationships between the fact and the dimension tables enable agile and efficient queries, compared with the E-R model.

Validation. In order to validate the functionality of the dimensional model of the data warehouse, we performed the Extraction, Transformation and Load (ETL) processes of the data, as well as we created an interactive dashboard that shows the indicators resulting from the queries to the data warehouse.

ETL processes. The required tables were extracted from the relational database, completed and stored in temporary tables using SQL code. Next, there is an example of the code that we used to perform the extraction of the data:

*select * into DW.dbo.paciente from mmanik_completa.dbo.paciente*

*select * into DW.dbo.municipio from mmanik_completa.dbo.municipio*

*select * into DW.dbo.estado from mmanik_completa.dbo.estado*.

After the extraction process, the extracted data were transformed using the tables and fields of interest and loaded into the tables of the dimensional database. An example of the code used to perform the transformation and load processes is the following:

select m.id,cedula,u.nombre,e.nombre as especialidad,cedula_validada insert into DW.mm.medico from medico as m left join especialidad as e on especialidad_id=e.id left join usuario as u on usuario_id=u.id.

Interactive dashboard. We designed and created an interactive TABLEAU [1] dashboard once the ETL processes were finished to validate the efficiency of the data warehouse design. This dashboard is connected with the data of the dimensional model and performs queries to extract the stored data in a rapid and simple way. The dashboard shows the diseases by stage of life presented by patients (Fig 5). The dashboard also shows the diseases by seasons of the year, with the same adjustments as the indicator of diseases by life stage. Additionally, the dashboard shows the diseases that are classified in the SUIVE in a geographical map.

Thus, the health professionals can observe the map with all the diseases or select one that is of his/her interest. All indicators can be filtered by state and municipality.

5 Conclusion

With our results it is possible to support the decision making related to epidemiological patterns, trends, areas of influence and other pathologies, based on the indicators that emerged in the business events of the proposed dimensional model.

This paper shows how the BEAM methodology can be applied to design a data warehouse based on EMR system. In the future, we plan to evaluate the use of the dashboard with health professionals to investigate the potential of the data warehouse in supporting decision making in a real-case scenario.

References

1. Berner, E. S., Detmer, D. E., Simborg, D.: Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. J. Am. Med. Informatics Assoc., 12(1) pp. 3–7 (2005)
2. Abidi, S.: Knowledge Management in Healthcare 63, pp. 5–18 (2001)
3. Kimball, R., Ross, M.: The data warehouse toolkit: the complete guide to dimensional modelling (2011)
4. Skarin, M., Kniberg, H.: Kanban y Scrum – obteniendo lo mejor de ambos (2010)
5. Corr, L., Stagnitto, J.: Agile Data Warehouse Design, 1 th. Leeds, UK: DecisionOne Press (2012)
6. Marco-Ruiz, L., Moner, D., Maldonado, J. A., Kolstrup, N., Bellika, J. G.: Archetype-based data warehouse environment to enable the reuse of electronic health record data, Int. J. Med. Inform., 84(9) pp. 702–714 (2015)
7. Laws, R., Gillespie, S., Puro, J., Van-Rompaey, S., Quach, T., Carroll, J., Chang Weir, R., Crawford, P., Grasso, C., Kaleba, E., McBurnie, M. A.:The Community Health Applied Research Network (CHARN) Data Warehouse: a Resource for Patient-Centered Outcomes Research and Quality Improvement in Underserved, Safety Net Populations, eGEMs (Generating Evid. Methods to Improv. patient outcomes), 2(3), pp. 10–14 (2014)
8. Ross,T. R., Ng, D., Brown, J. S., Pardee, R., Hornbrook, M. C., Hart, G., Steiner, J. F.: The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support

[1] https://www.tableau.com/

Collaboration, eGEMs (Generating Evid. Methods to Improv. patient outcomes), 2(1) (2014)
9. Dowie, J., Kjer-Kaltoft, M., Salkeld, G., Cunich, M.: Towards generic online multicriteria decision support in patient-centred health care, Heal. Expect., 18(5), pp. 689–702 (2015)
10. Pérez-Cuevas, R., Doubova, S. V., Suarez-Ortega, M., Law, M., Pande, A. H., Escobedo, J., Espinosa-Larrañaga, F., Ross-Degnan, D., Wagner, A. K.: Evaluating quality of care for patients with type 2 diabetes using electronic health record information in Mexico, BMC Med. Inform. Decis. Mak., 12(1) pp. 50 (2012)
11. Bailey, D., Weeks, J., Evans, E., Lowery, J., McFarland, L.: Technologies for Managing Virtual Data Warehouse Access and Identifying Appropriate Levels of Staffing at CHI Institute for Research and Innovation. Journal of Patient-Centered Research and Reviews, 4(3), pp. 197–198 (2017)