Bond University
ePublications@bond

Faculty of Society and Design Publications | Faculty of Society and Design

1-26-2015

Clinical data warehousing: A business analytics approach for managing health data

Lekha Narra
Queensland University of Technology

Tony Sahama
Queensland University of Technology

Peta Stapleton
Bond University, pstaplet@bond.edu.au

Follow this and additional works at: http://epublications.bond.edu.au/fsd_papers

 Part of the Databases and Information Systems Commons, and the Health Psychology Commons

Recommended Citation
Lekha Narra, Tony Sahama, and Peta Stapleton. (2015) "Clinical data warehousing: A business analytics approach for managing health data" 8th Australasian Workshop on Health Informatics and Knowledge Management. Sydney.Jan. 2015.

http://epublications.bond.edu.au/fsd_papers/227

This Conference Paper is brought to you by the Faculty of Society and Design at ePublications@bond. It has been accepted for inclusion in Faculty of Society and Design Publications by an authorized administrator of ePublications@bond. For more information, please contact Bond University's Repository Coordinator.

Clinical Data Warehousing

A Business Analytics approach for managing health data

Lekha Narra[1], Tony Sahama[1] and Peta Stapleton[2]

[1]School of Electrical Engineering and Computer Science, Science and Engineering Faculty
Queensland University of Technology (QUT), Brisbane, Queensland 4000, Australia.

[2]School of Psychology, Faculty of Society & Design, Bond University
Gold Coast, Queensland 4229, Australia

lekha.kolluri@connect.qut.edu.au; t.sahama@qut.edu.au & pstaplet@bond.edu.au

Abstract

Heterogeneous health data is a critical issue when managing health information for quality decision making processes. In this paper we examine the efficient aggregation of lifestyle information through a data warehousing architecture lens. We present a proof of concept for a clinical data warehouse architecture that enables evidence based decision making processes by integrating and organising disparate data silos in support of healthcare services improvement paradigms.

Keywords: Clinical Data Warehouse, Business Intelligence & Analytics, Health data, Obesity, Centralised data warehouse Architecture, Star schema

1 Introduction

Lifestyle trends are a dynamic process that may be located at the intersection of environmental contexts and genetic influences (Claassen et al. 2010). Some of these trends are contributing to an increase in chronic diseases as well as healthcare costs. As a result, focus on the changing relationship between lifestyle and quality of life as well as underlying health trends in a given population has gained prominence. However effective management of these aspects demands quality information for enabling informed decisions.

An unhealthy lifestyle significantly impacts the incidence of chronic diseases which adversely affect productivity and labour market participation. According to the Australian Chronic Disease Prevention Alliance, known and preventable risk factors including smoking, physical inactivity, obesity, poor nutrition and high blood pressure are accountable for up to one-third of all health problems (Productivity Commission 2006).An estimated $4 billion dollars in direct healthcare savings is achievable through better prevention and management of chronic disease according to the Productivity Commission report (Productivity Commission 2006). In this context, understanding the origin of expenditure on chronic disease management based on available and archived data will help policy makers to make appropriate decisions in order to improve the lifestyle, health and well-being of people.

According to IBM researchers, worldwide digital healthcare data will grow from 500 petabytes in 2012 to an estimated 25,000 petabytes in 2020 (Kuo et al. 2014). This predicted increase in the volume of data establishes the imperative to adopt dedicated technologies and processes with the capacity to process and analyse large volumes of data.

With the onset of Business Intelligence and Analytics (BI&A), enormous datasets can be effectively examined to gain invaluable insights which can improve the understanding of various health issues and may even predict future disease outbreaks. BI&A refers to the techniques, technologies, systems, tools, methodologies, and applications used for continuous iterative exploration of critical business data to better understand the business and market and make timely decisions.

BI & A is data driven; leveraging opportunities presented by large volumes of data as well as domain-specific analytics which is needed in many critical and high impact application areas (Chen, Chiang, Storey 2012). BI & A relies on a data warehousing approach where extraction, transformation and loading (ETL) is utilised for conversion and integration of enterprise-wide data. In this context, our experiment is to use the BI & A approach to identify the effect of individual lifestyle trends in health management scenarios.

2 Research scenario

A key goal of this paper is to offer a perspective of the antecedents for obesity in the Australian context with a focus on levels of physical activity, nutrition, sex and age. A major issue of global concern in recent decades, obesity is considered a significant factor contributing to cardiovascular disease, hypertension, type-II diabetes mellitus (DM), stroke, dyslipidaemia, osteoarthritis and some types of cancers (Burton and Foster 1985). According to Australian Bureau of Statistics (ABS), the

Copyright © 2015, Australian Computer Society, Inc. This paper appeared at the Eighth Australasian Workshop on Health Informatics and Knowledge Management (HIKM, 2015), Sydney, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 164 - A. Anthony Maeder and Jim Warren, Eds. Reproduction for academic, not-for-profit purposes permitted provided this text is included.

overall cost of obesity to Australian society and government was $58.2 billion in 2008 (ABS 2011a).

Changes in physical activity levels and food habits are recognised antecedents for obesity. In this scenario, there is much emphasis on detailed analysis of how exercise and nutrition affect obesity. On the whole, the following questions were of interest while developing the business analytics solution(s):

1 What is the relationship of weight to age and sex?
2 What is the relationship of weight to fruit and vegetable intake?
3 What is the relationship of weight to type of milk consumed?
4 What is the relationship of weight to amount of physical activity for fitness, recreation or sport?
5 What is the relationship of weight to number of days exercised for fitness, recreation or sport?

3 Data and information availability

Identification of health related trends requires analysing individual person's data over a population. The data used in this experiment is taken from the ABS (ABS 2011a, ABS 2011b) which presents overweight and obesity data from the 2007-08 National Health Survey (NHS) and the 1995 NHS in Microsoft Excel spreadsheets. The source files present data related to weight of adults in Australia, examining various factors including physical activity, nutrition, age and sex.

The source data in Table 1.1 of (ABS 2011a) was presented in summary form specifying the percentage of surveyed population (male or female adults or adult persons) in each weight category with respect to categories of nutrition and physical activity and gender while Table 1.1 and Table 1.2 of (ABS 2011b) presented the same with respect to each classification of age in 2007-08 and year 1995 respectively while the source data in Table 1.3 of (ABS 2011b) presented the same with respect to each classification of regions in Australia. The following tables Table 1 and Table 2 present the number of levels or categories present in each attribute as well as the nature of the data contained in each level/ category.

Factors	No. of categories	Range
Age	7	18+(years)
Weight	6	>18.5 to 30.0+ (BMI)
Gender	3	1~3

Table 1: Summary of the data sources

Factors	Levels of the Factors & Attributes			
	No. of levels	Sub-Levels	Nature of options	
			Check compliance with standards	Selections among pre-defined criteria
Nutrition	4	1	✓	
		2	✓	
		3	✓	
		4		✓
Physical Activity	3	1	✓	
		2		✓
		3		✓

Table 2: Summary of the data sources

The measured height and weight data from the 1995 ABS National Nutrition Survey (NNS) and the 2007-08 ABS NHS were used to calculate Body Mass Index (BMI) and classified people as underweight, normal weight, overweight and obese (ABS 2011a) as shown in Table 3.

Body Mass Index (BMI): It is defined as the weight in kilograms divided by the square of the height in metres (kg/m2) (ABS 2011a).

BMI (Adult)	Range of the BMI			
	< 18.5	18.5 ~ 25.0	25.0 ~ 30.0	> 30.0
Category	Underweight	Normal	Overweight	Obese

Table 3: Classification of adults according to BMI

4 Approach to the problem scenario

Various custom built tools and techniques were being used to analyse data and discover critical relationships between different aspects of an interested subject area. These conventional tools are generally appropriate to handle small to medium sized data. However, due to the large volume of health data, conventional tools lag in terms of capacity and performance. In addition, combining data from multiple sources in to a consistent form is challenging (Shepherd 2002) with such tools.

These constraints are motivating researchers to consider a data repository that will help information integration and support timely analysis. Research suggests that the use of a data warehousing approach is a promising start to overcome these constraints, hence in this paper; such techniques and tools were investigated.

A data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data that supports managerial decision making (Inmon 2005, 29). These features help us to easily access health data whenever required in a consistent form when we build the data warehouse according to the user requirements. Also the data is secure and enables evidence based and faster quality health care related decision making which is the main reason for adopting a BI & A approach.

In this case, as source data is in summary form, aggregated data was used to populate the data warehouse in this experiment. This aggregation is done because of restricted access to the raw data used to develop the source files in (ABS 2011a) and (ABS 2011b). Also, the available summary data was grouped combining various dimensions (e.g. in the form of data cubes) which cannot be used in the original form in order to populate the data warehouse. The aggregate data was synthesised in such a way that the averages of the aggregate data match the percentages in the original ABS spread sheets as closely as possible. This aggregated data was stored in Microsoft Excel spread sheets which were utilised as source data files to populate the data warehouse.

5 Tools and techniques used

Several data warehouse architectures were studied and matched in order to establish appropriate architecture for the current case from among those proposed in (Ponniah 2010, 32-34). After careful consideration of the nature of data used in this experiment, centralised data warehouse architecture is adopted. This is because of the absence of

significantly different subject areas in the data i.e. the main focus (or the subject area) is only on weight factor, which eliminated the need for separate data marts. Also the centralized data warehouse provides consistent, integrated and flexible source of data (Moody and Kortink 2000). With this architecture, queries and applications access data from central data warehouse itself for decision making processes.

The data warehouse is modelled using Dimensional modelling which comprises of a fact table and several dimension tables (Kimball and Ross 2002, 16). Each transaction i.e. each survey record in this case is recorded as a tuple in the Fact table and various aspects of each transaction are recorded in different dimensions. After careful study of various dimensional modelling techniques (eg. Star schema, Snowflake schema and Fact Constellation schema), Star schema was employed to build the data warehouse. Benefits like lesser query execution time due to less number of joins and foreign keys also guided the selection (Moody and Kortink 2000). A star schema consists of a fact table which mostly contains numeric data and foreign keys to connect with the dimension tables. Also there is no direct connection between the dimension tables (Weininger 2002). Finally, Star schemas offer the advantage of less complex queries and are easy to understand.

However, the disadvantage of using Star schema for the current case is that, though the fact table is in normalized form, the dimension tables are de-normalized which is deliberately adapted as storage is effectively cheap these days (Sahama and Croll 2007) and also due to the advantage of better performance (Sen and Sinha 2005).

The following figure presents the Star schema employed for the current case.

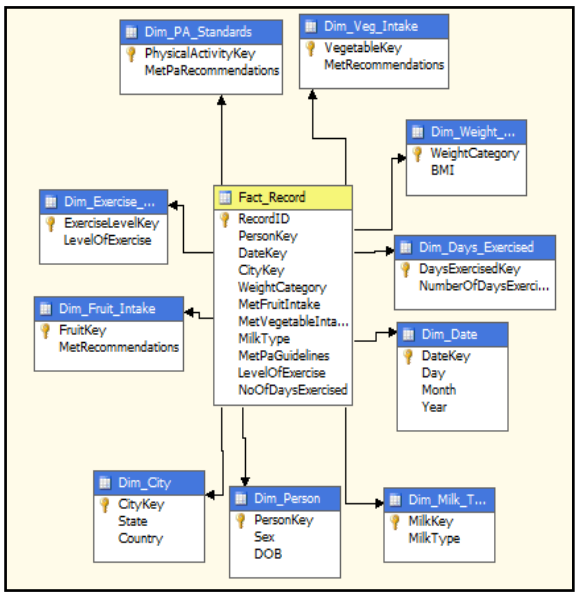

Figure 1: Star schema diagram for Obesity data warehouse

The levels indicated in Table 3 for Physical activity and Nutrition were mapped to individual dimensions as shown in Fig. 1. The other aspects like age and sex were mapped to single dimension Dim_Person. This schema allows us to analyse the weight factor with respect to all the attributes of interest as well as with respect to location and time.

Microsoft SQL Server 2012 was used for building the data warehouse in this experiment. The first step in building a data warehouse is to extract the data from source system. The ETL (Extraction, Transform and Load) approach is preferred over ELT (Extraction, Loading and Transformation) for this purpose. The basic difference between ETL and ELT is the order in which Loading and Transformation activities are performed. ETL is preferred because the source system is not a production system and performing the transformations before loading data in to the data warehouse wouldn't affect the performance of any production system. Also there were not many data transformations performed before loading the data due to the nature of the data involved and the analysis required. This enabled the ability to drill through the data to the required level of granularity without having a copy of the source data in the data warehouse (typically ELT).

SQL Server Integration Services (SSIS) was used for the ETL process. The data in the source files was first cleaned to remove erroneous data and necessary transformations were performed to match the source data with destination format and then loaded in to the data warehouse.

Later the SQL Server Analysis Services was used to build the cubes using the Obesity data warehouse as the data source. A single cube was built using all of the dimensions in the data warehouse. This is to allow for the analysis of weight factor with respect to each category of nutrition, physical activity, age and sex as well as location, time and individual. Building cubes enables the users to use OLAP (On-line Analytical Processing) tools for interactive analysis of multi-dimensional data at required granularity levels. Data cubes provide flexible access to summarised data i.e. data cubes can store pre-computed measures (like count(), sum() etc.) for varied combinations of data dimensions (Han and Kamber 2011) which enables faster query processing times. The count (number) of records is specified as a measure in the cube which can be analysed according to the categories in each dimension. This means that the number of records satisfying the criteria in a query will be counted as a result. Each dimension of this cube is related to individual dimensions in the Star schema.

This allows for the various categories of weight factor (based on the BMI) to be analysed with respect to individual as well as the combination of categories of nutrition, physical activity, age and sex shown in Tables 1 and 3. This cube structure allows us to answer the questions presented in section II of the paper. Later SQL Server Reporting Services can be used to query the cube and obtain the reports.

6 Discussion and conclusion

A data warehousing approach for studying health issues is an effective data management method crucial for addressing the increasing amount of digital health data across the globe. Data warehousing which is currently being used to study various diseases, effectiveness of various treatments, understand and analyse health care costs in hospitals can also be used to identify lifestyle

trends and their impact on individual's health. It helps to integrate data from multiple sources efficiently and in less time with dedicated tools like SSIS. The biggest challenge when integrating data from various sources is that the data structure and the attributes of persons available for analysis of a health aspect (e.g. obesity) usually differ with each individual source. However, with the advent of Electronic Health Records (EHR), such problems can be easily overtaken due to uniform data structure and data availability.

Having observed the implementation of data warehousing for health data, it can be said that this kind of implementation may be achieved using data available with health care organisations and government or private surveys over long periods. This will also enable identification of relationships between various aspects which can be only be possible with large amounts of data observed over a prolonged period, for example the relationship between food and/or lifestyle with various types of cancers.

In the current case, this approach enabled the study of relationship between weight with respect to physical activity, nutrition, sex, age, location as well as how these relationships and individual aspects (e.g. nutrition) change over time.

Disclaimer, discloser and Conflict of Interest

The authors declare that they have no conflict of interest to disclose. The information disseminated, scientific argument presented and analyses completed in this paper are the view of the authors and there is no contribution of the organisational data and information towards such findings.

Lekha Narra (LN) is responsible for the study conceptualisation. All three authors (LN, Tony Sahama—TS and Peta Stapleton—PS) were involved in the development of the information extraction, data validity check and scientific argument. LN is responsible for the data collection phase, TS and PS were responsible for results reporting. All authors were responsible due to having complete access to the study data and information supporting this publication.

7 References

Claassen, L., Henneman, L., Janssens, A.C., Wijdenes-Pijl, M., Qureshi, N., Walter, F.M. et al. (2010): Using family history information to promote healthy lifestyles and prevent diseases; a discussion of the evidence. *BMC Public Health*, **10**:248.

Productivity Commission 2006, *Potential Benefits of the National Reform Agenda, Report to the Council of Australian Governments*, Canberra.

Kuo, M.H., Sahama, T., Kushniruk, A.W., Borycki, E.M., Grunwell, D.K. (2014): Health Big Data Analytics: current perspectives, challenges and potential solutions. *International Journal of Big Data Intelligence*, 1(1/2), pp. 114—126.

Chen, H., Chiang, R. H., Storey, V. C. (2012): Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, **36**(4), 1165—1188.

Burton, B.T., Foster, W.R. (1985): Health implications of obesity: an NIH Consensus Development Conference. *Journal of the American Dietetic Association*, **85**(9):1117—1121.

Australian Bureau of Statistics 2011, *Overweight and Obesity in Adults in Australia: A Snapshot, 2007-08, Selected nutrition and physical activity characteristics by measured Body Mass Index(a), data cube: Excel spreadsheet*, cat. no. 4842.0.55.001, Canberra.

Australian Bureau of Statistics 2011, *Overweight and Obesity in Adults in Australia: A Snapshot, 2007-08, Measured Body Mass Index by demographic and socio-economic characteristics, data cube: Excel spreadsheet*, cat. no. 4842.0.55.001, Canberra.

Shepherd, M. (2007): Challenges in Health Informatics. *IEEE Proceedings of the 40th Hawaii International Conference on System Sciences*, Waikoloa, HI**,** DOI: 10.1109/HICSS.2007.123.

Inmon, W.H. (2005). From The Data Warehousing Environment. In *Building the data warehouse*. 29. Inmon, W.H. (eds) 4rth ed.. New York, Wiley.

Ponniah, P. (2010). From Data Warehouse: The Building Blocks. In *Data warehousing fundamentals for IT professionals*. 32-34. Ponniah, P.(eds). 2nd ed.. A John Wiley & Sons, Inc., publications.

Moody, D.L., Kortink, M.AR. (2000): From enterprise models to dimensional models: a methodology for data warehouse and data mart design. *DMDW*, p. **5**.

Kimball, R., Ross, M. (2002). From Dimensional Modeling Vocabulary. In *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. 16. Kimball, R., Ross, M. (eds). 2nd ed.. New York, Wiley Publishing Inc.

Weininger, A. (2002): Efficient execution of joins in a star schema. *Proc. ACM SIGMOD international conference on Management of data*, 542—545, ACM.

T. Sahama, P. R Croll, "A Data Warehouse Architecture for Clinical Data Warehousing", Proceedings Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007) CRPIT, **68**, 227—232, Ballarat, Victoria.

Sen, A., Sinha, A. P. (2005): A comparison of data warehousing methodologies. *Communications of the ACM*, **48**(3), 79—84.

Han, J., Kamber, M. (2011). From Data Cube Technology. In *Data Mining : Concepts and Techniques*. 187. Han, J., Kamber, M. (eds). 3rd ed. n.p. Morgan Kaufmann Publishers Inc.