Jaehwan Han

Data Warehousing Methodologies

Design

Data warehousing methodologies can be applied to a variety of industries, including data design, architectural design, business analytics, and healthcare. A clinical data warehouse (CDW) is designed to integrate heterogeneous clinical data stores (CDS), which contain blobs or ”islands” of disparate information into a single, subject-focused set. This information is combined and collected over time, enabling researchers to track an individual’s health status in response to various interventions.[1] While this heterogeneous information may seem unusable, they allow different information to be handled by employees, students, and researchers in the same vicinity.[2] This data integration is completed prior to analysis, which allows researchers to query and analyze the database without regard to processing. Proactive integration also facilitates stability, as the database does not change as a result of operational processes.[1] By contrast, traditional designs only process and integrate the information after a query is submitted, which limits the speed and sophistication of the database.

It is a monumental challenge to manipulate the various CDS data into a consistent and coherent format. Despite this complexity, the CDW offers a powerful and efficient mechanism to deliver quality patient care. When coupled with a data mart, performance is further increased with highly optimized data structures for specific tasks. CDW enhances timely analysis, increases the quality of real time decision making, and facilitates efficient storage.[1] Further, it allows healthcare providers to gain access to clinical data in the patient care process and dissolves currently existing barriers between public and private data. Recent advances in processing power and storage capability have made such a system feasible - a model of which is explored by Sahama and Croll.

Conventional methods were avoided in the design phase. Advances in processor speed and storage capabilities allowed Sahama and Croll to create a CDW that incorporates data duplication and ignores data normalization.[1] These two components would not have been possible using a traditional relational database format. Continuing their avoidance of convention, a business analytics approach was used in the design of the CDW. Data warehouses have been successfully implemented in business settings, which Sahama and Croll aimed to model in a healthcare context. Sahama and Croll state, “By effectively leveraging enterprise-wide data on labor expenditures, supply utilization, procedures, and medications prescribed, healthcare professionals can identify and correct wasteful prices and unnecessary expenditures.”[1] Not only was the data model concerned with patient outcomes, but cost efficiency of the care as well. A CDW has the potential to save healthcare organizations unnecessary expenses and duplicated records from disparate sources. Sahama and Croll supports this idea by saying that the CDW model is beneficial with patient management scenarios, especially with those that deal with admissions, discharges, and transfers.[1]

Data model

Sahama and Croll utilized a third-party system to facilitate data extraction from numerous sources into the new CDW. The data model was primarily based on existing warehousing methodologies proposed by Sen and Sinha. Sahama and Croll decided on a Distributed Data Warehouse Architecture with added data mart capability for improved efficiency with specialized operations. During this process, portability was an important consideration so that the warehouse could be easily shared and used in different locations. As part of this consideration, OnLine Analytical Processing (OLAP) was incorporated so that different teams could analyze the warehouse simultaneously.[1] OLAP integrates decision support into various data architectural frames. This alleviates the issue of complexity, as OLAP enables multidimensional analysis, which allows researchers to increase decreases the data aggregation, select and project the data, as well as re-orienting the data into a desired format.[3]

When designing CDW architecture, data integration tasks for medical information poses a challenge. For example, conducting data analysis on a patient with cancer (primarily quantitative), is very different than for one with a mental health disorder (primarily qualitative). In order to tackle the issue of qualitative vs quantitative patient management, Sahama and Croll designed two separate CDWs. The oncology-based CDW includes quantitative counts such as length of stay and discharge information. In contrast, the mental health CDW includes qualitative information such as client interviews and client outcomes.[1]

Implementation

The proposed data model was implemented using the SAS Warehouse Administrator. Data from disparate sources were saved in Microsoft Excel and imported into SAS. The the various data were then manipulated in SAS to adhere to a standardized table format that is consistent with most commercial databases. Operational Data Definitions were created and used to generate metadata for the various information.[1]

The use of cube metadata facilitates the execution of queries, due to the cube computation process.[4] This, in turn, allows for more efficient data expansion, which is an important consideration when dealing with a large set of data. To further facilitate performance, data tables that addressed specific issues, such as predictive analysis were loaded as well as data marts for report generation and targeted analysis. With all of the data organized in a CDW, SAS analytical techniques can be used on subsets of the large data set, as well as provide insights for further data mining. Without a structured CDW organizing this disparate data, it would not be possible to analyze all of the individual records.

Using SAS and an unconventional business analytics approach, Sahama and Croll were able to create a working model of a CDW that offers improved speed and analytical power for researchers. Sahama and Croll state that in a typical project, data preparation and integration comprises 90% of the required effort. Although the creation of a CDW is tedious, it allows researchers to focus more of their efforts on modeling and analysis, rather than independent data acquisition.

References

1. Sahama TR, Croll PR. A Data Warehouse Architecture for Clinical Data Warehousing. In: *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers - Volume 68*. ACSW '07. Darlinghurst, Australia, Australia: Australian Computer Society, Inc.; 2007:227–232. http://dl.acm.org/citation.cfm?id=1274531.1274560. Accessed December 5, 2018.

2. Flory A, Soupirot P, Tchounikine A. A Design and implementation of a data warehouse for research administration universities. In: ; 2001.

3. Chaudhuri S, Dayal U. An Overview of Data Warehousing and OLAP technology. *ACM SIGMOD Record*. 1997;26(1):65-74. doi:10.1145/248603.248616

4. Solodovnikova D, Niedrite L. An Approach to Handle Big Data Warehouse Evolution. *arXiv:180904284 [cs]*. September 2018. http://arxiv.org/abs/1809.04284. Accessed December 5, 2018.