

Query Performance and Architecture Gains after Adoption of a Data Warehouse Appliance for a Healthcare Data Warehouse

Vojtech Huser MD PhD, Nicholas R. Berger, Tuhina Kola MS, Cynthia S. Motszko MS, Justin B Starren MD PhD
Marshfield Clinic, Marshfield, WI, USA

Abstract

Use of parallel data architecture is a new paradigm in data warehousing. At our institution, we switched to using a Data Warehouse Appliance (DWA) for a large healthcare data warehouse. We present our positive experience with migration to DWA and compare a DWA to a traditional database.

Introduction

An Enterprise Data Warehouse (EDW) at a healthcare institution represents a key resource supporting business and management decisions, quality improvement programs or medical research. Exponential growth of Electronic Health Record (EHR) data presents complexity, size and performance challenges to existing database management systems (DBMSs). Understanding and maintaining today's DBMSs requires substantial expertise [1]. A category of data warehouse products known as a Data Warehouse Appliance (DWA) tries to address these limitations by use of parallel architectures and combining hardware and software solutions into one product. DWAs claim a significant increase in speed for complex queries as well as decrease in total cost of ownership. While DWA's presence in the DBMS market is growing, practical experiences of healthcare institutions transitioning from traditional DBMSs to a DWA have not been extensively reported in literature. We present our experience of migrating from a traditional DBMS (Oracle 10g) to a DWA platform (Netezza Performance Server 10100, 112 CPUs). At the time of our transition (2008), Netezza was the market leading DWA vendor. High administrative and maintenance costs and suboptimal query speed performance of our traditional data warehouse were the main reasons for this migration.

Methods

The following stages of our transition will be described: (1) *Pre-purchase testing* which included testing the DWA's performance for initial data load, table replication, incremental ETL (Extract, Transfer, Load) data processing, data querying as well as application compatibility testing; (2) *Regular use results* which compared the performance of the new DWA platform to the previous traditional DBMS

in numerous EDW processes after the purchase; and (3) *Qualitative analysis* which included collecting qualitative feedback from the EDW administrators as well as EDW customers (business and research analysts). *Institutional background:* Marshfield Clinic's (MC) EDW has 5.4 TB of data in 900+ tables. MC is a healthcare network in Wisconsin with 45 locations and 791 employed physicians.

Results

The pre-purchase testing indicated favorable results. The two key conclusions were: (1) limited data manipulation features beyond the ANSI SQL standard; and (2) the importance of performing any data transformation within the DWA after the data has been loaded (extract, load and then transform). The regular use experience showed major improvements in query speed and enabled daily re-builds of the entire warehouse thus offering an extremely flexible response to changing reporting requirements. The qualitative results indicated overall satisfaction with the migration to the DWA. The poster will include tables with detailed quantitative performance data for stages one and two. The major positive qualitative comments were: lower administration costs (hardware and database configurations), lower software costs (not a traditional CPU-based pricing model), fast query speed, and simpler maintenance and optimization (no complex indexing needed). DWA disadvantages, in comparison to the prior, traditional DBMS, included: no native data consistency and triggers functionality, slower single-record processing (e.g., querying or inserting a single row), and more restricted record-length and data types options.

Discussion

A DWA provides significant advantages and can be successfully used as a platform for a healthcare EDW. Our results are limited by the two compared vendors and the setting of MC. However, the difference in the two compared technologies and costs involved make a proper comparison unrealistic.

References

- [1] Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform* 2001; 34: 285-98