Decision Support Systems 95 (2017) 61–81

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Expediting analytical databases with columnar approach

Nenad Jukic [a,*], Boris Jukic [b], Abhishek Sharma [a], Svetlozar Nestorov [a], Benjamin Korallus Arnold [c]

[a] Loyola University Chicago, Quinlan School of Business, 16 E Pearson, 60611 Chicago, IL, United States
[b] Clarkson University, School of Business, Bertrand H. Snell Hall, Potsdam NY 13699-5790, United States
[c] University of Chicago, Graham School of Continuing Liberal and Professional Studies, 450 North Cityfront Plaza Drive, Chicago, IL 60611, United States

ARTICLE INFO

Article history:
Received 31 December 2015
Received in revised form 22 December 2016
Accepted 24 December 2016
Available online 6 January 2017

Keywords:
Data warehouses
Decision support
Big data
Performance
ETL
Columnar databases

ABSTRACT

The approaches and discussions given in this paper offer applicable solutions for a number of scenarios taking place in the contemporary world that are dealing with performance issues in development and use of analytical databases for the support of both tactical and strategic decision making. The paper introduces a novel method for expediting the development and use of analytical databases that combines columnar database technology with an approach based on denormalizing data tables for analysis and decision support. This method improves the feasibility and quality of tactical decision making by making critical information more readily available. It also improves the quality of longer term strategic decision making by widening the range of feasible queries against the vast amounts of available information. The advantages include the improvements in the performance of the ETL process (the most common time-consuming bottleneck in most implementations of data warehousing for quality decision support) and in the performance of the individual analytical queries. These improvements in the critical decision support infrastructure are achieved without resulting in insurmountable storage-size increase requirements. The efficiencies and advantages of the introduced approach are illustrated by showing the application in two relevant real-world cases.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

The importance of data in all aspects of business, science and government continues to grow. Data analysis underlies most if not all of the important decisions taken by corporate, academic and political leaders. Current technologies make available to analysts and managers a vast amount of structured and unstructured data from a variety of sources [1]. With the digitization of world commerce, the emergence of big data and the advance of analytical technologies, organizations have extraordinary opportunities to differentiate themselves through analytics [2]. Analytical processes that used to require month, days, or hours have been reduced to minutes, seconds, and fractions of seconds, with shorter processing times leading to higher expectations [3]. In fact, data and its analysis is increasingly required not only to guide and direct strategic plans but also to justify and explain tactical actions. In some ways, using analytics for these immediate operational needs can be more difficult than crafting long-term strategies. Whereas future strategies are typically iterated over time, operational decisions require precise and accurate insights to be available much more quickly: hence, the need for analytics speed [2]. This requirement, coupled with the explosion of data everywhere has led to the today's trend in which timely analytics capabilities are a must-have for many organizations.

As a result, firms are increasingly demanding immediate access to internal corporate data and external data, from an ever-growing list of sources, to develop insights and to generate actionable responses that produce measurable results in real-time [4]. To better harness the power of their own data, many firms have invested in data warehousing technology [5]. However, the increasing reliance on data analysis for everyday decisions has put a strain on the standard data warehousing technology. In particular the ETL (extract, transform, and load) process that brings data from the operational databases to the analytical data warehouse is required to run faster and more frequently than ever in order to make the most recent operational data available for analysis. For example, for many businesses detecting trends and patterns in their retail stores and supply chains is now a daily if not an hourly process that has critical implications on the bottom line [6].

Finally, the Big Data trend also plays a significant role in the increasing demands on the ETL process as many companies and organizations are shifting to data-driven decisions across all management levels. Capturing and processing data from a multitude of sensors [7] is helping companies include many external factors such as weather and traffic in their data warehouses and improve their analysis of customer patterns and trends.

In this paper, we examine a novel approach to expediting the ETL process for data warehousing by combining two key data warehousing

* Corresponding author at: 16 E. Pearson, Chicago, IL 60611, United States.
E-mail address: njukic@luc.edu (N. Jukic).

http://dx.doi.org/10.1016/j.dss.2016.12.002
0167-9236/© 2017 Elsevier B.V. All rights reserved.

methodologies that can enhance the ability to use massive amounts of data for more timely and comprehensive decision support and analysis. The first methodology is based on reducing the design complexity of the conceptual data model for the data warehouses by a process of denormalization of star schemas (described in detail in Section 5.1). The second methodology of our approach is based on the use of so-called columnar databases (described in detail in Section 4). Both of these methodologies have been widely used in the practice as well as academic research, however the theoretical foundations and implications of combining these methodologies, as well as practical cases based on the approach that combines them, have not so far been examined in details and presented in academic papers. In this paper, we offer a thorough examination of the synergies of the combination of denormalization and columnar databases, and analyze the requirements that justify the usage of this combination in the context of our proposed model, which we validate with a series of computational experiments. We also present two real-world projects that have taken advantage of this novel approach and describe the concrete benefits that include significantly faster ETL times, as well as improved performance of analytical queries. These benefits represent critical conditions for enabling wider use of massive amounts of data for timely decision support in near real-time tactical scenarios as well as in strategic decision making.

The key novel contribution of this approach is identifying a standard and widely accepted process of generation and usage of surrogate keys (to be explained in detail in the section below) as one of the critical bottlenecks of the ETL process and showing how our proposed approach eliminates the need for their utilization, resulting in better performance of large scale analytical depositories for timely decision making.

Another contribution of our paper is its focus on the logical implementation of the columnar database approach, and how it improves the utilization of large data depositories for decision making. There have been extensive reviews of the design and performance of column-oriented databases [8–11] for both operational and analytical databases. However, most of the articles up to the present point are focused on the physical implementation of the standard database operations and the optimal design of physical query plans. In fact, in many articles, the authors emphasized that the logical design of the database does not change and thus does not warrant any special attention. In this paper, we consider the interplay between the physical implementation of columnar databases and the logical implementation of the analytical database data models (star schemas) as a unified structure for comprehensive analytics contained within a single (denormalized) table.

The rest of this paper is organized as following. In Section 2 we give a brief overview of analytical databases, including the description of surrogate keys and slowly changing dimensions. In Section 3 we present a discussion of the role of analytical databases in the context of Big Data analytics. In Section 4 we give a brief overview of columnar databases. Section 5 gives a detailed description and discussion of the introduced columnar approach for analytical databases, providing a model based framework for performance evaluation of our approach, validated by a set of computational experiments. Section 6 describes two real-world implementations of the introduced approach, further illustrating its feasibility. And finally, Section 7 offers concluding remarks.

2. Analytical databases

Analytical databases, such as data warehouses and data marts, are databases that store and maintain analytical data separately from operational (transaction-oriented) databases. In the 1990's, it has gradually became apparent that analytical databases, designed to be used specifically for decision support in the context of organizational analytics, should be deployed logically and physically separately from day-to-day operational systems. Using this approach, the lengthy, and often unpredictable, business intelligence queries would not spoil the response time of standard operational systems. Additionally, data warehouses are based on the recognition that the conventional query optimization and execution engines do not work well on the large sets of analytical data. Hence, extension or modifications applied on separate analytical systems are required to achieve good performance [12].

Fig. 1 shows a high-level view of the architecture of an analytical database system. The analytical data in a corporate data warehouse or data mart is periodically retrieved from various data sources. Typical data sources are internal corporate operational databases that contain analytically-useful data, such as sales-transactions databases, marketing databases and HR databases. Other data sources can include external data, such as demographics data or stock-market data, and, increasingly, Big Data sources, such as machine logs, sensor data of every imaginable kind, data generated by social networks and blogs [13], etc. These extremely large amounts of unstructured and semi-structured data can be relatively straightforwardly processed using technologies such as Hadoop [14] into useful structured information.

The data from operational data sources, external sources and other repositories of interest is brought into the data warehouse or data mart via the process of extraction, transformation and load (ETL). The ETL process is responsible for the extraction of data from sources, their cleansing, customization and insertion into a data warehouse [15]. The ETL infrastructure extracts analytically useful data from the chosen data sources, transforms the extracted data so that it conforms to the structure of the data warehouse or data mart (while ensuring the completeness, consistency and other aspects of quality of the transformed data) and then loads the data into the data warehouse or data mart. After it is loaded, this data explicitly and implicitly reflects customer patterns and trends, business practices, organizational strategies, financial conditions, know-how, and other knowledge of great value to the organization [16].

Analytical databases are modeled and structured differently than operational databases. Fig. 2 shows an example of a database modeled for operational use, with larger number of tables whereby individual tables have smaller column count. Fig. 3 shows an example of an analytical database that contains the same data as the operational database shown in Fig. 2, but it is modeled differently, with model containing a smaller number of tables whereby individual tables have larger column count.

Typical operational database relational schema contains a number of connected tables, as shown in Fig. 2. As illustrated by this figure, operational databases are structured and modeled as so-called normalized relational databases, whereby the main goal is to minimize data redundancy, i.e. eliminate instances of storing of the same data more than once. If you observe data values in the bottom part of Fig. 2, you

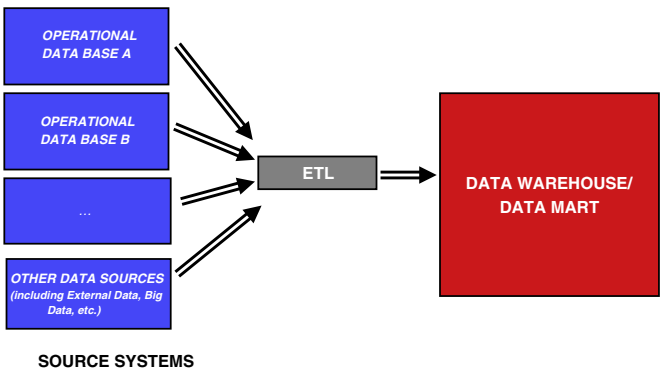

Fig. 1. Analytical database architecture.

Download English Version:

https://daneshyari.com/en/article/4972443

Download Persian Version:

https://daneshyari.com/article/4972443

Daneshyari.com