POSTER ABSTRACT

A Comparison of Two View Materialization Approaches for Disease Surveillance System [*]

Zhen Liu[1,2] Panos K. Chrysanthis[1,2] Fu-Chiang Tsui[2]
[1]The ADMT Laboratory, Department of Computer Science
[2]The RODS Laboratory, Center for Biomedical Informatics
University of Pittsburgh, Pittsburgh, PA 15260, USA
{liuzhen,panos}@cs.pitt.edu, tsui@cbmi.pitt.edu

ABSTRACT

The effectiveness of a disease surveillance system such as RODS depends heavily on the performance of the underlying data management components. Given that such system's core functionality is to support decision making and data mining, the response time of complex queries involving aggregation requires special attention. Traditionally, materialized views have been advocated to address this issue. The question that this paper examines is which approach implementing materialized views is more suitable in a disease surveillance environment. Our comparison focuses on the two most common approaches, namely, the Cache Tables approach and the Data Warehousing approach. Our evaluation shows that the choice of an approach is not limited by the cost-performance but by the cost-flexibility of usage as well.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Systems—*Query processing, Relational databases*; H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness); J.3 [**Life and Medical Sciences**]: Health, Medical Information Systems

General Terms

Design, Experimentation, Performance

Keywords

Data warehouse, materialized views, caching, public health

1. INTRODUCTION

The recent experience with the SARS epidemic and the threat of bioterrorism clearly show the increased importance of improving our capability to detect outbreaking of diseases as early as possible. The Real-time Outbreak and Disease Surveillance system (RODS) is a computer-based public health surveillance system developed for this purpose, i.e., for early detection of diseases, by the Center for Biomedical Informatics at the University of Pittsburgh [1, 2, 3]. It has been deployed in Western Pennsylvania since 1999, was used during the 2002 Winter Olympics and currently operates in three states – Ohio, Pennsylvania, and Utah.

RODS collects and processes data in real-time from clinical encounters including those found in admission, discharge and transfer records. Over 70 hospitals in the three states send data to RODS over virtual private networks and leased lines using the Health Level 7 (HL7) message protocol. RODS also processes sales of over-the-counter health care products. As opposed to clinical data, RODS receives such data in batch mode on a daily basis. All the data is stored in a relational database. Using univariate and multivariate statistical detection algorithms [4], RODS identifies anomalous patterns in the data and alerts the appropriate users accordingly. RODS also has a web interface that supports temporal and spatial analysis.

As the volume of data in the database increases (currently this is about ten thousands records per day), the execution time of queries also increases significantly. This increase in query response time is more pronounced in the case of online, statistical analysis, which often includes joins of several huge tables and aggregation over several attributes. An example of such a query is "computing daily percentage of patients with a particular prodrome (class of symptoms) in a region for one month period." Such queries are fundamental for both the detection and assessment of an outbreaking of a disease. Thus, the effectiveness of a disease surveillance system such as RODS depends heavily on the performance of the evaluation of such queries.

2. MATERIALIZED VIEWS

Besides using indexes and schema tuning, a common technique that reduces query response time is data caching, and in particular materializing views. Materialized views are summarized tables (group by queries) that cache precomputed aggregate query results and can be used to both accelerate the evaluation of queries as well as simplify the construction of complex, nested queries.

[*]Funded in part by the National Center for Disease Control, the Pennsylvania Department of Health Award ME-01-737.

SAC'04, March 14-17, 2004, Nicosia, Cyprus.
Copyright 2004 ACM 1-58113-812-1/03/04 ...$5.00.

Basically, there are two approaches in defining and using materialized views:

- *Cache Tables* (CT): Summarized tables are defined based on pre-defined users' queries and stored in the database as base tables containing the raw data.
- *Data Warehousing* (DW): The raw data is restructured into a centralized fact table in a data warehouse and summary tables are defined over the restructured data.

A clear first distinction between these two approaches is in the need of special data management tools and additional resources. The DT approach involves additional cost for acquiring the data warehouse management system (DWMS) and requires almost twice as much disk space compared to the CT approach, which does not replicate any of the raw data. In this paper, we focus only on their computational differences.

3. PERFORMANCE METRICS

The primary goal of our evaluation is to identify under which conditions each approach performs better. A secondary goal is to understand the limitations of the Cache Table approach which can operate in resource constraint environments and hence can be used to support a rapid deployment of RODS as part of a disaster management operation. These led us to evaluate the effectiveness of each materialized view approach in terms of *data freshness, responsiveness, flexibility* and *maintainability* using the following measurements, respectively:

- *Refresh Time:* which is the time to compute the aggregate results and store them in summary tables.
- *Query Response Time:* which is the time to execute a certain user's query on summary tables.
- *Instantiation Time*: which is the time to create and materialize a new summary table which is needed by the requirements of a new, possibly ad-hoc, query.
- *Aggregation Time:* which is the time to update a number of aggregate results in summary tables.

4. EXPERIMENTAL PLATFORM

All experiments ran in isolation (i.e., there was no other working load during the experiments) on a Penguin server containing dual AMD Athlon 1.6G CPU processors with 2Gbytes of memory and 800G of hard disk. The server ran Linux 7.1 and Oracle 9.2 and the same DBMS (same SID) was used to manage the raw tables, cache tables and the data warehouse.

Each Cache Table (Data Warehousing) experiment started with 1.15 million records in base tables (fact table). All data were real data coming from hospitals. In order to simulate a dynamic environment, a data feeder program was used to insert patient records into base table and fact table at the same rate as in the RODS production system.

The basic six queries in our evaluation were:

- Counts of patients with a given prodrome in a zip code
- Counts of patients with a given prodrome in a county
- Get counts of patients with a given prodrome in a state
- Counts of all patient visits in a zip code
- Counts of all patient visits in a county
- Counts of all patient visits in a state

Each of them is associated with a cache table.

5. SUMMARY OF FINDINGS

The CT approach offers the best performance with respect to freshness (i.e., lowest average refresh time). For example, for a fixed refresh interval, say 30 minutes, the refresh time of the CT approach is 3 times lower than the DW one.

The CT approach also offers the highest responsiveness (i.e., lowest average query response time). For example, for specific queries, the query response time in the CT approach is an order of magnitude lower than in the DW one. The reason is that cache tables are designed to store the counts in the specific intervals that are required by the queries. In the DW approach, each materialized view stores the counts for some basic interval, e.g., every 30 minutes. When a query requires count for a different interval, e.g., 24 hours, additional aggregation is required which increases the response time. Defining an additional materialized view for storing 24 hours counts can eliminate this additional computation but this requires more disk spaces and increases the average refresh time.

The DW approach is more flexible than the CT one, because a new summary table can be easily built based on fact tables, base level materialized views and dimension tables, in 1 minute. In contrast, the CT approacher requires a substantial amount of time to materialize a new summary table from scratch since it needs to access all the raw data.

DWMS provides a powerful function to update summary tables as the data arrive in any order with very low overhead during refresh time. On the other hand, the CT approach needs to monitor for data that arrive out of order and recompute the corresponding cache tables to correctly count for these "missing" (out-of-order) data. In unreliable environments, re-computation can be very expensive. Thus, in general, the DW approach offers better maintainability than the CT approach. In a relatively reliable environment with a small number of missing counts maintainability is not an issue and either approach can be used.

In conclusion, cache tables exhibit better performance with respect to freshness and responsiveness. On the other hand, the DW approach offers the best possible flexibility in supporting new queries and maintainability. Thus, for a disease surveillance system with a static functionality, the CT approach is more appropriate and practically free. For a disease surveillance system with more dynamic functionality, the DW approach is more appropriate and in some cases, the only choice.

6. REFERENCES

[1] http://www.health.pitt.edu/rods.

[2] Tsui FC, Espino JU, Dato VM, Gesteland PH, J Hutman, MM Wagner. Technical description of RODS: A real-time public health surveillance system, *Journal of American Medical Informatics*, 10(5): 399-408, 2003.

[3] Wagner MM, JM Robinson, FC Tsui, JU Espino, WR Hogan. Design of a national retail data monitor for public health surveillance. *Journal of American Medical Informatics Association*, 10(5): 409-418, 2003.

[4] Zhang J, FC Tsui, MM Wagner, and WR Hogan. Detection of outbreaks from time series data using wavelet transform. *Journal of the American Medical Informatics Association* (Supplement issue on the Proceedings of the Annual Fall Symposium of the American Medical Informatics Association), 2003.