Chapter 9

Saving DBMS Resources While Running Batch Cycles in Data Warehouses

Nayem Rahman
Enterprise Data Warehouse Engineering-ETL, Intel Corporation

ABSTRACT

In a large data warehouse, thousands of jobs run during each cycle in dozens of subject areas. Many of the data warehouse tables are quite large and they need to be refreshed at the right time, several times a day, to support strategic business decisions. To enable cycles to run more frequently and keep the data warehouse environment stable, the database system’s resource utilization must be optimal. This paper discusses refreshing data warehouses using a metadata model to make sure jobs under batch cycles run on an as-needed basis. The metadata model limits execution of the stored procedures in different analytical subject areas to source data changes in the source staging subject area tables, and then implements refreshes of analytical tables for which new data has arrived from the operational databases. The load is skipped if source data has not changed. Skipping unnecessary loads via this metadata driven approach enables significant database resources savings. The resource savings statistics based on an actual production data warehouse demonstrate an excellent reduction of computing resources consumption achieved by the proposed techniques.

INTRODUCTION

Data warehousing is a product of recent technological advances which fulfills the business needs of organizations (Wixom & Watson, 2001). It has appeared as a key platform to provide integrated management of decision support data in organizations (Shin, 2003). The data warehouse is used to hold historical and cross-functional data. Organizations use data warehouses as their integrated enterprise repository of data coming from disparate operational sources. As the business environment has become more global and competitive, the data warehouse has proved to be a very critical technology for an organization to better manage and leverage its information,
which in turn helps an organization to become more competitive, better understand its customers, and more rapidly meet market demands (Furlow, 2001; Wixom & Watson, 2001). Organizations use data warehouses for a variety of tasks such as planning, target marketing, decision making, data analysis, and customer services. They are changing the way business is conducted (Shin, 2003). Data warehousing continues to be very popular as many organizations are realizing its benefits (Furlow, 2001).

Until recently the data warehouses were usually refreshed after hours when business users went home. The business environment has become global, complex, and volatile and as a result nightly refresh is no longer practical. Business activities continue twenty four by seven as nighttime in one part of the globe is the day time of the other parts of the globe. The data warehouse users continue to look for up to date information more frequently. As a result we have to refresh data warehouses more frequently, every few hours. The good news is that through hardware advances such as massive parallel processing, and parallel database technology, it is now possible to load, maintain, and access databases of terabyte size (Wixom & Watson, 2001) in reasonable times. Thus data warehousing and other advances in information technology are now solving some of the very difficult technical problems and make it possible to organize, store, and retrieve huge volumes of information for a given decision (Cooper et al., 2000). In order to achieve this multiple facets needs to be considered. In addition to that data warehouse design, extract-transform-load (ETL) development, and load strategy need to be efficient. We need strategies as to how to save database management system (DBMS) resources during load processes in order to make the DBMS available to analytical tools and query processing while the load is running. All of these innovations are affecting how organizations conduct business, especially in sales and marketing, allowing companies to analyze the behavior of individual customers rather than demographic groups or product classes (Wixom & Watson, 2001).

In data warehouses the main users are the analytical community, namely, business people running reporting and analytical web tools. Data warehouse systems resources are designed for use by these tools, enabling business people to make all sorts of decisions based on data warehouse information. It is critical that enough computing resources be available for use by the analytical community to retrieve and process information into intuitive presentations (i.e., reads). In operational databases, the primary candidates to use computing resources are operational needs and requirements (i.e., writes). Any reporting and analytical tools get secondary considerations. However, in the case of data warehouses the analytical tools are primary candidates and get high priority in using computing resources. This means that the data warehouse batch processing should use the minimum resources possible. Data warehousing has evolved to hold huge volumes of historical as well as cross-functional data. Today, the knowledge workers such as business users, analysts and managers are more dependent on data warehouses for business information. These users’ information needs must be fulfilled on a priority basis by providing query results within a reasonable time in order for businesses to remain competitive.

Although an enterprise data warehouse has the potential to deliver competitive advantage, success is not necessarily guaranteed. The success of a data warehouse depends on data quality, data freshness, and accessibility to data by the analytical community with reasonable response times. In fact, a survey revealed that more than 60 percent of companies owning a data warehouse rated the system as having only limited success in meeting user expectations (Stedman, 1998). The survey by Watson and Wixom (2001) showed that the ability of many existing data warehouses to provide