Preliminary Project Proposal

CSci 8701: Overview of Database Research
University of Minnesota

Michael Goshey

Tuesday, October 3, 2006
Contents

PRELIMINARY PROPOSAL ...................................................................................................................................3

A Metadata-Driven Model for Benchmarking Data Warehouse Fitness .............................................................3

REFERENCES ............................................................................................................................................................4
A review of the database literature shows that the prevailing DBMS benchmarking schemes emphasize query execution time as a primary metric[1][2], a reasonable approach for a field dominated by OLTP systems required to swiftly record the state of large numbers of atomic transactions. In such a world it is meaningful to learn, for example, that under the same approximate conditions my database system processes more transactions than yours.

Data warehouses, however, are created with an altogether different purpose than OLTP systems: to facilitate decision-making by optimizing data for analytical and reporting operations[3][4][5]. This explains why data warehouses are commonly referred to as “decision support” systems [3][6]. While performance-based benchmarks have also been created specifically to evaluate decision support systems [2] and one concedes that faster is almost always considered better, the ultimate measure of success for a data warehouse is not its pure speed to respond to queries but rather the extent to which it actually supports decision making. In other words, the key benchmarking question for a decision support system should be framed as “How well does the warehouse support my decisions?” rather than (more generically) “How fast does it respond to my queries?”

To the extent that answering the first question above is a subjective exercise it may be unreasonable to assume that a set of metrics can be easily created to unambiguously benchmark the ability of a data warehouse to support decision making. However an interesting proxy for this ability might be to describe the relative fitness of the warehouse to those queries with which it is typically presented. If a particular data warehouse is well-suited to the particular set of queries it typically handles then it could be thought of as having a higher fitness score than one less well-suited to this purpose. This fitness measurement could be coupled with (and precede) performance-based benchmarking to more efficiently direct optimization work. Analogously, it would be inefficient to devote optimization effort to the mechanical system of a vehicle carrying a large, unnecessary payload. It would make more sense to decouple the payload and reduce the vehicle’s weight to its true need (improving its fitness) before concentrating on optimizing its performance.

A straightforward approach to establishing fitness metrics might be to utilize metadata about the queries submitted as well as metadata reflecting the utilization of data and other aspects of the warehouse.

This paper will pursue the articulation of a general purpose framework of metadata-driven metrics that evaluate the queries presented to the warehouse along with the utilization of the data and other features of the warehouse in order to derive an empirical score representing a data warehouse’s fitness benchmark.
References