Chapter XV

Data Warehouse Benchmarking with DWEB

Jérôme Darmont
University of Lyon (ERIC Lyon 2), France

ABSTRACT

Performance evaluation is a key issue for designers and users of Database Management Systems (DBMSs). Performance is generally assessed with software benchmarks that help, for example test architectural choices, compare different technologies, or tune a system. In the particular context of data warehousing and On-Line Analytical Processing (OLAP), although the Transaction Processing Performance Council (TPC) aims at issuing standard decision-support benchmarks, few benchmarks do actually exist. We present in this chapter the Data Warehouse Engineering Benchmark (DWEB), which allows generating various ad-hoc synthetic data warehouses and workloads. DWEB is fully parameterized to fulfill various data warehouse design needs. However, two levels of parameterization keep it relatively easy to tune. We also expand on our previous work on DWEB by presenting its new Extract, Transform, and Load (ETL) feature, as well as its new execution protocol. A Java implementation of DWEB is freely available online, which can be interfaced with most existing relational DBMSs. To the best of our knowledge, DWEB is the only easily available, up-to-date benchmark for data warehouses.

INTRODUCTION

Performance evaluation is a key issue for both designers and users of Database Management Systems (DBMSs). It helps designers select among alternate software architectures, performance optimization strategies, or validate or refute hypotheses regarding the actual behavior
of a system. Thus, performance evaluation is an essential component in the development process of efficient and well-designed database systems. Users may also employ performance evaluation, either to compare the efficiency of different technologies before selecting one, or to tune a system. In many fields including databases, performance is generally assessed with the help of software benchmarks. The main components in a benchmark are its database model and workload model (set of operations to execute on the database).

Evaluating data warehousing and decision-support technologies is a particularly intricate task. Though pertinent, general advice is available, notably on-line (Pendse, 2003; Greenfield, 2004a), more quantitative elements regarding sheer performance, including benchmarks, are few. In the late nineties, the OLAP (On-Line Analytical Process) APB-1 benchmark has been very popular. Henceforth, the Transaction Processing Performance Council (TPC) (1), a non-profit organization, defines standard benchmarks (including decision-support benchmarks) and publishes objective and verifiable performance evaluations to the industry.

Our own motivation for data warehouse benchmarking was initially to test the efficiency of performance optimization techniques (such as automatic index and materialized view selection techniques) we have been developing for several years. None of the existing data warehouse benchmarks suited our needs. APB-1’s schema is fixed, while we needed to test our performance optimization techniques on various data warehouse configurations. Furthermore, it is no longer supported and somewhat difficult to find. The TPC currently supports the TPC-H decision-support benchmark (TPC, 2006). However, its database schema is inherited from the older and obsolete benchmark TPC-D (TPC, 1998), which is not a dimensional schema such as the typical star schema and its derivatives that are used in data warehouses (Inmon, 2002; Kimball & Ross, 2002). Furthermore, TPC-H’s workload, though decision-oriented, does not include explicit OLAP queries either. This benchmark is implicitly considered obsolete by the TPC that has issued some draft specifications for its successor: TPC-DS (TPC, 2007). However, TPC-DS, which is very complex, especially at the ETL (Extract, Transform, and Load) and workload levels, has been under development since 2002 and is not completed yet.

Furthermore, although the TPC decision-support benchmarks are scalable according to Gray’s (1993) definition, their schema is also fixed. For instance, TPC-DS’ constellation schema cannot easily be simplified into a simple star schema. It must be used “as is”. Different ad-hoc configurations are not possible. Furthermore, there is only one parameter to define the database, the Scale Factor ($SF$), which sets up its size (from 1 to 100,000 GB). Users cannot control the size of dimensions and fact tables separately, for instance. Finally, users have no control on workload definition. The number of generated queries directly depends on $SF$.

Eventually, in a context where data warehouse architectures and decision-support workloads depend a lot on application domain, it is very important that designers who wish to evaluate the impact of architectural choices or optimization techniques on global performance can choose and/or compare among several configurations. The TPC benchmarks, which aim at standardized results and propose only one configuration of warehouse schema, are ill-adapted to this purpose. TPC-DS is indeed able to evaluate the performance of optimization techniques, but it cannot test their impact on various choices of data warehouse architectures. Generating particular data warehouse configurations (e.g., large-volume dimensions) or ad-hoc query workloads is not possible either, whereas it could be an interesting feature for a data warehouse benchmark.

For all these reasons, we decided to design a full data warehouse benchmark that would be able to model various configurations of database and