

The Need for Speed: Breaking the Data Aggregation Barrier

Ralph Kimball has argued that the single most dramatic way to affect performance in a large data warehouse is to provide a proper set of aggregates. In some cases, Kimball argues, this strategy can speed queries by a factor of 100 or even 1000.

Over the past ten years, businesses have made enormous investments in systems designed to transform operating and accounting data into business intelligence. They have created terabyte-sized data warehouses, developed powerful analytical programs to aggregate and query these huge data stores, and spent millions on multiprocessing servers powerful enough to process the resulting volumes.

As data volumes continue to grow, however, performance problems are proliferating. Even the most powerful database servers cannot always keep pace with the analytical demands imposed by scores of financial analysts, marketing researchers, strategic planners and customer relationship managers. As a result, many business organizations are literally choking on the terabytes of data they have collected. Complex queries can take hours and even days to aggregate and process. System performance slows to a crawl, and the complaints begin to roll in.

Poor Processing Performance: A Barrier to Business Intelligence

The impact of degraded processing performance can be subtle and far-reaching. Business intelligence is not really a property of an information system, but of the interaction between the data in the system and the intelligence, curiosity and expertise of analysts and decision-makers. To get the right answer, one must ask the right question. Unfortunately, the right question usually makes its appearance at the end of a long line of wrong, or at least partially wrong, questions. Good business intelligence, like good science, proceeds by a process of trial and error. The analyst must almost always work through repeated iterations to get to real insight.

Unfortunately, the opportunity for this type of iteration is a direct function of processing performance. The analyst or decision-maker who can initiate and receive timely responses to a dozen or more queries in one morning has a reasonable chance of unearthing something genuinely useful to the enterprise. The analyst/decision-maker who must wait overnight is fighting against much higher odds. This person must constantly cut corners, reduce sample populations and neglect promising alternatives along with the off-the-wall speculations that sometimes pay off in revolutionary understanding.

Performance degradation may also completely rule out some processing-intensive strategies such as data mining, where information "prospectors" run open-ended queries against large data sets hoping that something interesting will somehow pop out of the mix. Data mining techniques rely on the ability of database systems to run repeated comparisons of entire populations along every possible pair of coordinates in the database. Many of today's large database installations are simply not up to the task.

Despite the impact of diminishing performance, database administrators have had few alternatives for improving processing performance on very large databases. They could adopt a "bigger hammer" strategy, demanding large capital allocations for the latest server hardware or the newest versions of database management systems. Or they could adopt a less costly and more labor-intensive approach, attempting to optimize the performance of existing infrastructure by "tuning" such elements as database table structures or the sequences in which query instructions are executed. Both of these methods are short-term solutions at best, and because neither truly addresses the root of the problem, the next round of user demands and additional volumes of data are always right around the corner.

Data Aggregation: A New Alternative

The real solution lies in addressing performance issues at their source: **the data aggregation bottleneck**. For some years, Ralph Kimball has argued that the single most dramatic way to affect performance in a large data warehouse is to provide a proper set of aggregates. In some cases, Kimball argues, this strategy could speed queries by a factor of 100 or even 1000.

HyperRoll's patented data aggregation technology is making good on Kimball's projections. HyperRoll combines patented aggregation and caching algorithms with advanced load-balancing administrative facilities to deliver immediate and substantial improvements in database performance. HyperRoll provides lightning-fast access to large data volumes while drastically reducing the need for additional hardware or storage, and works seamlessly within customers' existing OLAP or relational systems to dramatically improve load times by as much as 100 times while enabling access and analysis of data in unparalleled detail.

HyperRoll resides on the existing database server and is accessible as an existing database table. Once installed, HyperRoll can be referenced by standard SQL queries and existing applications and OLAP tools, and is completely transparent to business intelligence users and their applications. HyperRoll replaces the summary tables that traditional relational database management systems must generate to aggregate data. HyperRoll emulates multidimensional as well as traditional relational summary tables and can be accessed by many different kinds of tools including object linking and embedding database (OLE-DB), open database connectivity (ODBC), and Java database connectivity (JDBC).

HyperRoll in Action

At one large telecommunications company, HyperRoll cut data aggregation time on a multi-million record database from nine-and-a-half hours to twelve minutes, a performance improvement of 4,700 percent. HyperRoll also yielded a five-fold reduction in the time required to generate reports once data was aggregated.

In another benchmark test, an industrial firm maintained a database with 250,000 records organized along six multi-hierarchical dimensions, with as many as twelve hierarchies in each dimension. Full data aggregation on this complex database used to take sixteen hours; installation of HyperRoll reduced aggregation time to four minutes. In another setting, HyperRoll reduced on-demand aggregation time for a five million-record, ten-dimensional database to two minutes, and yielded average random query response times of five seconds.

The remarkable performance improvements produced by HyperRoll translate directly into significant improvements in business intelligence functions, with no added demands on end users. HyperRoll can be integrated into an existing database infrastructure quickly and easily, and does not require extensive installation downtime or application re-tooling. Analysts and decision-makers can continue to use the same tools, applications and routines they have been using – only faster, better, and more effectively.