The Evolution of the Massively Parallel Processing Database in Support of Visual Analytics

Ian A. Willson, The Boeing Company, USA

ABSTRACT

This article explores the evolution of the Massively Parallel Processing (MPP) database, focusing on trends of particular relevance to analytics. The dramatic shift of database vendors and leading companies to utilize MPP databases and deploy an Enterprise Data Warehouse (EDW) is presented. The inherent benefits of fresher data, storage efficiency, and most importantly accessibility to analytics are explored. Published industry and vendor metrics are examined that demonstrate substantial and growing cost efficiencies from utilizing MPP databases. The author concludes by reviewing trends toward parallelizing decision support workload into the database, ranging from within database transformations to new statistical and spatial analytic capabilities provided by parallelizing these algorithms to execute directly within the MPP database. These new capabilities present an opportunity for timely and powerful enterprise analytics, providing a substantial competitive advantage to those companies able to leverage this technology to turn data into actionable information, gain valuable new insights, and automate operational decision making.

Keywords: Decision Support Workload, Enterprise Analytics, Enterprise Data Warehouse (EDW), Massively Parallel Processing (MPP) Database, Visual Analytics

Analytics, including the growing field of visual analytics, has an opportunity to significantly expand its effectiveness by leveraging advances in Massively Parallel Processing (MPP) relational databases. Despite a 25-year commercial history, it has been only recently that capabilities have emerged that analytical applications can utilize in effective and economic ways, particularly based upon the Enterprise Data Warehouse (EDW) architecture. Many companies face growing decision support system expenditures, shrinking budgets, increasingly volatile and complex operating environments and yet are unable to integrate and analyze their data to make timely operational, tactical and strategic decisions. The focus of this article is on advances in MPP database technology, representing the decision support system component within the broader Systems of Systems architecture that spans data and information throughout the enterprise. Much of the history of MPP, relational database technology and scientific computing applications of MPP will not be considered, some of which may be better known to readers than MPP database advances. Our focus is on advances in

DOI: 10.4018/irmj.2011100101
data warehouse technology that represent an historic opportunity for those in the analytics field, who until recently have relied on a variety of purpose-built stand-alone tools or methods of limited capability that increasingly compete for scarce Information Technology (IT) resources. To achieve the full potential of analytics on large or complex data, an examination of recent MPP database trends is in order.

This article presents a brief history of MPP databases and their role in decision support, with an unavoidable focus on Teradata Corporation, which commercialized MPP and remains a market leader (Feinberg & Beyer, 2008). Significant downward trends in price per unit of performance, both absolute and relative to transactional non-MPP databases are examined. The synergy of MPP technology and the activated Enterprise Data Warehouse (EDW) make it possible to support wide-ranging analytical queries on a single instance of time-referenced data. The EDW provides an authoritative source of reference ideally suited to analytics, particularly the emerging trend of ‘analytics inside’, or pushing down parallelized algorithms based on temporal, spatial and statistical primitives executed directly within the MPP database, greatly increasing the accessibility and potential value of these analytical methods. The result is a cost effective delivery system providing powerful analytics to the enterprise, generating business value from IT investments.

THE MASSIVELY PARALLEL PROCESSING DATABASE

With a focus on analytics, most of us have not been concerned with the precise details of the systems that host our data. To start off with, let us define our domain of interest to be broad enterprise-scale relational databases focused on supporting analytics and queries of all types, rather than transaction processing. The focus of this article is on performing a wide variety of valued analytics against such databases using an architecture specifically created to process this workload. MPP technology has been successfully deployed in large database systems for more than 25 years, supporting, with varying degrees of automation, many of the types of analytics we perform today. The author has been working with very large relational databases for decision support and analytical applications since 1986. In an early example, an entire IBM® 4081 mainframe was used to host a Human Resources decision support system. Since that time, what constitutes large has changed, as has the cost per unit of storage or unit of power, but surprisingly the fundamental method of storing and retrieving structured data has changed little.

The storage and retrieval of normalized data in tables first contemplated by Dr. Ted Codd (1970) is increasingly practical now that there is a suitably powerful hardware platform and database architecture available. The various design mitigations previously required for performance are fading in prevalence. The one fundamental change since 1970 is the commercial introduction of the shared-nothing or Massively Parallel Processing (MPP) database by Teradata Corporation, which shipped its first production system in 1984. One important distinction in this application of MPP to relational databases is that the parallelization is being applied to Structured Query Language (SQL) operators, a fundamentally functional language. Better known applications of MPP, at least 20 years ago, were in the scientific computing realm, where massive parallelization was thought to offer the opportunity to speed up calculations for a wide range of high value scientific computing problems.

Thinking Machines Corporation was one of the most widely-known parallel supercomputer manufacturers of the 1980s and early 1990s until its bankruptcy in 1994 (Taubes, 1995). In order to execute scientific calculations in parallel, special language extensions were required to parallelize computing on these MPP systems, such as C* and CM Fortran. Ultimately this approach was not cost-effective, with more economical designs for moderate parallelism utilizing specialized clusters of SMP-based computers succeeding in the commercial mar-
Related Content

Measuring Information Propagation and Processing in Biological Systems
[www.igi-global.com/chapter/measuring-information-propagation-processing-biological/27796?camid=4v1a](www.igi-global.com/chapter/measuring-information-propagation-processing-biological/27796?camid=4v1a)

An Analysis of Student Persistence in Online Education
[www.igi-global.com/chapter/analysis-student-persistence-online-education/22729?camid=4v1a](www.igi-global.com/chapter/analysis-student-persistence-online-education/22729?camid=4v1a)

On-Line Analytical Processing at Washtenaw Mortgage
[www.igi-global.com/article/line-analytical-processing-washtenaw-mortgage/44616?camid=4v1a](www.igi-global.com/article/line-analytical-processing-washtenaw-mortgage/44616?camid=4v1a)
Knowledge Flow Identification

[www.igi-global.com/chapter/knowledge-flow-identification/13908?camid=4v1a](http://www.igi-global.com/chapter/knowledge-flow-identification/13908?camid=4v1a)