Invited Talk:

Distributed and Parallel Computing Issues in Data Warehousing*

Hector Garcia-Molina, Wilburt J. Labio, Janet L. Wiener, Yue Zhuge
Stanford University
{hector,wilburt,wiener,zhuge}@cs.stanford.edu
http://www-db.stanford.edu/warehousing/warehouse.html

A data warehouse is a repository of data that has been extracted and integrated from heterogeneous and autonomous distributed sources. For example, a grocery store chain might integrate data from its inventory database, sales databases from different stores, and its marketing department's promotions records. Warehouse applications differ from traditional database applications in several key features. The quantity of data is often much larger, between 100 Gb and multiple Tb, since warehouses combine and archive data from multiple data stores. Second, the warehouse must solve new distributed consistency problems, since the sources are autonomous and previous consistency algorithms rely on cooperation between sources. Third, the integration software is distinct from both the sources and the warehouse. It can be both distributed and parallelized to improve performance. In addition, it requires new resumption from failure algorithms, since integration may take hours and traditional algorithms would start over. Fourth, portions of the warehouse are often replicated as local data marts; data mart maintenance also requires distributed algorithms. In this talk we overview our work on warehouse creation and maintenance, highlighting the distributed and parallel aspects of the problem and in our solutions.

The paper corresponding to this invited talk is available from the above URL.

*This research was funded by Rome Laboratories under Air Force Contract F30602-94-C-0237, by the Massive Digital Data Systems (MDDS) Program sponsored by the Advanced Research and Development Committee of the Community Management Staff, and by Sagent Technologies, Inc.