The Ins and Outs (and Everything in Between) of Data Warehousing

Phil Fernandez
Red Brick Systems
philf@redbrick.com

Donovan Schneider
Red Brick Systems
donovan@redbrick.com

Tutorial Description

Data warehousing is the latest "hot topic" in the industry. With market projections of $8 billion by the year 2000, vendors of all flavors are claiming the suitability and superiority of their products for this market segment. This has led to a great deal of confusion, with terms such as OLAP, ROLAP, MDDB, decision support systems (DSS) and data warehousing being defined, re-defined, and sometimes even used interchangeably.

Although data warehousing is currently in the spotlight, it is not a new activity. Rather, companies have been successfully deploying data warehouses to enable strategic decision making for over 10 years.

This tutorial seeks to provide clarity to the current confusion over data warehousing by focussing on the core aspects of a complete data warehousing solution: architectures, data modeling, loading the warehouse, client analysis/queries, and storage and query processing to support client analysis. These are briefly summarized below.

Warehouse Architectures and Applications

centralized/distributed, data marts, "closed-loop"

Data Modeling and Schema Design

normalization, star schemas, differences from OLTP systems, metadata management

Data Extraction, Preparation and Loading

source data handling/conversion, integration, cleansing, referential integrity, consolidation/aggregation, load "windows", load recovery

SIGMOD '96 6/96 Montreal, Canada

Data Storage within the Warehouse

partitioning (time based, others), storage methods/ requirements, hierarchical storage, compression, backup/restore, etc.

Data Access with the Warehouse

indexing, dimensional queries, join methods, parallelism, query optimization, large memories, read-mostly workload, etc.

Client Analysis

extracting data from warehouse to the client, client queries/analysis (dimensional analysis), "drill down", etc.

Similarities and differences with OLTP systems will be emphasized in the tutorial, as will examples from customers. Research opportunities will be highlighted.

Instructors

Phil Fernandez is the Vice-President of Research and Development at Red Brick System, Inc., in Los Gatos, CA. From March 1989 to Nov. 1991, he was the Director of Systems Software Development at Metaphor Computer Systems. Prior to that time, he was employed by Stanford University from March 1986 to March 1989, most recently as Director of Data and Technology Resources. Mr. Fernandez holds a B.A. in History from Stanford University. He has over seven years of experience in the data warehousing industry.

Donovan Schneider is a Senior Software Engineer at Red Brick Systems, Inc. where he is working on their next generation data warehouse server. From Sept. 1990 to Nov. 1994, he was a Member of Technical Staff at Hewlett-Packard Laboratories in Palo Alto, CA. Dr. Schneider received a Ph.D. in Computer Science from the University of Wisconsin, Madison in 1990. He has published over 20 research papers, served on numerous program committees, and was the editor for SIGMOD '95. He received the Best Paper Award at SIGMOD in 1990.