

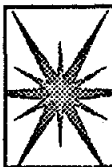
Star Schema Trek:

“Where No One Has Gone Before”



“Why Be Normal?”--Kretzman '91

- Normalized DB Optimized for OLTP
- Entity/Relationship Data Model
- Database split into a large number of tables; each with few columns
- Each element stored once
- Advantages:
 - 1) minimum space,
 - 2) update accuracy, and
 - 3) fast retrieval of single rec.



So What's the Problem? Complexity!!

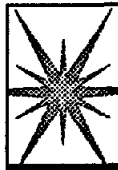
- Too complex for users to understand. The large number of joins for analytical work generate errors and varying results
- Software can't handle the load generated with reasonable performance.



To Say It Again!!!

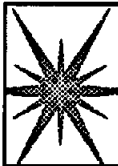
“Entity relation data models are a disaster for querying because they cannot be understood by users and they cannot be navigated usefully by DBMS software. Entity relation models cannot be used as the basis for enterprise data warehouses.”

-Ralph Kinball, The Data Warehouse Toolkit, p. 9



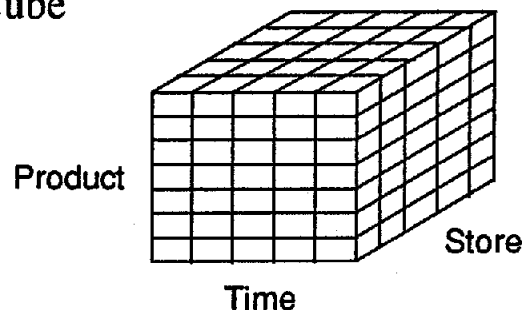
“Why Be Normal?”--Why Indeed!!!

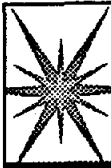
- Need a DB Optimized for OLAP
- Dimensional Data Model
- Data is kept together the way it is used
- Not de-normalized for performance.
- Need a general rule or theory to follow in Query Optimized Database Design!



The Query Optimized Database

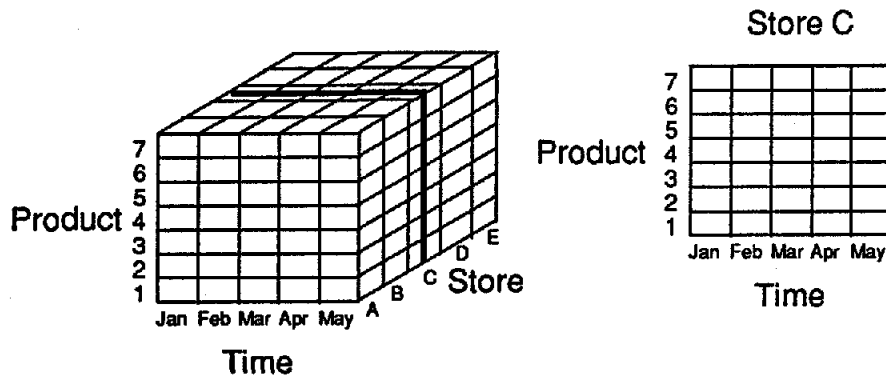
- Star Join Schema
- Dimensional Model
- Data Cube





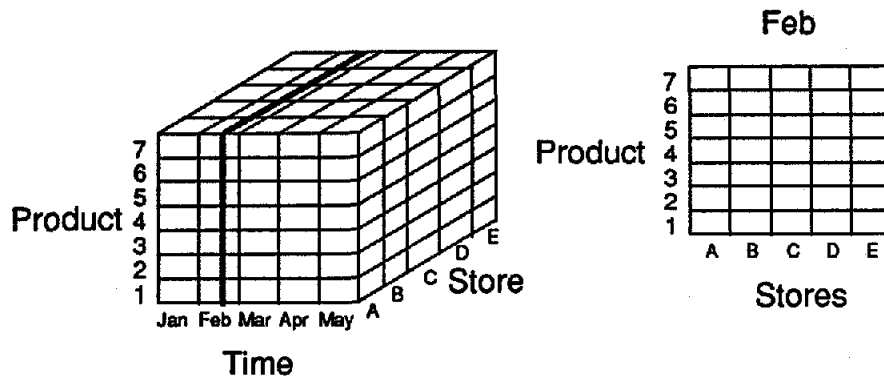
Three Basic Data Cube Reports

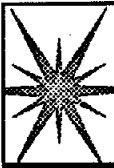
Time Series Report of Products



Three Basic Data Cube Reports

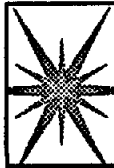
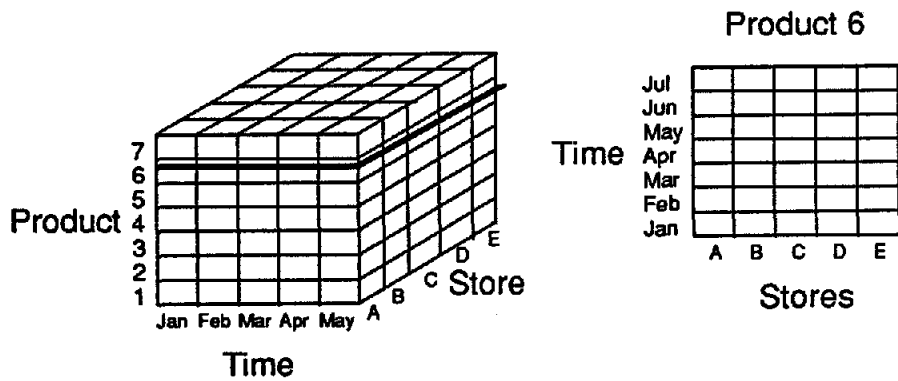
Cross Section Report of Products and Stores





Three Basic Data Cube Reports

Cross Section of Time Series Report of Stores



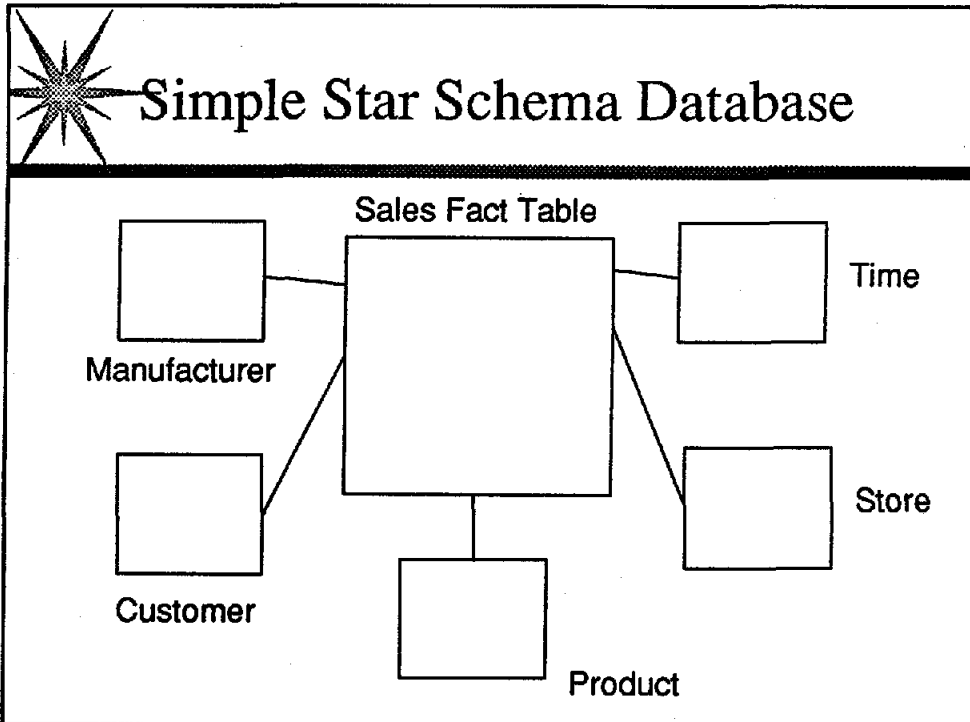
Dimensional Hierarchies

Product
Brand
Category

Store
District
State
Region

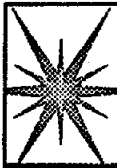
Day
Week
Month
Quarter
Year

Customer
Age
Gender



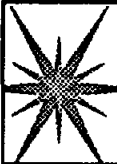
Fact Table: Measurements

- The Fact Table is where the numerical measurements are stored. It contains the cells of the data cube.
- Fact Table Issues:
 - Granularity
 - Continuous values
 - Additive measures
 - Sparsely



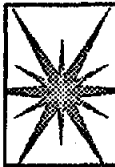
Dimension Tables: Attributes

- The Dimension Tables at the points of the Star Schema contain hierarchical attribute information about the items in the fact table. The 'People, Places, Things' important to our analysis and shown at the corners of the data cube.
- Features:
 - Text Items
 - Discrete
 - Used as Query Constraints



Standard SQL Query

```
SELECT      c.Customer_ID, Customer_Name, Address, sum(Quantity_Sold)
FROM        Sales s, Time t, Product p, Customer c
WHERE       Month = 3
AND         Year = 1996
AND         UPC_Code = 12678754390
AND         s.Time_ID = t.Time_ID
AND         s.Customer_ID = c.Customer_ID
AND         s.Product_ID = p.Product_ID
GROUP BY    c.Customer_ID
ORDER BY    c.Customer_ID
```



OLAP Concepts and Issues

- Codd Article coined the term in 1993
- FASMI (Nigel Pendse and Richard Creeth)

Fast: 1-20 Second Response, 5Sec Average

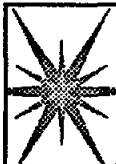
Access: Can cope with any business logic statistical analysis; keep it easy

Shared: Secure multiple concurrent access

Multidimensional: Conceptional view to the user, includes hierarchies

Information: all of the data and derived information needed, wherever and however much is needed by the application

- Architecture: MDDDB vs. ROLAP vs. ???



More OLAP Issues

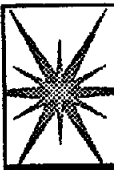
- Join Query Optimization
- Indexing Techniques
 - Join Indexes
 - Bit-Mapped Indexes:

Gender: m f f m f m ?

M: '1 0 0 1 0 1 0'h

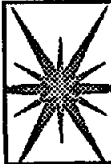
F: '0 1 1 0 1 0 0'h

M or F: '1 1 1 1 1 1 0'h



Even More Issues

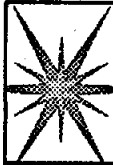
- Aggregation, Compute/Store or As Called?
- Computational Efficiency
- Loading/Managing--Administration Tools
- Parallelization of Hardware



So Why Is This Important to Me?

- SAS MDDB Provided as Part of Ver. 6.12
How will we organize our data to make use of it?
- We SAS Software users have been in the query and analysis world for years.
- It is better to follow proven design principles than to have everyone inventing something ad hoc.

Let's do it right!!!



References

- The Data Warehouse Toolkit, Ralph Kimbal, John Wiley & Sons, Inc, 1996
- "Stars: A Pattern Language for Query-Optimized Schemas", Stephen Peterson, in Pattern Languages of Program Design, James O. Coplien and Douglas C. Schmidt, 1995
- "Maximizeing Your Warehouse", Neil Raden, Information Week, March 18, 1996, pp. 42-48
- "Boost Warehouse Performance", Neil Raden, Information Week, March 25, 1996, pp. 77-89
- "Providing OLAP(On-line Analytical Processing) to User-Analysts: An IT Mandate", E. F. Codd, S. B. Codd, and C. T. Salley, Arbor Software, 32 pp.
- "Understanding the Need for On-Line Analytical Servers", Richard Finkelstein, Arbor Software, 10 pp.
- "Multi-Table Joins Through Bitmapped Join Indices", Patrick O'Neil, and Goetz Graefe, SIGMOD Record, September 1995, pp. 8-11
- "The Truth About OLAP", Maurice Frank, DBMS Magazine, August 1995, pp. 40-48

- "Managing Multidimensional Data: Harnessing the Power", Jay-Louise Weldon, Database Programming & Design, August 1995, pp. 24-33
- "Modeling the Data Warehouse", Neil Raden, Information Week, January 29, 1996.
- Several pages on the WWW have interesting information. Search on OLAP
- The USENET newsgroup comp.databases.olap often contains useful information.