Data Warehousing

What You Should Know Before You Begin

Bill Langston, Business Intelligence Specialist, langston@ngsi.com

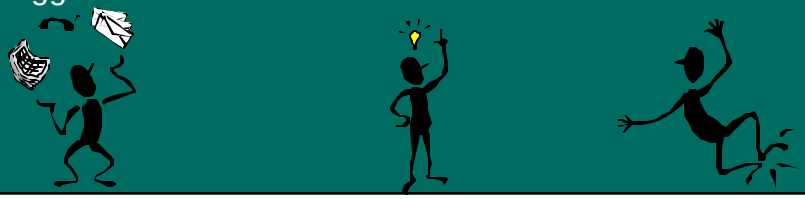
NGS™ NewGeneration Software, Inc.

Member of IBM. PartnerWorld for Developers
www.developer.ibm.com

(800) 824-1220 - World Wide Web: www.ngsi.com

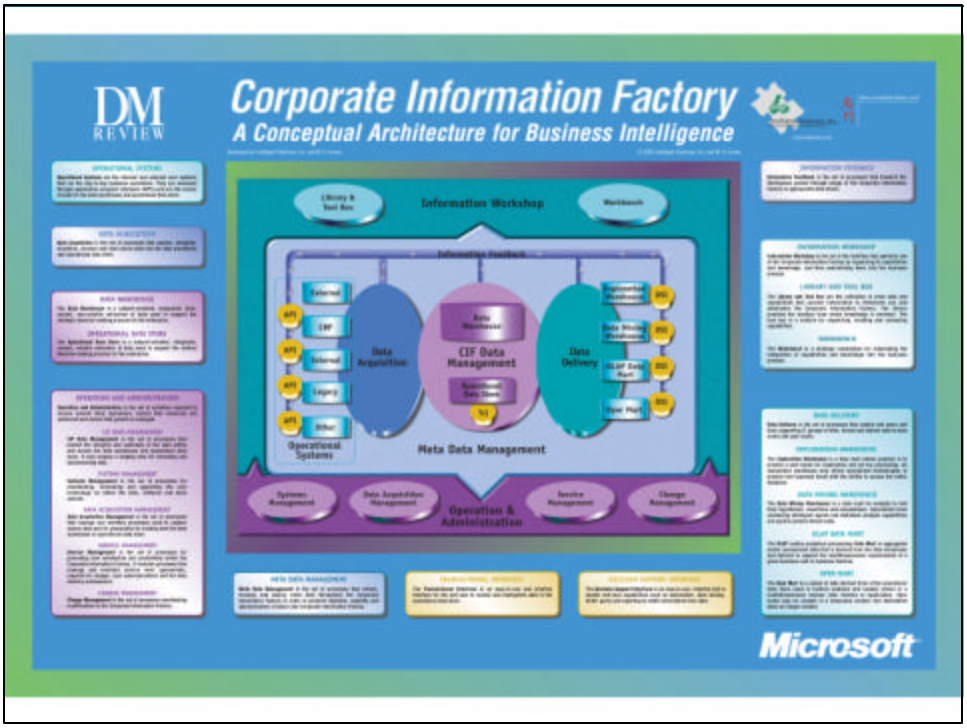
What is Data Warehousing?

- methodology designed to help you leverage the potentially strategic value of the information in your corporate data bases
- extension of traditional query, report writing, decision support and executive information systems
- software designed to simplify and improve end-user data base access
- a great way for IBM to convince your company to buy a bigger iSeries 400



Data Warehousing Conceptual Background

- extract, reorganize, rename and index your data so users can readily find the information they want
- isolate the query, reporting and analytical processing workload
- provide powerful new software that helps users look at data multi-dimensionally
- support sophisticated statistical and mathematical processing to find hidden relationships



Data Warehousing Vocabulary

- **Data Warehouse:** Enterprise-level data repository of cleansed, transformed relational data.
- **Data Mart:** Topic specific data base of either relational or OLAP/MDD architecture.
- **Three-Tier Data Warehouse:** Architecture encompassing the OLTP data base which is used as a source of new data for the relational data warehouse which is in turn used as the source for multiple topic-specific data marts.
- **OLTP Database:** The production or "on-line transaction processing database."
- **Star Schema Database:** Relational data base architecture including a central Fact table and associated Dimension tables designed specifically to support multi-dimensional analysis.

Data Warehousing Vocabulary

- **Multi-Dimensional Database:** A specialized data base where the data base designer creates multi-dimensional models where pre-calculated summary values are stored to support rapid query response time.
- **On-Line Analytical Processing (OLAP):** The ability to organize and look at data from a multi-dimensional perspective. OLAP views of data may be achieved through the database model or at the time of request in different vendor solutions.
- **Sparse Matrix:** Another way of describing a multi-dimensional data base. The cube or matrix is described as sparse since it does not have a value in every cell.
- **ROLAP:** Business intelligence model where multi-dimensional views of data are built on-demand from relational tables rather than from an OLAP data base.
- **Operational Data Store:** A cross-application relational data base holding current transactional data to support production reporting better than the OLTP data base. May serve as the feeder to the data warehouse.

Data Warehousing Vocabulary

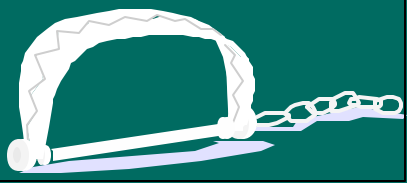
- **Data Transformation:** The process of de-normalizing an OLTP data base to simplify end-user comprehension and access.
- **Data Cleansing:** The process of modifying data for consistency and accuracy.
- **Data Propagation/Replication:** Data propagation is the movement of data from a source database to a target database, usually with some level of transformation and aggregation. Data replication is the real-time copying or "mirroring" of a database usually to support high availability or to move production reporting and other read-only workload out of the production environment.
- **Front-End / Back-End Tools:** In the context of software tools, front-end refers to tools operated by end-users while back-end refers to those used by the IT staff. Front-end tools are also sometimes called visualization tools.
- **ETL Tool:** "Extraction, Transformation and Loading" application designed to support database design and maintenance.

Data Warehousing Vocabulary

- **Surrogate Key:** Unique "key" value generated by data warehouse/data mart designer to enable the joining of records in one file to records in another when no such unique value exists in the source database.
- **Meta-Data:** Descriptive information (documentation) associated with individual data base elements to assist developers and end-user's in locating, interpreting, recreating and using data quickly and correctly.
- **Data Mining:** Process of searching through a data base to discover statistically significant relationships between variables. Queries and OLAP reports are not data mining. Data mining is a much more complex form of data analysis using statistical and mathematical algorithms to find associations, sequential patterns, classifications, clustering, time series forecasts and data mapping models.

Data Warehousing Challenges

- Data Quality
- Data Availability
- Incorrectly Defining Data Model and Granularity
- Management and End-User Buy-in
- Focusing Too Much on the OLAP Tools
- Lack of IT Staff Involvement
- Letting Project Get Too Large, Too Fast



Data Cleansing Challenges

- Underestimating CPU Memory and Disk Requirements
- Do you have the data you need?
- Is the data usable?
- Will anyone take responsibility for data quality?
- Can anyone correctly decipher the historical data?
- Does management understand this task will never end?

Insurance Data Modeling Example

Possible Levels of Granularity for Analyzing Earned and Unearned Premiums by Period

Policy #
Product
Location
Annual Statement Line of Business
Subline
Class
Coverage Code

Choosing coverage code instead of class as our lowest level of granularity increases data mart size by 13X!

Sizing a Shipping Fact Table in a Star Schema Model

- **Assumptions:**
  - 3 years of history
  - 25,000 invoices per month
  - 15 line items per invoice
  - 1 “deal” per line item on each invoice
  - 1 ship from instruction per line item, per invoice
  - 1 shipping mode instruction per line item, per invoice
  - = 13,500,000 records in the fact table
  - Assume 19 fact fields, 8 key fields
  - Fact Table Size:
  - 13.5 million records X 27 fields X 5 bytes = 1.82 GB

Architectural Considerations

- Scale - enterprise or departmental solution?
- Where does the OLTP data originate?
- How much data will you need to consolidate?
- How current must the data warehouse be?
- How dynamic is your company, industry?
- What are your multi-dimensional modeling requirements?
- How much load time can you afford?
- What is the architecture of the solution that best satisfies your end-users needs?

What is an OLAP/MDD?

Example of a Relational Data Base File

Product	Region	State	Units Sold
Nuts	East	NY	50
Nuts	West	CA	60
Nuts	Central	IL	100
Screws	East	MA	40
Screws	West	NV	70
Screws	Central	WI	80
Bolts	East	RI	90
Bolts	West	OR	120
Bolts	Central	MN	140
Washers	East	VT	20
Washers	West	WA	10
Washers	Central	KS	30

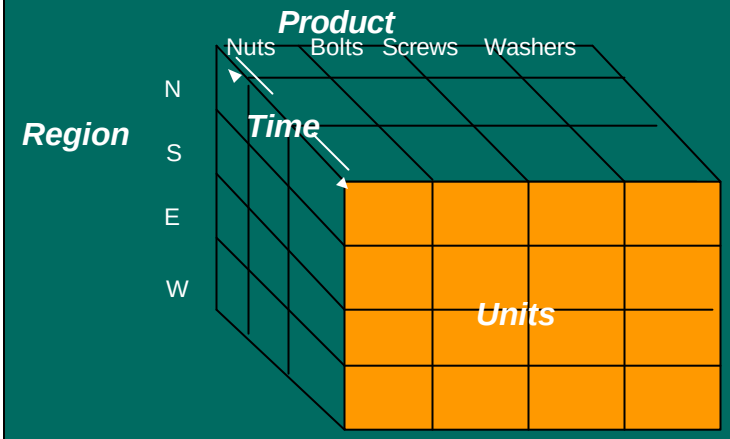
What is an OLAP/MDD?

Example of Same Data Represented in a Two Dimensional Format

Product	East	West	Central
Nuts	50	60	100
Screws	40	70	80
Bolts	90	120	140
Washers	20	10	30

What is an OLAP/MDD?

- **OLAP View:** Units by Product and Region over Time



OLTP vs. Data Warehouse/Mart

- OLTP is designed for add, update, delete, single record retrieval.
- Data warehouse/mart is designed for retrieving and summarizing large volumes of records on the fly.
- OLTP is not structured the way managers think.
- Data warehouse/mart is designed for analytical users.
- OLTP data bases are process oriented, e.g., order entry
- Data marts are subject oriented, e.g., sales analysis
- OLTP storage is relational, accessible to off-the-shelf tools.
- Data marts may be stored in OLAP/MDD format.

Relational vs MDD OLAP

- OLAP/MDD's depend on extensive pre-aggregation of the data.
- OLAP/MDD's require architect to pre-define and load in the aggregations, dimensions, data elements, etc.
- OLAP/MDD's require longer load times during update.
- Relational data bases are more scalable, dynamic, flexible.
- OLAP/MDD models are usually front-ends to relational data repository
- Relational model still required for traditional reporting and single record access.
- OLAP/MDD vendors moving to hybrid OLAP models with less pre-calculation.

Do You Need an OLAP/MDD Server?

- Is there a solution that satisfies your needs that is built around an OLAP/MDD data base?
- How fast will your data marts grow?
- Is your goal strategic planning support or management reporting and report analysis?
- Do you need compatibility with legacy query, reporting or OLAP tools?
- Is your management willing to invest in the training and consulting?
- Do you have the IT staff to maintain the OLAP/MDD?

Do You Need a Separate Data Warehouse Server?

- **Yes** -
  - if your data is already on multiple CPU's
  - if upgrading your iSeries 400 would cost more than buying a server and data warehousing software
  - if query performance is already a problem
  - if the data warehouse is going to lead to a significant increase in batch workload
  - if you intend to do data mining

How Much Does It Cost?

Rough Pricing Guidelines for DW Software and Hardware

- Data Transformation, Cleansing Software: $2K to $150K
- Data Propagation Software: $5K to $50K
- Data Replication Software: $20K to $100K+
- iSeries 400 OLAP/MDD: $20K to $100K
- Query/Reporting Software: $1000 to $2000/user
- Client OLAP Software: $500 to $2000 per user
- Consulting, Training: $ to $$$$$$$$$
- iSeries Server: $20K+
- Vendor Packaged Solutions available for almost any budget.

Market Trends, Directions

- More Industry-specific or application-specific data models and analytical processes
- Integration of CRM into BI/Data Warehousing projects.
- Integration of Data Warehousing/BI/CRM into ERP systems.
- Web portals as end-user point of entry.

For More Info

- www.dw-institute.com - Data Warehousing Institute
  (sponsors educational seminars and publications)
- http://www-4.ibm.com/software/data/bi/ - IBM Business Intelligence home page
- http://www.dwinfocenter.org - Data Warehouse Information Center
- http://www.datawarehousingonline.com/- web portal to numerous DW/BI resources
- www.intelligententerprise.com - Intelligent Enterprise magazine
- www.dmreview.com - DM Review magazine's web site
- www.datawarehouse.com - multi-vendor sponsored site with DW/BI resources
- www.ngsi.com - New Generation Software's web site
- www.olapreport.com - reviews of OLAP software products (ANNUAL FEE)
- Authors to Look for: Ralph Kimball, W.H. "Bill" Inmon, Claudia Imhoff

Upcoming Seminars, Conferences

"Data – Your Path to Profits",
Sponsored by
New Generation Software, Inc.
Visit www.ngsi.com for a list of the upcoming events.

- **NGS Data Warehouse Development Training Class**
  Call 1-800-824-1220 for details
- **Data Warehousing Institute Seminars and Conferences**
  http://www.dw-institute.com
- **COMMON Spring 2003**
  Indianapolis, Indiana
  March 9-13, 2003
  www.common.org