INTERNATIONAL JOURNAL OF NETWORK MANAGEMENT, VOL. 8, 42–56 (1998)
Int. J. Network Mgmt., vol. 8, 42–56 (1998)

An Introduction to Data Warehousing: What Are the Implications for the Network?

Data warehousing is an information systems environment, rather than a product. It has emerged as an essential business entity for sophisticated analysis of data. This article presents a clear overview of the implications of data warehousing for business. © 1998 by John Wiley & Sons, Ltd.

*By Dr. Katherine Jones**

The need for analysing corporate data for trends and strategic information is hardly new. In an ideal world, the analyst could simply query the operational database in a corporation to ascertain the long-term effects of a policy or the fiscal implications of a marketing decision. In reality, few operations environments have the bandwidth—literally—to allow information seekers to create the elaborate queries necessary for sophisticated analysis of data. Hence, the development of data warehouses as an essential—but separate—business entity has emerged. The split of operational and informational databases occurred for several reasons:

- The data serving operational needs is physically different from that serving informational or analytic needs.
- The supporting technology for operational processing is fundamentally different from that for informational or analytic needs.
- The user community of operational data is different from those using informational or analytical data.
- The processing characteristics for the operational environment and the informational environment are fundamentally different.
- Data warehousing is an information systems environment, rather than a product. As an IS environment, it is based on the organization's information needs, competitive requirements and investment criteria. Therefore, it is a subset of the corporate enterprise strategy, along with integrated business applications, on-line transaction processing, decision support systems, and on-line analytical processing.
- Data warehousing is concerned with informational and analytic processing. Unlike operational data which deals with the day-to-day conduct of business, this information serves the needs of management in the decision-making process. Often called DSS (decision support systems) or OLAP (on-line analytical processing), this analytical processing looks across broad vistas of data to detect trends. Instead of looking at one or two

**Correspondence to: Dr Katherine Jones, 2 Alden Street, Newton Center, MA 02159, USA.*
Email: Katherine_Jones@DGC.ceo.dg.com

 CCC 1055–7148/97/010042–15$17.50

records of data as is the case for operational processing, analytical processing deals with many records, collected over time.

A large-scale, enterprise data warehouse stores an organization's business data in a single, integrated relational database. A distributed data warehouse strategy may deploy a combination of relational and multidimensional databases:

- A data warehouse is separate from the operational data of an organization, though its data derives from the operational data.
- It provides a historical perspective on information.
- It represents a blend of technologies (database, metadata management, data loading, etc.).
- It is the proper foundation for decision support systems (DSS)
- It is not a gateway between existing databases
- It is not a master index for existing data
- It is not a replication of the operational databases
- Its sole purpose is support of decision support applications and business querying
- Data warehouse components include:
  - –The data store for the warehouse—the main consolidator system for decision support data
  - –Extraction programs or other software for loading data programs for managing metadata (information about the data stored in the warehouse)
  - –Optionally, a data mart or the departmental DSS—which has a section (subset) of the data in the warehouse
  - –Various decision support applications, query and reporting tools, desktop applications (spreadsheets, databases, and analysis tools).

What are some of the differences between operational and analytical databases?

Operations	*Decision support systems*
Constants updates	Updates are rare, if at all
Deals with few records	Deals with many records
Rapid response times (seconds)	Slower response time required (30 minutes–24 hours)
Many users	Fewer users—analysts, managers
Large network	Smaller network

Data warehousing implies a three-tier architecture because processing takes place at three levels: PC clients, the data warehouse server, and the production systems (see Figure 1).

The Data Warehouse Environment

The data warehouse is the foundation for decision support systems (DSS) and analytical processing. Because there is a single integrated source of data in the data warehouse, and because the stable data are accessible, the job of the DSS analyst in the data warehouse environment is more straightforward than in the classical operations environment. At the conceptual level, the data warehouse is in between the production data and the information displayed by front-end tools (Figure 2).

What is a Data Warehouse?

The data in a data warehouse is made up of snapshots of an enterprise's multiple operational databases. It consists of hardware and software optimized for executive information systems (EIS) and decision support systems and is integrated to support on-line analytical processing, rather than the on-line transaction processing which characterizes the operations world.

—A Data Warehouse is More than a Database!—

The data warehouse is inseparable from the complex tools and programs which load it with

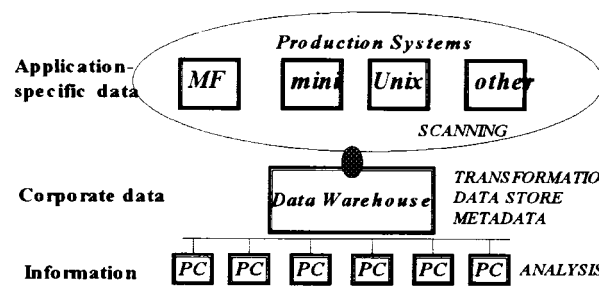

Figure 1. Data warehousing as a three-tier architecture.

Data warehousing is an information systems environment, rather than a product.

data and synchronize and synthesize views of the multiple operational databases. It consists of:

(1) Loading programs
- Transforming data extracted from production systems
- Combining mapping and the data
- Inserting the data

(2) Metadata and its management
- The structure of the data
- The algorithms used for transformation
- The mapping of the operational data structures to the warehouse data structures

(3) Current and detailed aged data

(4) Lightly to highly summarized historical data

A data warehouse has been defined as a collection of data in support of management decisions which is:

- Subject oriented
- Integrated
- Nonvolatile
- Time variant

Definitions are as follows:

- *Subject oriented* Operational systems are organized around the applications that support the business functions of a company. 'Subject' refers to the queryable category of interest to the researcher. Data on a subject will likely come from multiple operational systems.
  Example: In an insurance company, applications are auto, health, life, and casualty. The subject areas for query might be customer,

Going from Data to Information

—Operational Data (Primitive Data)—

- Supports day-to-day operations
- Detailed information
- Transaction driven
- Application oriented (financials, order entry, payroll)
- High availability a clear requirement
- Accurate, as of the moment of access
- Serves the clerical community
- Can be updated
- Can be run repetitively
- Requirements for processing are understood *a priori*
- Compatible with the system development life cycle
- Performance sensitive
- Accessed a unit at a time
- Control of update is major concern in terms of ownership
- Managed in its entirety
- Nonredundancy (to ensure data integrity)
- Static structure; variable contents
- Amount of data used in a process is small
- Probability of access is high

—Decision Support Data (Derived Data)—

- Supports executive decision making
- Summarized information
- Analysis driven
- Analysis oriented
- High availability not always required
- Represents values over time
- Serves the managerial community
- Is not updated
- Is run heuristically
- Requirements for processing are **not** understood *a priori*
- Completely different life cycle
- Performance relaxed
- Accesses a set of information at a time
- Control of update is not an issue
- Manager by subsets
- Redundancy (to increase performance)
- Flexible structure
- Amount of data used in a process is large
- Modest to low probability of access

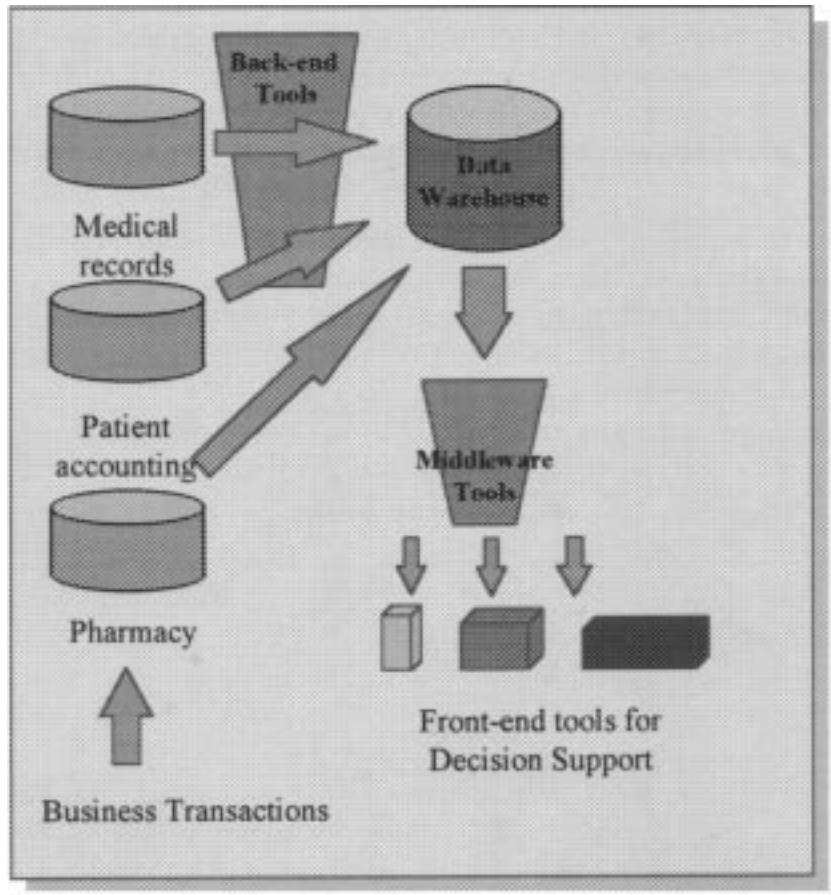

Figure 2.

The data in a data warehouse is made up of snapshots of an enterprise's multiple operational databases.

policy, premium, and claim.
Typical subjects are:
Customer
Product
Transaction or activity
Policy
Claim
Account

- *Integrated* Data from all relevant database-based applications is amalgamated so it can be accessed by the DSS professional. This is the most important aspect of a data warehouse. There is no point in bringing data over from the operational environment into the data warehouse **without** integrating it; then it could not be used to support a corporate view of data, which is the whole point of the decision support system in the first place.
  Example: Databased data from the auto, health, life, and casualty applications is all consolidated so all information is available to answer the questions posed by the analyst.
- *Nonvolatile* Operational data is volatile data. There is record-by-record manipulation of that data—it may be edited, deleted, and updated as well as accessed. In the data warehouse, there is a mass load of the information, and access is the only function performed on the data.
  Example: In the operations environment, a new policy is drawn up for a second car. The buyer's policy is updated. In the query

environment, the DSS professional looks for all the policies in New England where the second car is insured for more than $50,000.

- *Time variant*: Operational data usually has a 60–90 day time horizon. Update of records is the primary task at hand. The key structure may or may not contain an element of time. In the analytical data warehousing environment, the time horizon may be 5–10 years. It is a sophisticated snapshot of data at a point in time—which can be vital for future decision making.

Example: payments received are noted on a daily basis. The DSS analyst, however, doesn't care who paid on what day; he or she may be more interested in the general late payment trends during December and January over the past five years.

—Goals of Data Warehousing—

From the customer's point of view, his or her goals for data warehousing include:

- Getting information for operational management
- Analyzing information for strategic decision-making and planning
- Garnering information that would be impractical or nearly impossible otherwise
- Achieving a faster path to the information needed for competitive advantage
- Leveraging their company's investments in information technology for better business

—Metadata—

Metadata in a data warehouse plays the role of a card catalog in a library. It is metadata that allows an organization to track and understand where its data is. When an organization has an effective metadata structure, the DSS analysts can effectively find and analyze data. Product support of metadata should be in the same DBMS as the data warehouse. Functions of metadata include:

- Identification of what the contents of the data warehouse are
- Identification of the legacy source of the data in the warehouse
- Specification of the integration and transformation of the logic connecting the data warehouse to the legacy systems environment (when applicable)
- 'Versioning' of the changes to the metadata so that changes to metadata are trapped
- Identification of alias information
- Support of both computer and business terms
- Timing of refreshment so that the end user can determine when any unit of data was last refreshed
- Metrics, so the end user can know prior to the submission of a request whether the request will be consuming an inordinate amount of resources.

Operational Development versus the Data Warehousing Development Model

Bill Inmon states the following differences between the methodologies for a production system and a data warehouse:

Operations development model
(1) Requirements gathering
(2) Analysis
(3) Design
(4) Programming
(5) Testing
(6) Integration
(7) Implementation

Data warehouse development model
(1) Implement warehouse
(2) Integrate data
(3) Test for bias
(4) Program against data
(5) Design DSS system
(6) Analyze results
(7) Understand requirements

Data warehouses are not built all at once. Rather, they are designed and populated a step at a time, and as such are evolutionary, not revolutionary. The costs of building a data warehouse all at once, the resources required, and the possible disruption to the environment all dictate that the data warehouse be built in an orderly, iterative, one-step-at-a-time way.

Cautionary note: Because the data warehouse is a copy of operational data, your customer may think that traditional operational concerns do not apply to it. Not true! The following common concerns still pertain:

- Overall system reliability
- Availability
- System manageability
- Maintenance

The data warehouse is a unique and complete data source in the enterprise. The consolidation and integration which occurs during the data engineering process creates unique data elements, and scrubs and rationalizes those data elements found elsewhere in the enterprise's databases. Hence it is a very valuable resource in the enterprise. Most operational evaluation criteria you would apply to an on-line transaction processing (OLTP) system apply equally to the data warehouse. As organizations rely more upon their data warehouses, high availability becomes more of a requirement.

On-line Analytical Processing

—OLAP: What Is It?—

On-line analytical processing is a data analysis technology that accomplishes the following:

- It presents a multidimensional, logical view of data to the end user with no requirements as to how the data is stored.
- It sorts, forecasts, tracks trends, and performs other complex analyses.
- It lets users move from one query to another and get results quickly and easily.

Examples:
A non-OLAP query: 'How many widgets did we sell last month?'
A query requiring OLAP: 'How many red widgets with polka dots did we sell last month in each region, Europe, and Asia/Pacific, compared to the same month last year, actual versus budget?'

The system development life cycle for the data warehouse environment is almost the exact opposite of that for an operations environment.

Data Warehousing and Technology

—What Are the Technology Requirements for the Data Warehousing Environment?—

The first and easily the most important technological requirement for the data warehouse is the ability to manage large amounts of data. This really means two points: the capacity to manage

Multidimensional OLAP products		Relational OLAP products	
(Contain their own DBMS)		(Run against major relational DBMS)	
Vendor	Product	Vendor	Product
Kenan Technologies Cambridge, MA	Acumate ES	Information Advantage Minnetonka, MN	Axsys
Arbor Software Sunnyvale, CA	Essbase	Prodea Software Eden Prairie, MN	Beacon
Oracle Redwood Shores, CA (through IRI Software)	Express	MicroStrategy Vienna, VA	DSS Agent
Holistic Systems Edison, NJ	Holos	Stanford Technology San Francisco, CA	Metacube
Dun & Bradstreet Framingham, MA	Pilot	Sagent Technology Menlo Park, CA	Unnamed yet

Table 1. Two categories of OLAP products

very large amounts of data at all and the ability to manage that data well—through addressability, indexing, extensions of data, efficient management of overflow, and the like. Any technology that purports to support the data warehouse must satisfy both requirements for capability and efficiency.

The first and easily the most important technological requirement for the data warehouse is the ability to manage large amounts of data.

In addition to the requirement for managing massive amounts of data efficiently and cost effectively, the technology underlying the data warehouse needs to be able to handle multiple storage media. It is insufficient to manage a mature data warehouse on DASD alone (Table 2).

The very essence of the data warehouse is the flexible and unpredictable access of data. Thus, required is the ability to access data quickly and easily. If data is not efficiently indexed and users cannot access data rapidly, the data warehouse will not succeed.

In addition, the data in the data warehouse needs to be able to be monitored at will. The cost of monitoring data cannot be so high and the complexity of monitoring data cannot be so hard that a monitoring program cannot be run whenever necessary. (Monitoring is a means of determining that a reorganization needs to be done, if an index is poorly structured, if data is in overflow, if space is still available, and the like.) The data warehouse also needs to be able both to receive data from and pass data to a wide variety of technologies. The technology supporting the data warehouse is practically worthless if there are major constraints in the passing of data to and from the data warehouse.

This interface to the data has to be both efficient and easy to use. It also has to support batch mode. Operating in on-line mode is interesting but not terribly useful.

Summary of Technical Requirements

(1) The ability to manage large volumes of data
(2) The ability to manage multiple storage media
(3) Support for easy and efficient indexing and monitoring of data
(4) Capability to interface with a wide variety of technologies—for both import and export of data
(5) Ability to support optimal programmer activity in the design and development stage when physically placing the data—at the block/page level
(6) Ability to manage data in parallel
(7) Support for solid metadata control
(8) Provision of a rich language interface to the data warehouse
(9) Ability to load the warehouse efficiently
(10) Support for efficient use of indexes
(11) Ability to store data compactly
(12) Ability to support compound keys
(13) Ability to manage variable-length data efficiently
(14) Ability to turn the lock manager on and off at will—that is, control of the lock manager at the programmer level
(15) Ability to conduct index-only processing

Media	Speed	Cost
Main memory	Very fast	Very expensive
Expanded memory	Very fast	Expensive
Cache	Very fast	Expensive
DASD	Fast	Moderate
Optical disk	Not slow	Not expensive
Fiche	Slow	Cheap

Table 2. Managing multiple media

(16) Capability to restore data from a bulk medium quickly and completely

What Runs Where: Components of Data Warehousing

—Sources of Data—

Data such as batch operational data or transaction operational data may be stored on any of the following plus others:

- IMS
- IDMS
- VSAM
- DB2
- AS400
- Adabas
- Informix
- Oracle
- Sybase
- External databases

—Data Warehouse Management Software—

Information is extracted from these databases, transformed and maintained through vendor solutions such as the following:

- Prism, with Prism Warehouse Manager and Prism Directory Manager
- ETI
- Carleton

—The Data Warehouse—

The extracted data is then housed in another database on, for example, Data General's AViiON NUMA (non-uniform memory architecture) servers, using Oracle, Sybase, or Informix. The metadata—the data about the data—is also housed here.

- Oracle RDBMS
- Sybase RDBMS
- Informix RDBMS

—Data Access Tools—

At the desktop, a variety of products provide decision support for the PC user: NewEra, ViewPoint, Business Objects, Information Advantage, MicroStrategy, Stanford Technology Group, and others.

The Data Warehousing Market

'Ninety per cent of all information processing organizations will be pursuing a data warehouse strategy in the next three years' (The Meta Group, 1994). The Meta Group surveyed 175 attendees at its Data Warehouse conference (1995) on several questions relating to data warehousing. Questions concerned the estimated number of data warehouse users, their data warehouse size and the time spent in their respective corporations on creating the data warehouses. Results were as follows:

Estimated number of data warehouse users

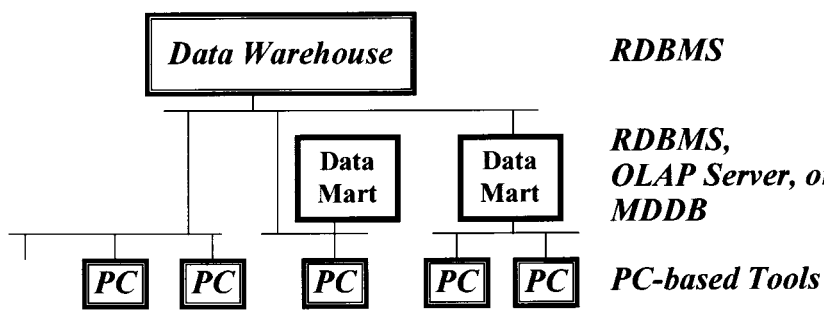

Figure 3. Three-tier data warehousing architecture with Data Mart.

Current users today:

Fewer than 10	33% of the respondents
10–49	46%
50–99	13%
100–499	6%
500–1000	1%
Over 1000	1%

Projected in 12–18 months:

Fewer than 10	0% of the respondents
10–49	17%
50–99	25%
100–499	40%
500–1000	11%
Over 1000	7%

What this tells us is that almost half of this population had between 10 and 49 people actively using data warehousing today, with almost the same percentage anticipating between 100 and 499 users within the next year. Once in production, data warehouses may be supporting upwards of 500 users.

The Meta Group also looked at these respondents current and projected warehouse database size. Results are as follows:

Current database size:

<5 GB	24% of the respondents
5–9 GB	20%
10–19 GB	18%
20–49 GB	19%
50 GB+	19%

Database size projected in 12 to 18 months:

<5 GB	2% of the respondents
5–9 GB	11%
10–19 GB	11%
20–49 GB	16%
50 GB+	59%

Important to note here is the massive projections in expected size of the database within the next year. This is exceedingly critical in having realistic expectations about future size requirements.

—Time Spent on Data Warehouses to Date—

How long has this audience been working on the implementation of a data warehouse? Attendees at the Meta Group's conference reported the following:

No project to date	15%
6 months or less	37%
6–12 months	16%
12–18 months	11%
18–24 months	6%
2 years or more	15%

The conclusion here is that for this particular audience, at least, data warehousing projects were quite new—within 6 months for a large percentage. Interesting is the fact that fully as many had been working in this area for 2 years as had not begun projects at all.

In another report, Metagroup cited that 47% will spend up to $500,000 on data warehousing software in 1995; 11% will top $3 million. Many of the over-$3 million clients were financial services, manufacturing, and banking.

—Market Issues Addressed by Data Warehousing—

- Improves access to integrated data
  - –Eliminates 'spiderweb' effect of legacy systems
  - –Eliminates impact on production systems
  - –Supports open systems environment
  - –Creates historical perspective on data
- Ensures data integrity and quality
- Lowers information technology costs

—Benefits to Customers—

- More effective decision making
- Improved business intelligence
- Enhanced customer interaction
  - –Sales—revenue
  - –Service—satisfaction

  Key application areas for data warehousing
- Greater productivity, profitability, and quality
- Better asset/liability management
- Greater insights for successful and effective business process re-engineering

Query	Application development	Relational OLAP	MDDB OLAP
Cognos (Power Play) Business Objects (Business Objects) Andyne (GQL)	Informix (NewEra) Sybase (Power Builder) (Delphi) Gupta (Centura) Microsoft (Visual BASIC)	Informix (Meta Cube) MicroStrategy (OSS Agent) Information Advantage (IA Suite) Cognos (Impromptu)	Arbor (Essbase) Oracle (Express) Kenan (Acumate ES)

Table 3. Sample data access tools

Data in millions	1999 ($)	CAGR (%)
Total market size	6960	34.7
Data extraction movement	210	26.4
Administration	450	114.1
RDBMS	1100	30.7
Hardware	3950	29.7
Consulting services	1250	57.3

Table 4.

Data Warehouse Market Segment Revenue Forecast

The Gartner Group predicts total market sizes by 1999 and the compound annual growth rate of each of these components of data warehousing as shown in Table 4 (this includes hardware, software, and systems integration services (see also Figure 4).

Potential Applications (Subject Areas) of Data Warehousing

- Analysis of sales results, channels, and marketing programs
- Customer and product profitability
- Integrating customer databases
- Risk management
- Procurement
- Asset management
- Human resources
- Quality assurance
- Actuarial and statistical analysis

And What About The Network?

Discussion with both implementors and vendors who support data warehouse environments report that the best way to ascertain the network

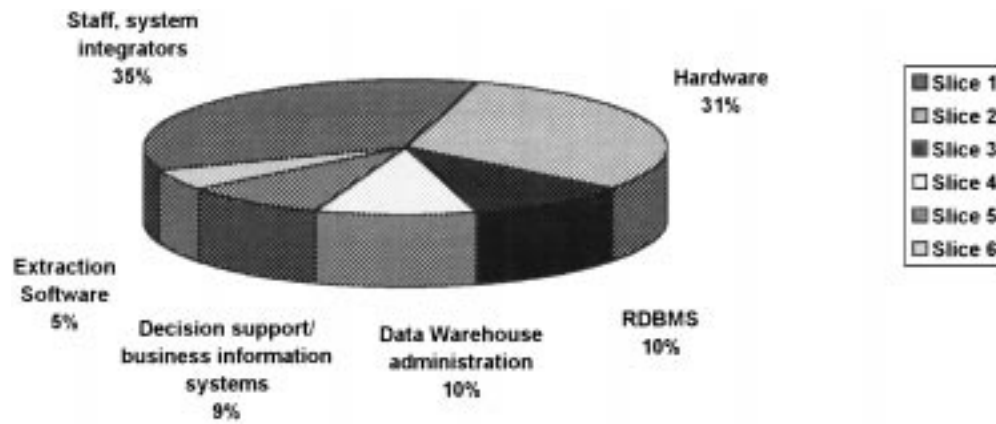

Figure 4. Where the data warehouse dollars go.

needs in supporting a data warehouse is that it is the same as the network required for supporting the operational environment. Now clearly if the corporate operational environment is a WAN connecting international sites and the analysts seeking data for strategic use are all in headquarters, the corporate network is not recreated. The point is that the bandwidth and network selection criteria can be the same as that which is adequate for the operational data network.

Managing the data warehouse environment can be accomplished through the same or similar tools as the corporate operations environment as well: database management tools, SNMP-based network and device management tools are most frequently used.

Glossary

Access path The path chosen by a database management system to retrieve the requested data.

***Ad-hoc* query** Any query which cannot be determined prior to the moment the query is issued. A query that consists of dynamically constructed SQL, which is constructed by desktop-resident query tools.

***Ad-hoc* query tool** An end-user tool that accepts an English-like or point-and-click request for data and constructs an *ad-hoc* query to retrieve the desired result.

Administrative data In a data warehouse, the data that helps a warehouse administrator manage the warehouse. Examples are user profiles and order history data.

Aggregate data Data that is the result of applying a process to combine data elements. Data that is taken collectively or in summary form.

Alerts A notification from an event that has exceeded a pre-defined threshold.

Atomic data Data elements that represent the lowest level of detail. For example, in a daily sales report, the individual items sold would be atomic data, while the rollups such as invoice and summary totals from invoices are aggregate data.

Base tables The normalized data structures maintained in the target warehousing database. Also known as detail data.

Bi-directional extracts The ability to extract, cleanse, and transfer data in two directions among different types of databases, including hierarchical, networked, and relational databases.

Braking mechanism A software mechanism which prevents users from querying the operational database once transaction loads reach a certain limit.

Bulk data transfer A software-based mechanism designed to move large data files. It supports compression, blocking, and buffering to optimize transfer times.

Business data Information about people, places, things, business rules, and events, which is used to operate the business. It is not metadata. (Metadata defines and describes business data.)

Business transaction A unit of work acted upon by a data capture system to create, modify, or delete business data. Each transaction represents a single valued fact describing a single business event.

Catalog A component of a data dictionary that contains a directory of its DBMS objects as well as attributes of each object.

Central warehouse A database created from operational extracts that adheres to a single, consistent, enterprise data model to ensure consistency of decision-support data across the corporation.

Change data capture The process of capturing changes made to a production data source. Change data capture is typically performed by reading the source DBMS log. It consolidates units of work, ensures data is synchronized with the original source, and reduces data volume in a data warehousing environment.

Collection A set of data that resulted from a DBMS query.

Consumer An individual, group or application that accesses data/information in a data warehouse.

Consumer profile Identification of an individual, group or application and a profile of the data they request and use: the kinds of warehouse data, physical relational tables needed, and the required location and frequency of the data (when, where, and in what form it is to be delivered).

Copy management A process that takes all or a snapshot of data from a source environment and copies that data to a new target environment.

Crosstab A process or function that combines and/or summarizes data from one or more sources into a concise format for analysis or reporting.

Currency data The date the data is considered

effective. It is also known as the 'as of' date or temporal currency.

Data Items representing facts, text, graphics, bit-mapped images, sound, analog or digital live-video segments. Data is the raw material of a system supplied by data producers and is used by information consumers to create information.

Data access tools An end-user oriented tool that allows users to build SQL queries by pointing and clicking on a list of tables and fields in the data warehouse.

Data analysis and presentation tools Software that provides a logical view of data in a warehouse. Some create simple aliases for table and column names; others create data that identify the contents and location of data in the warehouse.

Data dictionary A database about data and database structures. A catalog of all data elements, containing their names, structures, and information about their usage. A central location for metadata. Normally, data dictionaries are designed to store a limited set of available metadata, concentrating on the information relating to the data elements, databases, files and programs of implemented systems.

Data element The most elementary unit of data that can be identified and described in a dictionary or repository which can not be subdivided.

Data extraction software Software that reads one or more sources of data and creates a new image of the data.

Data loading The process of populating the data warehouse. Data loading is provided by DBMS-specific load processes, DBMS insert processes, and independent fastload processes.

Data management software Software that converts data into a unified format by taking derived data to create new fields, merging files, summarizing and filtering data; the process of reading data from operational systems. It is also known as data extraction software.

Data mapping The process of assigning a source data element to a target data element.

Data mart A small subset of a data warehouse used by small number of users. A mini mart is a very focused slice of a larger data warehouse. It is regularly updated or extracted from to analyze data for a project, for example. This is the typical use for a MDDB.

Data mining Data mining is the process of sifting through large amounts of data to produce data content relationships; usually the user does not know exactly what he or she is looking for, but is searching for particular patterns or trends. It is a set of analytical techniques for discovering relationships in data which were previously unknown. Data mining requires a data warehouse—or 'mine'—of some kind to get started. It is a niche application which complements other DSS tools. Sometimes called 'quarrying' or 'data surfing'.

Data model A logical map that represents the inherent properties of the data independent of software, hardware or machine performance considerations. The model shows data elements grouped into records, as well as the association around those records.

Data modeling A method used to define and analyze data requirements needed to support the business functions of an enterprise. These data requirements are recorded as a conceptual data model with associated data definitions. Data modeling defines the relationship between data elements and structures.

Data partitioning The process of logically and/or physically partitioning data into segments that are more eaily maintained or accessed. Current RDBMS systems provide this kind of distribution functionality. Partitioning of data aids in performance and utility processing.

Data pivot A process of rotating the view of multidimensional data. It is performed with an OLAP tool.

Data propagation The distribution of data from one or more source data warehouses to one or more local access databases, according to propagation rules.

Data replication The process of copying a portion of a database from one environment to another and keeping the subsequent copies of the data in sync with the original source. Changes made to the original source are propagated to the copies of the data in other environments.

Data repository A database designed for storage of metadata and access by end users and administrators.

Data scrubbing The process of filtering, merging, decoding, and translating source data to create validated data for the data warehouse.

Data store A place where data is stored; data at rest. A generic term that includes databases and flat files.

Data surfing A technique using software tools geared for the user who typically does not know exactly what he or she is searching for, but is looking for particular patterns or trends. Data surfing is the process of sifting through large amounts of data to product data content relationships. This is also known as data mining.

Data warehouse A central repository for storing and analyzing vast amounts of data (historical and reference) from a number of different sources, enabling users to gain insights into corporate performance and customer behavior that cannot be gained using disconnected operational systems. A data warehouse forms the core of a decision support system of a corporation.

Data warehouse engines Relational databases (RDBMS) and multi-dimensional databases (MDDB). Data warehousing engines require strong query capabilities, fast load mechanisms, and large storage requirements.

Data warehousing A process of assembling disparate data, transforming it into a consistent state for business decision making, and empowering users by providing them with fast, flexible access to this information. This process integrates data from a variety of source databases into one target database that is optimally designed for decision support.

Decision support system (DSS) A set of tools and data to help employees make informed decisions quickly. The data may include historical, demographical, cost and competitive information. The tools may support such functions as scenario building, data mining, presentation and proposal builders, or OLE automation. Sample DSS tools are Cognos' PowerPlay, Andyne's Pablo, and BusinessObjects from Business Objects, among others.

Delta update Only the data that was updated between the last extraction or snapshot process and the current execution of the extraction or snapshot.

Derived data Data that is the result of a computational step applied to reference or event data. Derived data is the result of relating two or more elements of a single transaction (such as an aggregation), or of relating one or more elements of a transaction to an external algorithm or rule.

Drill down A method of exploring detailed data that was used in creating a summary level of data. Drill down levels depend on the granularity of the data in the data warehouse. All OLAP and MDDB products support drill down.

Drill-thru Related to drill-down, drill-thru allows the user to drill all the way down to a particular row in a SQL database. Oracle Express supports drill-thru to an Oracle RDBMS.

Dynamic queries Dynamically constructed SQL that is usually constructed by desktop-resident query tools. Queries that are not pre-processed and are prepared and executed at run time.

Enterprise modeling The development of a common consistent view and understanding of data elements and their relationship across the enterprise.

Entity relationship diagramming A process that visually identifies the relationships between data elements.

Executive information system (EIS) A business monitoring system that presents summary data about business functions in a timely and easily understood manner. Usually involves heavy use of status indicators (traffic lights) to indicate parameters in normal or undesirable states.

Extraction engines Software which scans files of databases and performs selection criteria for records to be copied, then transports them to another file or database.

Extract specifications The standard expectations of a particular source data warehouse for data extracts from the operational database system-of-record. A system-of-records uses an extract specification to retrieve a snapshot of shared data, and formats the data in the way specified for updating the data in the source data warehouse. An extract specification also contains extract frequency rules for use by the data access environment.

Gateway A software product that allows SQL-based applications to access relational and non-relational data sources.

Host-driven A processing method in which the host computer controls the session. A host-driven session typically includes terminal emulation, front ending, or client/server types of connections. The host determines what is displayed on the desktop, receives user input from the desktop, and determines how the application responds to the input.

Increment Data warehouse implementation can be broken down into segments or increments. An increment is a defined data warehouse implementation project that has a specified beginning and

end. An increment may also be referred to as a departmental data warehouse within the context of an enterprise.

Inverted file indexes A more efficient method to access data in an *ad-hoc* or analysis environment. It maintains indexes to all values contained in an indexed field. Those values, in turn, can be used in any combination to identify records that contain them, without actually scanning them from disk.

Metadata Metadata is data about data. Examples of metadata include data element descriptions, data type descriptions, attribute / property descriptions, range / domain descriptions, and process / method descriptions. The repository environment encompasses all corporate metadata resources: database catalogs, data dictionaries, and navigation services. Metadata includes things like the name, length, valid values and descriptions of a data element. Metadata is stored in a data dictionary and repository. It insulates the data warehouse from changes in the schema of operational systems.

Middleware A communications layer that allows applications to interact across hardware and network environments.

Multidimensional database Multidimensional database system (MDDB). These products use their own data stores—often proprietary—into which data must be copied or moved before it can be analysed. They are high performing because they use pre-aggregated data, or data that has been summarized or precalculated in some other way. However, preaggregation limits query flexibility, and they bog down approaching 50GB. They tend to lack the security and administration features of major RDBMSs—which is important as MDDs grow in importance in an organization. They also require additional training and expertise to set up and administer. They have limited ability to drill down into data and show users how a conclusion was reached. MDDBs are ideal for data marts. Examples are IRI / Oracle's Express and Probit.

Normalization The process of reducing a complex data structure into its simplest, most stable structure. In general, the process entails the removal of redundant attributes, keys, and relationships from a conceptual data model.

OLAP OLAP (on-line analytical processing) refers to the storage and manipulation of data warehouse data similar to DSS applications and EIS (see above).

On-line complex processing This refers to complex queries involving multiple table updates and user-defined data types. Used with mission critical applications, one may find simultaneous queries and transactions against the same database. This is a blend of OLTP and heavy DSS in one database—and is contrary to the definition and purpose of a data warehouse.

OLTP OLTP (on-line transaction processing) updates the production or operations systems of a corporation. Usually terminal or client / server based, these transactions are typically predefined and updates usually touch only a few files.

Operational data store An ODS is an integrated database of operational data. Its sources include legacy systems and it contains current or near term data. An ODS may contain 30 to 60 days of information, while a data warehouse typically contains years of data.

Production data Source data which is subject to change. It is a data capture system, often on a corporation's mainframe.

Production system Unlike a data warehouse, a production system deals with operational data, maintained in real time. It runs the day-to-day operations of a corporation, keeping data for 30-90 days as current data. It is typically updated by transaction processing. Typical productions are legacy-based mainframes or minis, or more recently, Unix servers with a standard relational database.

Propagated data Data that is transferred from a data source to one or more target environments according to propagation rules. Data propagation is normally based on transaction logic.

Protocol A set of conventions that govern the communications between processes. Protocol specifies the format and contents of messages to be exchanged.

Relational OLAP Technology which provides multidimensional analysis against data that remains in a relational database management system. This category of OLAP works with wares from Sybase, Oracle, Informix, etc. There is no need to copy or move data into a specialized DBMS, no need for preaggregation, and no additional administration. Sites retain full control over access and security; the only limit to the database's size is that of the RDBMS. On the downside,

relational OLAP products run more slowly than MDDB OLAP products. The RDBMs accessed by the OLAP tool can be a data warehouse or a data mart.

Roll up queries Queries that summarize data at a level higher than the previous level of detail.

Scalability The ability to scale to support larger or smaller volumes of data and more or less users. The ability to increase or decrease size or capacity in cost-effective increments with minimal impact on the unit cost of business and the procurement of additional services.

Schema The logical and physical definition of data elements, physical characteristics, and inner-relationships, the schema is described by meta-data.

Slice and dice A term used to describe a complex data analysis function provided by RDBMS tools.

Source database An operational, production database or a centralized warehouse that feeds into a target database.

SQL Structured query language for accessing relational, ODBC, DRDA, or non-relational compliant databases.

SQL query tool An end user tool that accepts SQL to be processed against one or more relational databases.

Standard query A stored procedure of a recently executed query. Technically, a standard query may be stored on the desktop as 'canned' SQL and passed as dynamic SQL to the server database to execute. This is undesirable unless the stored query is seldom executed.

Static query A stored, parameterized procedure, optimized for access to a particular data warehouse.

Subject oriented databases Rather than build one massive, centralized data warehouse, some companies are building numerous subject-oriented warehouses to serve the needs of different divisions. Also referred to as data marts.

Summarization tables These tables are created along commonly used access dimensions to speed query performance, although the redundancies increase the amount of data in the warehouse.

Syntactic mapping The mapping required to unravel the syntax of information.

Target database The database in which data will be loaded or inserted.

Tool encyclopedias Encyclopedias, repositories or dictionaries used by application development tools. The non-definable 'repository' used by a tool.

Transformers Rules applied to change data.

Triggering data Data that selects and loads data on a scheduled basis.

Unit of work consolidation The process of consolidating multiple updates to a single row image into a single update.

Versioning The ability for a single definition to maintain information about multiple physical instantiations.

References

1. W. H. Inmon, *Building the Data Warehouse*, John Wiley, New York, 1992. ■

