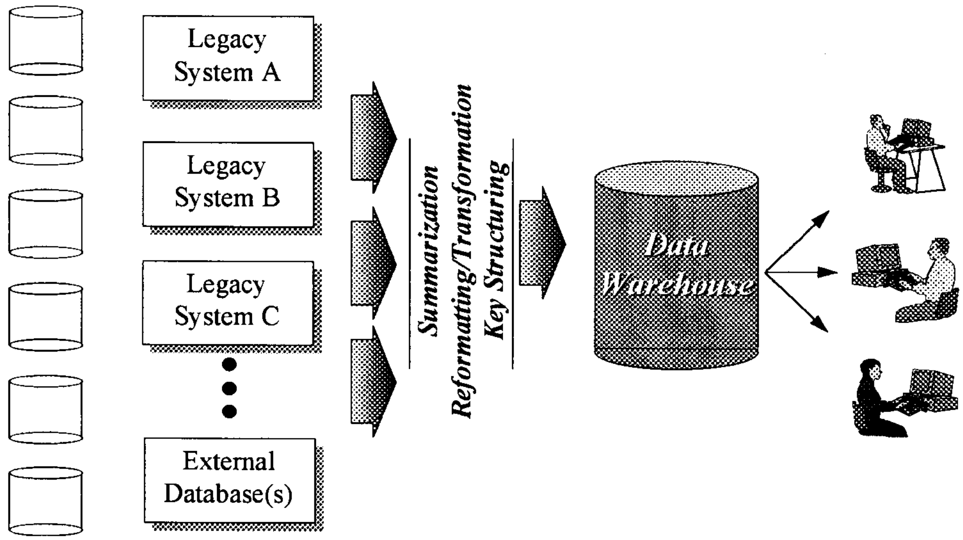
Data Warehouse Overview

Legacy transaction processing applications comprise the primary data source for the data warehouse in most implementations. However, data from external sources is increasingly being included in data warehouses to provide visibility to information not available with in-house systems. The figure below shows a representation of a typical data warehouse.



As information flows into the warehouse it usually undergoes significant transformation. The detail-level information stored in disparate legacy-based systems is summarized and organized along a time continuum and reformatted to ensure consistency and conformance with standardized business terminology. External data is transformed to align with internal data and terminology.

A primary objective of the data warehouse is to provide an integrated information delivery mechanism supporting historical perspectives, trending analysis and consistent business measures. The warehouse does not provide integration to the operational data within the legacy applications themselves, nor does it give users real-time access to that information. A successfully architected data warehouse, however, can revitalize legacy system application data by supporting easy, integrated access to it, leveraging the enterprise's legacy investment. Additionally, by separating decision support processing from transaction processing, organizations can achieve greater flexibility in information access and usage without affecting on-line transaction processing.

Operational Data Store

A recent trend is to utilize the same philosophies of the data warehouse in to the world of operational systems by creating an operational data store (ODS). The ODS provides a foundation to achieve integrated operational results to organizations struggling with the operational aspects of unintegrated operational systems. Structurally, the ODS is designed to support a real-time transactional data model, while the data warehouse data model is designed to support historical and decision support data. Architecturally, the ODS fits *between* the legacy systems and the data warehouse. However, the ODS *is not* part of the data warehouse and under no conditions should the two ever be combined—it is clearly in the operational domain. It typically contains current or near-current information. Through the effective implementation of the operational data store, re-architecting transaction processing systems becomes more straightforward. Migration towards a client/server technology architecture is facilitated by an ODS because the data foundation on which new systems will be built is stabilized.

Key Concepts

Data warehousing introduces design concepts which differ considerably from other systems development approaches due to the unique way a warehouse organizes, stores and delivers information. An awareness of these key concepts helps in understanding how a data warehouse achieves its objective of an integrated information delivery vehicle.

Data Quality

A critical issue for a data warehouse is the quality of the data. Data quality is measured in terms of:

- Data integrity (how the data ties together and relates)
- Data consistency (ensuring elements from various source systems are uniformly defined)
- Data completeness (ensuring fields are completely populated with correct values).

The degree to which data quality is achieved directly influences the data warehouse's level of credibility with its users. This re-enforces the need for transaction data from operational legacy systems to be extracted, cleansed and transformed, turning data into information as it is loaded into the warehouse.

Data Cleansing

Consolidating and integrating information from multiple sources presents significant challenges in achieving the data quality necessary for an effective data warehouse. The process of resolving inconsistencies and discrepancies between source systems is called data cleansing. Reference codes, terminology, valid spellings, field lengths, etc., are often different among systems for the same data elements. These differences must be recognized and addressed during development of the data warehouse. For example, consider two separate operational systems that contain customer account data. In one system, data on the Wal-Mart account is stored as "Wal-Mart" while the second system stores data on this same account under the code "WalMart." In this case, cleansing rules would need to be established to consolidate and summarize data for customers stored with multiple variations in the account name spelling.

Data cleansing rules not only ensure consistency across multiple source systems, but also enforce compatibility with the approved data warehouse data model. Data cleanliness should be assessed early in the development cycle to allow time for any changes to legacy systems required to address major data issues which are not able to be resolved with cleansing rules during the load of the data warehouse.

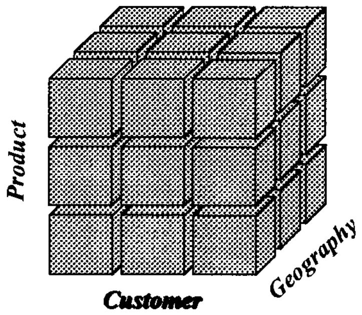
Uniform Business Definitions

One of the objectives and benefits of implementing a data warehouse is establishing uniform business definitions across the enterprise. Many corporations have inconsistent and redundant business terms that have evolved over time. One business term may have different meanings based on the department using the term. For example, net sales in the accounting department may measure sales prior to point-of-sale discounts while net sales in the sales department is reported including discounts. Similarly, some business areas may use different terms for the same business measure. Standard calculations and measures should be defined to ensure consistency in reporting. It is critical in positioning the data warehouse as a key component in an enterprise-wide information delivery system that consistent terminology and performance measures be used.

Dimensions, Measures and Facts

The primary decision support information delivered through the warehouse takes the form of business measures and dimensions. This information is often represented graphically by a cube as shown below.

Data Warehouse Dimensions



A business measure is a unit of information that provides an accurate and well-defined performance indicator to be used in decision making. A "fact" is the physical instance of a business measure stored in a warehouse database and is generally used synonymously with business measure. Examples of measures are gross sales dollar amount, net sales units, inventory level, etc. Some business measures are derived, that is, composed of other measures, usually as a result of a calculation (e.g., net income, loss ratio, product profitability). A business measure may not be directly reflected in the logical data model used to define the warehouse, but may actually be the combination of several data model attributes.

Dimensions are views of the business used to organize how business measures will be reported. Common dimensions are customer, product, geography, and time. Most companies sell products, to customers, within some defined geographic boundary, during some period of time. Dimensions often follow the word "by" in a report description. For example, reporting gross sales by region displays the business measure gross sales by the geographic dimension of region. Dimensions usually contain occurrences which relate to one another in a hierarchical manner, enabling business measures to be rolled-up or aggregated at various reporting levels. With the geography dimension, for example, the roll-up sequence might be county, state, region, country. A powerful decision support technique supported by data warehouses is multi-dimensional analysis. In this case, business measures are viewed relative to more than one dimension simultaneously, resulting in a very focused perspective of the data.

Summarization

A significant amount of transformation occurs as data passes from legacy systems to the data warehouse. To support decision making focused around trending and historical comparisons, information is stored at various levels of summarization in the warehouse. Data is typically provided at the following levels:

- Older detail (usually archived)
- Current detail
- Lightly summarized (e.g., sales summary by region by product line)
- Highly summarized (e.g., sales summary by country for all products)

Subject Orientation

A data warehouse is subject oriented. Data is stored recognizing the shift from an application-orientation which is designed to support operational transaction processing to a decision making focus. Subject-oriented data provides a comprehensive view of key aspects of the business such as customer, product, line of business, etc. which typically span multiple applications and business areas.

Time Variance

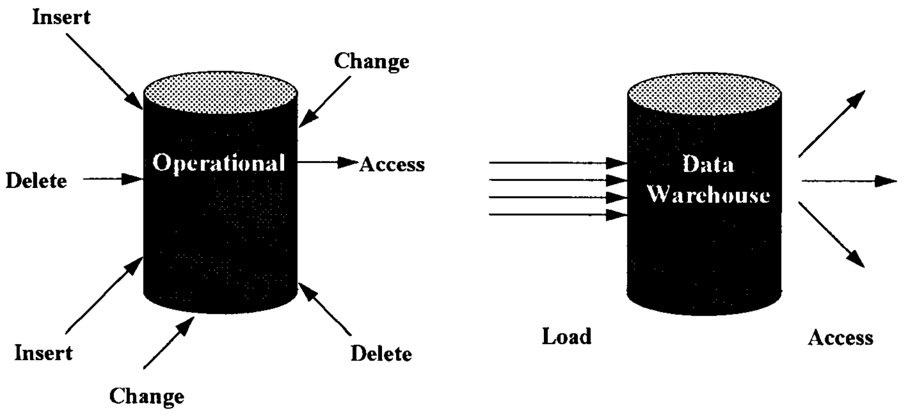
Operational data is generally valid only at the time it is captured and therefore has a short effective life. For example, an inventory quantity increases and decreases as production is completed and orders are committed. This makes the window to determine actual inventory available measured in minutes, or less. Informational data within a warehouse, by contrast, has a time dimension. Each business measure in the warehouse is associated with a point in time that allows for comparison and trending along time-oriented perspectives. A data warehouse retains data over time as it was originally captured. Another characteristic which differentiates a data warehouse is the time horizon for data retention. The time horizon for the data warehouse is significantly longer, up to five years of data retention, while operational systems do not typically retain on-line data for longer than 60-90 days, and rarely for more than one year.

Data Mining

Data mining is an approach that extends and enhances the value of a data warehouse. Specialized tools and techniques enable the warehouse user to recognize and analyze patterns in the information stored in the warehouse. These patterns may point to customer purchasing behavior, product movement trends, or may reveal other relevant data relationships not immediately obvious with traditional data analysis techniques. Data mining capabilities can also support users in analyzing information across multiple data warehouses, and often enhance advanced decision making activities such as forecasting and business modeling.

Non-Volatile Data

Operational data supporting on-line transaction processing is regularly accessed, manipulated and updated. Conversely, information in the warehouse is loaded, accessed by users for analysis and reporting, but generally not updated. The figure below illustrates the differences in operational and data warehouse data volatility. New data is constantly being cleansed, consolidated and added to the data warehouse, but existing warehouse data is not updated. The focus of the warehouse is to store data snapshots in time to support trend and pattern analysis.



Record-by-Record Manipulation

Mass-Load Access of Data

Data Warehouse Terms and Concepts

Data Cleansing - The process of resolving data inconsistencies and discrepancies prior to data warehouse loading.

Data Integrity - The property of data that ensures that data within the data warehouse is accurate, consistent, and relationships among data elements are preserved.

Data Mart - A physical subset of a data warehouse or separate instance of the data warehouse database, usually created to enhance information access and retrieval performance or security control for a specific user base or business area.

Data Mining - The application of data analysis techniques and tools to identify patterns in data, often undetectable with conventional data analysis procedures. Also, the process of sub-setting large unmanageable sets of information to narrow in on the data set to be analyzed. For example, claim transactions to identify hundreds of physicians performing high cost procedures.

Data Transformation - The process of changing data from its source system state to the format required in the data warehouse.

Data View - A preliminary design technique where business measures and dimensions are grouped into logical clusters. Data views list measures, dimensions, calculations, time dimensions, etc. and are used to assist prototype development.

Data Warehouse - A single, integrated source of decision support information formed by collecting data from multiple sources, internal to the organization as well as external, and transforming and summarizing this information to enable improved decision making. A data warehouse is designed for easy access by users to large amounts of information, and data access is typically supported by specialized analytical tools and applications. Data is structured around subject areas, for example, customer, and each unit of data is relevant to some fixed period of time. Clear definitions and clarifying descriptions about the information are contained within a data warehouse in a component called metadata. *A data warehouse is not an on-line transaction processing system and typically does not generate data for other applications.* A data warehouse is driven from an analytical data model, is based upon standardized business terminology, and reduces the amount of resources required to capture, consolidate and disseminate decision support information.

Decision Support System (DSS) - Application/system with the purpose and function to provide key information, detailed and summarized, to support decision making.

Decision Maker - End user of the data warehouse.

Denormalization - The process of reorganizing normalized data into physical tables to optimize query performance, typically resulting in data redundancy.

Derived Business Measures - Business measures that are composed of other business measures, usually as the result of a calculation (e.g., loss ratio, net income).

Dimensions - Aspects of a business which provide desired information perspectives to enable business performance analysis and decision making. Dimensions contain attributes which often relate to one another in a hierarchical manner (e.g., Country, Region, State, County). Dimensions often follow the word 'by' in a report description.

Drill Down - The process of transitioning from an information presentation at a given dimensional level to the next lower dimensional level.

Drill Up - The process of transitioning from an information presentation at a given dimensional level to the next higher dimensional level.

Executive Information System (EIS) - Analysis system designed for high-level business executives, featuring easy to use interface including drill down analysis, trend analysis, graphs, and tables.

Fact - The physical instance of a business measure in a data base.

Integrated Data - Data that is consistent in type, formatting, and encoding conventions. Data coming from disparate systems must be integrated to establish data consistency.
Interactive Decision Support - an information intensive analysis process which is performed in an iterative fashion with each analysis cycle taking place in seconds or minutes. Discoveries relative to data contents or key indicators are made in each analysis cycle which direct the user to dynamically form successive data warehouse queries or information requests. Data surfing, drill down, drill across, and data pivoting are examples of key data warehouse application features which enable interactive decision support.

Legacy System - Existing operational systems typically hosted on a mainframe or other transaction processing platform.

Metadata - Information about data within the data warehouse. Contains a description of the quality, structure, content, keys, indexes, and sources for data contained within the data warehouse. Metadata is analogous to a card catalog that allows the user to find the information they need in the data warehouse. It contains no data taken from the operational environment. It is:

a directory to help the DSS analyst locate contents of the data warehouse

Retroactivity - The process of restating data for a given time period using a given as-of date.
Snapshot - A row in the data warehouse containing an as-of date is considered a snapshot. Data is added to the warehouse as of a moment time and is not updated.

Source System - The operational source for a given data warehouse data element.

Staging Area - An intermediate storage area to hold data during the transformation and load process prior to loading data into the data warehouse.

Subject Area - A grouping of logically related entities.

Subject-Oriented - The characteristic of relating to subjects of an organization. The data warehouse is subject oriented. Examples of subjects are: customer, product, part, vendor.

Summarization - The process of calculating and aggregating business measures based on dimensional hierarchies.

Transformation - See Data Transformation.

Transaction Based System - See On-Line Transaction Processing System (OLTP).

Trend Analysis (Trending) - The process of looking at a business measure(s) over a spectrum of time.

Uniform Business Definitions - An unambiguous set of clearly defined definitions used to describe key aspects of the business. Uniform Business Definitions apply to dimensions, business measures, and other data warehouse data elements.

User Class - Identifies a group of data warehouse users which have similar data access needs and security level. A user class can, and probably will, span across functional departments.