Introduction to Data warehousing

Data, Information & Knowledge

- Data is composed of observable and recordable facts that are often found in operational or transactional systems. Data are any facts, numbers or text that can be processed by a computer.

- Information is an integrated collection of facts and is used as the basis for decision making. The patterns, associations or relationships among all data can provide information.

- Information can be converted into knowledge about historical patterns and future trends.

Operational & Informational Processing

- Operational processing (transaction processing) captures, stores and manipulates data to support daily operations.

- Information processing is the analysis of data or other forms of information to support decision making.

Operational Vs Information system

Comparison of Operational and Informational Systems

Characteristic	*Operational Systems*	*Informational Systems*
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance throughput, availability	Ease of flexible access and use
Volume	Many, constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

What is A Data Warehouse?

- The term "data warehouse" refers to a special type of database that acts as the central repository for company data. It can be thought of as a database archive that is segregated from the operational databases, and used primarily for reporting and data mining purposes.

History

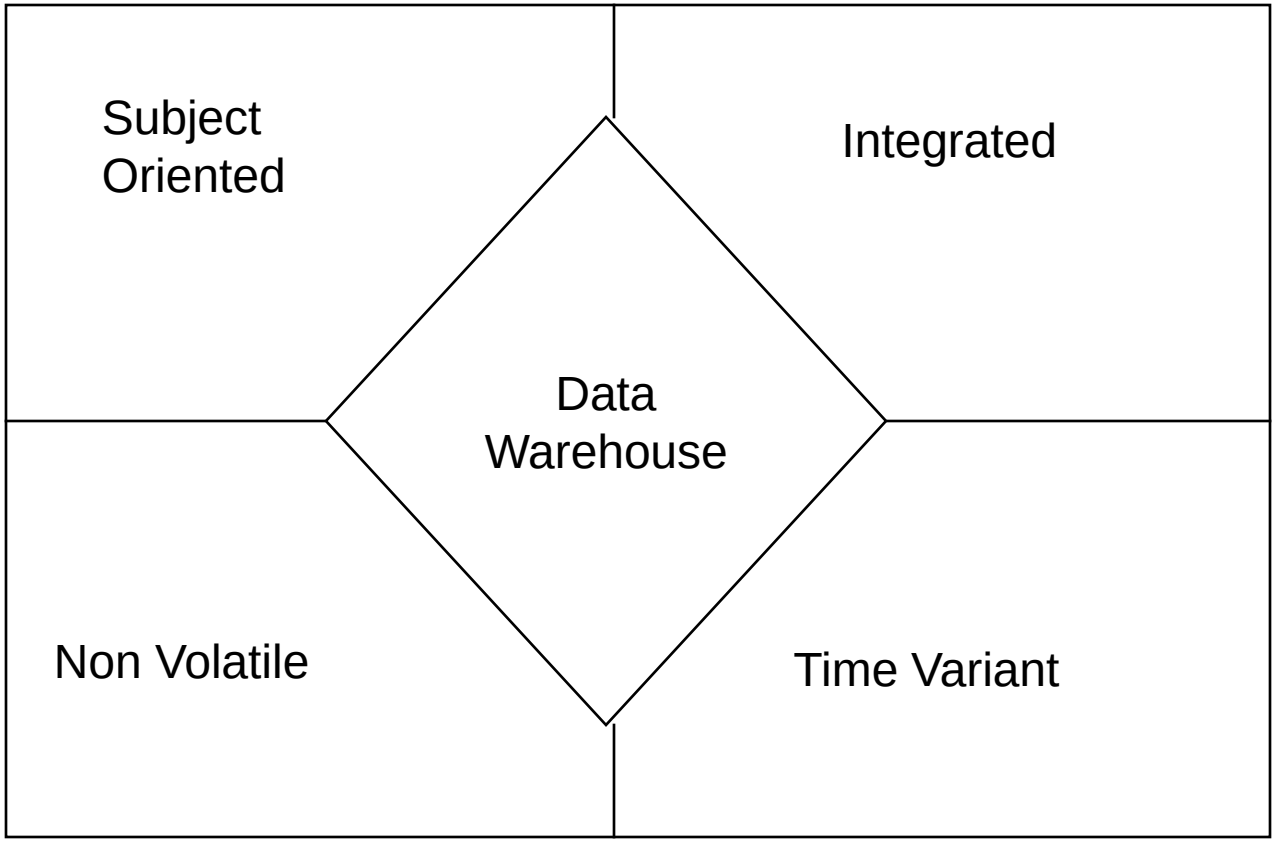
- Data warehouses were first developed in the 1980s in response to the growing demand for management information analysis, which operational databases could not perform without drastically affecting response time.

DATA WAREHOUSE

- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—W. H. Inmon

- Data warehousing is the process of constructing and using data warehouses

Data Warehouse Properties



Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"

Data Warehouse—Nonvolatile

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*

Other Definitions

Data Warehouse: A data structure that is optimized for distribution. It collects and stores integrated sets of historical data from multiple operational systems and feeds them to one or more data marts. It may also provide end-user access to support enterprise views of data.

Data Mart: A data structure that is optimized for access. It is designed to facilitate end-user analysis of data. It typically supports a single, analytic application used by a distinct set of workers.

Staging Area: Any data store that is designed primarily to receive data into a warehousing environment.

Operational Data Store: A collection of data that addresses operational needs of various operational units.

OLAP (On-Line Analytical Processing): A method by which multidimensional analysis occurs.

Multidimensional Analysis: The ability to manipulate information by a variety of relevant categories or “dimensions” to facilitate analysis and understanding of the underlying data. It is also sometimes referred to as “drilling-down”, “drilling-across” and “slicing and dicing”.

Hypercube: A means of visually representing multidimensional data.

Star Schema: A means of aggregating data based on a set of known dimensions. It stores data multi-dimensionally in a two dimensional Relational Database Management System (RDBMS), such as Oracle.

Snowflake Schema: An extension of the star schema by means of applying additional dimensions to the dimensions of a star schema in a relational environment.

Multidimensional Database: Also known as MDDB or MDDBS. A class of proprietary, non-relational database management tools that store and manage data in a multidimensional manner, as opposed to the two dimensions associated with traditional relational database management systems.

OLAP Tools: A set of software products that attempt to facilitate multidimensional analysis. Can incorporate data acquisition, data access, data manipulation, or any combination.

CHARACTERISTICS OF DATA WAREHOUSE

- It is a database designed for analytical tasks, using data from multiple applications.
- It supports a relatively small number of users with relatively long interactions.
- Its usage is read intensive.
- Its content is periodically updated.
- It contains current and historical data to provide a historical perspective of information.
- It contains a few large tables.
- Each query frequently results in a large result set and involves frequent full table scan and multiple joins.

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

OLTP vs. OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

DATAWAREHOUSE SOFTWARE

Company	Software
IBM	IBM Netezza
Microsoft	MS SQL Parallel Data Warehouse
Oracle	Oracle Exadata
101 Data Solutions	101Data
VMWARE	Cetas

Why Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

DATA MARTS

- A data mart contains a subset of wide data that is of value to a specific group of users.
- It is a data store that is a subsidiary to a datawarehouse of integrated data.
- It is a set of summarized or aggregated data.
- The data contents in data marts tends to be summarized .
- They are usually implemented on low cost departmental servers (UNIX, windows NT)
- The implementation cycle of a data mart is measured in weeks rather than month or years.
- Depending on source of data, data marts can be categorized as independent or dependent.

INDEPENDENT DATA MARTS :

- These are sourced from data captured from one or more operational systems or external information providers.
- Each independent data marts makes its own assumptions about how to consolidate the data and the data across several data marts may not be consistent.

DEPENDENT DATA MARTS:

- It is sourced directly from enterprise datawarehouse.

Problems with data marts

- Scalability in situations where an initial small data mart grows quickly in multiple dimensions
- Data Integration
- Situations where independent data marts are use
- Extremely urgent user requirements.
- The absence of a budget for a full datawarehouse
- The absence of a sponsor for an enterprise decision support strategy.