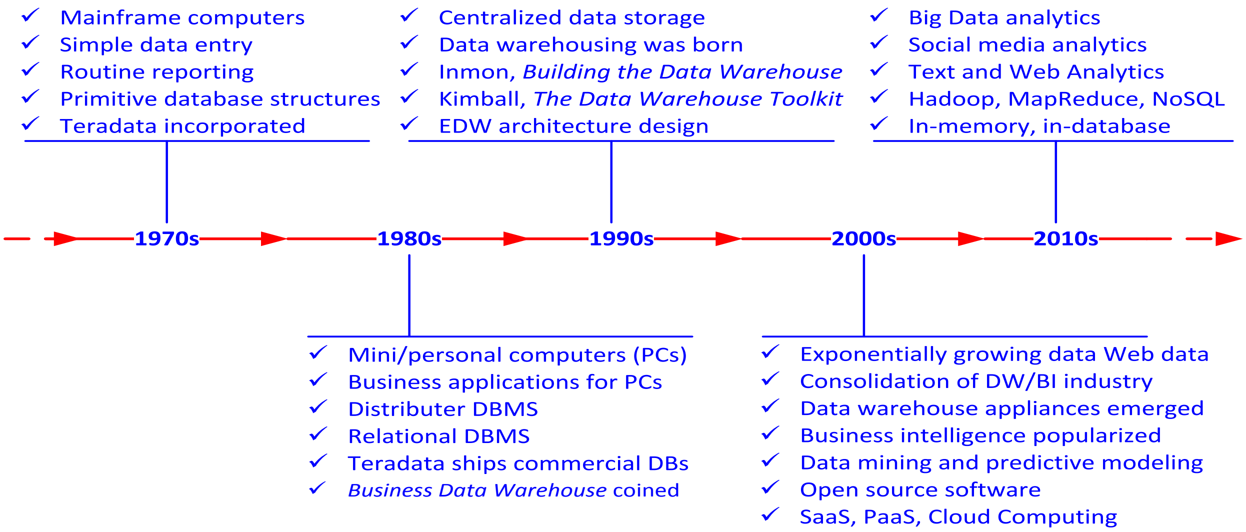
# Chapter 3:

## Data Warehousing

### What is a Data Warehouse?

- A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format
- "The data warehouse is a collection of integrated, subject-oriented databases designed to support DSS functions, where each unit of data is non-volatile and relevant to some moment in time"

### A Historical Perspective to Data Warehousing



1970s

- ✓ Mainframe computers
- ✓ Simple data entry
- ✓ Routine reporting
- ✓ Primitive database structures
- ✓ Teradata incorporated

1980s

- ✓ Mini/personal computers (PCs)
- ✓ Business applications for PCs
- ✓ Distributer DBMS
- ✓ Relational DBMS
- ✓ Teradata ships commercial DBs
- ✓ *Business Data Warehouse* coined

1990s

- ✓ Centralized data storage
- ✓ Data warehousing was born
- ✓ Inmon, *Building the Data Warehouse*
- ✓ Kimball, *The Data Warehouse Toolkit*
- ✓ EDW architecture design

2000s

- ✓ Exponentially growing data Web data
- ✓ Consolidation of DW/BI industry
- ✓ Data warehouse appliances emerged
- ✓ Business intelligence popularized
- ✓ Data mining and predictive modeling
- ✓ Open source software
- ✓ SaaS, PaaS, Cloud Computing

2010s

- ✓ Big Data analytics
- ✓ Social media analytics
- ✓ Text and Web Analytics
- ✓ Hadoop, MapReduce, NoSQL
- ✓ In-memory, in-database

### Characteristics of DWs

- Subject oriented
- Integrated
- Time-variant (time series)
- Nonvolatile
- Summarized
- Not normalized
- Metadata
- Web based, relational/multi-dimensional

Client/server, real-time/right-time/active

## Data Mart

A departmental small-scale "DW" that stores only limited/relevant data

- Dependent data mart

  A subset that is created directly from a data warehouse
- Independent data mart

  A small data warehouse designed for a strategic business unit or a department

## Other DW Components

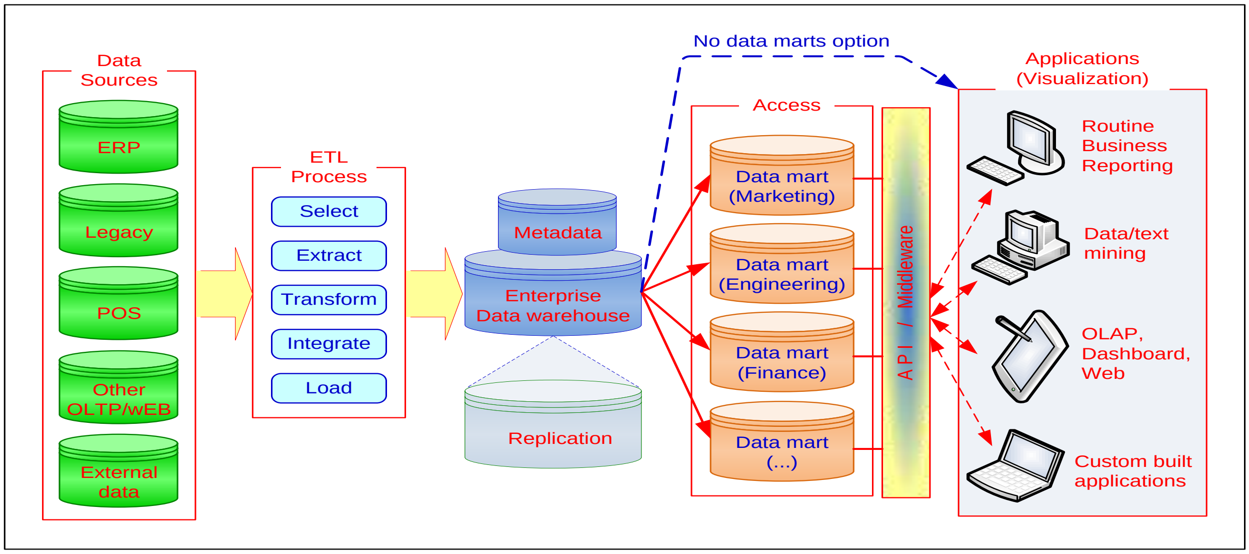
- Operational data stores (ODS)

  A type of database often used as an interim area for a data warehouse
- Oper marts - an operational data mart.
- Enterprise data warehouse (EDW)

  A data warehouse for the enterprise.
- Metadata: Data about data.

  In a data warehouse, metadata describe the contents of a data warehouse and the manner of its acquisition and use

## A Generic DW Framework



## DW Architecture

- Three-tier architecture
  1. Data acquisition software (back-end)
  2. The data warehouse that contains the data & software
  3. Client (front-end) software that allows users to access and analyze data from the warehouse

- Two-tier architecture

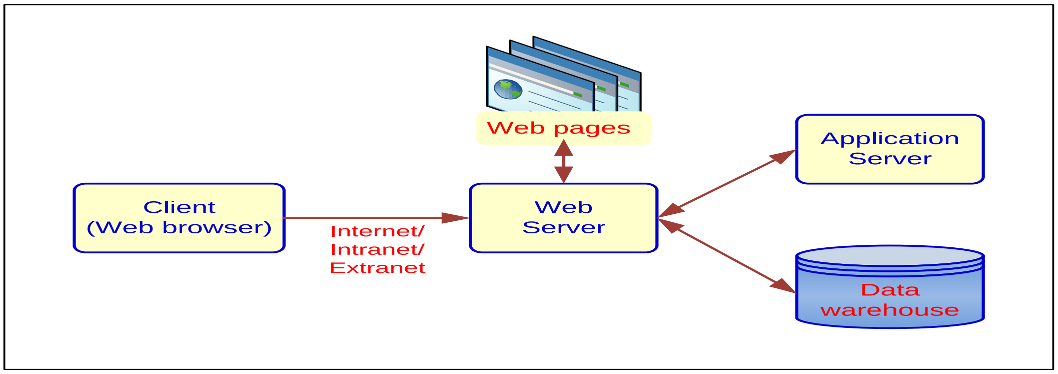
  First two tiers in three-tier architecture is combined into one

  ... sometimes there is only one tier?

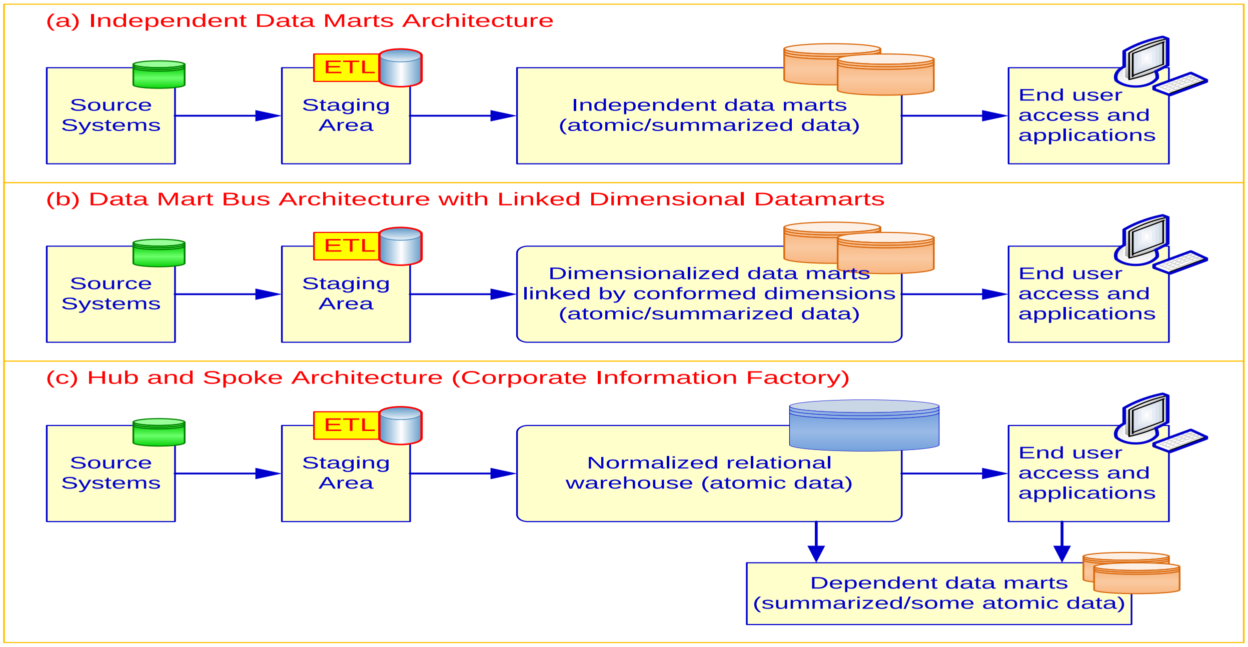
## Data Warehousing Architectures

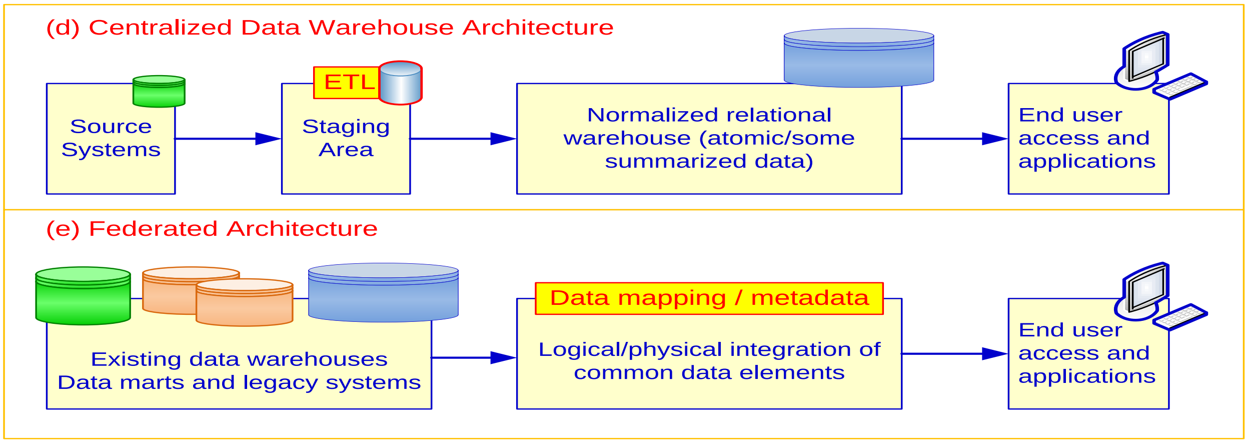
- Issues to consider when deciding which architecture to use:
  - Which database management system (DBMS) should be used?
  - Will parallel processing and/or partitioning be used?
  - Will data migration tools be used to load the data warehouse?
  - What tools will be used to support data retrieval and analysis?

## A Web-Based DW Architecture



## Alternative DW Architectures


(a) Independent Data Marts Architecture
Source Systems
ETL
Staging Area
Independent data marts (atomic/summarized data)
End user access and applications
(b) Data Mart Bus Architecture with Linked Dimensional Datamarts
Source Systems
ETL
Staging Area
Dimensionalized data marts linked by conformed dimensions (atomic/summarized data)
End user access and applications
(c) Hub and Spoke Architecture (Corporate Information Factory)
Source Systems
ETL
Staging Area
Normalized relational warehouse (atomic data)
End user access and applications
Dependent data marts (summarized/some atomic data)

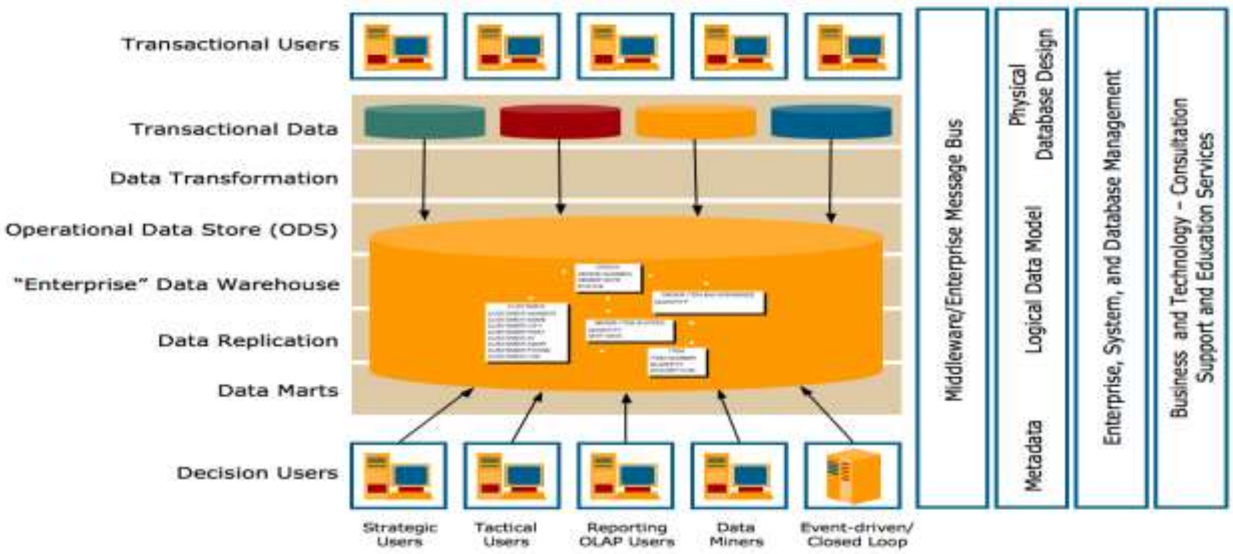

(d) Centralized Data Warehouse Architecture
Source Systems
ETL
Staging Area
Normalized relational warehouse (atomic/some summarized data)
End user access and applications
(e) Federated Architecture
Existing data warehouses Data marts and legacy systems
Data mapping / metadata
Logical/physical integration of common data elements
End user access and applications

- Each architecture has advantages and disadvantages!
- Which architecture is the best?

## Ten factors that potentially affect the architecture selection decision

1. Information interdependence between organizational units
2. Upper management's information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors

## Teradata Corp. DW Architecture


Transactional Users
Transactional Data
Data Transformation
Operational Data Store (ODS)
"Enterprise" Data Warehouse
Data Replication
Data Marts
Decision Users
Strategic Users
Tactical Users
Reporting OLAP Users
Data Miners
Event-driven/ Closed Loop
Middleware/Enterprise Message Bus
Physical Database Design
Logical Data Model
Metadata
Enterprise, System, and Database Management
Business and Technology - Consultation Support and Education Services

## Data Integration and the Extraction, Transformation, and Load Process

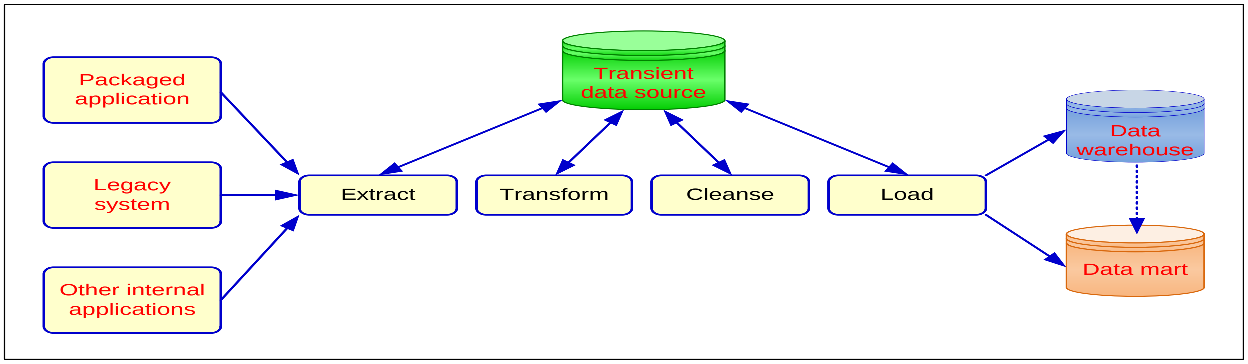
- ETL = Extract Transform Load
- Data integration

  Integration that comprises three major processes: data access, data federation, and change capture.
- Enterprise application integration (EAI)

  A technology that provides a vehicle for pushing data from source systems into a data warehouse
- Enterprise information integration (EII)

  An evolving tool space that promises real-time data integration from a variety of sources, such as relational or multidimensional databases, Web services, etc.

## Data Integration and the Extraction, Transformation, and Load Process



## ETL (Extract, Transform, Load)

- Issues affecting the purchase of an ETL tool
  - Data transformation tools are expensive
  - Data transformation tools may have a long learning curve
- Important criteria in selecting an ETL tool
  - Ability to read from and write to an unlimited number of data sources/architectures
  - Automatic capturing and delivery of metadata
  - A history of conforming to open standards
  - An easy-to-use interface for the developer and the functional user

## Data Warehouse Development

Data warehouse development approaches

- Inmon Model: EDW approach (top-down)
- Kimball Model: Data mart approach  (bottom-up)
- Which model is best?

- Table 3.3 provides a comparative analysis between EDW and Data Mart approach
- One alternative is the hosted warehouse

## Additional DW Considerations Hosted Data Warehouses

- Benefits:
  - Requires minimal investment in infrastructure
  - Frees up capacity on in-house systems
  - Frees up cash flow
  - Makes powerful solutions affordable
  - Enables solutions that provide for growth
  - Offers better quality equipment and software
  - Provides faster connections
  - ... more in the book

## Representation of Data in DW

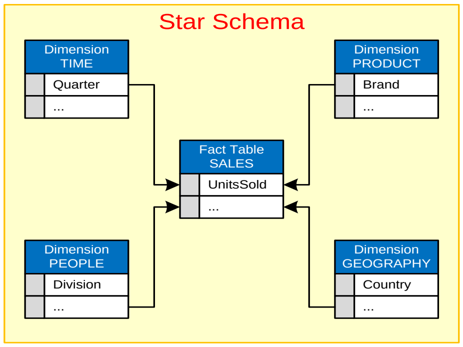
- Dimensional Modeling
  - A retrieval-based system that supports high-volume query access
- Star schema
  - The most commonly used and the simplest style of dimensional modeling
  - Contain a fact table surrounded by and connected to several dimension tables
- Snowflakes schema
  - An extension of star schema where the diagram resembles a snowflake in shape

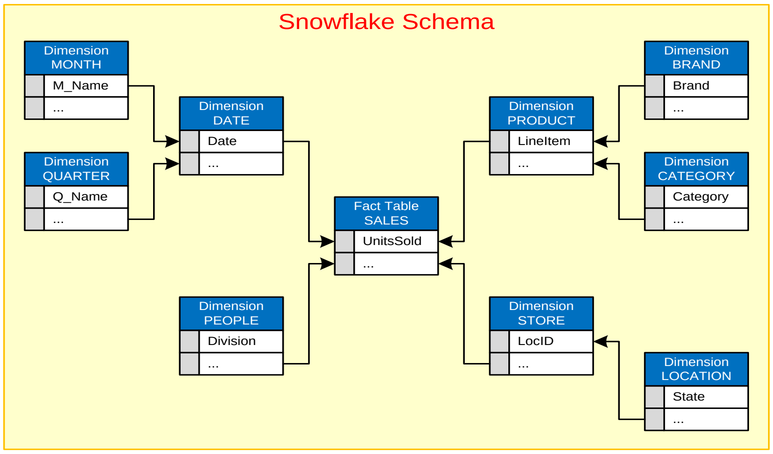
## Multidimensionality

The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)

- Multidimensional presentation
  - Dimensions: products, salespeople, market segments, business units, geographical locations, distribution channels, country, or industry
  - Measures: money, sales volume, head count, inventory profit, actual versus forecast
  - Time: daily, weekly, monthly, quarterly, or yearly

## Star versus Snowflake Schema





## Analysis of Data in DW

- OLTP vs. OLAP…
    - OLTP (online transaction processing)
  - Capturing and storing data from ERP, CRM, POS, …
  - The main focus is on efficiency of routine tasks
    - OLAP (Online analytical processing)
  - Converting data into information for decision support
  - Data cubes, drill-down / rollup, slice & dice, …
  - Requesting ad hoc reports
  - Conducting statistical and other analyses
  - Developing multimedia-based applications
  - …more in the book

## OLAP vs. OLTP

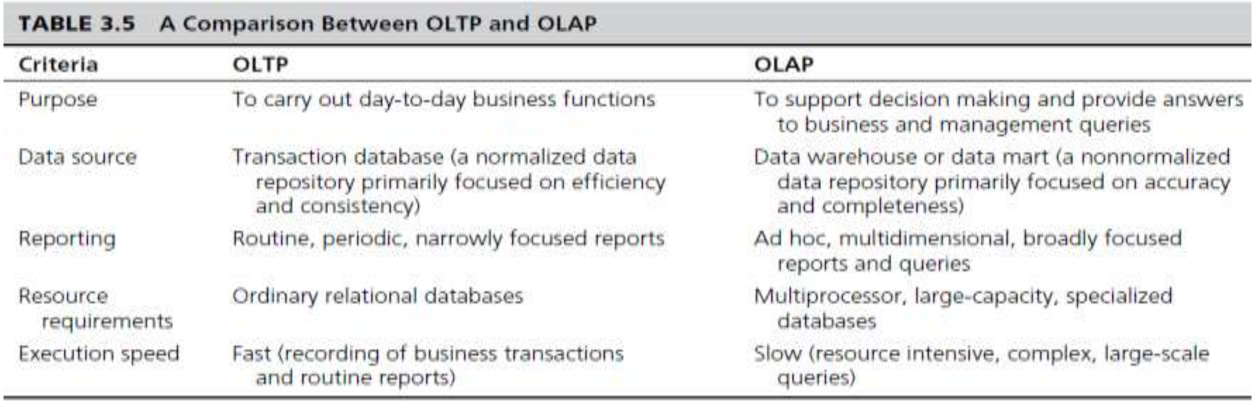


**TABLE 3.5 A Comparison Between OLTP and OLAP**

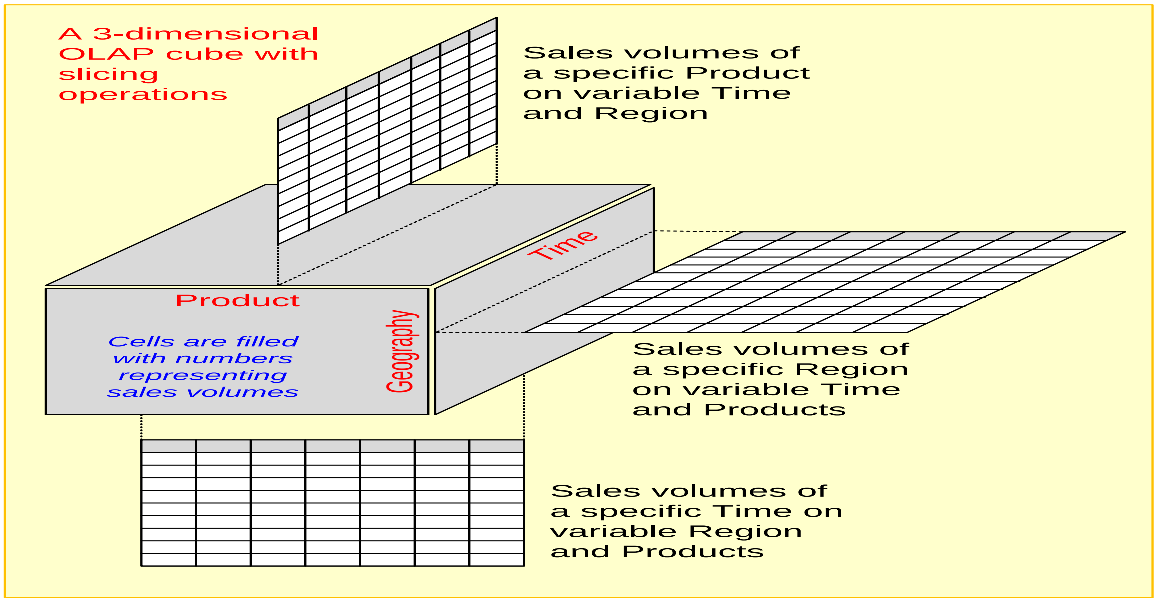
| Criteria | OLTP | OLAP |
|---|---|---|
| Purpose | To carry out day-to-day business functions | To support decision making and provide answers to business and management queries |
| Data source | Transaction database (a normalized data repository primarily focused on efficiency and consistency) | Data warehouse or data mart (a nonnormalized data repository primarily focused on accuracy and completeness) |
| Reporting | Routine, periodic, narrowly focused reports | Ad hoc, multidimensional, broadly focused reports and queries |
| Resource requirements | Ordinary relational databases | Multiprocessor, large-capacity, specialized databases |
| Execution speed | Fast (recording of business transactions and routine reports) | Slow (resource intensive, complex, large-scale queries) |

## OLAP Operations

- Slice - a subset of a multidimensional array
- Dice - a slice on more than two dimensions
- Drill Down/Up - navigating among levels of data ranging from the most summarized (up) to the most detailed (down)
- Roll Up - computing all of the data relationships for one or more dimensions
- Pivot - used to change the dimensional orientation of a report or an ad hoc query-page display

## OLAP

### Slicing Operations on a Simple Tree-Dimensional Data Cube


A 3-dimensional OLAP cube with slicing operations
Sales volumes of a specific Product on variable Time and Region
Time
Product
Cells are filled with numbers representing sales volumes
Geography
Sales volumes of a specific Region on variable Time and Products
Sales volumes of a specific Time on variable Region and Products

## Variations of OLAP

- Multidimensional OLAP (MOLAP)

OLAP implemented via a specialized multidimensional database (or data store) that summarizes transactions into multidimensional views ahead of time

- Relational OLAP (ROLAP)

The implementation of an OLAP database on top of an existing relational database Database OLAP and Web OLAP (DOLAP and WOLAP); Desktop OLAP,…

## DW Implementation Issues

- Identification of data sources and governance
- Data quality planning, data model design
- ETL tool selection
- Establishment of service-level agreements
- Data transport, data conversion
- Reconciliation process
- End-user support
- Political issues

## Successful DW Implementation : Things to Avoid

- Starting with the wrong sponsorship chain
- Setting expectations that you cannot meet
- Engaging in politically naive behavior
- Loading the data warehouse with information just because it is available
- Believing that data warehousing database design is the same as transactional database design

- Choosing a data warehouse manager who is technology oriented rather than user oriented

## Failure Factors in DW Projects

- Lack of executive sponsorship
- Unclear business objectives
- Cultural issues being ignored
    - Change management
- Unrealistic expectations
- Inappropriate architecture
- Low data quality / missing information
- Loading data just because it is available

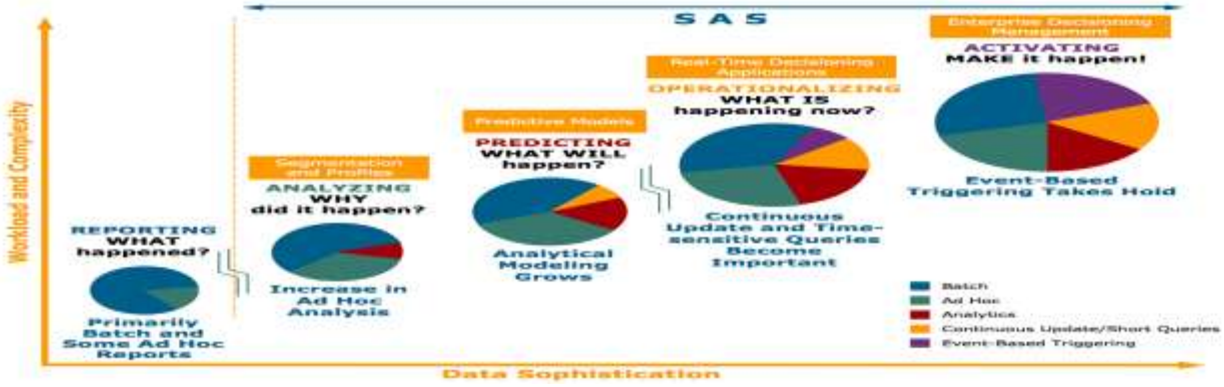
## Massive DW and Scalability

- Scalability
    - The main issues pertaining to scalability:
        - The amount of data in the warehouse
        - How quickly the warehouse is expected to grow
        - The number of concurrent users
        - The complexity of user queries
    - Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse

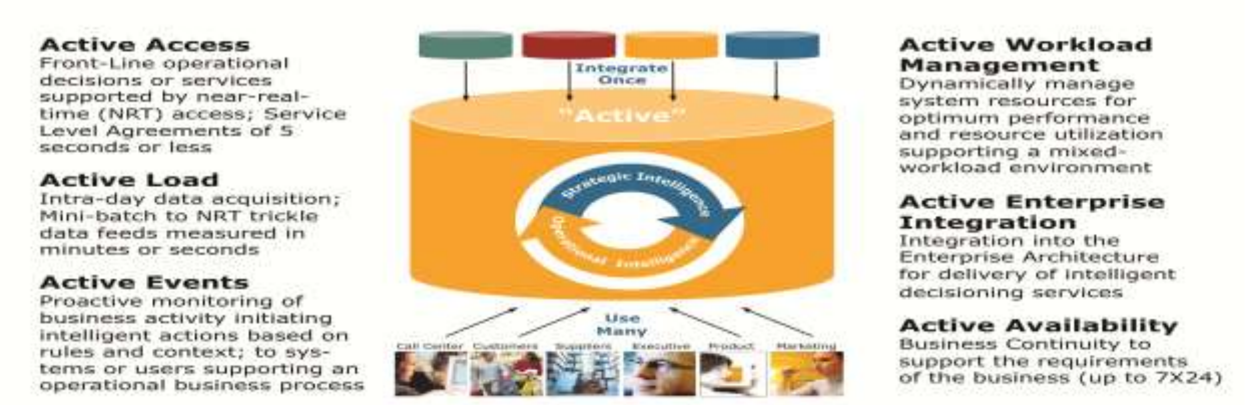
## Real-Time/Active DW/BI

- Enabling real-time data updates for real-time analysis and real-time decision making is growing rapidly
    - Push vs. Pull (of data)
- Concerns about real-time BI
    - Not all data should be updated continuously
    - Mismatch of reports generated minutes apart
    - May be cost prohibitive
    - May also be infeasible

## Enterprise Decision Evolution and Data Warehousing



## Real-Time/Active DW at Teradata



**Active Access**
Front-Line operational decisions or services supported by near-real-time (NRT) access; Service Level Agreements of 5 seconds or less

**Active Load**
Intra-day data acquisition; Mini-batch to NRT trickle data feeds measured in minutes or seconds

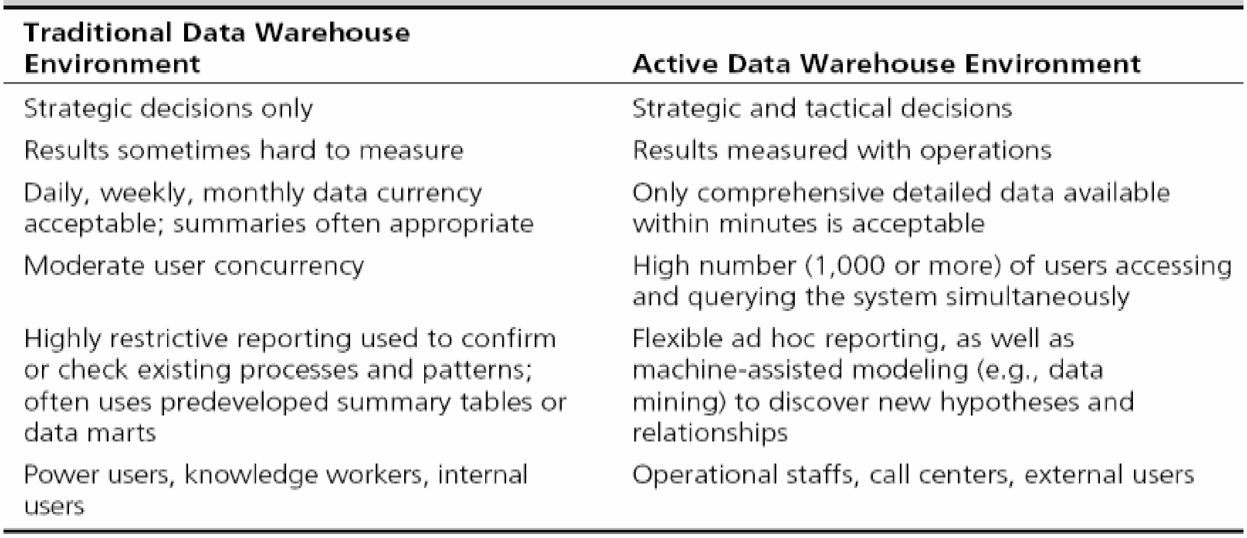
**Active Events**
Proactive monitoring of business activity initiating intelligent actions based on rules and context; to systems or users supporting an operational business process

**Active Workload Management**
Dynamically manage system resources for optimum performance and resource utilization supporting a mixed-workload environment

**Active Enterprise Integration**
Integration into the Enterprise Architecture for delivery of intelligent decisioning services

**Active Availability**
Business Continuity to support the requirements of the business (up to 7X24)

## Traditional versus Active DW



| Traditional Data Warehouse Environment | Active Data Warehouse Environment |
|---|---|
| Strategic decisions only | Strategic and tactical decisions |
| Results sometimes hard to measure | Results measured with operations |
| Daily, weekly, monthly data currency acceptable; summaries often appropriate | Only comprehensive detailed data available within minutes is acceptable |
| Moderate user concurrency | High number (1,000 or more) of users accessing and querying the system simultaneously |
| Highly restrictive reporting used to confirm or check existing processes and patterns; often uses predeveloped summary tables or data marts | Flexible ad hoc reporting, as well as machine-assisted modeling (e.g., data mining) to discover new hypotheses and relationships |
| Power users, knowledge workers, internal users | Operational staffs, call centers, external users |

## DW Administration and Security

- Data warehouse administrator (DWA)
  - DWA should...
    - have the knowledge of high-performance software, hardware and networking technologies
    - possess solid business knowledge and insight
    - be familiar with the decision-making processes so as to suitably design/maintain the data warehouse structure
    - possess excellent communications skills
- Security and privacy is a pressing issue in DW
  - Safeguarding the most valuable assets
  - Government regulations (HIPAA, etc.)
  - Must be explicitly planned and executed

## The Future of DW

- Sourcing...
  - *Web, social media, and Big Data*
  - Open source software
  - SaaS (software as a service)
  - Cloud computing
- Infrastructure...
  - Columnar
  - Real-time DW
  - Data warehouse appliances
  - Data management practices/technologies
  - In-database & In-memory processing New DBMS
  - Advanced analytics