Chapter 2:Data Warehousing

By: Shohreh Ajoudanian

2014 (1392)

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"

Characteristics of DW

- Subject oriented
- Integrated
- Time-variant (time series)
- Nonvolatile
- Summarized
- Not normalized
- Metadata
- Web based, relational/multi-dimensional
- Client/server
- Real-time and/or right-time (active)

Data Mart

A departmental data warehouse that stores only 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

Data Warehousing Definitions

- Operational data stores (ODS)

  A type of database often used as an interim area for a data warehouse

- Oper marts

  An operational data mart

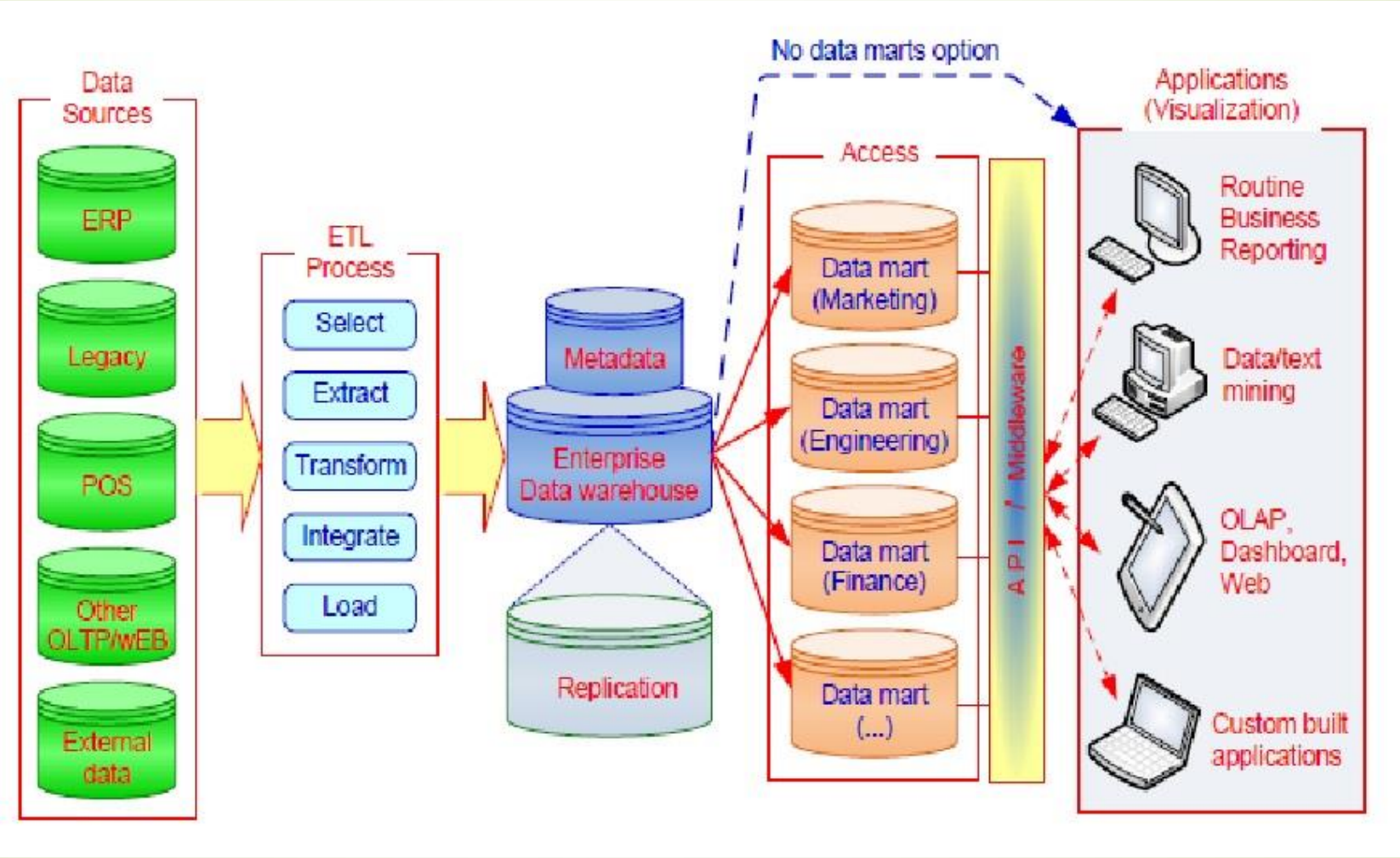
- Enterprise data warehouse (EDW)

  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

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

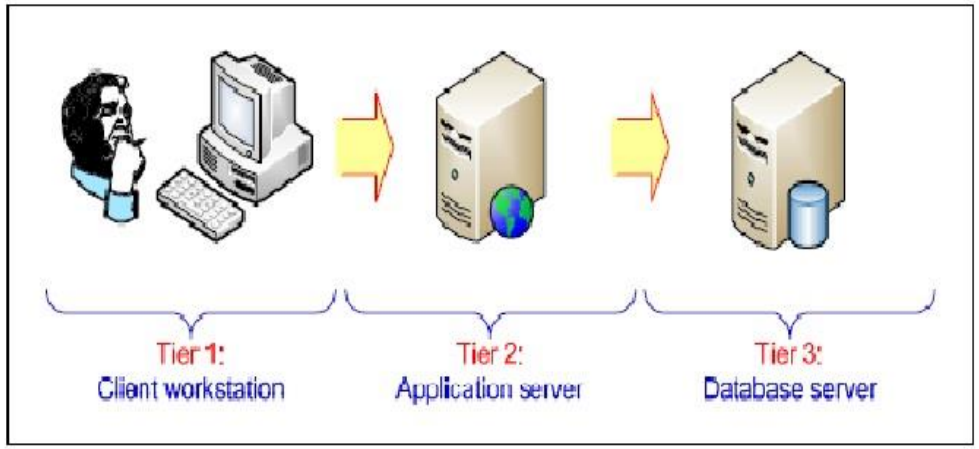
- Two-tier architecture

  First 2 tiers in three-tier architecture is combined into one

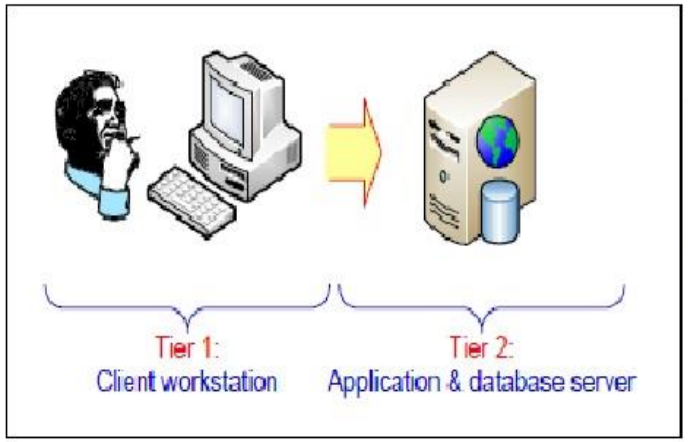
Sometimes there is only one tier

DW Architectures

3-tier architecture

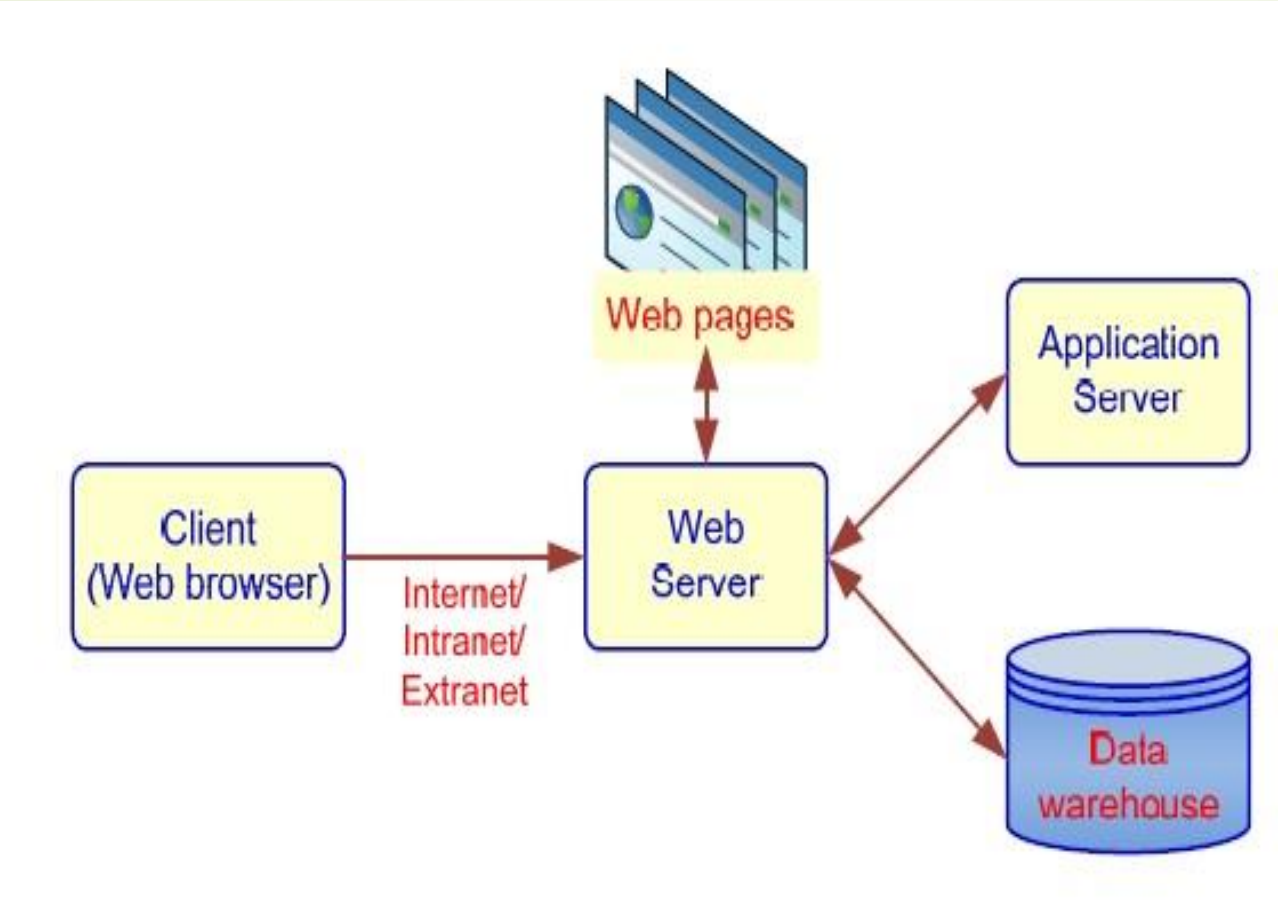


2-tier architecture

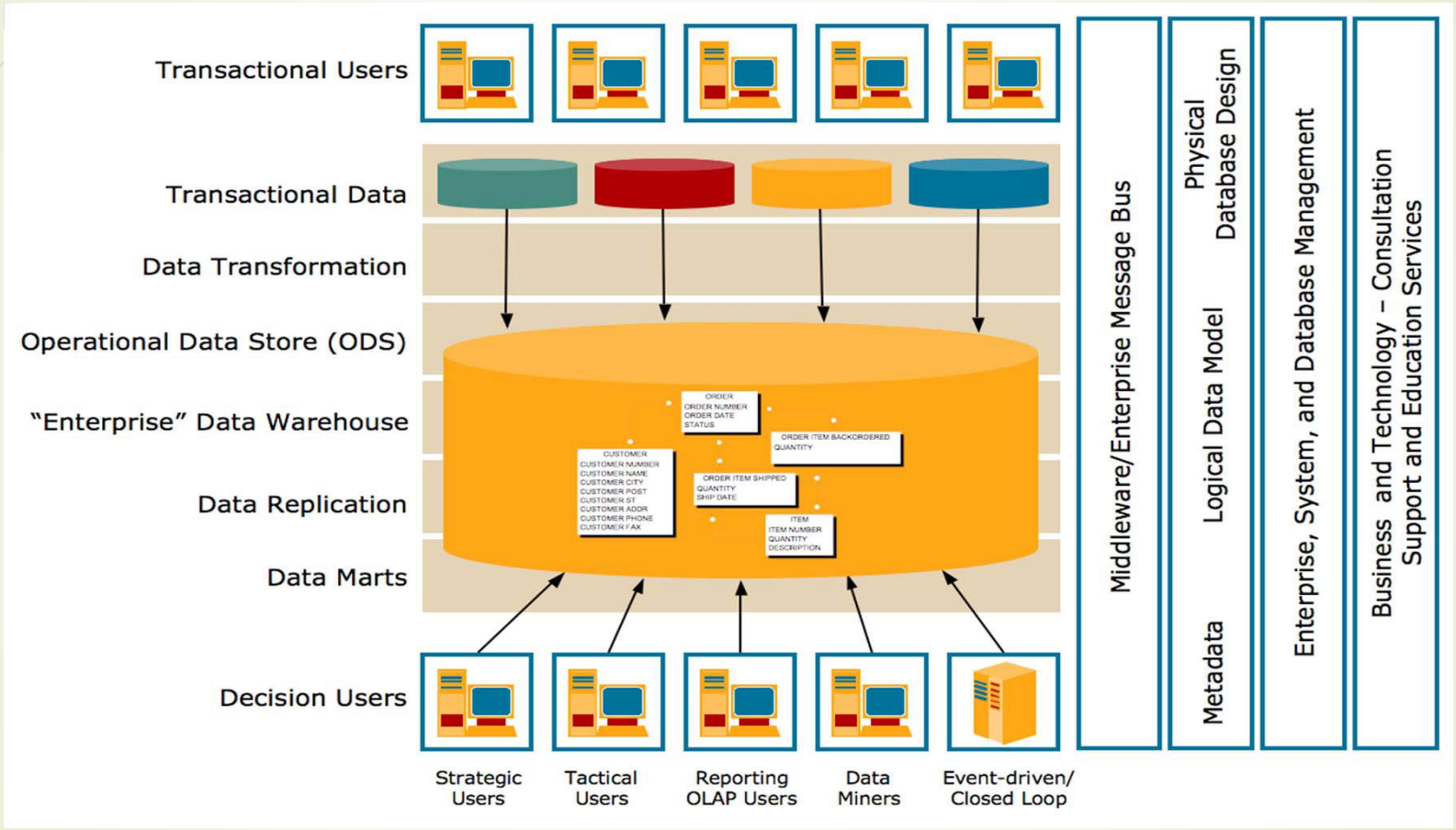


1-tier Architecture

A Web-based DW Architecture



Teradata Corp. DW Architecture



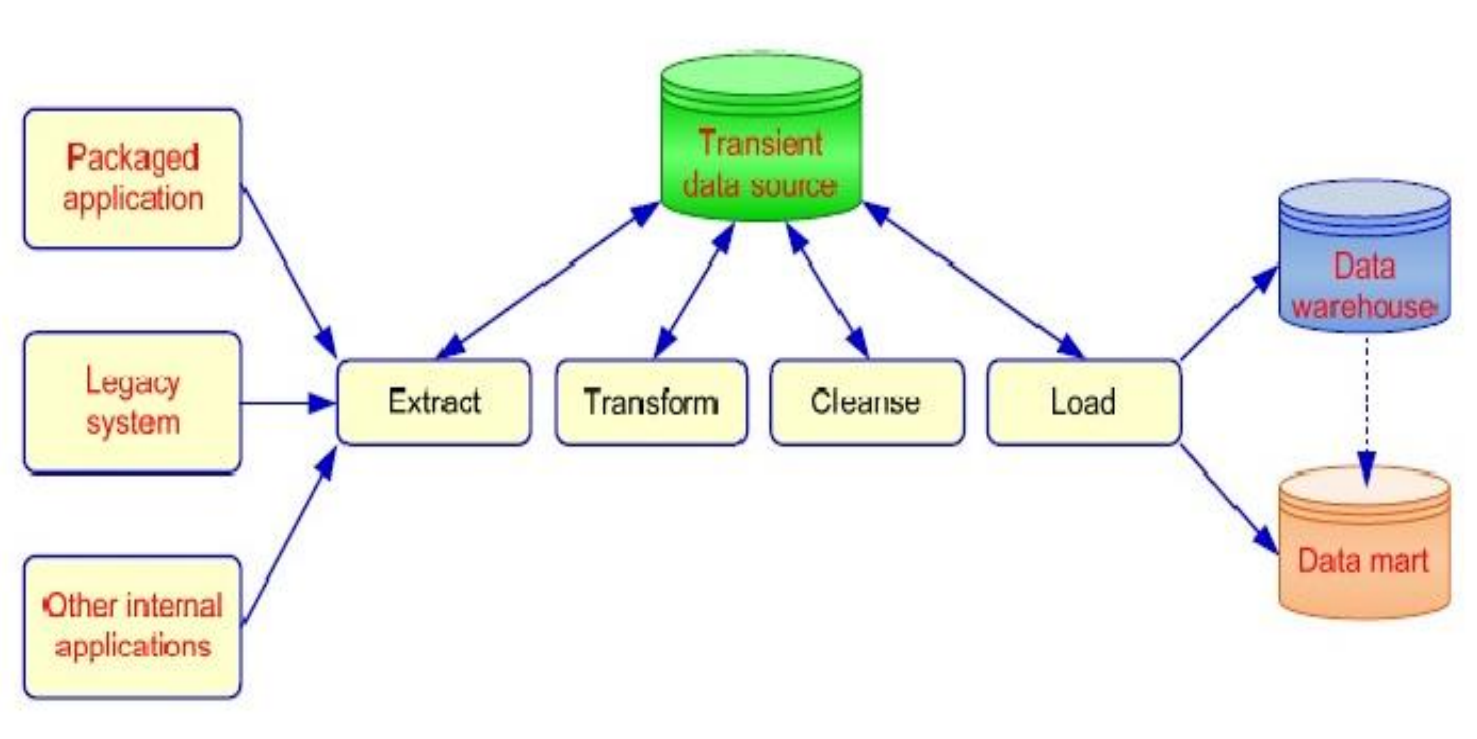
Data Warehousing Architectures

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

Data Integration and the Extraction , Transformation , and Load (ETL) Process

Extraction, transformation, and load (ETL)



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
  - Fact table contains the descriptive attributes (numerical values) needed to perform decision analysis and query reporting
  - Dimension tables contain classification and aggregation information about the values in the fact table
- Snowflakes schema – an extension of star schema where the diagram resembles a snowflake in shape

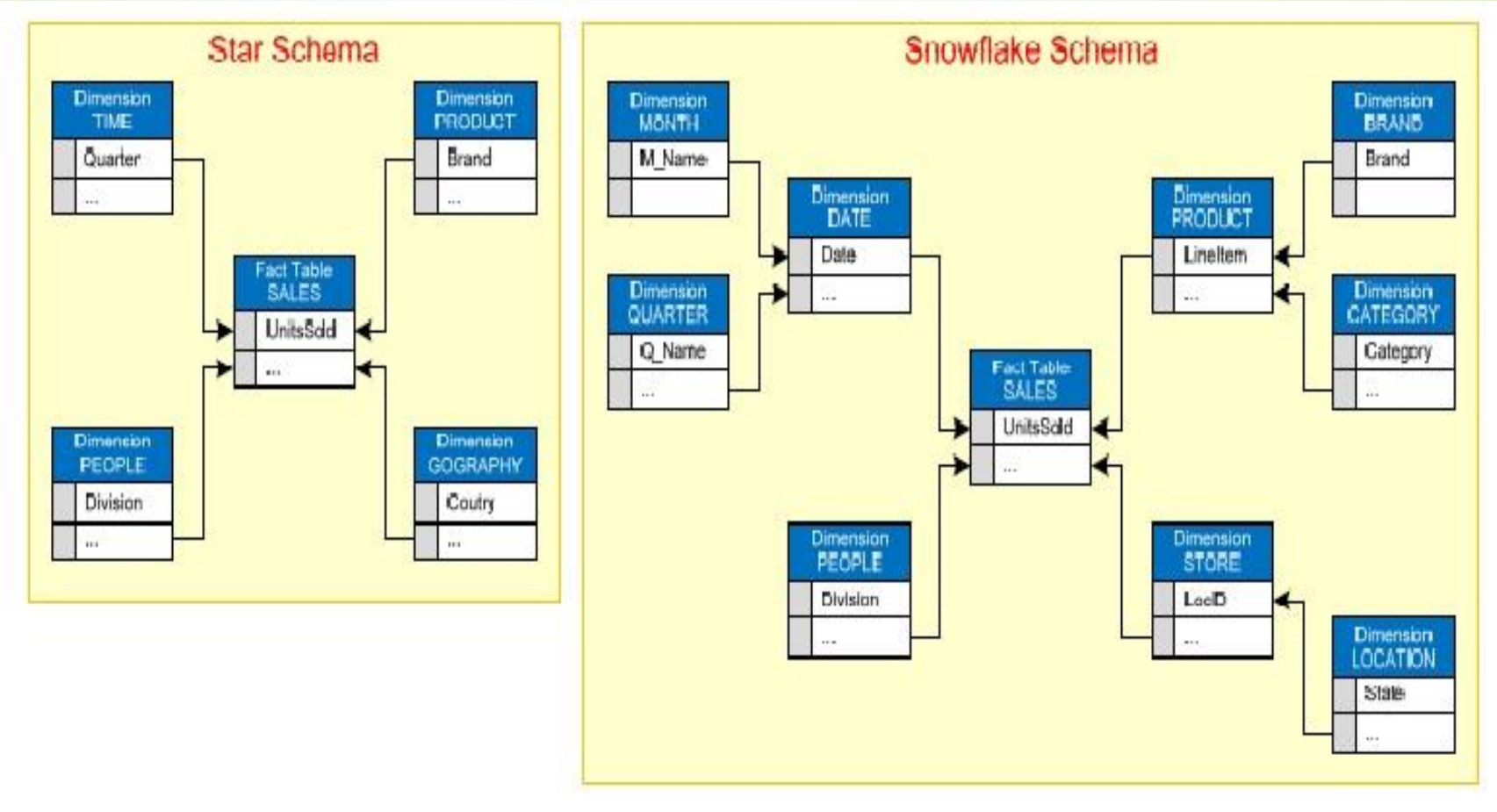
Multidimensionality

- 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 vs Snowflake Schema



Analysis of Data in DW

- Online analytical processing (OLAP)
  - Data driven activities performed by end users to query the online system and to conduct analyses
  - Data cubes, drill-down / rollup, slice & dice, …
- OLAP Activities
  - Generating queries (query tools)
  - Requesting ad hoc reports
  - Conducting statistical and other analyses
  - Developing multimedia-based applications

Analysis of Data Stored in DW OLTP vs. OLAP

- OLTP (online transaction processing)
  - A system that is primarily responsible for capturing and storing data related to day-to-day business functions such as ERP, CRM, SCM, POS,
  - The main focus is on efficiency of routine tasks
- OLAP (online analytic processing)
  - A system is designed to address the need of information extraction by providing effectively and efficiently ad hoc analysis of organizational data
  - The main focus is on effectiveness

OLAP vs. OLTP

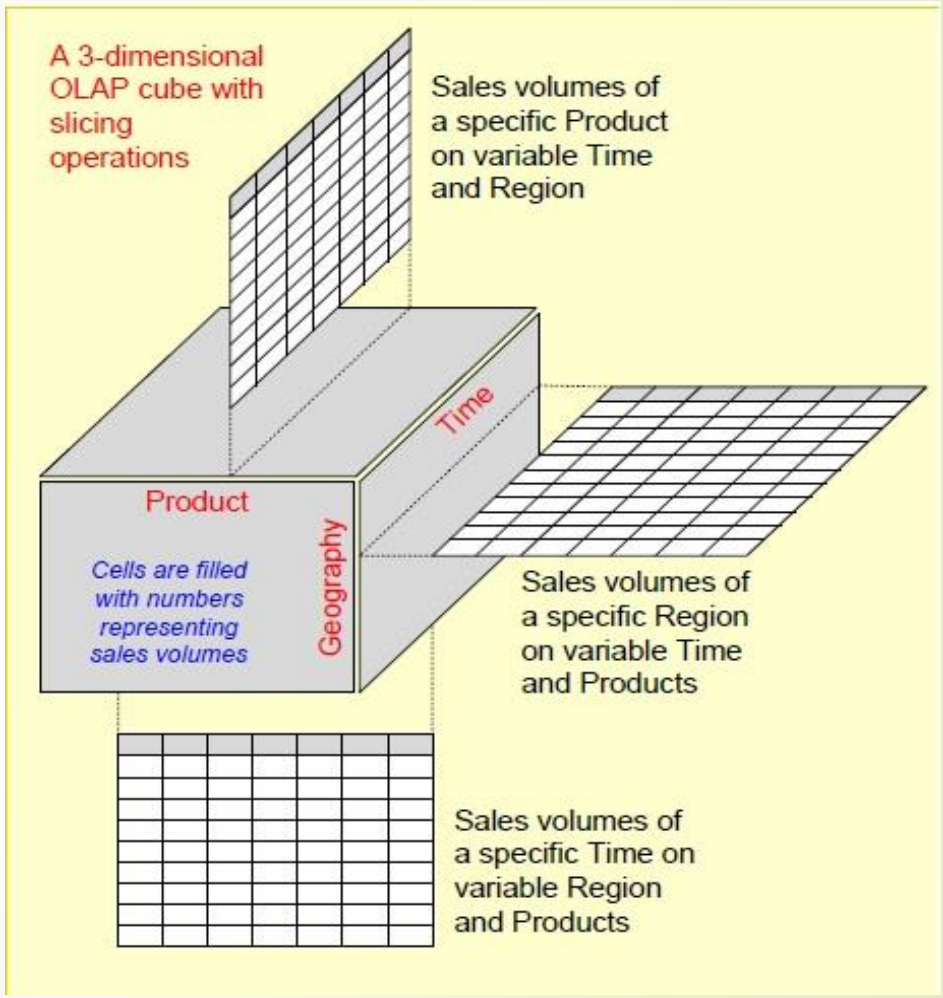
Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions.	To support decision making and provide answers to business and management queries.
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency).	Data warehouse or data mart (a non-normalized 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



DW Implementation Issues

- 11 tasks for successful DW implementation
  - Establishment of service-level agreements and data-refresh requirements
  - Identification of data sources and their governance policies
  - Data quality planning
  - Data model design
  - ETL tool selection
  - Relational database software and platform selection
  - Data transport
  - Data conversion
  - Reconciliation process
  - Purge and archive planning
  - End-user support

DW Implementation Guidelines

- Project must fit with corporate strategy & business objectives
- There must be complete buy-in to the project by executives, managers, and users
- It is important to manage user expectations about the completed project
- The data warehouse must be built incrementally
- Build in adaptability, flexibility and scalability
- The project must be managed by both IT and business professionals
- Only load data that have been cleansed and are of a quality understood by the organization
- Do not overlook training requirements

- Be politically aware

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

Successful DW Implementation Things to Avoid Cont.

- Focusing on traditional internal record oriented data and ignoring the value of external data and of text, images, etc.
- Delivering data with confusing definitions
- Believing promises of performance, capacity, and scalability
- Believing that your problems are over when the data warehouse is up and running
- Focusing on ad hoc data mining and periodic reporting instead of alerts

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

Real-time/Active DW/BI

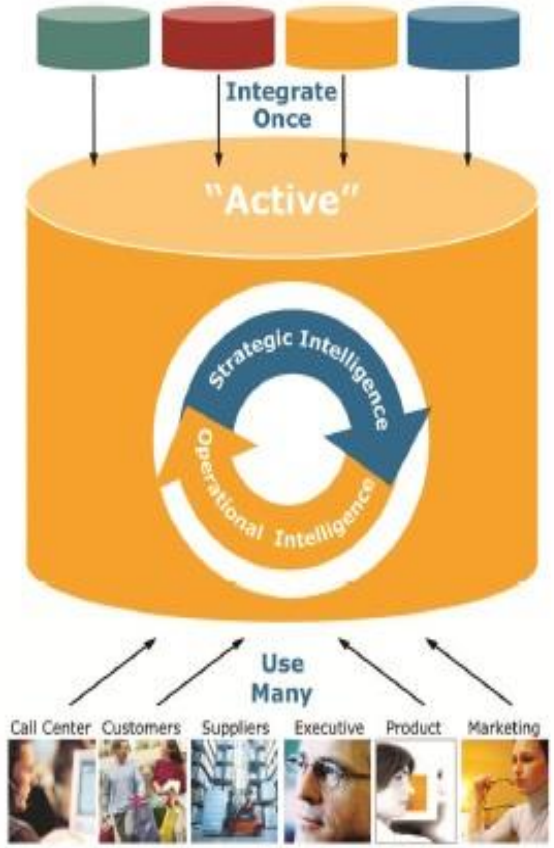
- Enabling real-time data updates for real-time analysis and real-time decision making is growing rapidly
  - Push vs. Pull (of data)
- Concerns about real-time BI
  - Not all data should be updated continuously
  - Mismatch of reports generated minutes apart
  - May be cost prohibitive
  - May also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

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 sys-tems or users supporting an operational business process

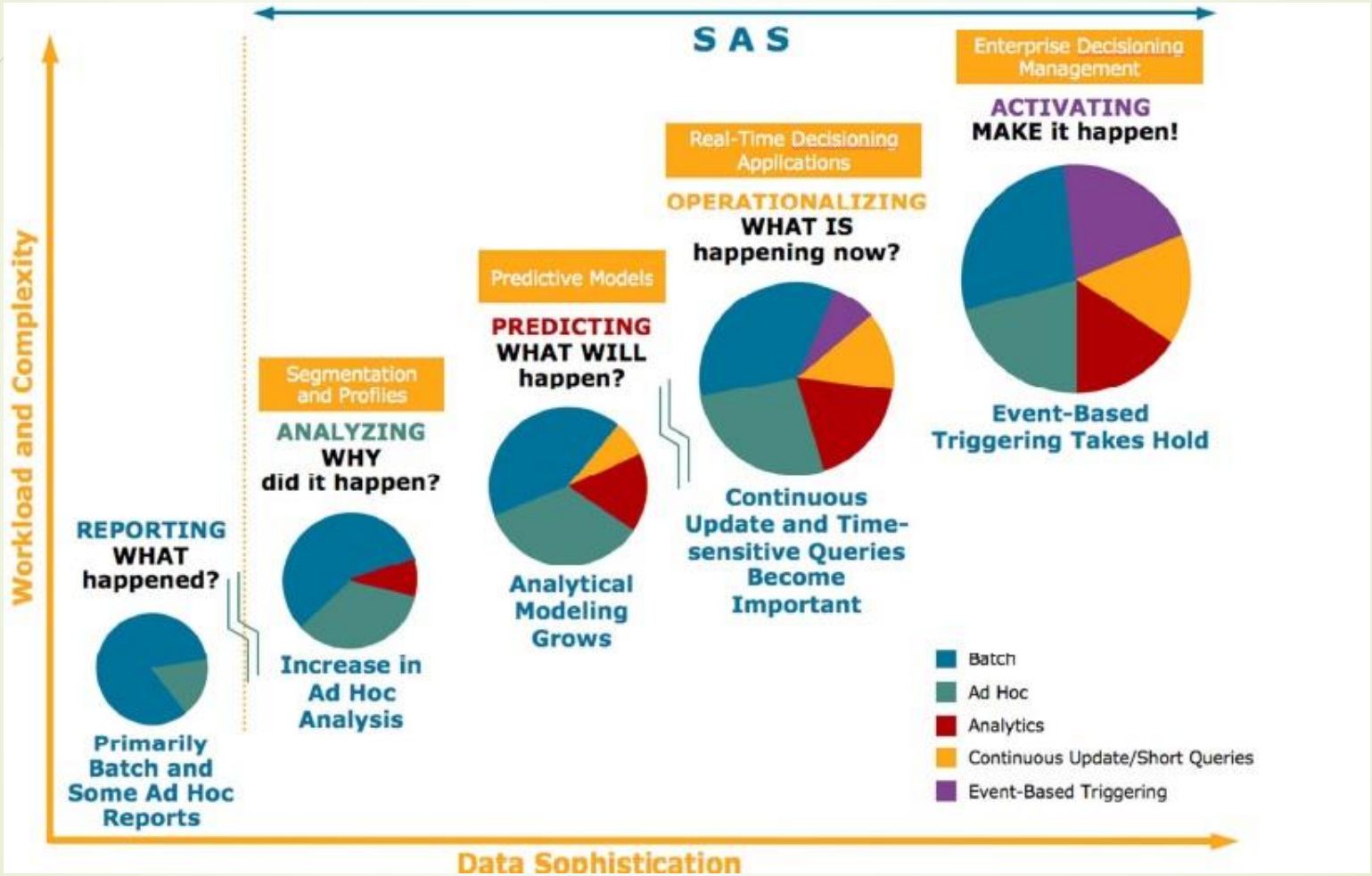


Active Workload Management
Dynamically manage system resources for optimum performance and resource utilization supporting a mixed-workload environment

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)

Enterprise Decision Evolution and DW

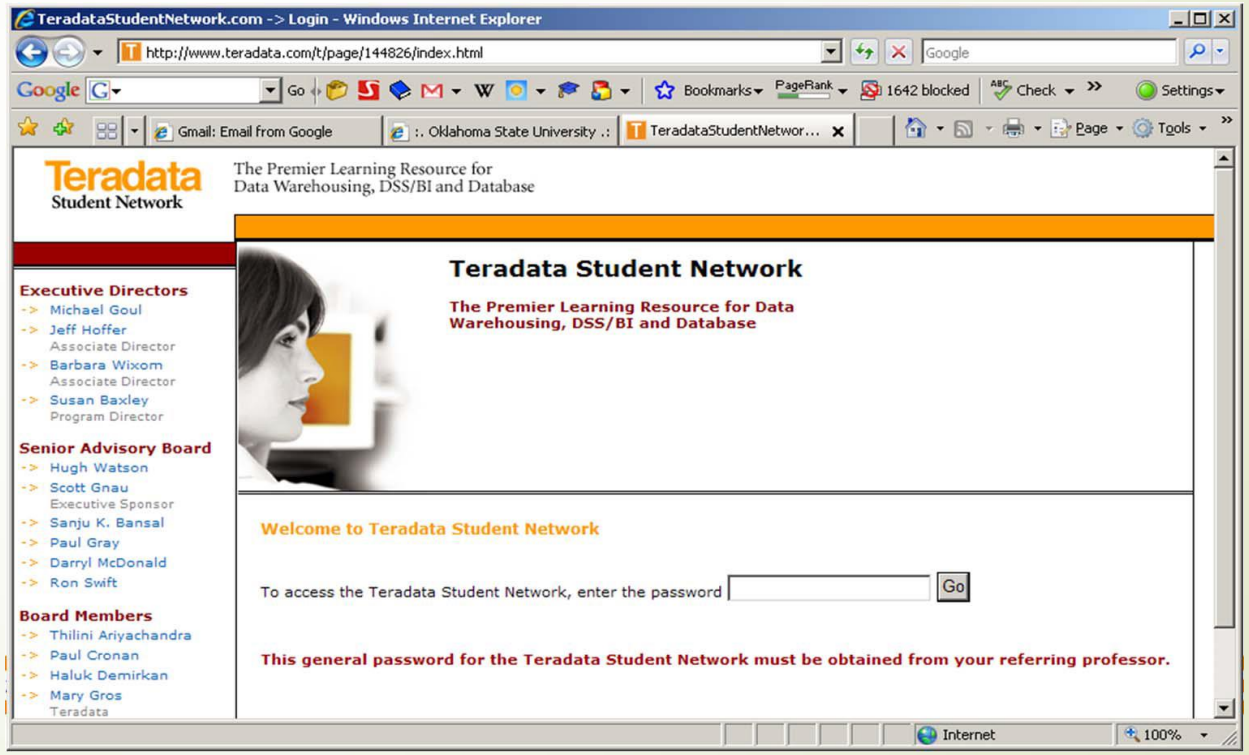


The Future of DW

- Sourcing…
  - Open source software
  - SaaS (software as a service)
  - Cloud computing
  - DW appliances
- Infrastructure…
  - Real-time DW
  - Data management practices/technologies
  - In-memory processing (“super-computing”)
  - New DBMS
  - Advanced analytics

BI / OLAP Portal for Learning

- MicroStrategy, and much more…
- www.TeradataStudentNetwork.com
- Pw: <check with TDUN>



www.shohrehajoudanian.ir
shajoudanian@yahoo.com