DATA WAREHOUSE

EGCO321 DATABASE SYSTEMS

KANAT POOLSAWASD
DEPARTMENT OF COMPUTER ENGINEERING
MAHIDOL UNIVERSITY

CHARACTERISTICS

- Data warehouse is a central repository for summarized and integrated data from operational databases and external data sources.
- The processing requirement of decision support applications have led to four distinguishing characteristics for data warehouses, as described in the following:
  - Subject-Oriented
  - Integrated
  - Time-Variant
  - Nonvolatile

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.

INTEGRATED

- Constructed by integrating multiple, heterogeneous data sources
  - Relational databases, flat files, on-line transaction records, webs (html), etc.
- 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, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

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”.

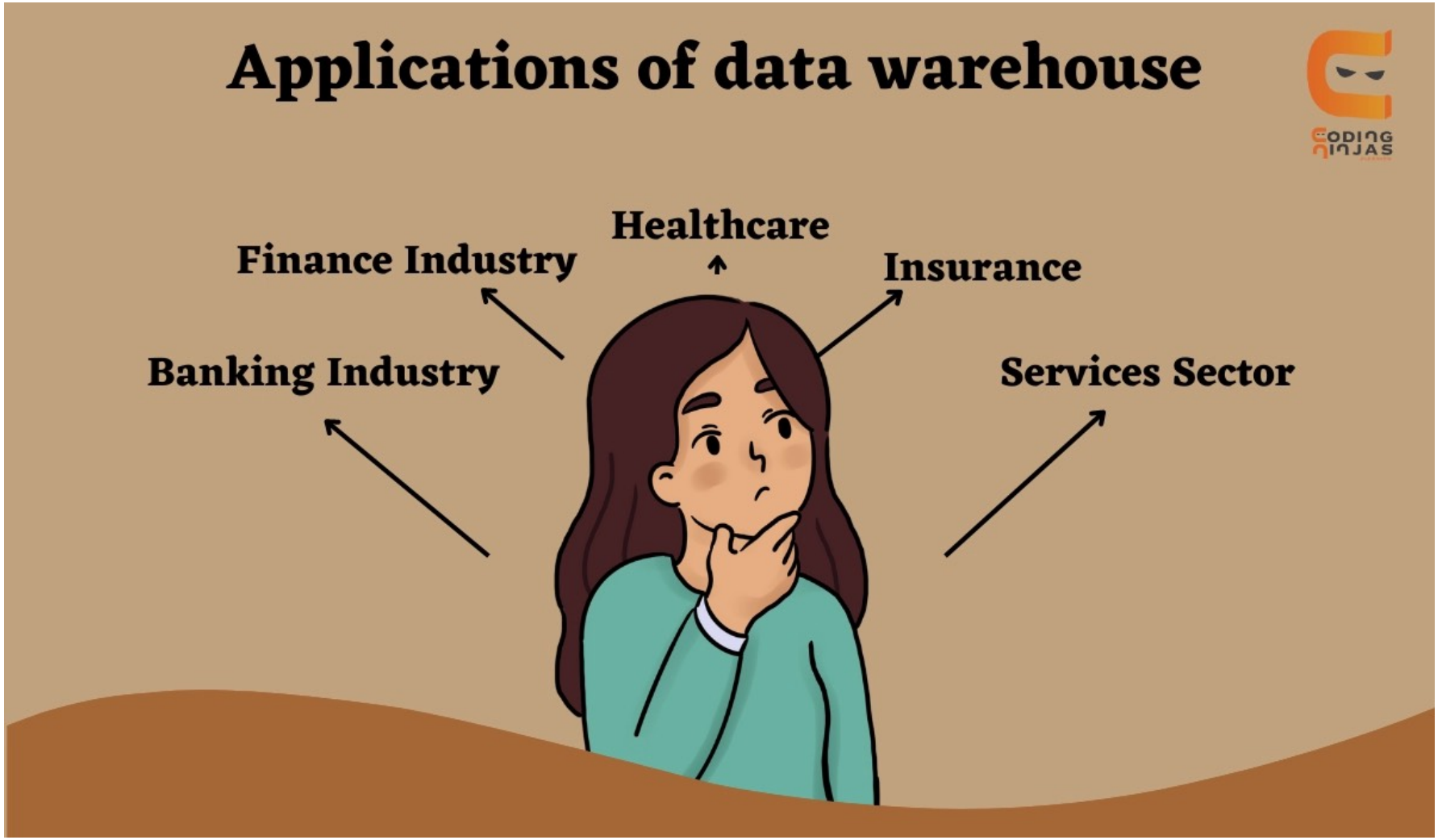
NON-VOLATILE

- 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*.

COMPARISON OF OPERATIONAL DATABASES AND DATA WAREHOUSES

Characteristic	Operational Database	Data Warehouse
Currency	Current	Historical
Detail level	Individual	Individual and summary
Orientation	Process orientation	Subject orientation
Number of records processed	Few	Thousands
Normalization level	Mostly normalized	Frequent violations of BCNF
Update level	Volatile	Nonvolatile (refreshed)
Data model	Relational	Relational model with star schemas and multidimensional model with data cubes

APPLICATION OF DATA WAREHOUSES

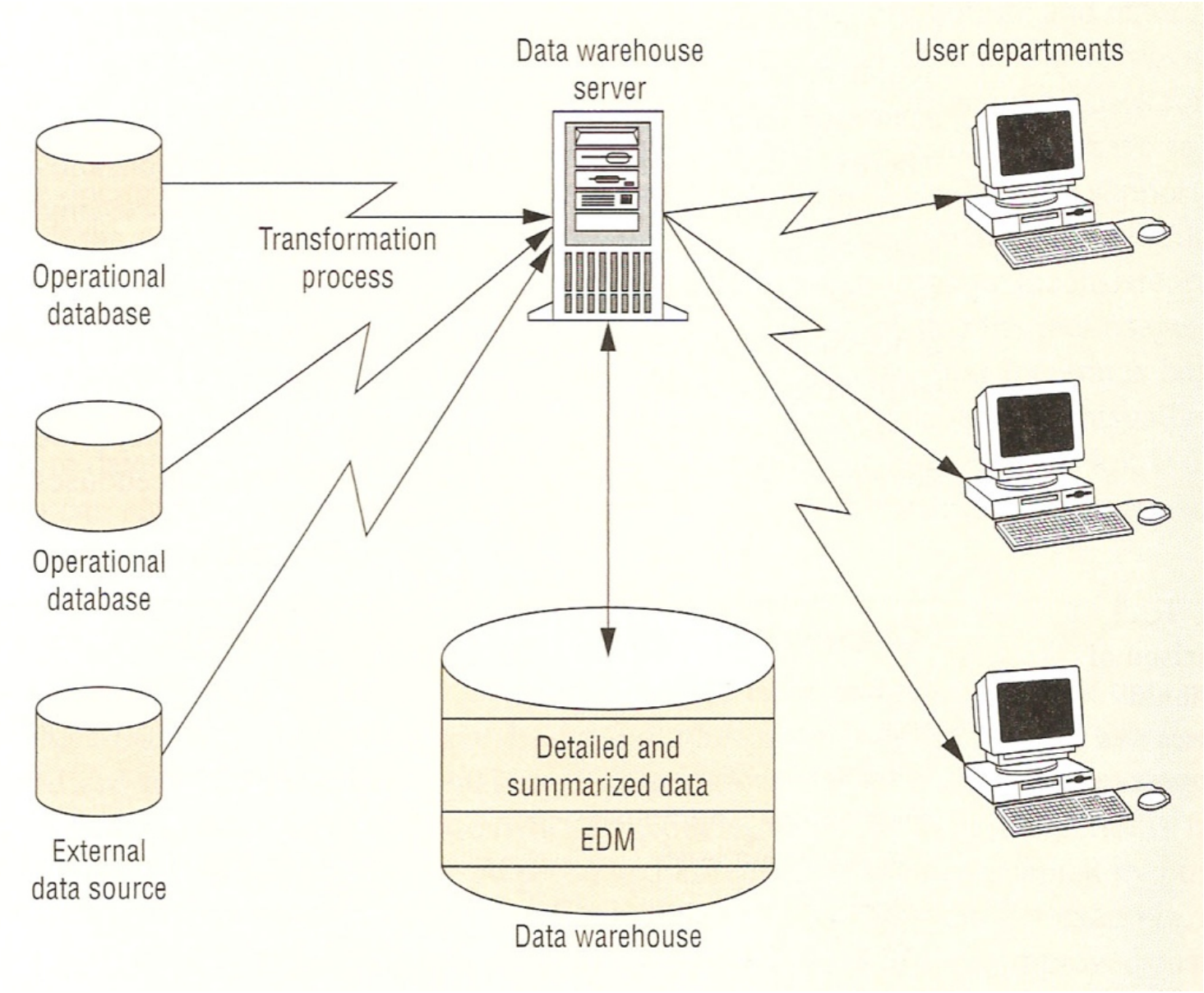


* https://www.codingninjas.com/studio/library/applications-of-data-warehousing

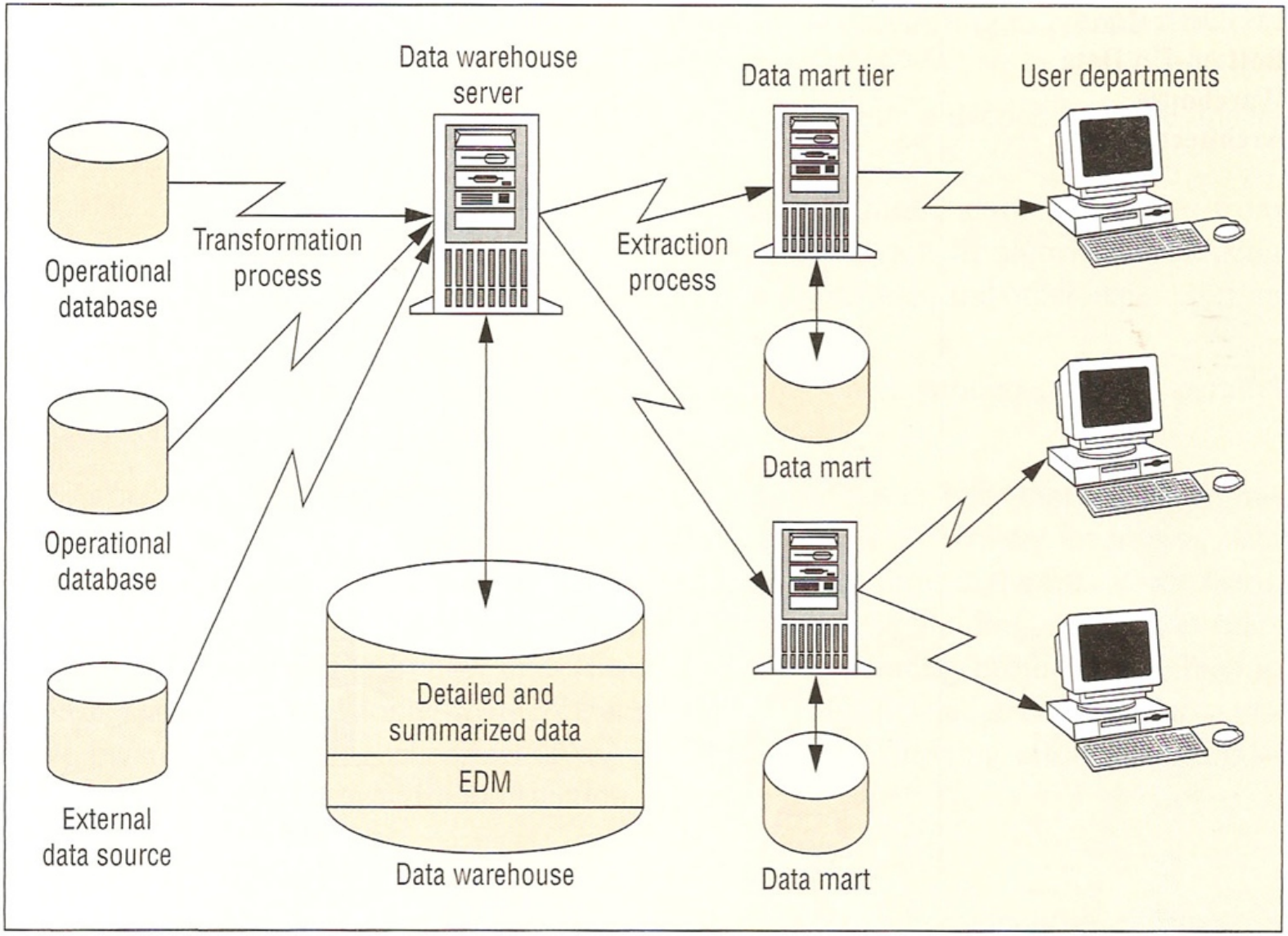
DATA WAREHOUSE ARCHITECTURE

- Enterprise Data Model is a conceptual data model of the data warehouse defining the structure of the data warehouse and the metadata to access and transform operational databases and external data sources.

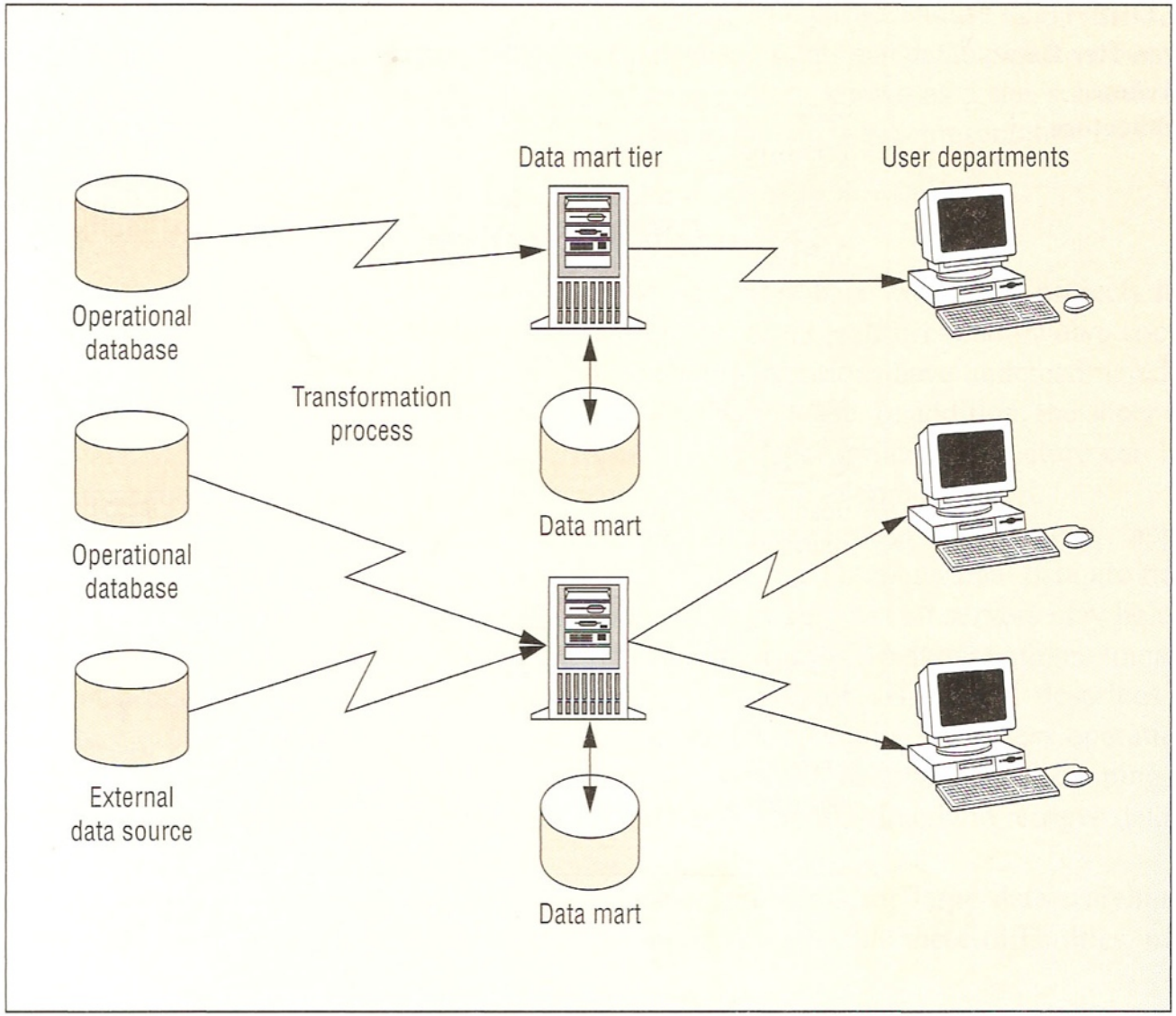
TWO-TIER DATA WAREHOUSE ARCHITECTURE



THREE-TIER DATA WAREHOUSE ARCHITECTURE



BOTTOM-UP DATA WAREHOUSE ARCHITECTURE



DATA MART

- Data mart is a subset or view of a data warehouse, typically at a department or functional level, that contains all data required for decision support tasks of that department.

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

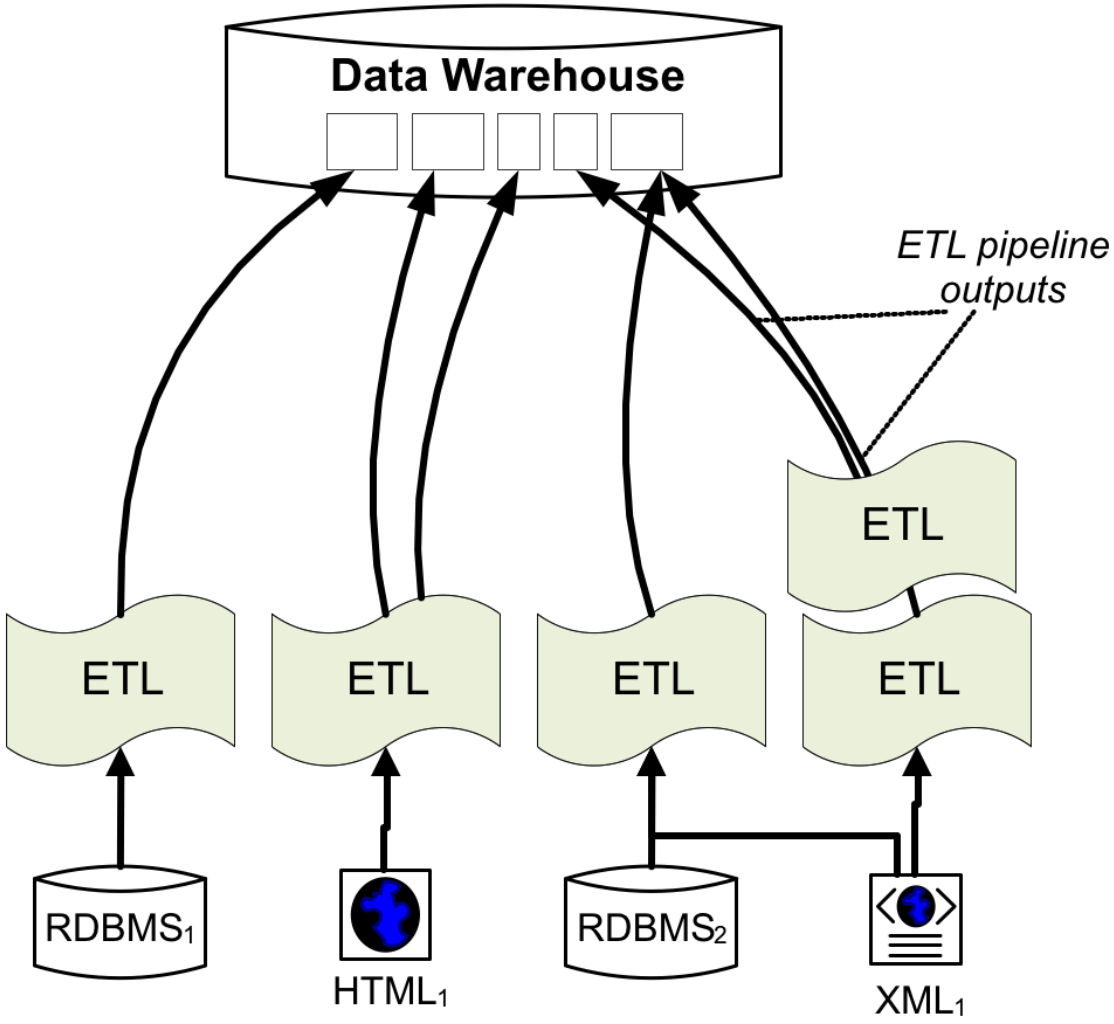
EXTRACT TRANSFORM LOAD (ETL)

- Data moved from source to target data bases.
- Focus: Preparing the data for analysis/reporting
  - Extract: get data from source(s) as efficiently as possible
  - Transform: perform calculations / map data / clean data.
  - Load: load data to target storage.

ETL TOOLS (1)

- To support complexity of data warehouse maintenance, software product known as Extraction, Transformation, and Load (ETL) tools have been developed.
- ETL tools are software tools for extraction, transformation, and loading of change data from data sources to a data warehouse.
- ETL tools eliminate the need to write custom coding for many data warehouse maintenance tasks.

ETL TOOLS (2)


Data Warehouse
ETL pipeline outputs
ETL
ETL
ETL
ETL
ETL
RDBMS1
HTML1
RDBMS2
XML1

ETL TOOLS (3)

- ETL tools are the equivalent of schema mappings in virtual integration, but are more powerful
- Arbitrary pieces of code to take data from a source, convert it into data for the warehouse:
  - import filters – read and convert from data sources
  - data transformations – join, aggregate, filter, convert data
  - de-duplication – finds multiple records referring to the same entity, merges them
  - profiling – builds tables, histograms, etc. to summarise data
  - quality management – test against master values, known business rules, constraints, etc.

DATA CLEANING - PARSING

- Parsing: locates and identifies individual data elements in the source file and then isolates these data elements in the target files.

Name
Harry Johnson
Mary J. Blair
Dr. Chris Young
Martin K. Brown II

Title	First Name	Middle Name	Last Name	Suffix
	Harry		Johnson	
	Mary	J.	Blair	
Dr.	Chris		Young	
	Martin	K.	Brown	II

- Combing: locates and identifies individual data elements in the source file and then combines these data elements in the target files.

Date	Month	Year
9	May	2016
06	04	2015

Date
2016-05-09
2015-04-06

DATA CLEANING - CORRECTING

- Correcting parsing individual data components using sophisticated data algorithms and secondary data sources
- Correct data according to data rules
- Example includes converting the combined date into a standard date format.

DATA CLEANING - STANDARDIZING

- Standardizing: applies conversion routines to transform data into its preferred (and consistent) format using standard and custom data rules.

Sex
Male
F
M
Female
NULL
f
m

Sex
1
2
1
2
0
2
1

DATA CLEANING - MATCHING

- Searching and matching records within and across the parsed, corrected and standardized data based on predefined data rules to eliminate duplications, sequences.

Pregnancy Number	Outcome Date	Outcome
1	2011-06-07	Twin

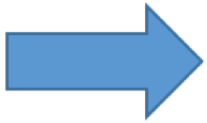
DoB	Name
2011-06-07	Child1
2011-06-07	Child2

Pregnancy Number	Birth Date	Name
1	2011-06-07	Child1
1	2011-06-07	Child2

DATA CLEANING - CONSOLIDATING

- Analyzing and identifying relationships between matched records and consolidating / merging them into correct representation.

Migration Dates
2006-05-09
1995-06-06

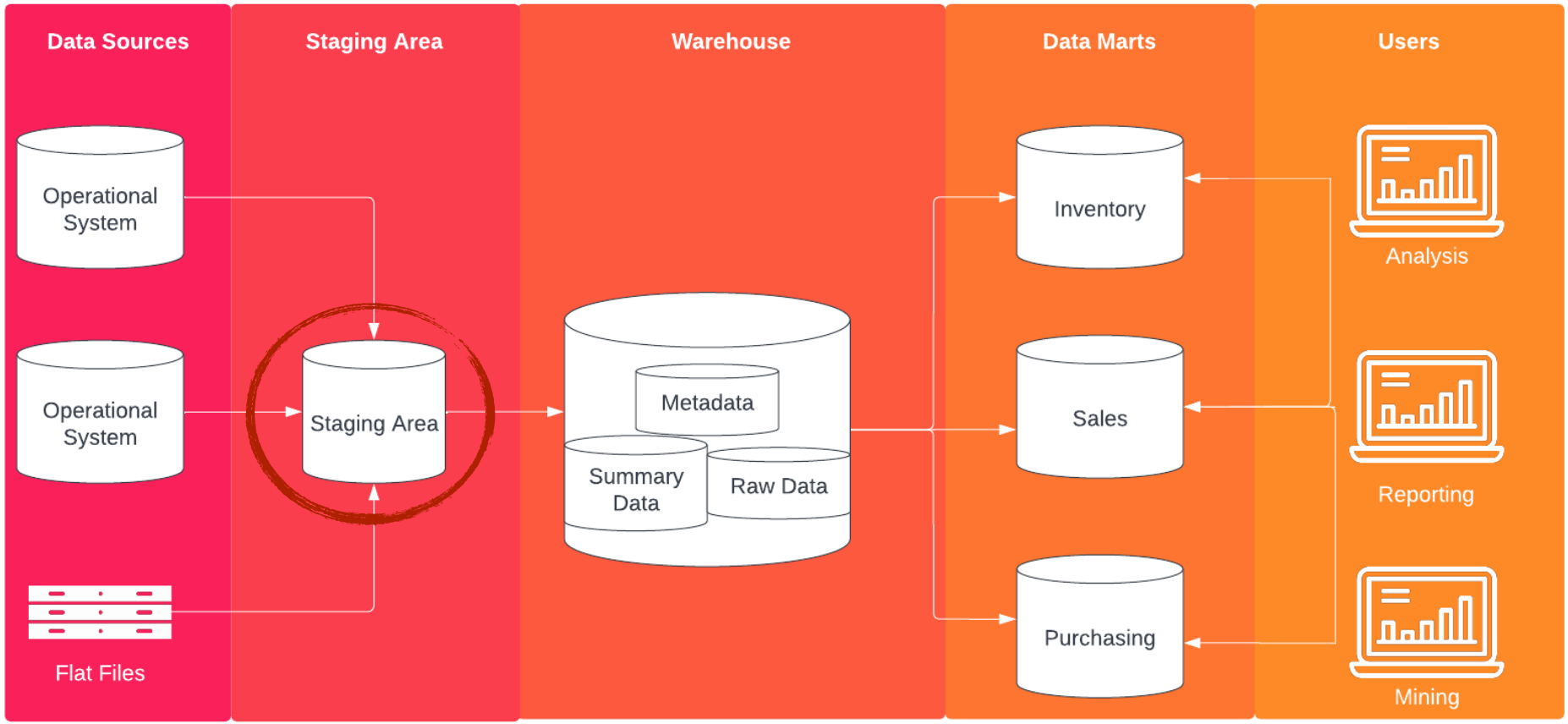


Sequence	Event	Date
1	Inmigration	1995-06-06
2	Outmigration	2006-05-09

DATA STAGING (1)

- Often used as an interim step between data extraction and later step.
- Accumulates data from asynchronous sources using native interfaces, flat files, FTP session, or other processes.
- Data in the staging file is transformed and located to the warehouse
- There is no end user access to the staging file
- An operational data store may be used for data staging

DATA STAGING (2)



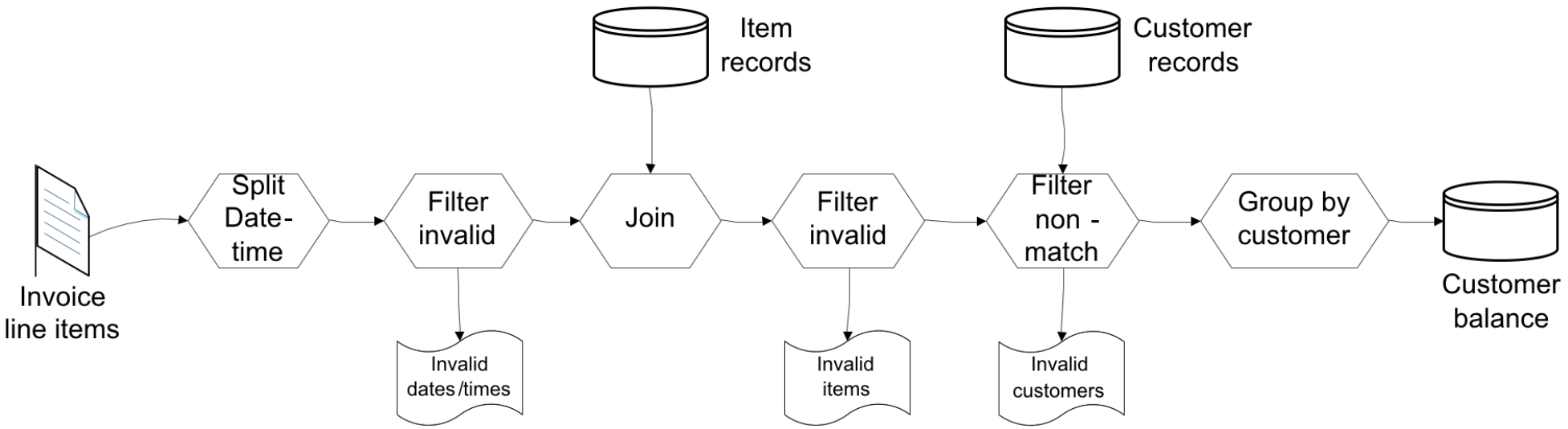
DATA TRANSFORMING

- Transforms the data in accordance with the business rules and standards that have been established
- Example include: format changes, deduplication, splitting up fields, replacement of codes, derived values, and aggregates

DATA LOADING

- Data are physically moved to the data warehouse
- The loading takes place within a “load window”
- The trend is to near real time updates of the data warehouse as the warehouse is increasingly used for operational applications

EXAMPLE ETL TOOL CHAIN



- This is an example for e-commerce loading
- Note multiple stages of filtering (using selection or join-like operations), logging bad records, before we group and load

APPLICATION OF DATA WAREHOUSES

Industry	Key Applications
Airline	Yield management, route assessment
Telecommunications	Customer retention, network design
Insurance	Risk assessment, product design, fraud detection
Retail	Target marketing, supply-chain management

MULTIDIMENSIONAL DATA

- The multidimensional data model supports data representation and operations specifically tailored for decision support processing in data warehouses.

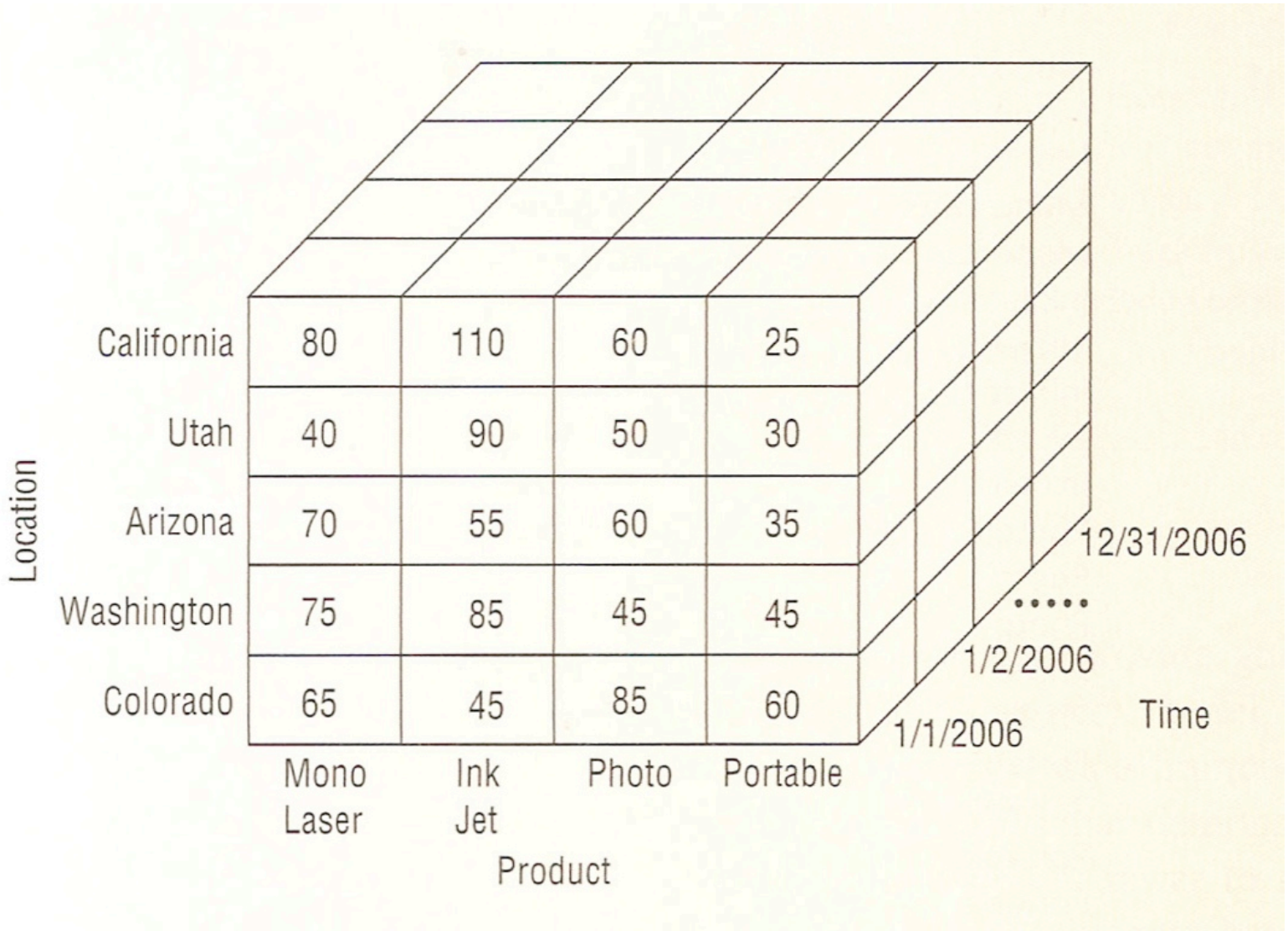
RELATIONAL REPRESENTATION OF SALES DATA

Product	Location	Sales
Mono Laser	California	80
Mono Laser	Utah	40
Mono Laser	Arizona	70
Mono Laser	Washington	75
Mono Laser	Colorado	65
Ink Jet	California	110
Ink Jet	Utah	90
Ink Jet	Arizona	55
Ink Jet	Washington	85
Ink Jet	Colorado	45
Photo	California	60
Photo	Utah	50
Photo	Arizona	60
Photo	Washington	45
Photo	Colorado	85
Portable	California	25
Portable	Utah	30
Portable	Arizona	35
Portable	Washington	45
Portable	Colorado	60

MULTIDIMENSIONAL REPRESENTATION OF SALES DATA

	Product			
Location	**Mono Laser**	**Ink Jet**	**Photo**	**Portable**
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

A THREE-DIMENSIONAL DATA CUBE



MULTIDIMENSIONAL REPRESENTATION OF SALES DATA WITH ROW TOTALS

	Product				
Location	Mono Laser	Ink Jet	Photo	Portable	Totals
California	80	110	60	25	275
Utah	40	90	50	30	210
Arizona	70	55	60	35	220
Washington	75	85	45	45	250
Colorado	65	45	85	60	255

DATA CUBE

- Data cube is a multidimensional format in which cells contain numeric data called measures organised by subjects called dimensions.
- A data cube is sometimes known as a hypercube because conceptually it can have an unlimited number of dimensions.

MULTIDIMENSIONAL TERMINOLOGY

- Data cube or hypercube generalizes the two-dimensional and three-dimensional representations show in the previous page.
- A data cube consist of cells containing measures and dimensions to label or group numeric data.
- Each dimension contains values known as members.
- Dimension can have hierarchies composed of levels.
- Cells in a data cube contain measures such as the sales values.

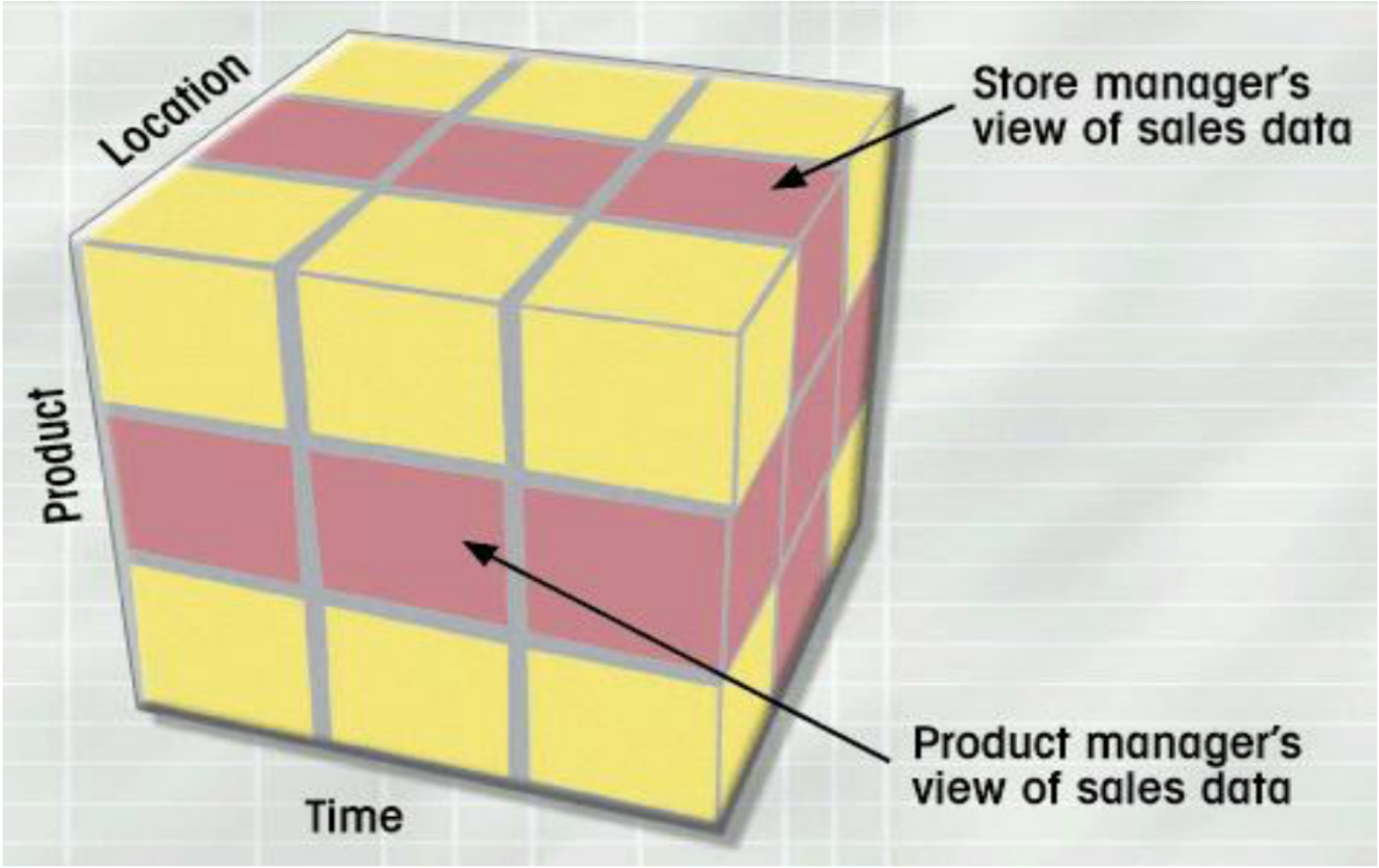
TIME-SERIES DATA

- Time is one of the most common dimensions in a data warehouse and is useful for capturing trends, making forecasts, and so forth.
- The following list shows typical properties for a time series.
  - Data Type
  - Start Date
  - Calendar
  - Periodicity
  - Conversion

DATA CUBE OPERATIONS (1)

- **Slice**
  Because a data cube can contain a large number of dimensions, uses often need to focus on a subset of the dimensions to gain insights.
- **Dice**
  Because individual dimensions can contain a large number of members, uses need to focus on a subset of members to gain insights.

DATA CUBE OPERATIONS (2)



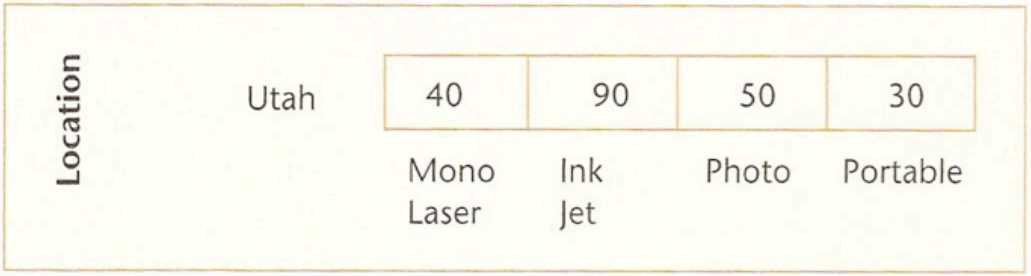
DATA CUBE OPERATIONS (3)

- **Drill-Down**
  Uses often want to navigate among the levels of hierarchical dimensions. The drill-down operator allows uses to navigate from a more general level to a more specific level.
- **Roll-Up**
  Roll-up (also called drill-up) is the opposite of drill-down. Roll-up involves moving from a specific level to a more general level of a hierarchical dimension.
- **Pivot**
  The pivot operator supports rearrangement of the dimensions in a data cube.

EXAMPLE SLICE AND DICE OPERATION

	Product			
Location	**Mono Laser**	**Ink Jet**	**Photo**	**Portable**
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

	Time			
Location	**1/1/2006**	**1/2/2006**	. . .	**Total Sales**
California	400	670	. . .	16,250
Utah	340	190	. . .	11,107
Arizona	270	255	. . .	21,500
Washington	175	285	. . .	20,900
Colorado	165	245	. . .	21,336



DRILL-DOWN OPERATION FOR THE STATE OF UTAH

	Product			
Location	Mono Laser	Ink Jet	Photo	Portable
California + Utah	80	110	60	25
Salt Lake	20	20	10	15
Park City	5	30	10	5
Ogden	15	40	30	10
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

SUMMARY OF THE DATA CUBE OPERATIONS

Operator	Purpose	Description
Slice	Focus attention on a subset of dimensions	Replace a dimension with a single member value or with a summary of its measure values
Dice	Focus attention on a subset of member values	Replace a dimension with a subset of members
Drill-down	Obtain more detail about a dimension	Navigate from a more general level to a more specific level of a hierarchical dimension
Roll-up	Summarize details about a dimension	Navigate from a more specific level to a more general level of a hierarchical dimension
Pivot	Allow a data cube to be presented in a visually appealing order	Rearrange the dimensions in a data cube

RELATIONAL DATA MODELING FOR MULTIDIMENSIONAL DATA (1)

- When using a relational database for a data warehouse, a new data modeling technique is needed to represent multidimensional data.
- A star schema is a data modeling representation of multidimensional data cubes.
- In a relational database, a star schema diagram looks like a star with one large central table, called the fact table, at the centre of the star that is linked to multiple dimensional tables in a radial manner.

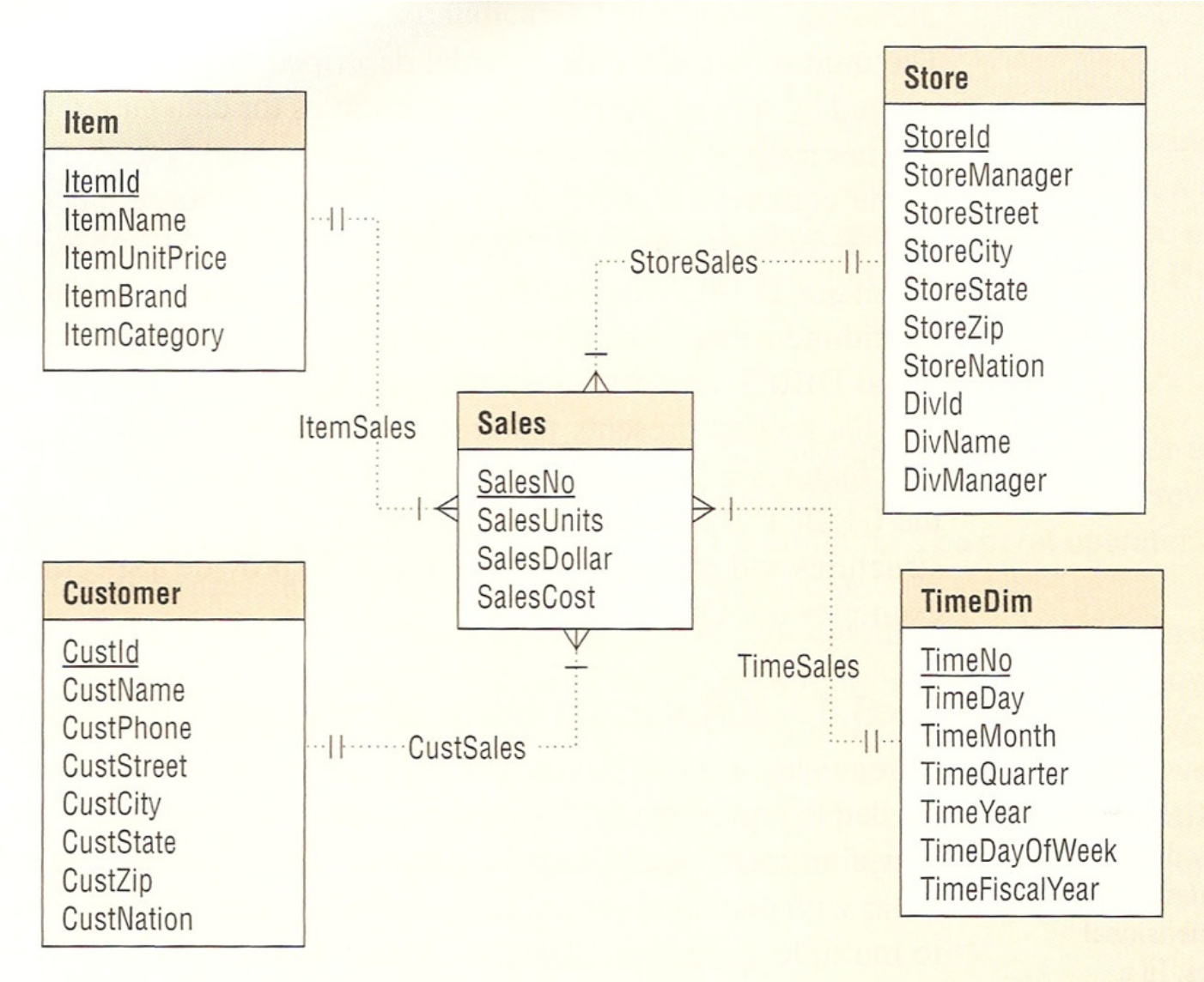
RELATIONAL DATA MODELING FOR MULTIDIMENSIONAL DATA (2)

- The fact table stores numeric data (facts), such as sales results, while the dimension tables store descriptive data corresponding to individual dimensions of data cube such as product, location, and time.
- There is a 1-M relationship from each dimension table to fact table.

THE (CLASSIC) STAR SCHEMA

- A relational model with a one-to-many relationship between dimension table and fact table.
- A single fact table, with detail and summary data
- Fact table primary key has only one key column per dimension
- Each dimension is a single table, highly denormalized

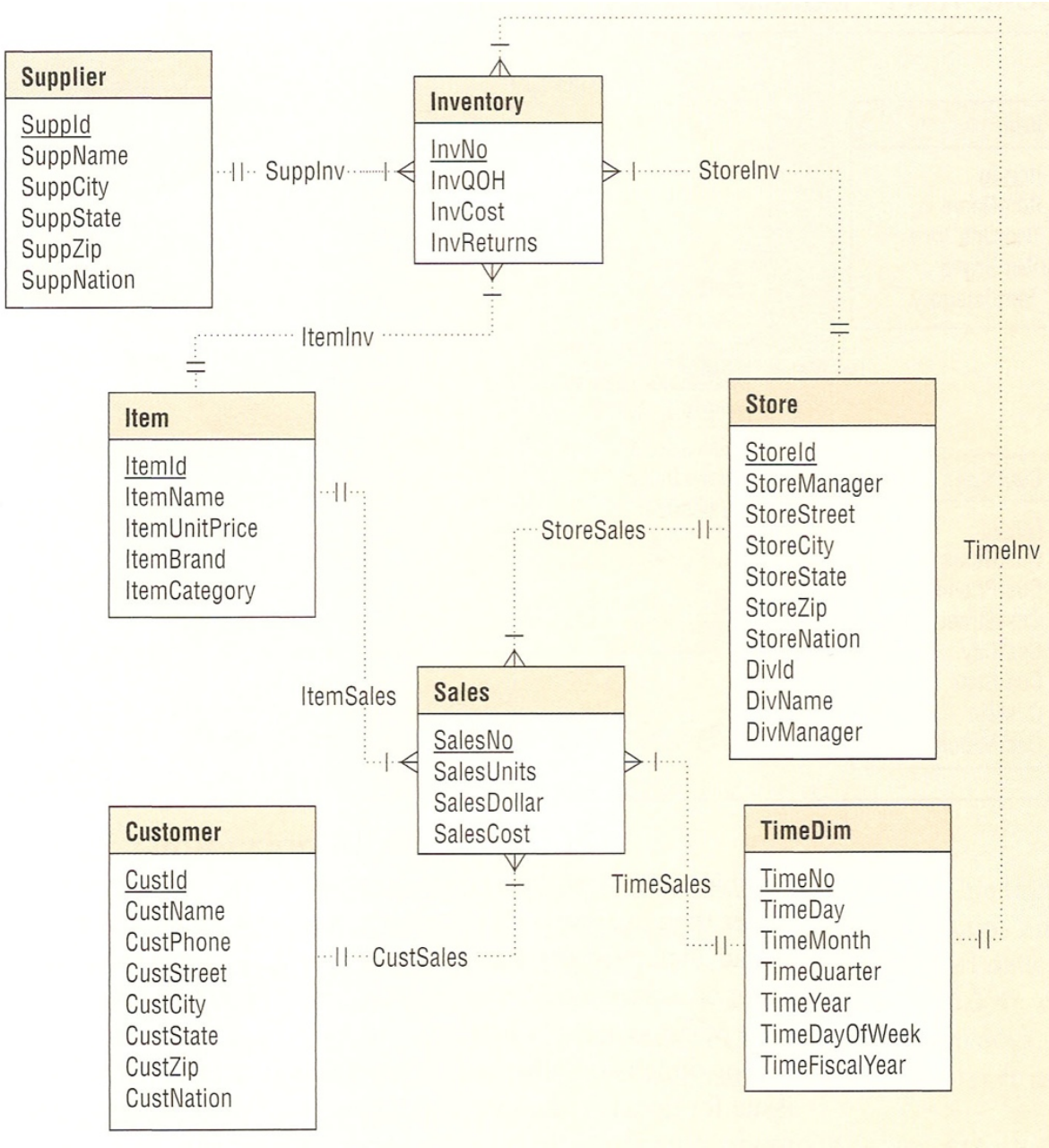
ERD STAR SCHEMA EXAMPLE



CONSTELLATION SCHEMA

- Constellation schema is a data modeling representation for multidimensional databases.
- In a relational database, a constellation schema contains multiple fact tables in the centre related to dimension tables.
- Typically, the fact tables share some dimension tables.

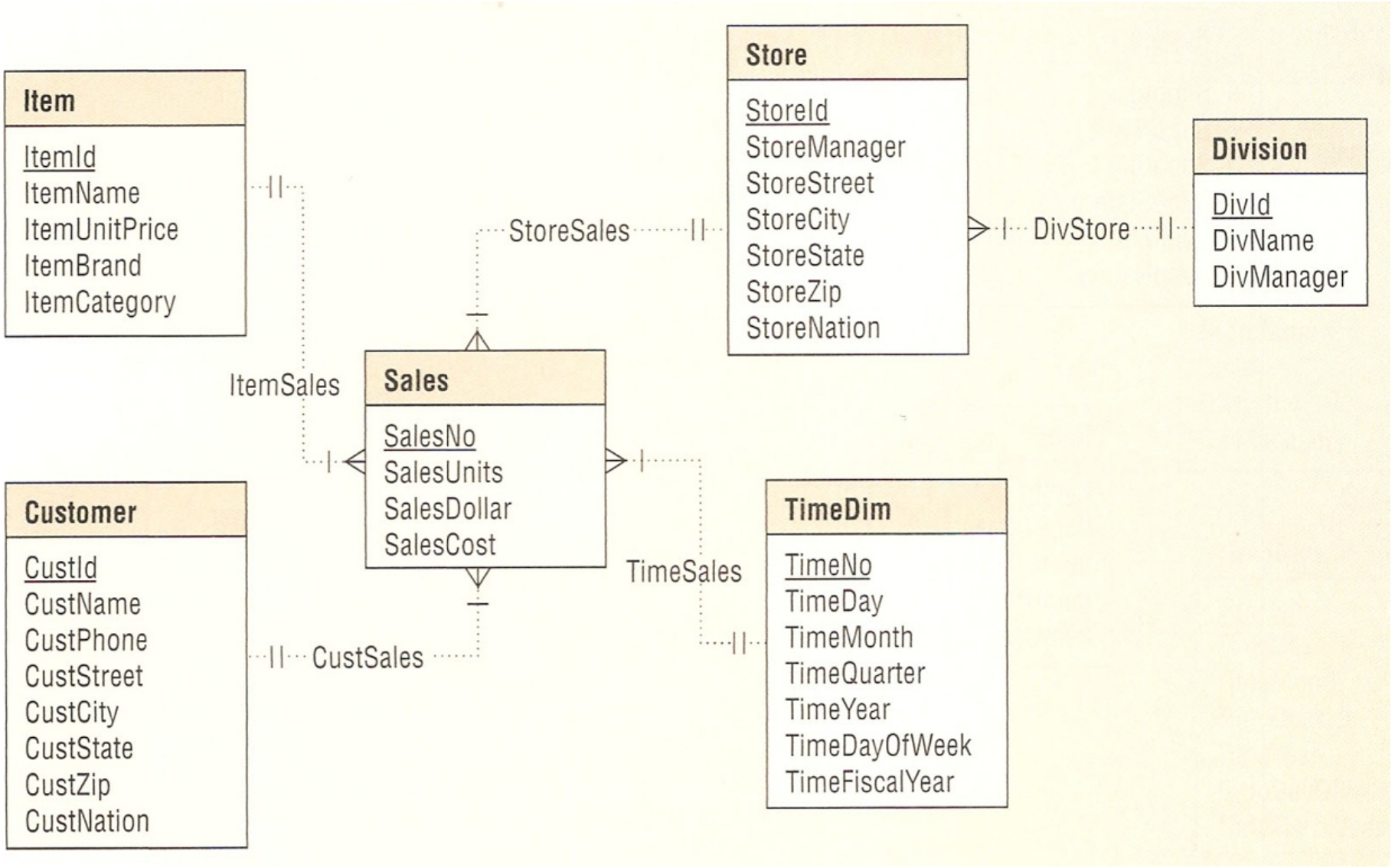
ERD CONSTELLATION SCHEMA EXAMPLE



SNOWFLAKE SCHEMA

- Snowflake schema is a data modeling representation for multidimensional databases.
- In a relational database, a snowflake schema has multiple levels of dimension tables related to one or more fact tables.
- You should consider the snowflake schema instead of the star schema for small dimension tables that are not in 3NF.

ERD SNOWFLAKE SCHEMA EXAMPLE



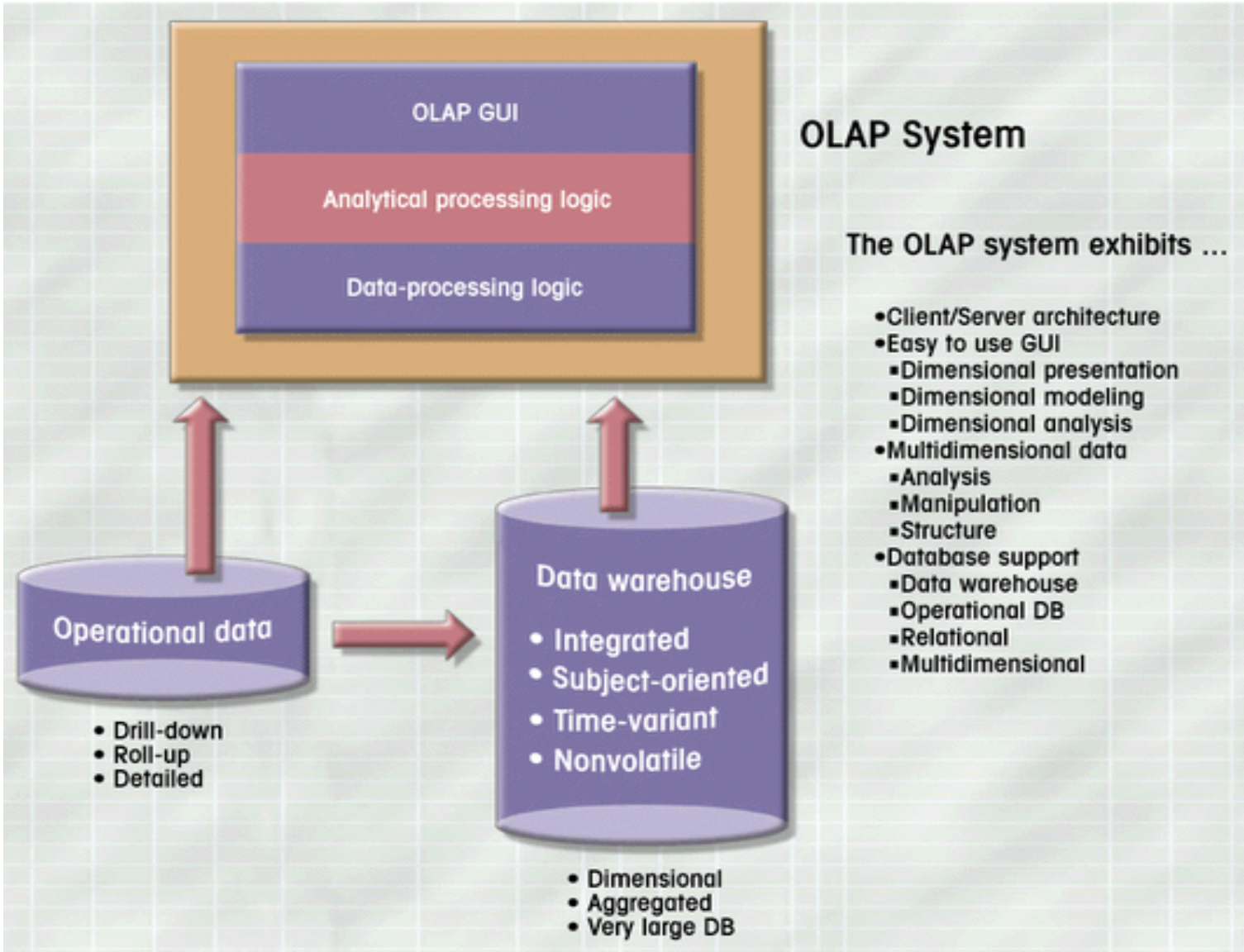
STORAGE AND OPTIMISATION TECHNOLOGIES

- Several storage technologies have been developed to provide multidimensional data capabilities.
- The storage technologies support On Line Analytic Processing (OLAP), a generic name applied to decision support capabilities for data cubes.

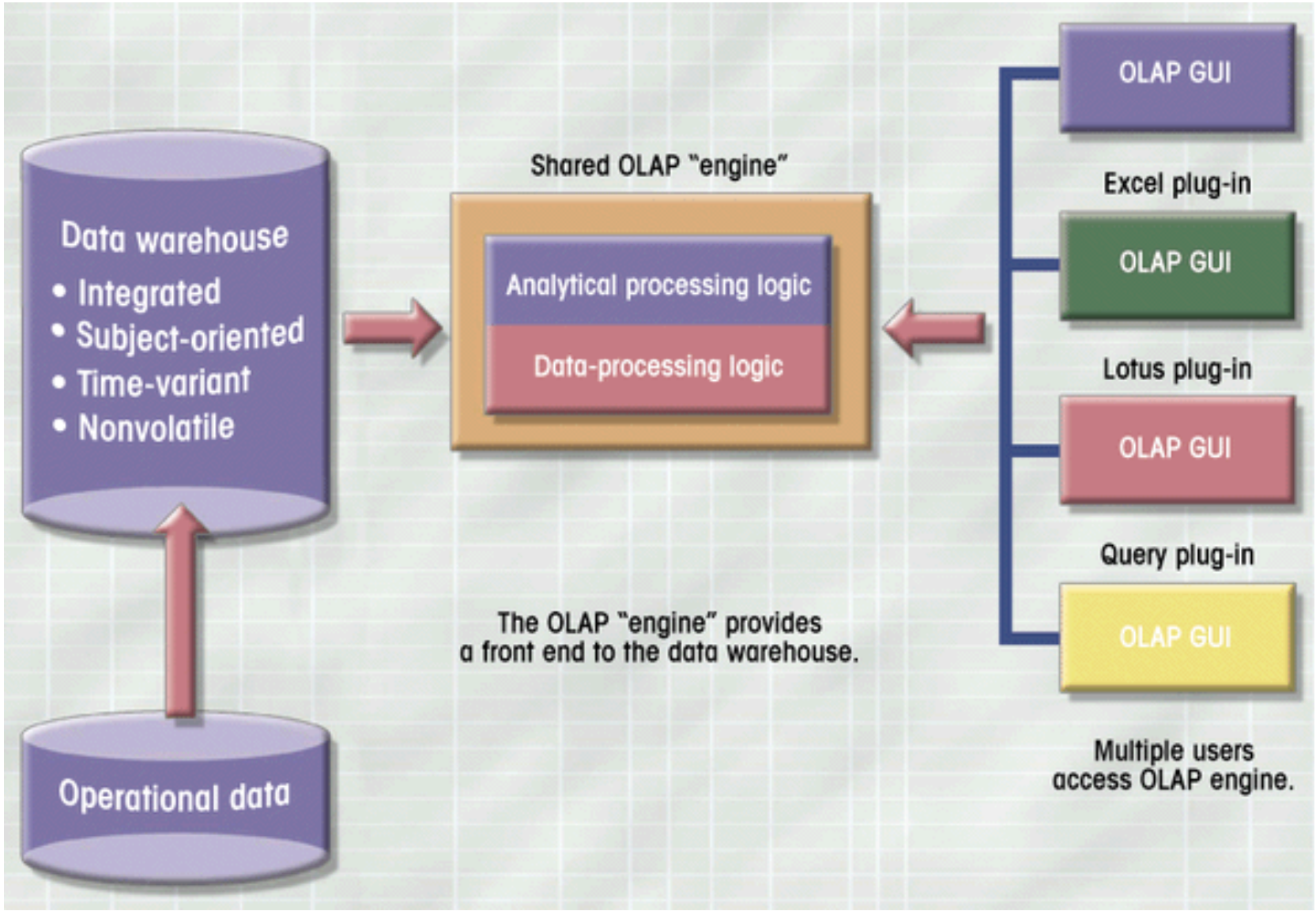
ONLINE ANALYTICAL PROCESSING (OLAP)

- Advanced data analysis environment
- Supports decision making, business modeling, and operations research activities
- Characteristics of OLAP
  - Use multidimensional data analysis techniques
  - Provide advanced database support
  - Provide easy-to-use end-user interfaces
  - Support client/server architecture

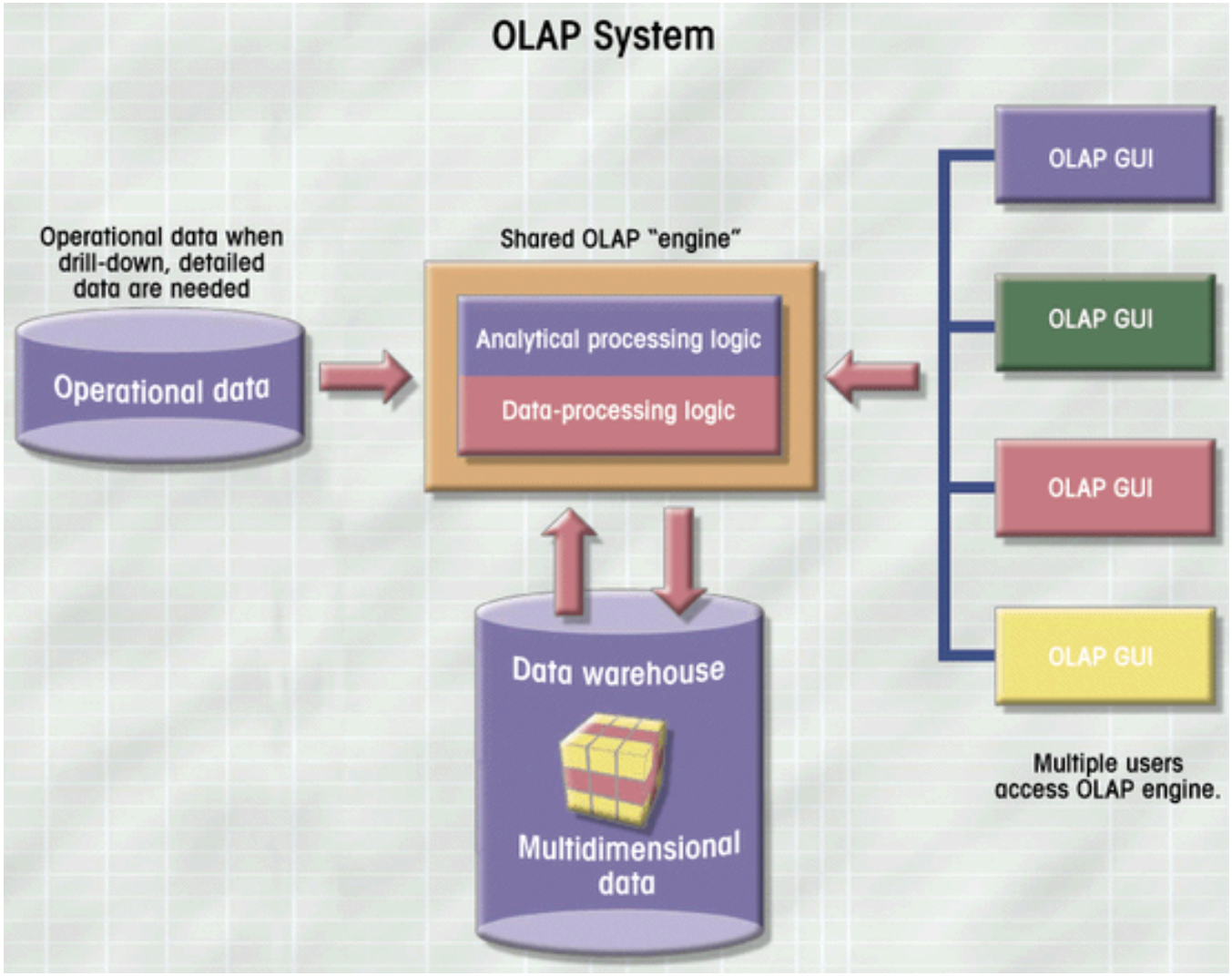
OLAP CLIENT/SERVER ARCHITECTURE



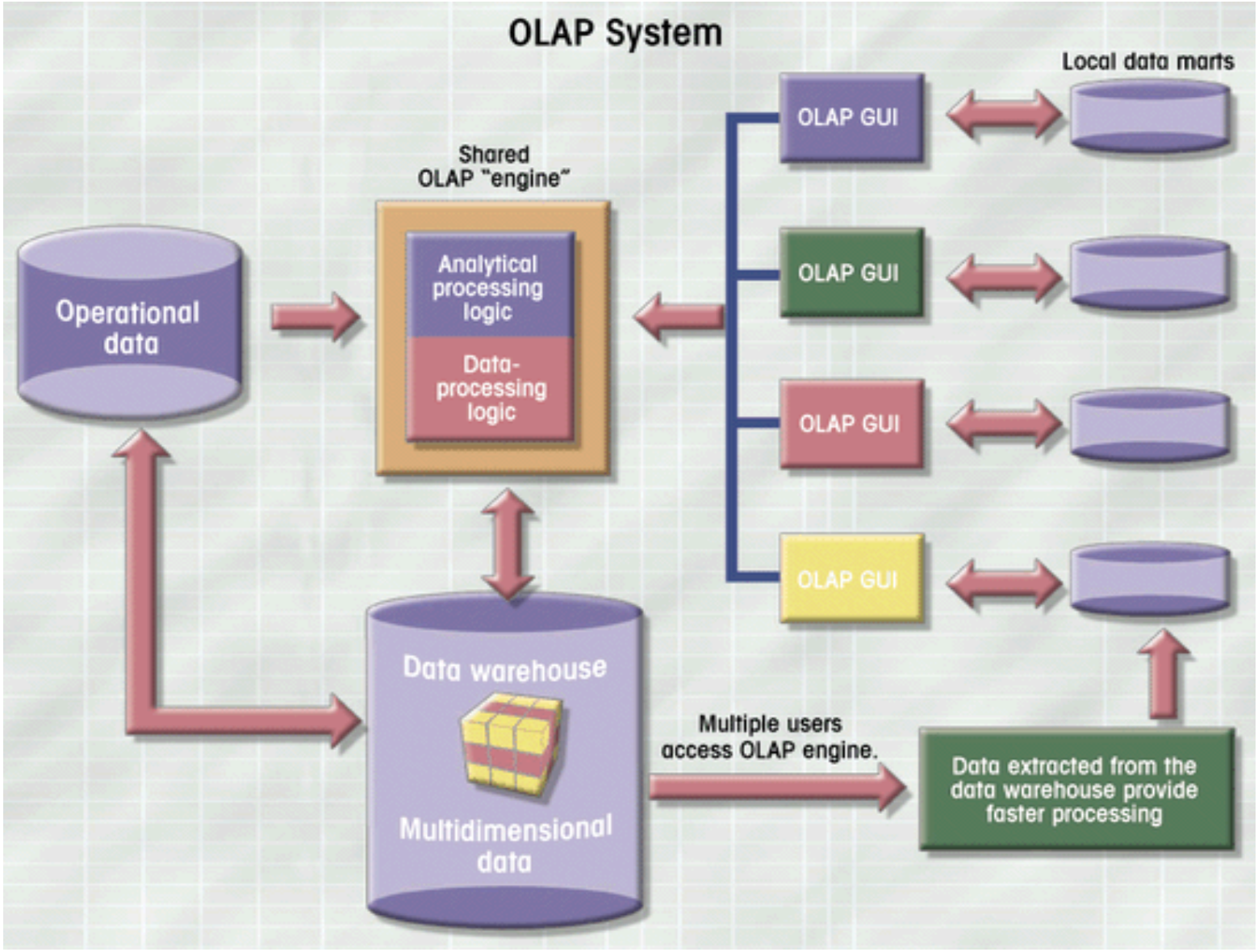
OLAP SERVER ARRANGEMENT



OLAP SERVER WITH MULTIDIMENSIONAL DATA STORE ARRANGEMENT



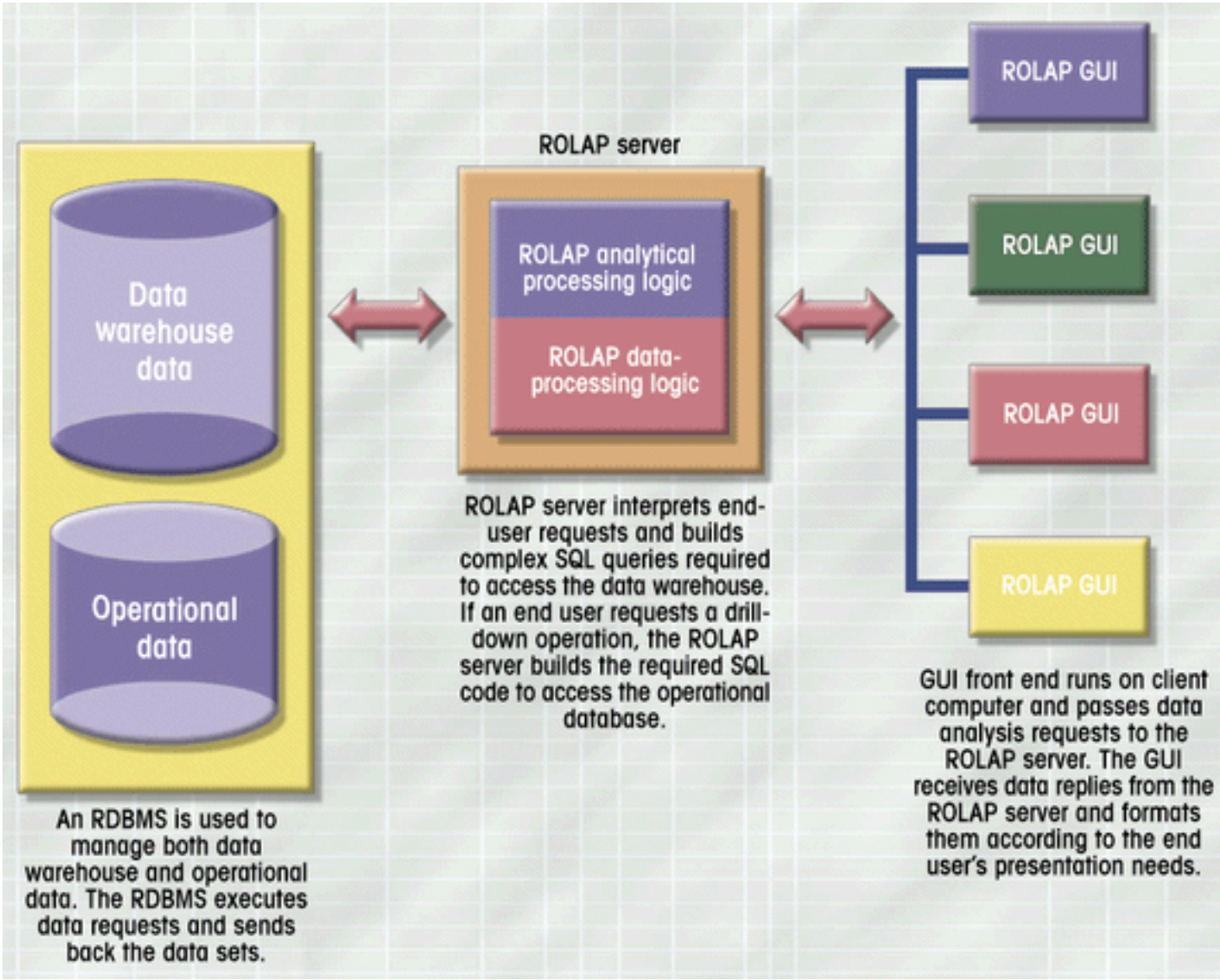
OLAP SERVER WITH LOCAL MINI DATA MARTS



ROLAP

- ROLAP is a relational DBMS extensions to support multidimensional data.
- ROLAP engines support a variety of storage and optimisation techniques for summary data retrieval.

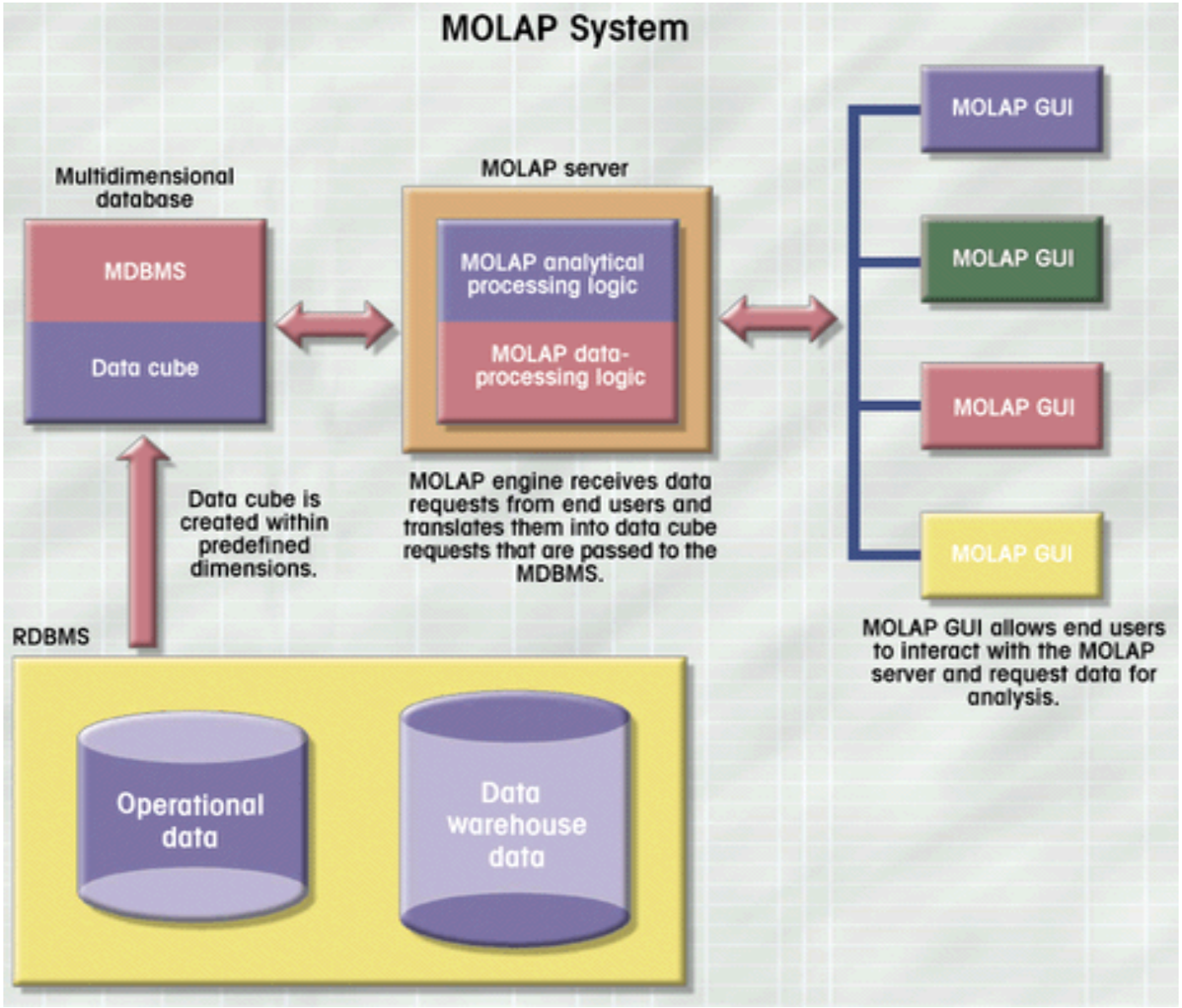
TYPICAL ROLAP CLIENT/SERVER ARCHITECTURE



MOLAP

- MOLAP is a storage engine that directly stores and manipulates data cubes.
- MOLAP engines generally offer the best query performance but place limits on the size of data cubes.

MOLAP CLIENT/SERVER ARCHITECTURE



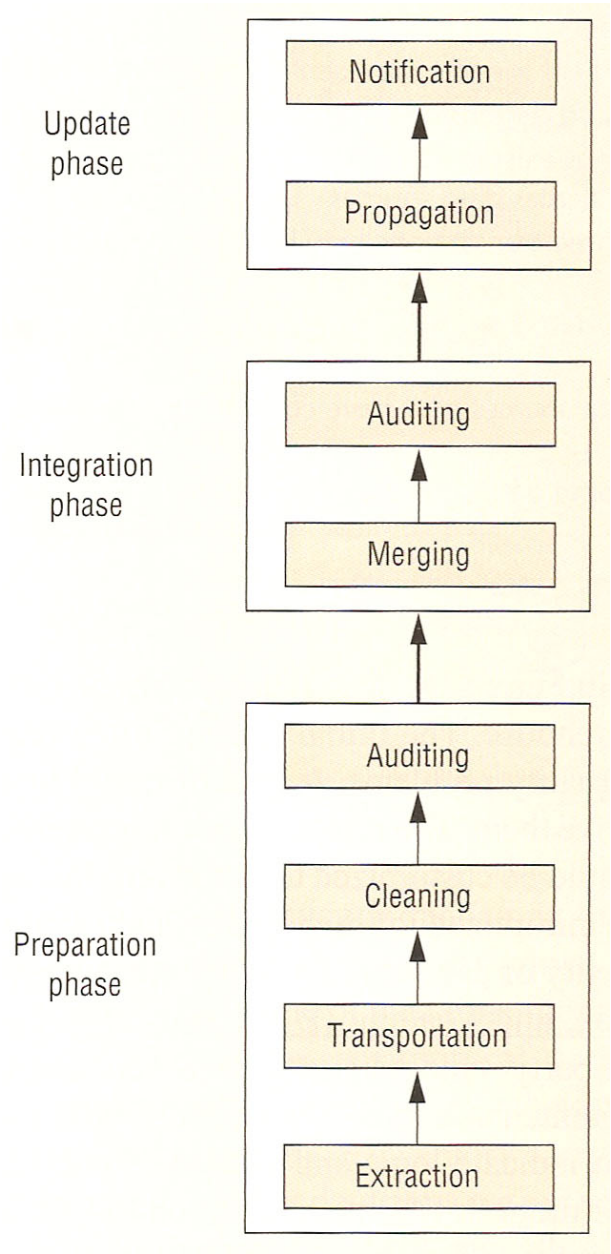
HYBRID OLAP (HOLAP)

- HOLAP is a storage engine for data warehouse that combines ROLAP and MOLAP storage engines.
- HOLAP involves both relational and multidimensional data storage as well as combing data from both relational and multidimensional sources for data cube operators.

MAINTAINING A DATA WAREHOUSE

- Although data warehouse largely contain replicated data, maintaining a data warehouse is much more difficult than simply copying from data sources.

WORKFLOW FOR DW MAINTENANCE



OVERVIEW OF THE DATA WAREHOUSE REFRESH PROCESS