Data Warehousing

Learning Objectives

- Understand the basic definitions and concepts of data warehouses
- Describe data warehouse architectures (high level).
- Describe the processes used in developing and managing data warehouses
- Explain data warehousing operations
- Explain the role of data warehouses in decision support

Learning Objectives

- Explain data integration and the extraction, transformation, and load (ETL) processes
- Describe real-time (active) data warehousing
- Understand data warehouse administration and security issues

Data Warehousing
Definitions and Concepts

- **Data warehouse**

  A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format

Data Warehousing Definitions and Concepts

- Characteristics of data warehousing
  - Subject oriented
  - Integrated
  - Time variant (time series)
  - Nonvolatile
  - Web based
  - Relational/multidimensional
  - Client/server
  - Real-time
  - Include metadata

Data Warehousing Definitions and Concepts

- **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 and Concepts

- **Operational data stores (ODS)**

  A type of database often used as an interim area for a data warehouse, especially for customer information files

Data Warehousing
Definitions and Concepts

- **Enterprise data warehouse (EDW)**

  A technology that provides a vehicle for pushing data from source systems into a data warehouse

- **Metadata**

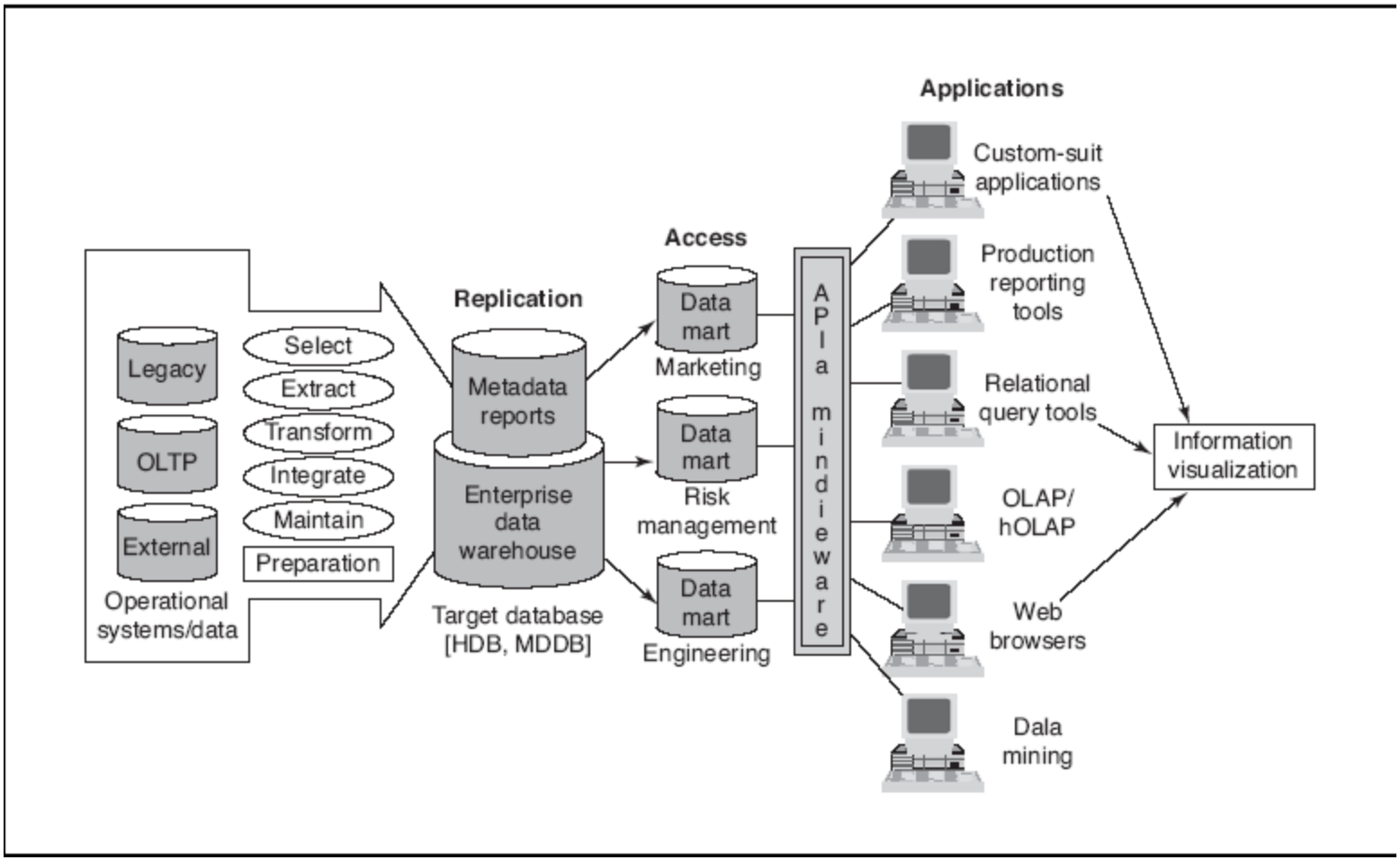
  Data about data. In a data warehouse, metadata describe the contents of a data warehouse and the manner of its use

Data Warehousing Process Overview

- Organizations continuously collect data, information, and knowledge at an increasingly accelerated rate and store them in computerized systems
- The number of users needing to access the information continues to increase as a result of improved reliability and availability of network access, especially the Internet

Data Warehousing Process Overview

FIGURE 2.1 Data Warehouse Framework and Views



Data Warehousing Process Overview

- The major components of a data warehousing process
  - Data sources
  - Data extraction
  - Data loading
  - Comprehensive database
  - Metadata
  - Middleware tools

Data Warehousing Architectures

- Three parts of the data warehouse
  - The data warehouse that contains the data and associated software
  - Data acquisition (back-end) software that extracts data from legacy systems and external sources, consolidates and summarizes them, and loads them into the data warehouse
  - Client (front-end) software that allows users to access and analyze data from the warehouse

Data Warehousing Architectures

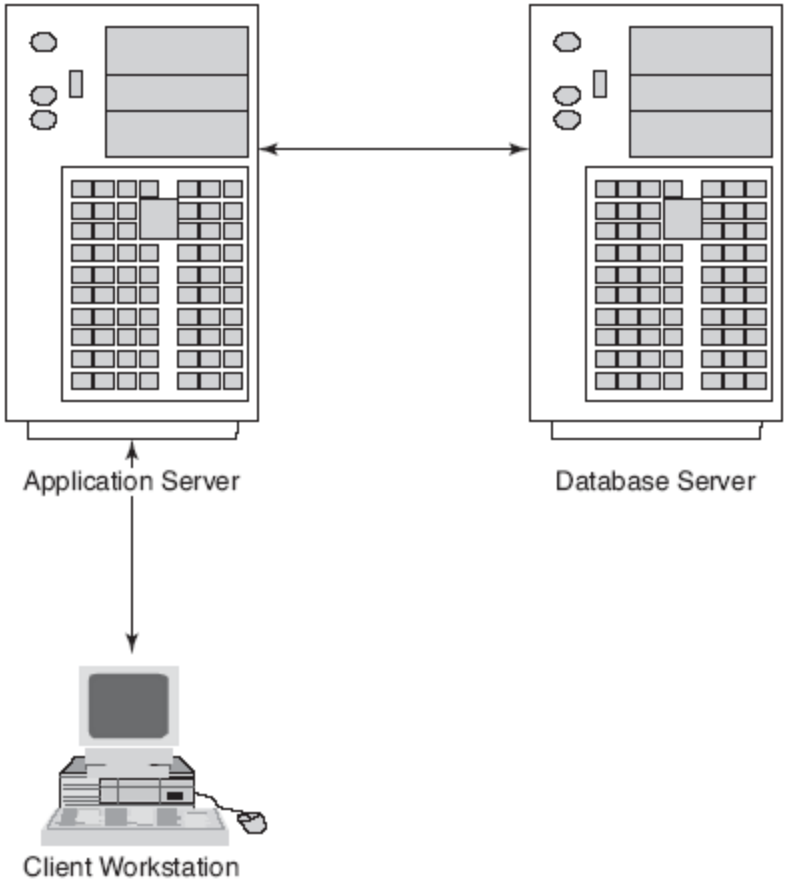


FIGURE 2.2 Architecture of a Three-Tier Data Warehouse

Data Warehousing Architectures

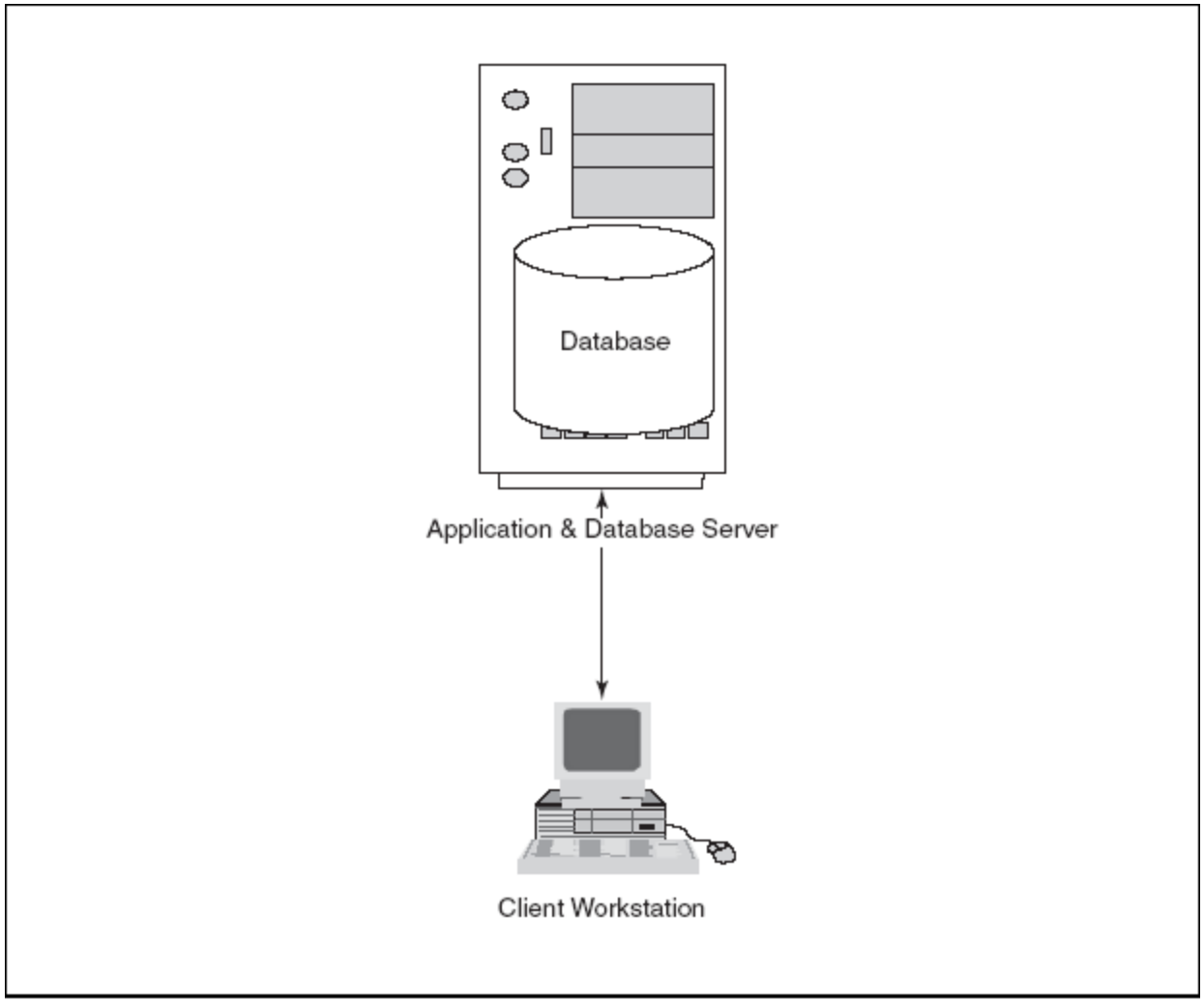
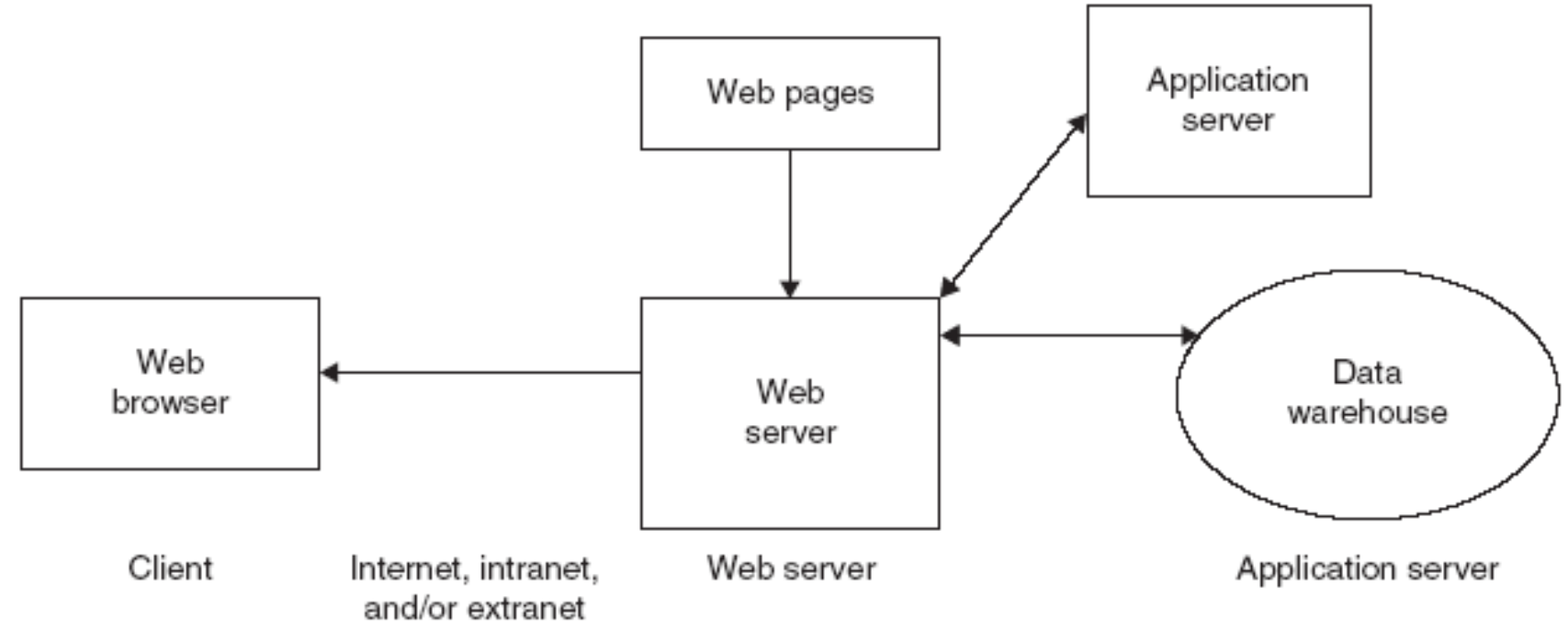


FIGURE 2.3 Architecture of a Two-Tier Data Warehouse

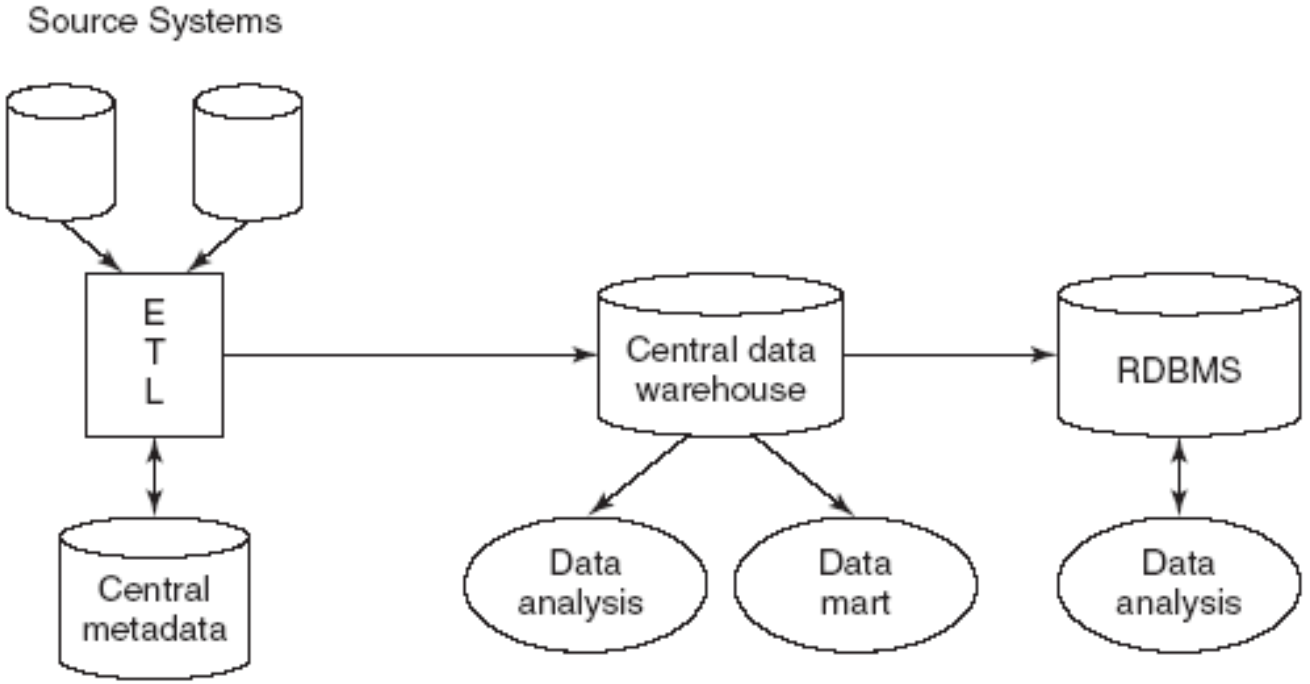
Data Warehousing Architectures

FIGURE 2.4 Architecture of Web-Based Data Warehousing



Data Warehousing Architectures

FIGURE 2.5 Alternative Data Warehouse Architectures



2.5a Enterprise Data Warehousing Architecture

Data Warehousing Architectures

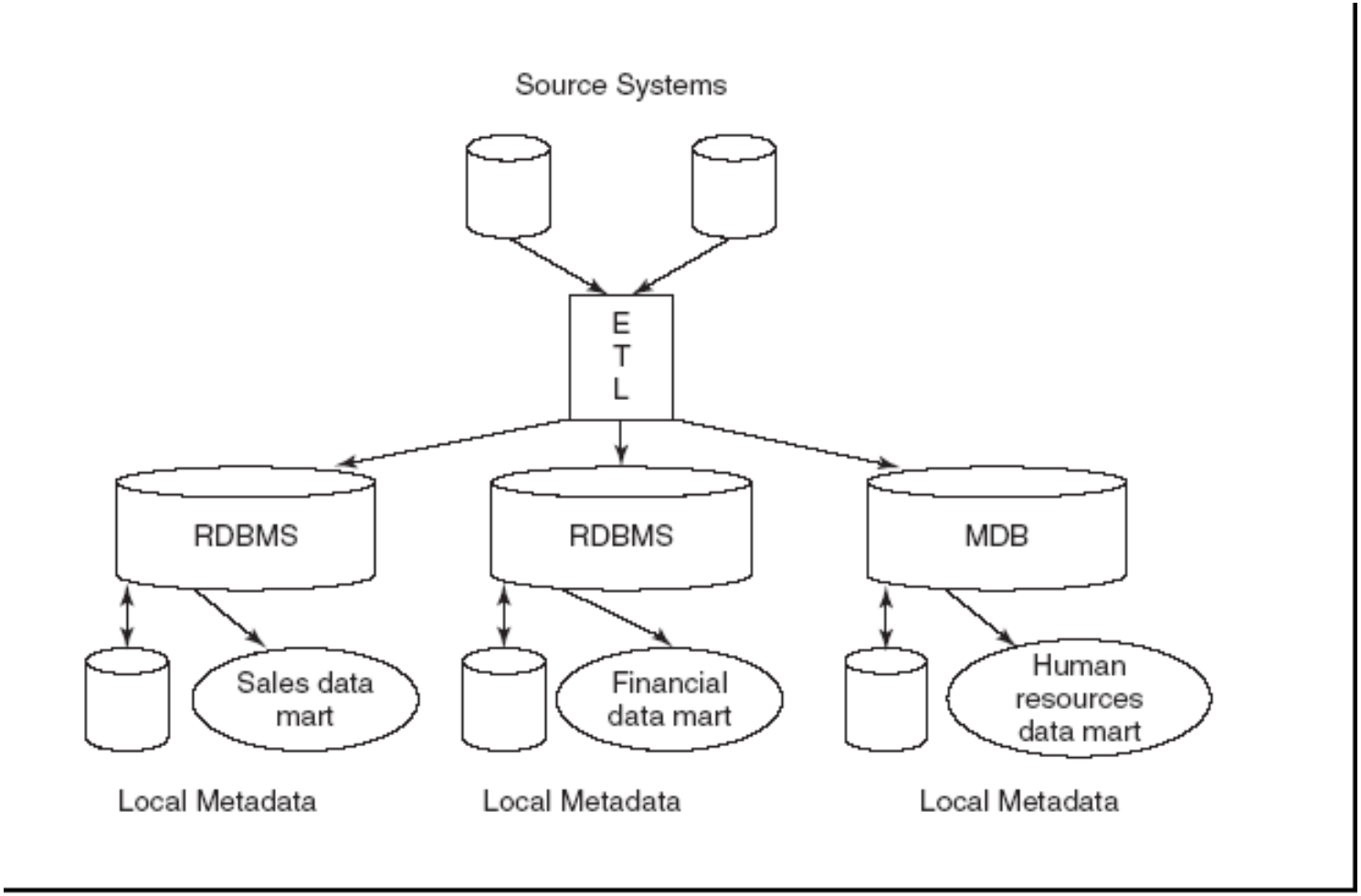


FIGURE 2.5b Data Mart Architecture

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

- **Data integration**

  Integration that comprises three major processes: data access, data federation, and change capture. When these three processes are correctly implemented, data can be accessed and made accessible to an array of ETL and analysis tools and data warehousing environments

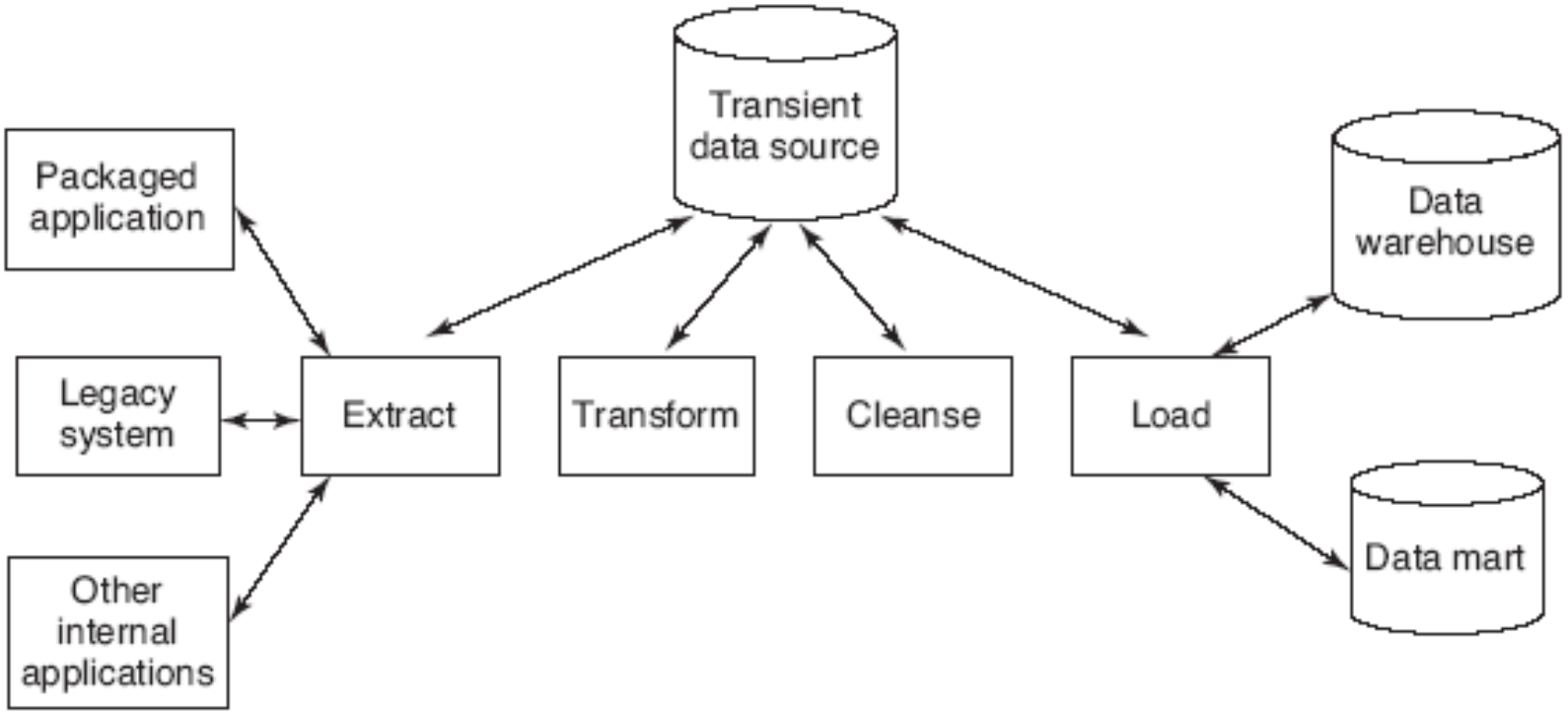
Data Integration and the Extraction, Transformation, and Load (ETL) Process

- **Extraction, transformation, and load (ETL)**

  A data warehousing process that consists of extraction (i.e., reading data from a database), transformation (i.e., converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database), and load (i.e., putting the data into the data warehouse)

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

FIGURE 2.8 The ETL Process



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

- Issues affect whether an organization will purchase data transformation tools or build the transformation process itself
  - Data transformation tools are expensive
  - Data transformation tools may have a long learning curve
  - It is difficult to measure how the IT organization is doing until it has learned to use the data transformation tools

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

- Important criteria in selecting an ETL tool
  - Ability to read from and write to an unlimited number of data source 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

- Direct benefits of a data warehouse
  - Allows end users to perform extensive analysis
  - Allows a consolidated view of corporate data
  - Better and more timely information A
  - Enhanced system performance
  - Simplification of data access

Data Warehouse Development

- Indirect benefits result from end users using these direct benefits
  - Enhance business knowledge
  - Present competitive advantage
  - Enhance customer service and satisfaction
  - Facilitate decision making
  - Help in reforming business processes

Data Warehouse Development

- Data warehouse vendors
  - Six guidelines to considered when developing a vendor list:
    1. Financial strength
    2. ERP linkages
    3. Qualified consultants
    4. Market share
    5. Industry experience
    6. Established partnerships

Data Warehouse Development

- Data warehouse development approaches
  - Inmon Model: EDW approach
  - Kimball Model: Data mart approach
- Which model is best?
  - There is no one-size-fits-all strategy to data warehousing
  - One alternative is the hosted warehouse

Data Warehouse Development

- Data warehouse structure: The Star Schema
  - **Dimensional modeling**

    A retrieval-based system that supports high-volume query access
  - **Dimension tables**

    A table that address *how* data will be analyzed

Data Warehouse Development

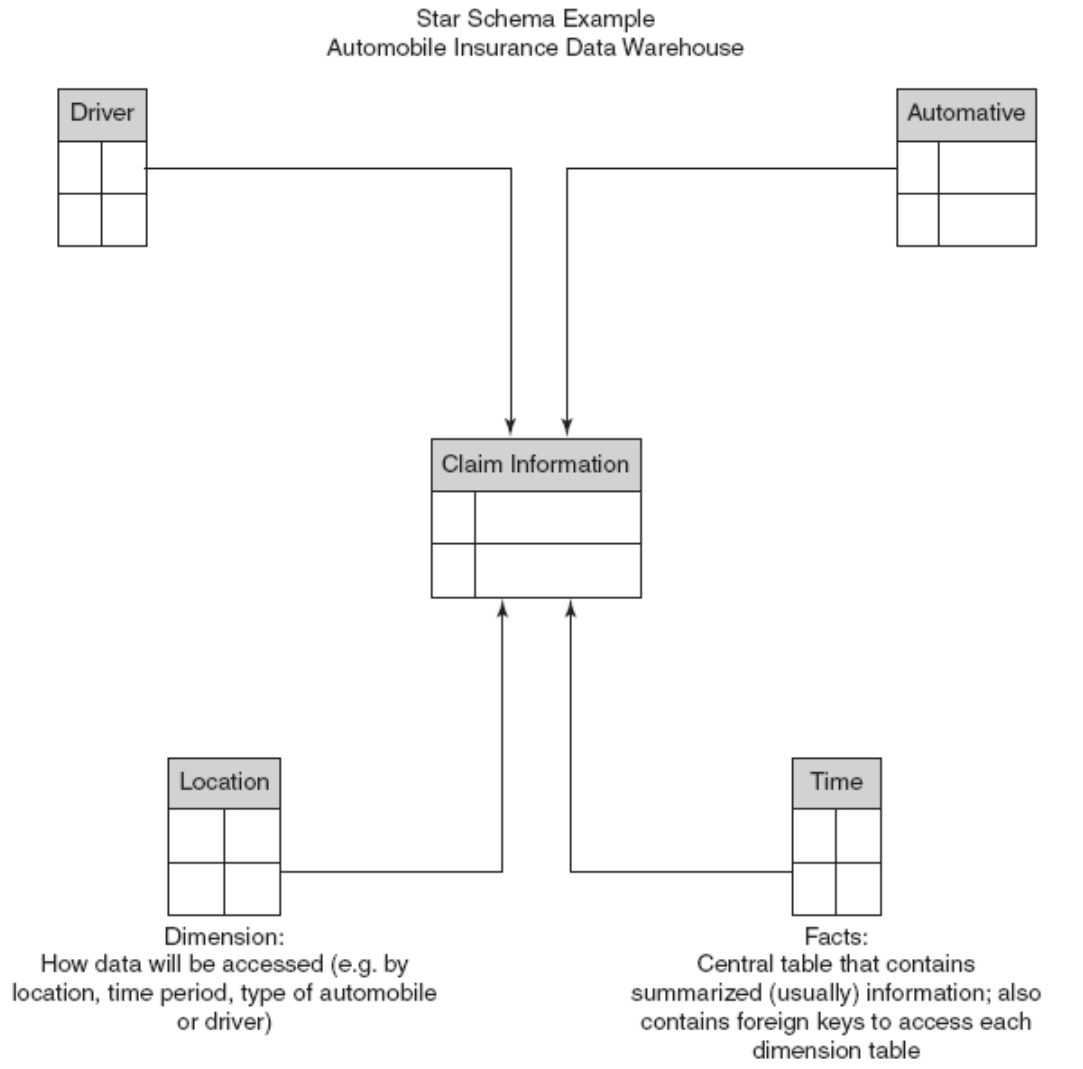


FIGURE 2.9 Star Schema

Data Warehouse Development

- **Grain**

  A definition of the highest level of detail that is supported in a data warehouse

- **Drill-down**

  The process of probing beyond a summarized value to investigate each of the detail transactions that comprise the summary

Data Warehouse Development

- Data warehousing implementation issues
  - Implementing a data warehouse is generally a massive effort that must be planned and executed according to established methods
  - There are many facets to the project lifecycle, and no single person can be an expert in each area

Data Warehouse Development

Eleven major tasks that could be performed in parallel for successful implementation of a data warehouse (Solomon, 2005) :

1. Establishment of service-level agreements and data-refresh requirements
2. Identification of data sources and their governance policies
3. Data quality planning
4. Data model design
5. ETL tool selection
6. Relational database software and platform selection
7. Data transport
8. Data conversion
9. Reconciliation process
10. Purge and archive planning
11. End-user support

Data Warehouse Development

- Some best practices for implementing a data warehouse (Weir, 2002):
  - Project must fit with corporate strategy and 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

Data Warehouse Development

- Some best practices for implementing a data warehouse (Weir, 2002):
  - The project must be managed by both IT and business professionals
  - Develop a business/supplier relationship
  - Only load data that have been cleansed and are of a quality understood by the organization
  - Do not overlook training requirements
  - Be politically aware

Data Warehouse Development

- Failure factors in data warehouse projects:
  - Cultural issues being ignored
  - Inappropriate architecture
  - Unclear business objectives
  - Missing information
  - Unrealistic expectations
  - Low levels of data summarization
  - Low data quality

Data Warehouse Development

- Issues to consider to build a successful data warehouse:
  - Starting with the wrong sponsorship chain
  - Setting expectations that you cannot meet and frustrating executives at the moment of truth
  - Engaging in politically naive behavior
  - Loading the warehouse with information just because it is available

Data Warehouse Development

- Issues to consider to build a successful data warehouse:
  - 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
  - Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images, and, perhaps, sound and video

Data Warehouse Development

- Issues to consider to build a successful data warehouse:
  - Delivering data with overlapping and 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

Data Warehouse Development

- Implementation factors that can be categorized into three criteria
  - Organizational issues
  - Project issues
  - Technical issues
- User participation in the development of data and access modeling is a critical success factor in data warehouse development

Data Warehouse Development

- Massive data warehouses and 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 Data Warehousing

- **Real-time (active) data warehousing**

  The process of loading and providing data via a data warehouse as they become available

Real-Time Data Warehousing

- Levels of data warehouses:
  1. Reports what happened
  2. Some analysis occurs
  3. Provides prediction capabilities,
  4. Operationalization
  5. Becomes capable of making events happen

Real-Time Data Warehousing

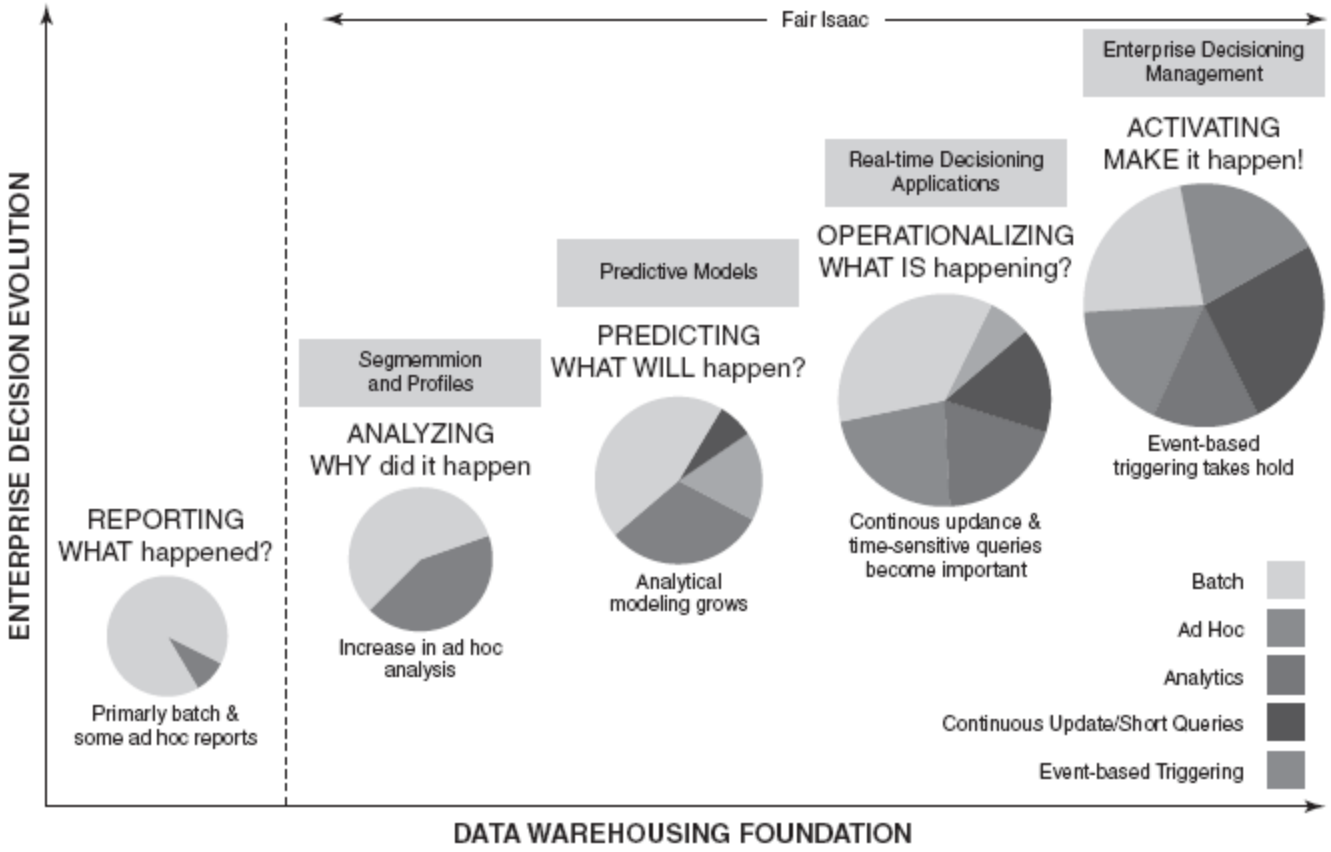


FIGURE 2.10 Enterprise Decision Evolution

Real-Time Data Warehousing

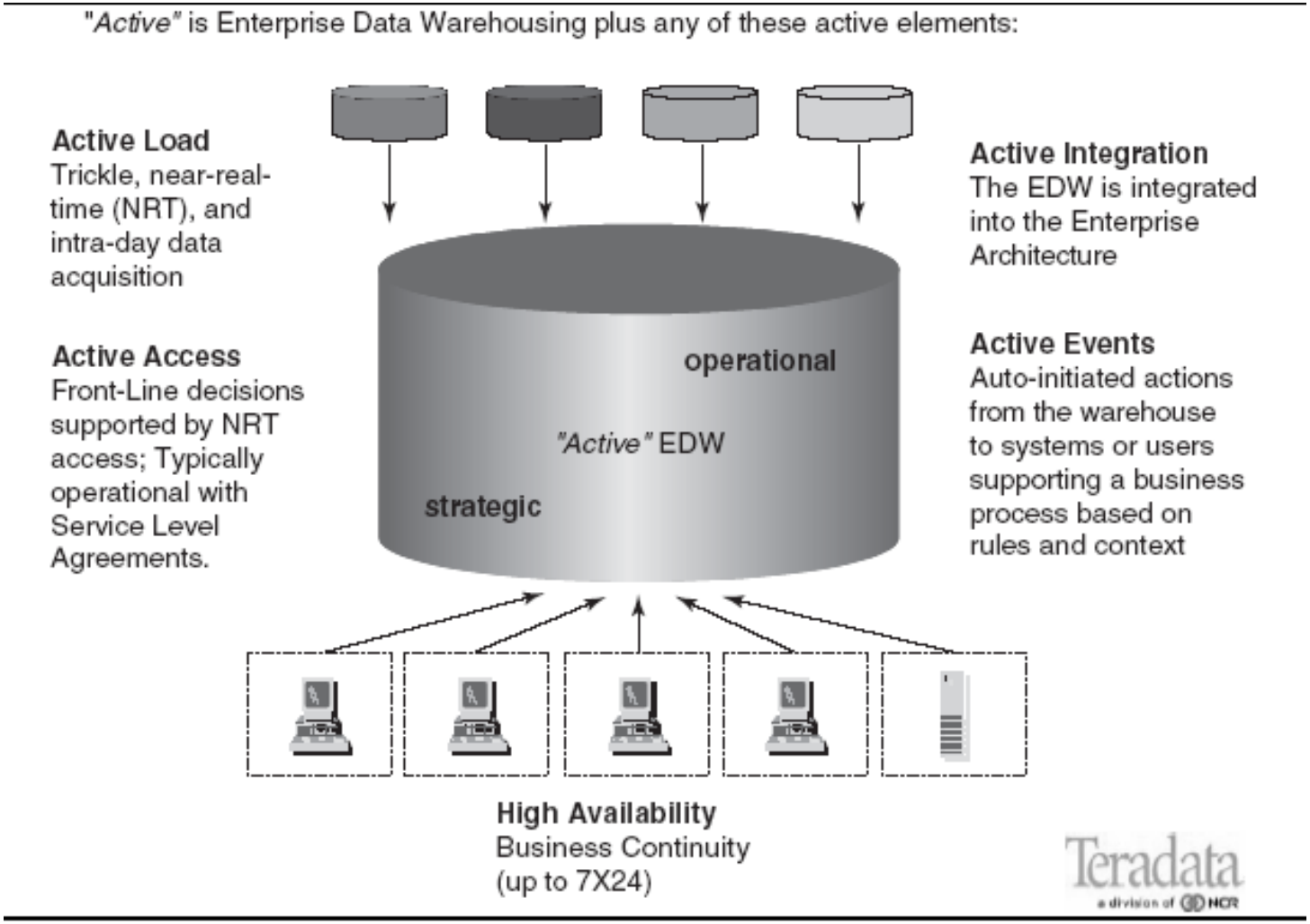


FIGURE 2.11 The Teradata Active EDW

Real-Time Data Warehousing

- The need for real-time data
  - A business often cannot afford to wait a whole day for its operational data to load into the data warehouse for analysis
  - Provides incremental real-time data showing every state change and almost analogous patterns over time
  - Maintaining metadata in sync is possible
  - Less costly to develop, maintain, and secure one huge data warehouse so that data are centralized for BI/BA tools
  - An EAI with real-time data collection can reduce or eliminate the nightly batch processes

Data Warehouse Administration and Security Issues

- **Data warehouse administrator (DWA)**

  A person responsible for the administration and management of a data warehouse

Data Warehouse Administration and Security Issues

- Effective security in a data warehouse should focus on four main areas:
  - Establishing effective corporate and security policies and procedures
  - Implementing logical security procedures and techniques to restrict access
  - Limiting physical access to the data center environment
  - Establishing an effective internal control review process with an emphasis on security and privacy

Why Metadata

- Before touches the keyboard by the prominent(superlative) users.
- Several questions come to their mind.
  - What are various data in the data warehouse?
  - Is there information about unit sale and costs by product?
  - How old is the data in the warehouse.
  - When was the last time fresh data was brought in?
  - Are there any summaries by month and

Why Meta data contd.,

- These questions and several more information are very valid and pertinent.
- What are answers? Where are the answers? Can ur user see the answers? How easy to get the answers?

What is Meta data?

- Metadata in a data warehouse contains the answer to questions about the data in the data warehouse.
- Keep the answer in a place called the metadata repository.

Different definitions for metadata

- Data about the data.
- Table of contents for the data.
- Catalog for the data.
- Data warehouse roadmap.
- Data warehouse directory.
- Tongs to handle the data.
- The nerve center.

So, what exactly is metadata

- It tells u more.
- It gives more than the explanation of the semantics and the syntax.
- Describes all the pertinent(relevant) aspect of the data in the data warehouse.
- Pertinent to whom?
- The answer is – primarily to the user and also developer and even project team.

Meta data

Entity Name:	Customer
Alias Names:	Account, Client

Definition:	A person or an organization that purchases goods or services from the company.
Remarks:	Customer entity includes regular, current, and past customers.
Source Systems:	Finished Goods Orders, Maintenance Contracts, Online Sales.

Create Date:	January 15, 1999
Last Update Date:	January 21, 2001
Update Cycle:	Weekly
Last Full Refresh Date:	December 29, 2000
Full Refresh Cycle:	Every six months
Data Quality Reviewed:	January 25, 2001
Last Deduplication:	January 10, 2001
Planned Archival:	Every six months
Responsible User:	Jane Brown

Figure 9-1 Metadata element for *Customer* entity.

Why it is important?

- Metadata is necessary for using, building and administering your data warehouse.

For using the data warehouse

- Without the metadata support, user cannot get an information every time they needed(user create their own ad hoc report).
- Today data warehouses are much larger in size, wide in scope.
- User critically need metadata.

For building the data warehouse(extraction team)

- Need to know the structure and data content in the data warehouse.
- Mapping and data transformations.
- Logical structure of data warehouse database.

For administering the data warehouse

- Impossible to administer the data warehouse(size)
- Large no of questions and queries.
- Cannot administer without answers for these questions.
- Metadata must address these issues.

List of questions relating to data warehouse

Data Extraction/Transformation/Loading

- How to handle data changes?
- How to include new sources?
- Where to cleanse the data? How to change the data cleansing methods?
- How to cleanse data after populating the warehouse?
- How to switch to new data transformation techniques?
- How to audit the application of ongoing changes?

Data from External Sources

- How to add new external data sources?
- How to drop some external data sources?
- When mergers and acquisitions happen, how to bring in new data to the warehouse?
- How to verify all external data on ongoing basis?

Data Warehouse

- How to add new summary tables?
- How to control runaway queries?
- How to expand storage?
- When to schedule platform upgrades?
- How to add new information delivery tools for the users?
- How to continue ongoing training?
- How to maintain and enhance user support function?
- How to monitor and improve ad hoc query performance?
- When to schedule backups?
- How to perform disaster recovery drills?
- How to keep data definitions up-to-date?
- How to maintain the security system?
- How to monitor system load distribution?

Figure 9-2 Data warehouse administration: questions and issues.

Who needs metadata?

- Without metadata, Datawarehouse is like such a filling cabinet without information.
- Go through the image column(it address the our question)

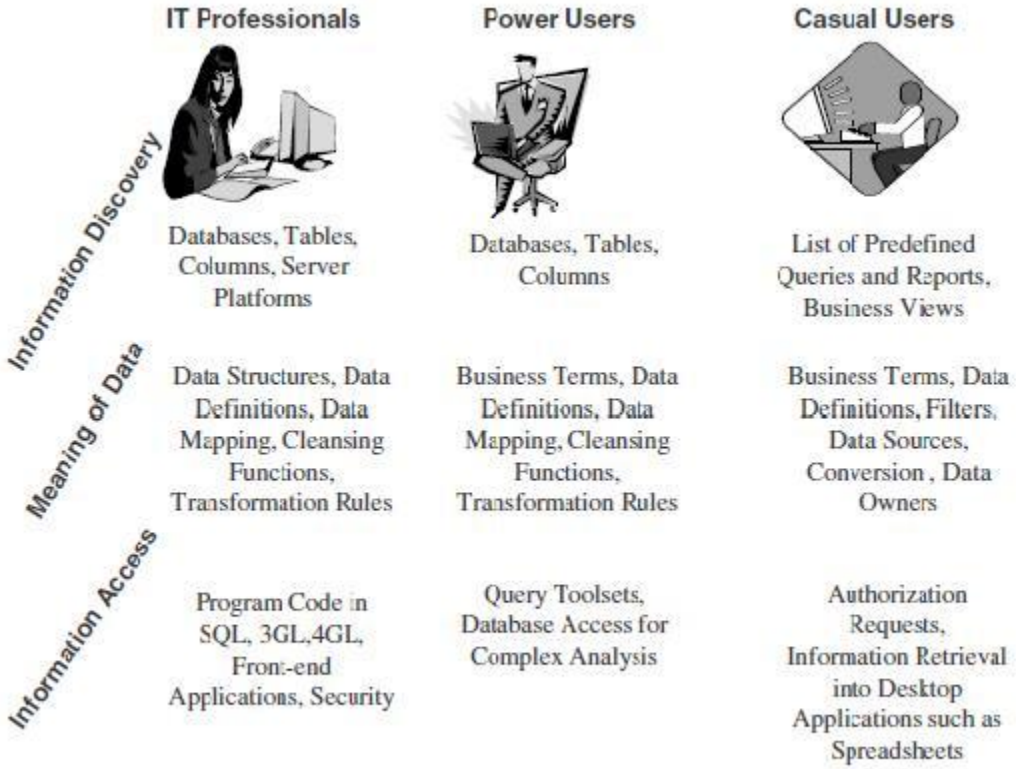


Figure 9-3 Who needs metadata?

Like a nerve center

- Metadata placed in a key position and enables communication among various processes.

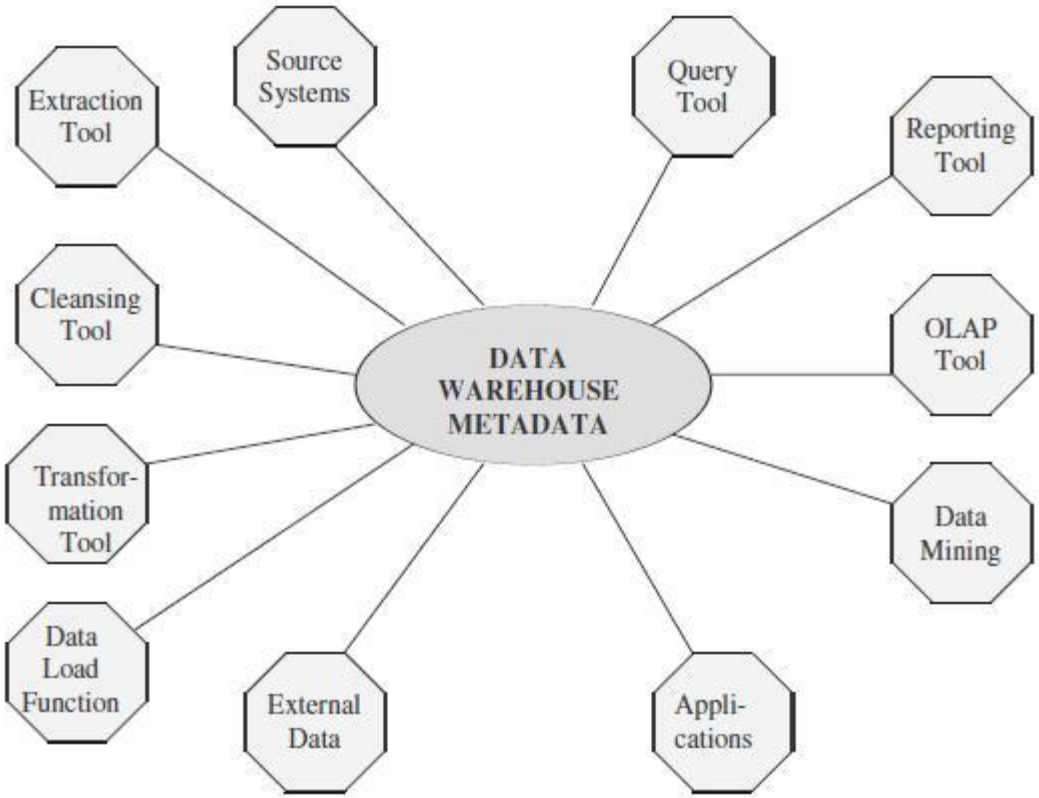


Figure 9-4 Metadata acts as a nerve center.

Why metadata is essential for IT

- Development and deployment of data warehouse is joint effort between users and IT staff.
- Technical issue-IT primarily responsible design and ongoing administration.
- Metadata essential for IT
- beginning with the data extraction and ending with information delivery.
- It drives and record the each processes.

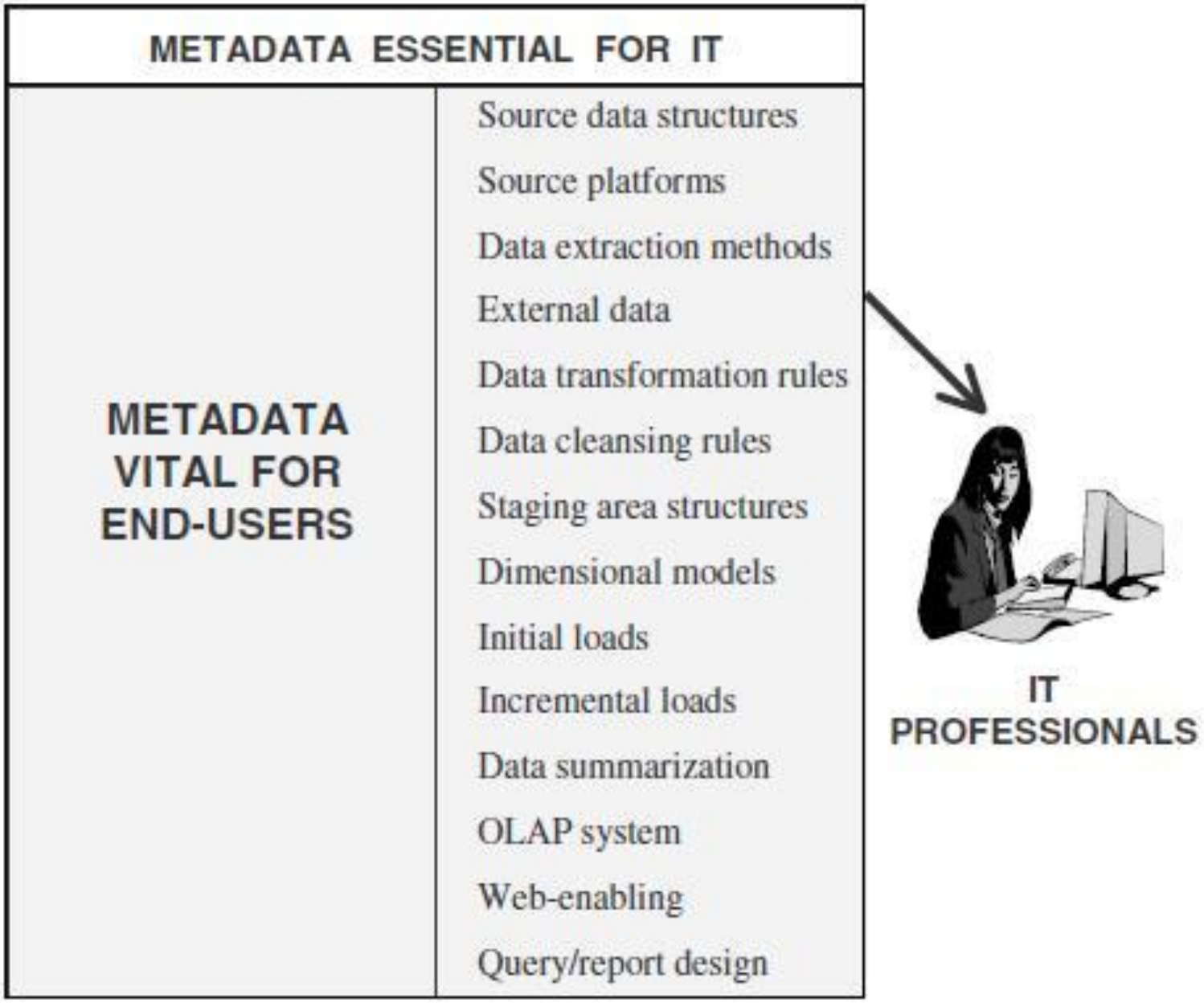


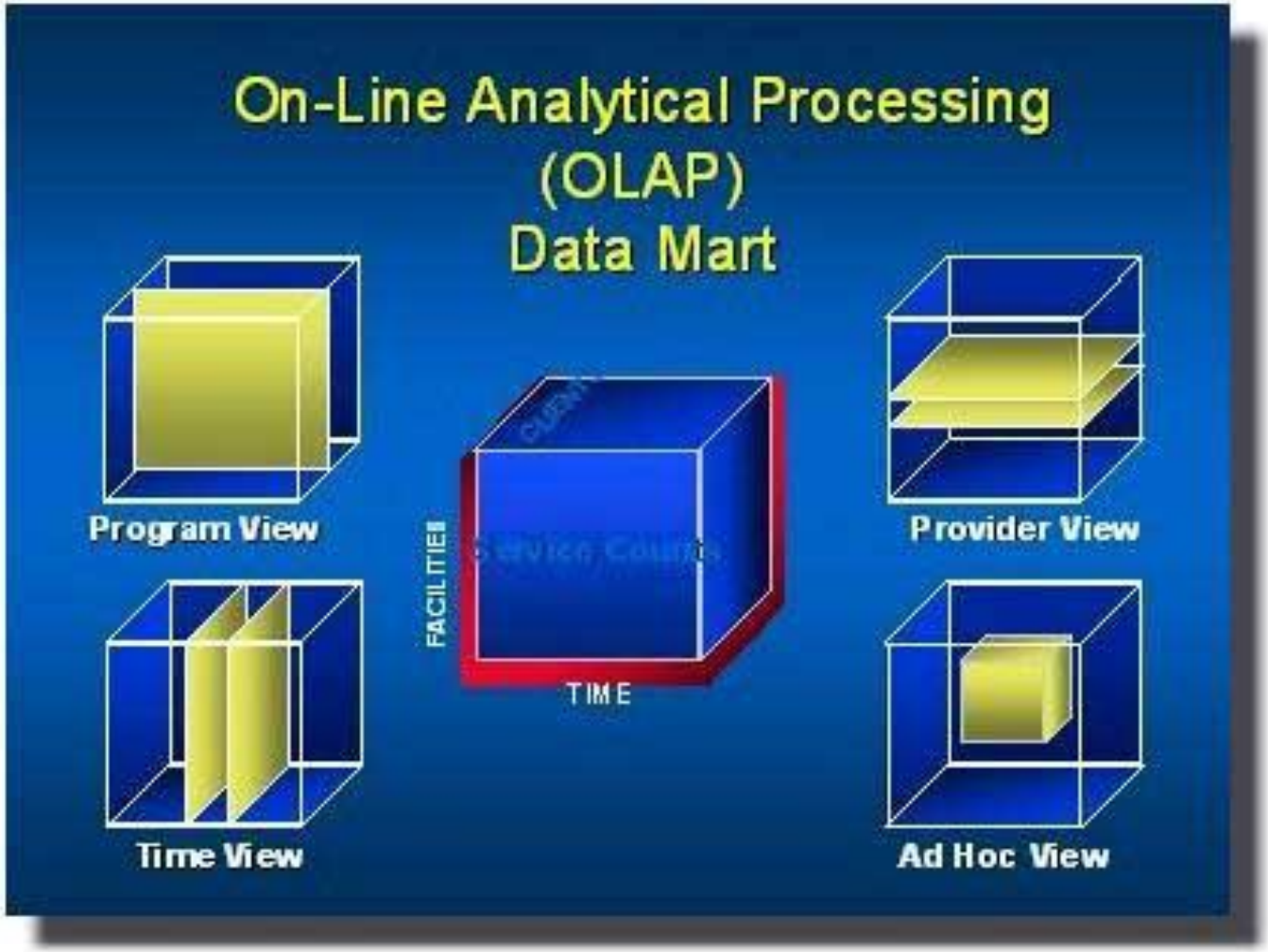
Figure 9-6 Metadata essential for IT.

OLAP

- Online Analytical Processing - coined by EF Codd in 1994 paper contracted by Arbor Software
- Generally synonymous with earlier terms such as Decisions Support, Business Intelligence, Executive Information System
- OLAP = Multidimensional Database

OLAP

- Online analytical processing refers to such end user activities as DSS modelling using spreadsheets and graphics that are done online.
- OLAP involves many different data items in complex relationships.
- Objective of OLAP is to analyze complex relationships and look for patterns, trends and exceptions.


On-Line Analytical Processing
(OLAP)
Data Mart
Program View
FACILITIES
TIME
Provider View
Time View
Ad Hoc View

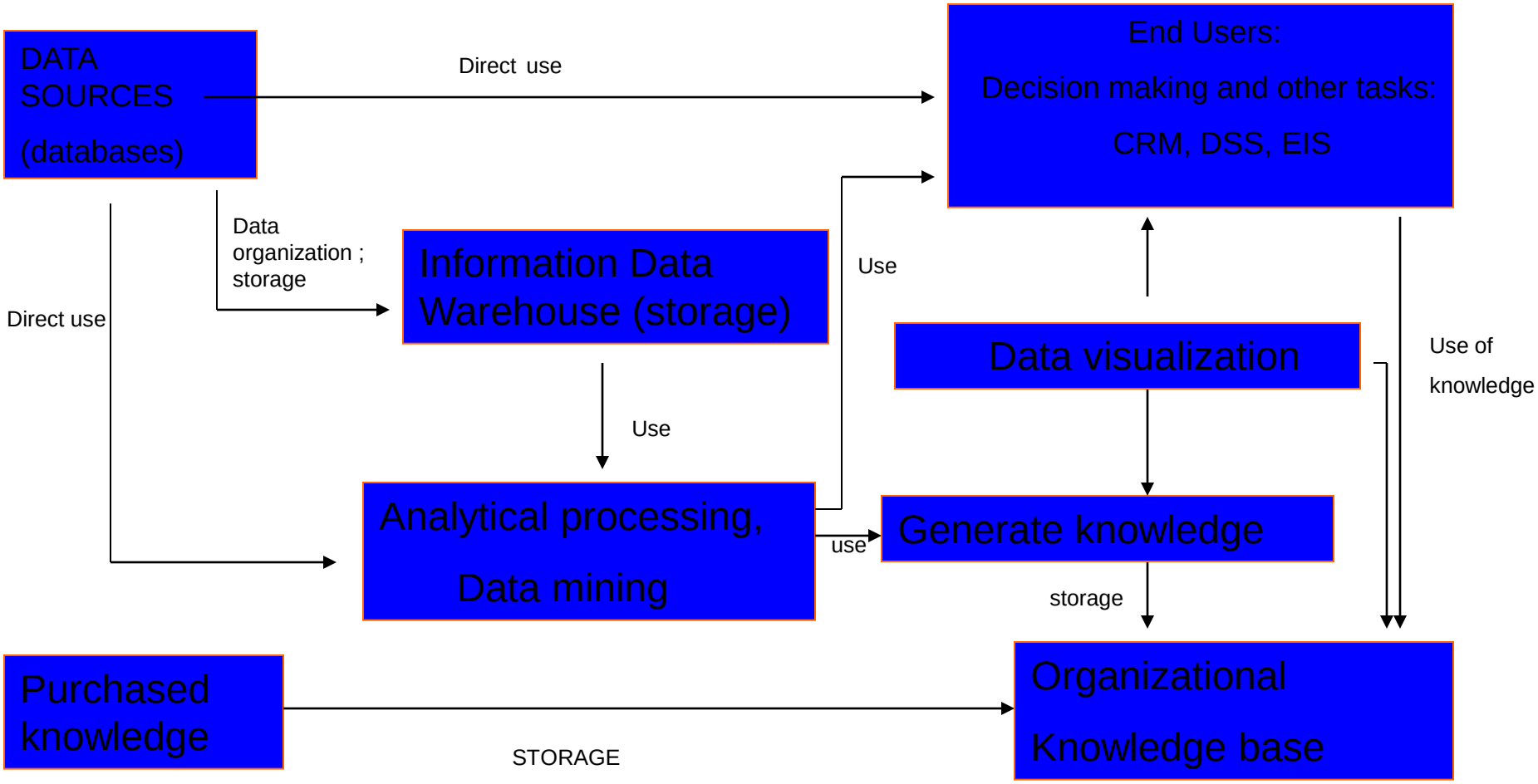
OLAP Is FASMI

- Fast
- Analysis
- Shared
- Multidimensional
- Information

Strengths of OLAP

- It is a powerful visualization paradigm
- It provides fast, interactive response times
- It is good for analyzing time series
- It can be useful to find some clusters and outliers
- Many vendors offer OLAP tools such as brio.com, cognus.com, microstrategy.com etc and it is possible to access an OLAP database from web.

Data warehousing integration



- Businesses run on information and the knowledge of how to put that information to use.

- Knowledge is not readily available, it is continuously constructed from data and/or information, in a process that may not be simple or easy.

- The transformation of data into knowledge may be accomplished in several ways

Data collection from various sources stored in simple databases

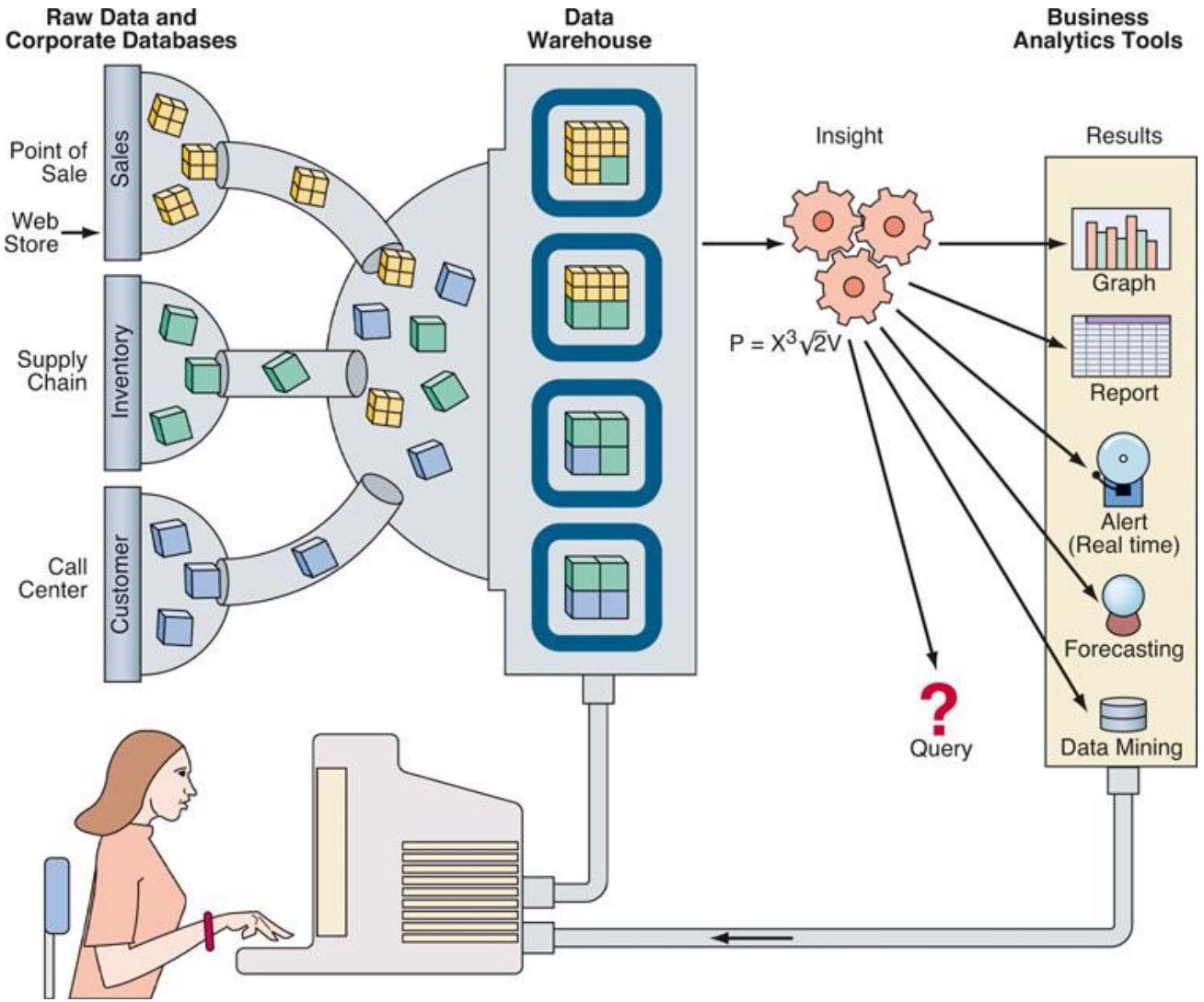
- Data can be processed, organized, and stored in a data warehouse and then analyzed (e.g.) by using analytical processing) by end users for decision support.
- Some of the data are converted to information prior to storage in the data warehouse, and some of the data and/or information can be analyzed to generate knowledge. For example, by using data mining, a process that looks for unknown relationships and patterns in the data, knowledge regarding the impact of advertising on a specific group of customers can be generated.
- This generated knowledge is stored in an organizational knowledge base, a repository of accumulated corporate knowledge and of purchased knowledge.
- The knowledge in the knowledge base can be used to support less experienced and users, or to support complex decision making.

Both the data and the information, at various times during the process, and the knowledge derived at the end of the process, may need to be presented to users.

Data Warehouse for Decision Support

- Putting Information technology to help the knowledge worker make faster and better decisions
- Used to manage and control business
- Data is historical or point-in-time
- Optimized for inquiry rather than update
- Use of the system is loosely defined and can be ad-hoc
- Used by managers and end-users to understand the business and make judgments

Business intelligence and data warehousing



Business Intelligence

- One ultimate use of the data gathered and processed in the data life cycle is for business intelligence.
- Business intelligence generally involves the creation or use of a data warehouse and/or data mart for storage of data, and the use of front-end analytical tools such as Oracle's Sales Analyzer and Financial Analyzer or Micro Strategy's Web.
- Such tools can be employed by end users to access data, ask queries, request ad hoc (special) reports, examine scenarios, create CRM activities, devise pricing strategies, and much more.

How business intelligence works?

- The process starts with raw data which are usually kept in corporate data bases. For example, a national retail chain that sells everything from grills and patio furniture to plastic utensils had data about inventory, customer information, data about past promotions, and sales numbers in various databases.
- Though all this information may be scattered across multiple systems-and may seem unrelated-business intelligence software can being it together. This is done by using a data warehouse.
- In the data warehouse (or mart) tables can be linked, and data cubes are formed. For instance, inventory information is linked to sales numbers and customer databases, allowing for deep analysis of information.

- Using the business intelligence software the user can ask queries, request ad-hoc reports, or conduct any other analysis.
- For example, deep analysis can be carried out by performing multilayer queries. Because all the databases are linked, one can search for what products a store has too much of, determine which of these products commonly sell with popular items, bases on previous sales. After planning a promotion to move the excess stock along with the popular products (by bundling them together, for example), one can dig deeper to see where this promotion would be most popular (and most profitable). The results of the request can be reports, predictions, alerts, and/or graphical presentations. These can be disseminated to decision makers to help them in their decision-making tasks.

More advanced applications of business intelligence include outputs such as

- financial modeling
- budgeting
- resource allocation
- and competitive intelligence.

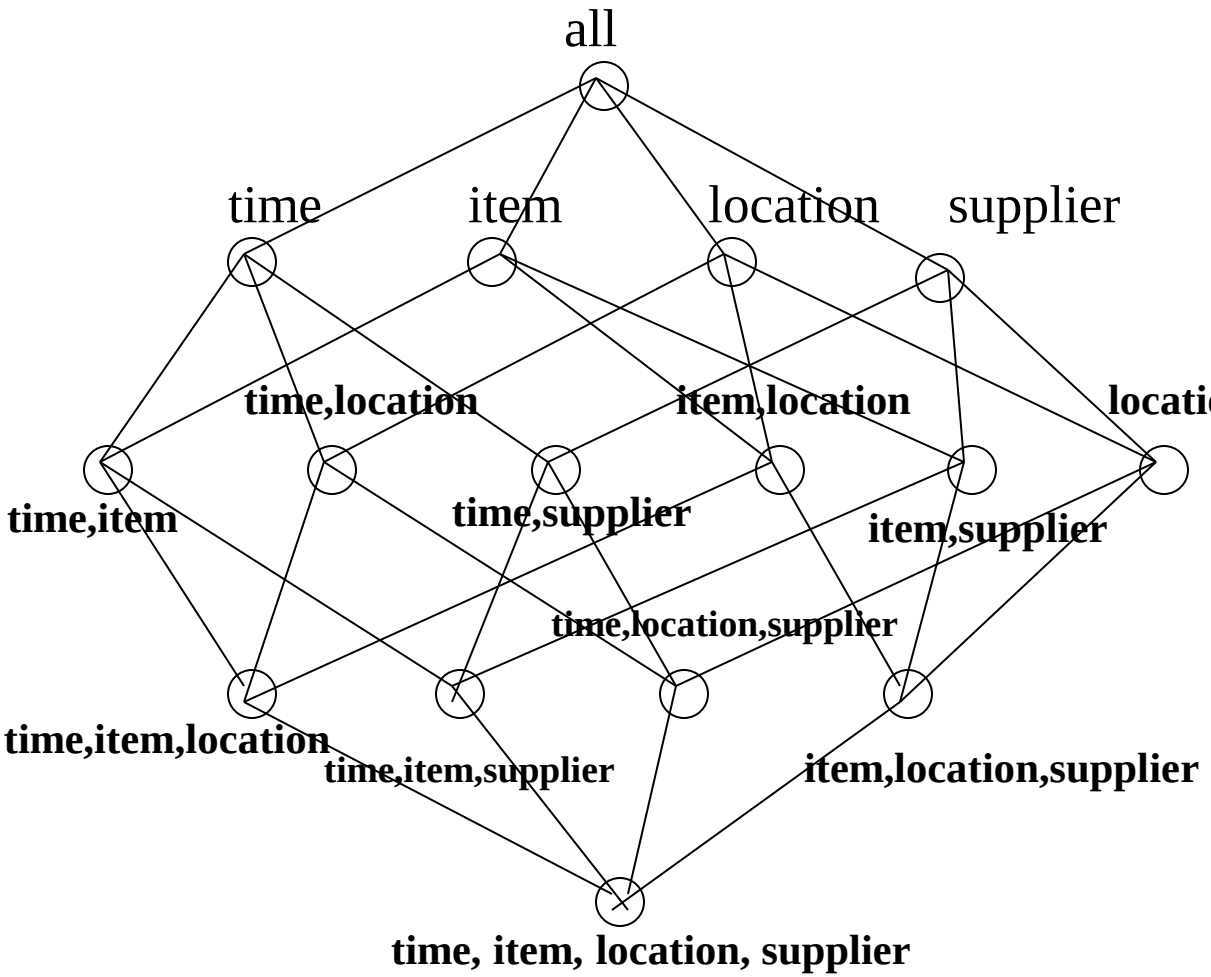
OLAP Operations: Data Cube

- The key operation of a OLAP is the formation of a **data cube**
  - A data cube is a multidimensional representation of data, together with all possible aggregates.
  - Aggregates: similar to class attribute
    - result by selecting a proper subset of the dimensions and summing over all remaining dimensions.
    - Cached to improve speed and support online computation
  - For example,
    - if we choose the species type dimension of the Iris data and
      - sum over all other dimensions,
      - the result will be a one-dimensional entry with three entries,
  - each of which gives the number of flowers of each type.

From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables

Cube: A Lattice of Cuboids



0-D(apex) cuboid

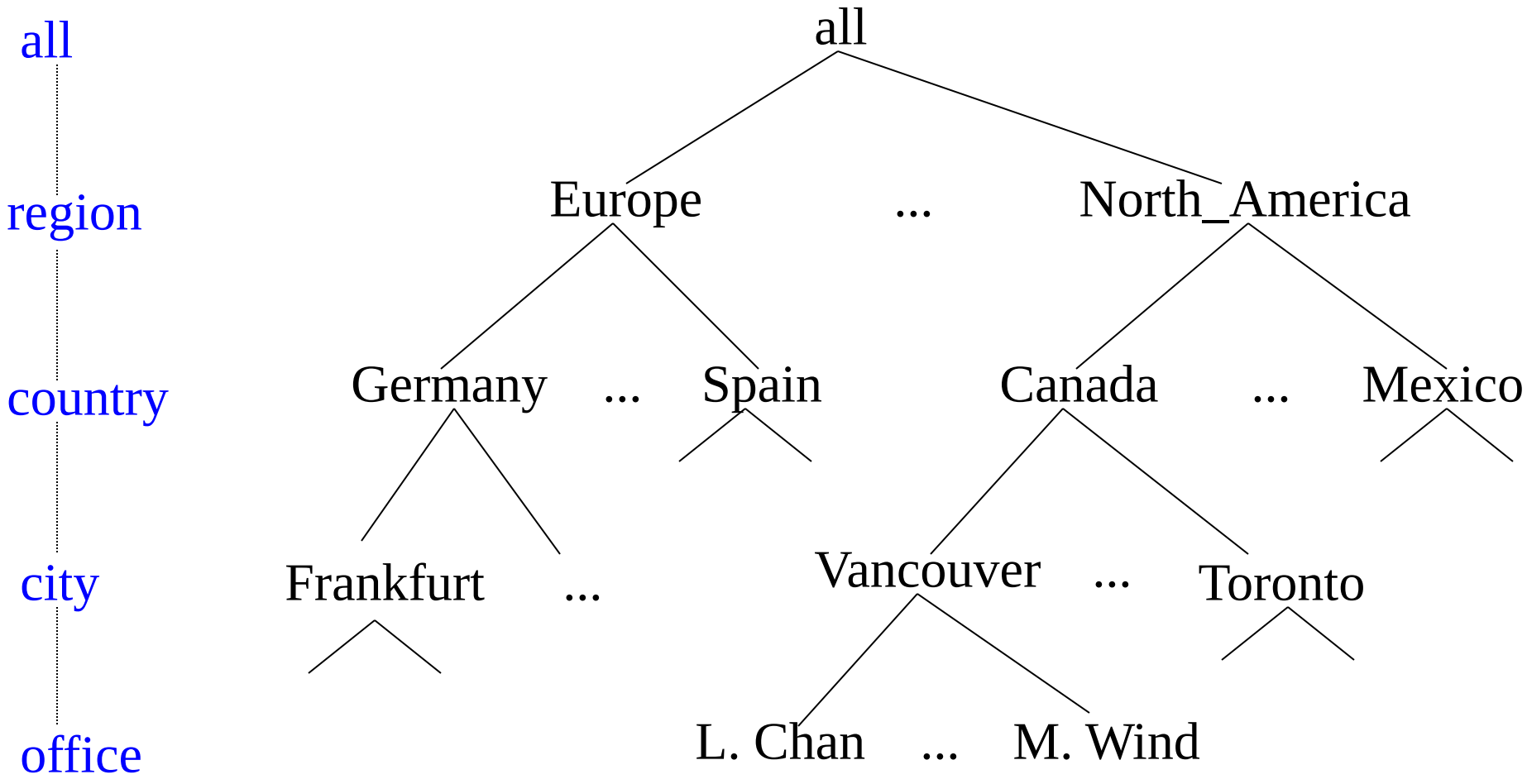
1-D cuboids

2-D cuboids

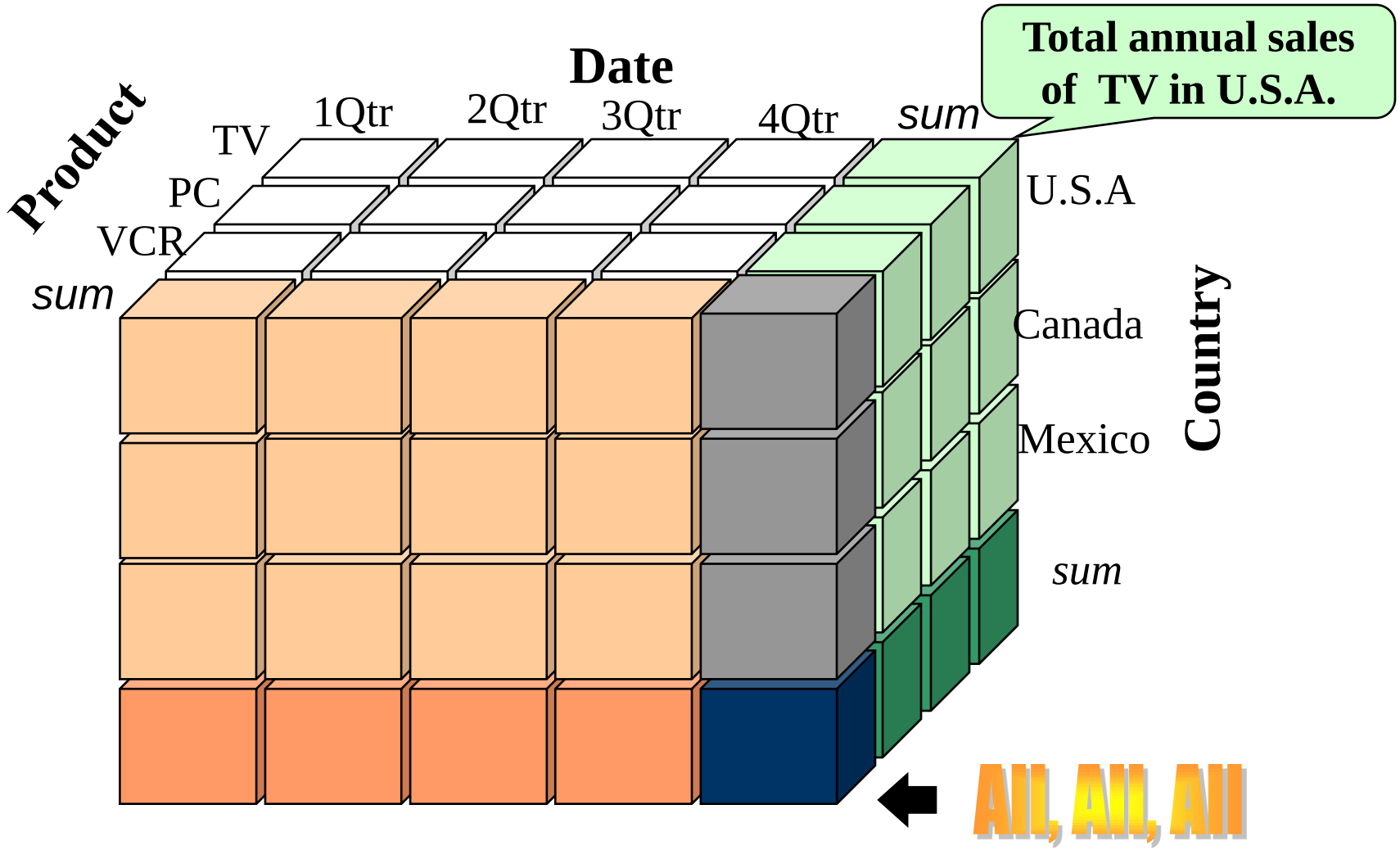
3-D cuboids

4-D(base) cuboid

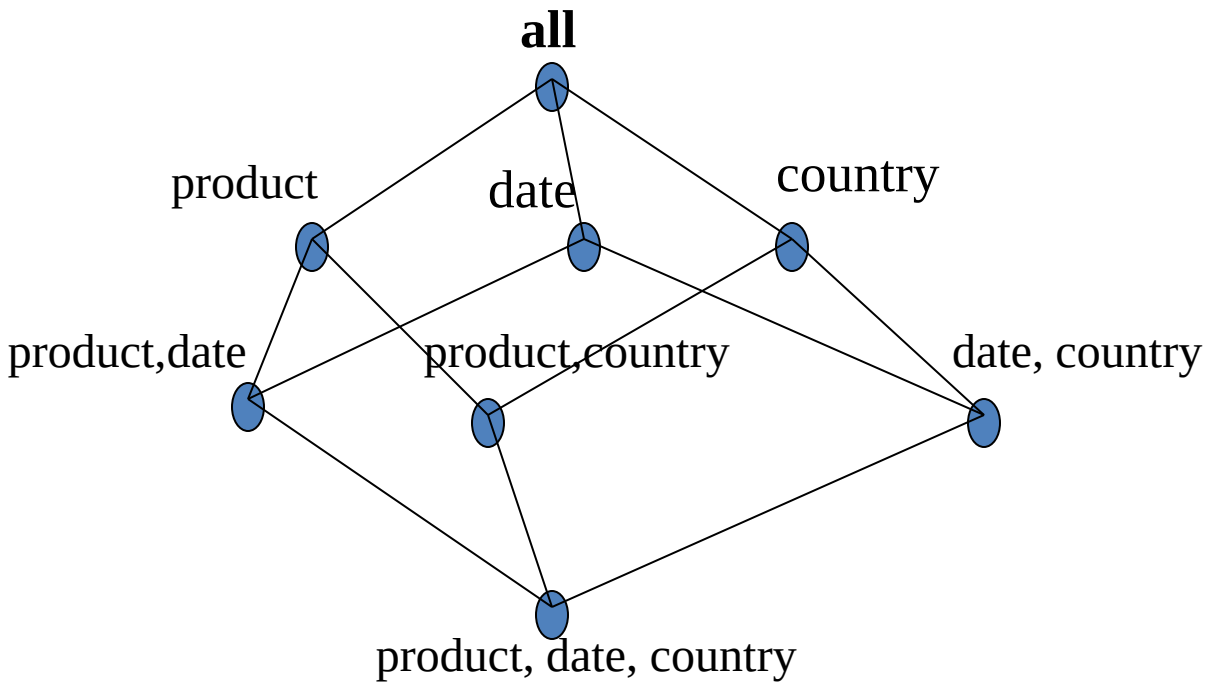
A Concept Hierarchy: Dimension (location)



A Sample Data Cube



Cuboids Corresponding to the Cube



0-D(apex) cuboid

1-D cuboids

2-D cuboids

3-D(base) cuboid

Data Cube Example (continued)

- The following figure table shows one of the two dimensional aggregates, along with two of the one-dimensional aggregates, and the overall

product ID \ date	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127

OLAP Operations: Slicing and Dicing

- **Slicing** is selecting a group of cells from the entire multidimensional array by specifying a specific value for one or more dimensions.
- **Dicing** involves selecting a subset of cells by specifying a range of attribute values.
  - This is equivalent to defining a subarray from the complete array.
- In practice, both operations can also be accompanied by aggregation over some dimensions.

OLAP Operations: Roll-up and Drill-down

- This hierarchical structure gives rise to the **roll-up and drill-down** operations.
  - For sales data, we can aggregate (roll up) the sales across all the dates in a month.
  - Conversely, given a view of the data where the time dimension is broken into months, we could split the monthly sales totals (drill down) into daily sales totals.
  - Likewise, we can drill down or roll up on the location or product ID attributes.