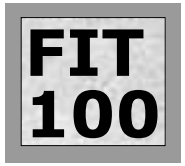
Fluency With Information Technology
CSE100/IMT100



Larry Snyder & Mel Oyler, Instructors

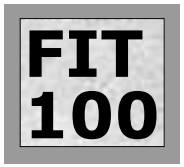
Ariel Kemp, Isaac Kunen, Gerome Miklau &
Sean Squires, Teaching Assistants

University of Washington, Autumn 1999

FIT 100

Enterprise Database Systems Learning Objectives

- Understand needs and concepts of decision support systems
- Understand concepts and issues of the Data Warehouse
- Understand concepts of On-Line Analytical Processing (OLAP)
- Understand concepts of data mining



Data Warehousing Overview

- The Need for Data Analysis
- Decision Support Systems
- The Data Warehouse
- On-Line Analytical Processing (OLAP)
- Data Mining

FIT 100

The Need for Data Analysis

Executives compete at decision making

- **Constant pressure from external and internal forces requires prompt tactical and strategic decisions.**
- **The decision-making cycle time is reduced, while problems are increasingly complex with a growing number of internal and external variables.**
- **Managers need support systems for facilitating quick decision making in a complex environment.**
- **Decision support systems (DSS).**

FIT 100 Decision Support Systems (DSS)

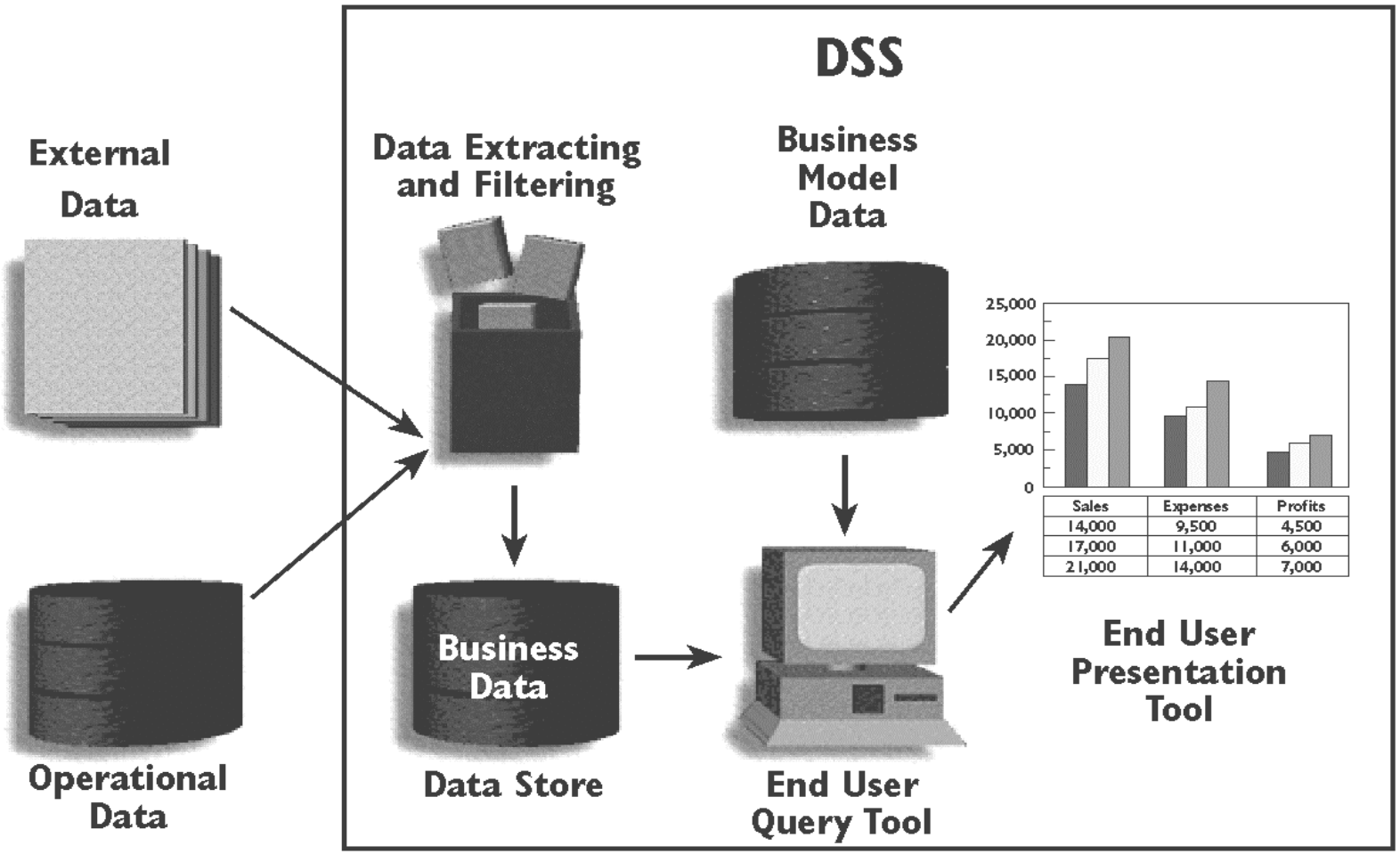
- Decision Support **is a methodology (or a series of methodologies) designed to extract information from data and to use such information as a basis for decision making.**
- **A** decision support system **(**DSS**) is an arrangement of computerized tools used to assist managerial decision making within a business.**
  - A DSS usually requires extensive data “massaging” to produce information.
  - The DSS is used at all levels within an organization and is often tailored to focus on specific business area or problems.
  - The DSS is interactive and provides *ad hoc* query tools to retrieve data and to display data in different formats.

FIT 100 Decision Support Systems

- Four Components of a DSS
  - The **data store** component is a basically a DSS database.
  - The **data extraction** and **filtering** component is used to extract and validate the data taken from the operational database and the external data sources.
  - The **end user query tool** is used by the data analyst to create the queries that access the database.
  - The **end user presentation tool** is used by the data analyst to organize and present the data.

FIT 100

Main Components of a DSS



FIT 100 Decision Support Systems

Three Main Areas in Which DSS Data Differ from Operational Data

- **Time span**
  - Operational data represent current (atomic) transactions.
  - DSS data tend to cover a longer time frame.
- **Granularity**
  - Operational data represent specific transactions that occur at a given time.
  - DSS data must be presented at different levels of aggregation.
- **Dimensionality**
  - Operational data focuses on representing atomic transactions.
  - DSS data can be analyzed from multiple dimensions.

FIT 100 Decision Support Systems

- **End-User Analytical Interface**
  - The DSS DBMS must support advanced data modeling and data presentation tools, data analysis tools, and query generation and optimization components.
  - The end user analytical interface is one of the most critical components.
- **Database Size Requirements**
  - The DBMS must be capable of supporting very large databases (VLDB).

FIT 100 The Data Warehouse

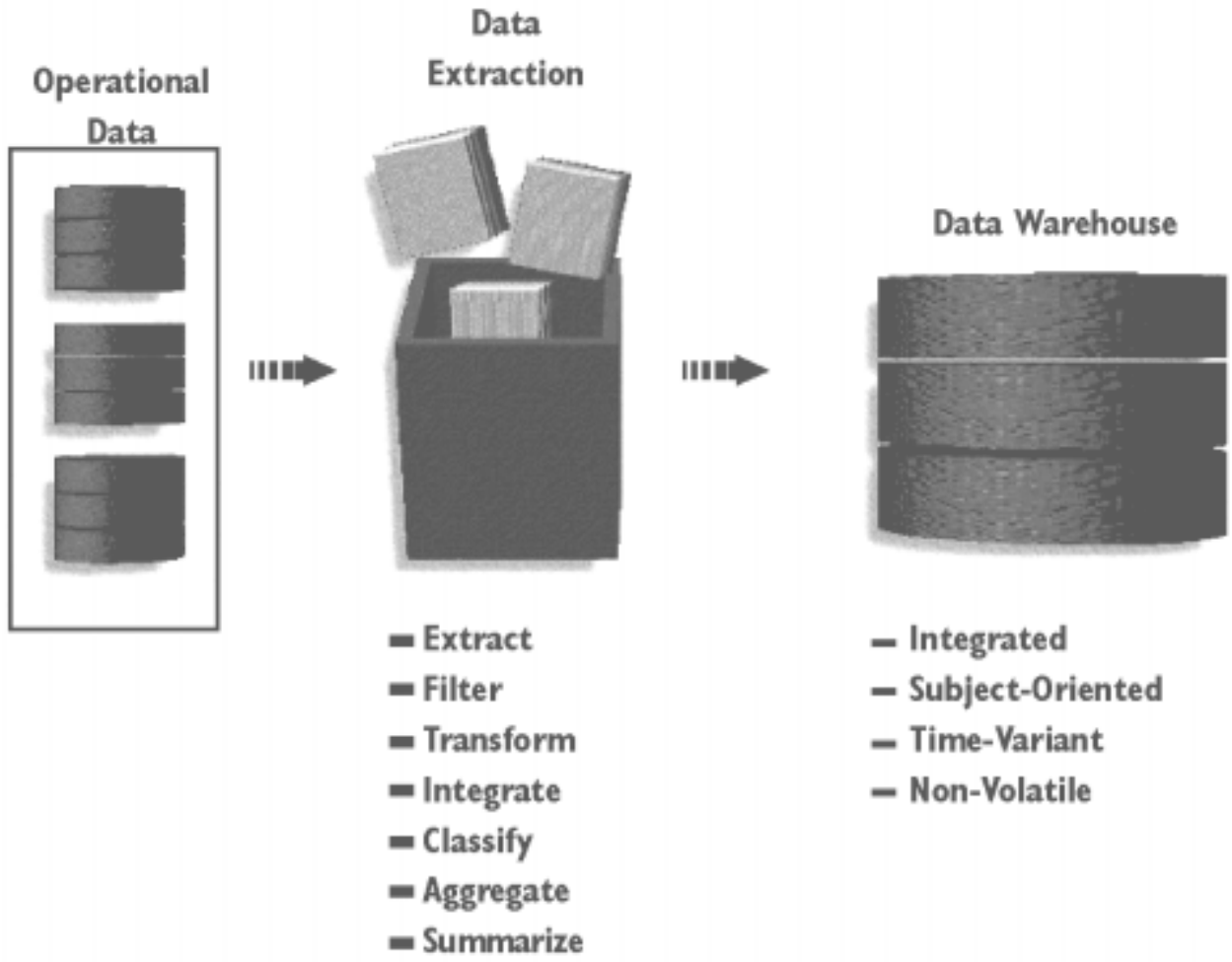
- **The** Data Warehouse **is an integrated, subject-oriented, time-variant, non-volatile database that provides support for decision making.**
  - **Integrated**
    - The Data Warehouse is a centralized, consolidated database that integrates data retrieved from the entire organization.
  - **Subject-Oriented**
    - The Data Warehouse data is arranged and optimized to provide answers to questions coming from diverse functional areas within a company.

FIT 100 The Data Warehouse

- **Time Variant**
  - The Warehouse data represent the flow of data through time. It can even contain projected data.
- **Non-Volatile**
  - Once data enter the Data Warehouse, they are never removed.
  - The Data Warehouse is always growing.

FIT 100

Creating a Data Warehouse



FIT 100 The Data Mart

- Data Mart
  - A Data Mart is a small, single-subject Data Warehouse subset that provides decision support to a small group of people.
  - Data Marts can serve as a test vehicle for companies exploring the potential benefits of Data Warehouses.
  - Data Marts address local or departmental problems, while a Data Warehouse involves a company-wide effort to support decision making at all levels in the organization.

FIT 100 The Data Warehouse

- Twelve Rules That Define a Data Warehouse
  1. The Data Warehouse and operational environments are separated.
  2. The Data Warehouse data are integrated.
  3. The Data Warehouse contains historical data over a long time horizon.
  4. The Data Warehouse data are snapshot data captured at a given point in time.
  5. The Data Warehouse data are subject-oriented.
  6. The Data Warehouse data are mainly read-only with periodic batch updates from operational data. No online updates are allowed.
  7. The Data Warehouse development life cycle differs from classical systems development. The Data Warehouse development is data driven; the classical approach is process driven.

FIT 100 The Data Warehouse

8. The Data Warehouse contains data with several levels of detail; current detail data, old detail data, lightly summarized, and highly summarized data.
9. The Data Warehouse environment is characterized by read-only transactions to very large data sets. The operational environment is characterized by numerous update transactions to a few data entities at the time.
10. The Data Warehouse environment has a system that traces data resources, transformation, and storage.
11. The Data Warehouse's metadata are a critical component of this environment. The metadata identify and define all data elements. The metadata provide the source, transformation, integration, storage, usage, relationships, and history of each data element.
12. The Data Warehouse contains a charge-back mechanism for resource usage that enforces optimal use of the data by end users.

FIT 100 On-Line Analytical Processing

- On-Line Analytical Processing **(**OLAP**) is an advanced data analysis environment that supports decision making, business modeling, and operations research activities.**
- Four Main Characteristics of OLAP
  - Use multidimensional data analysis techniques.
  - Provide advanced database support.
  - Provide easy-to-use end user interfaces.
  - Support client/server architecture.

FIT 100 On-Line Analytical Processing

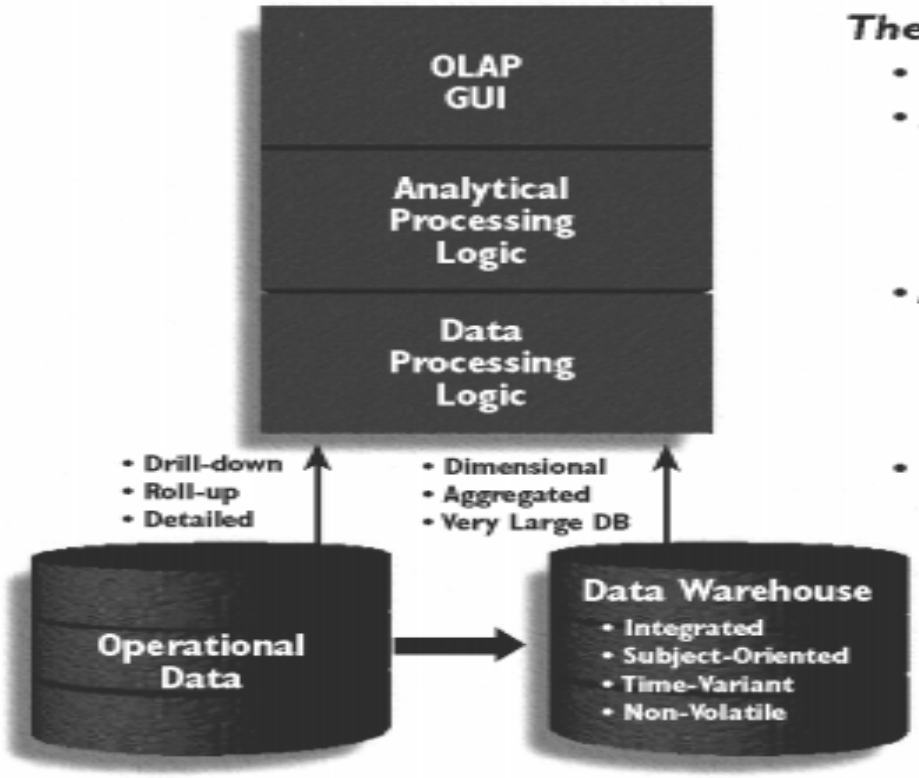
- Additional Functions of Multidimensional Data Analysis Techniques
  - Advanced data presentation functions
  - Advanced data aggregation, consolidation, and classification functions
    - Customer segment analysis on eCommerce web sites
  - Advanced computational functions
  - Advanced data modeling functions

FIT 100 On-Line Analytical Processing (OLAP)

- OLAP Architecture
  - Three Main Modules
    - OLAP Graphical User Interface (GUI)
    - OLAP Analytical Processing Logic
    - OLAP Data Processing Logic
  - OLAP systems are designed to use both operational and Data Warehouse data.

FIT 100

OLAP Client/Server Architecture

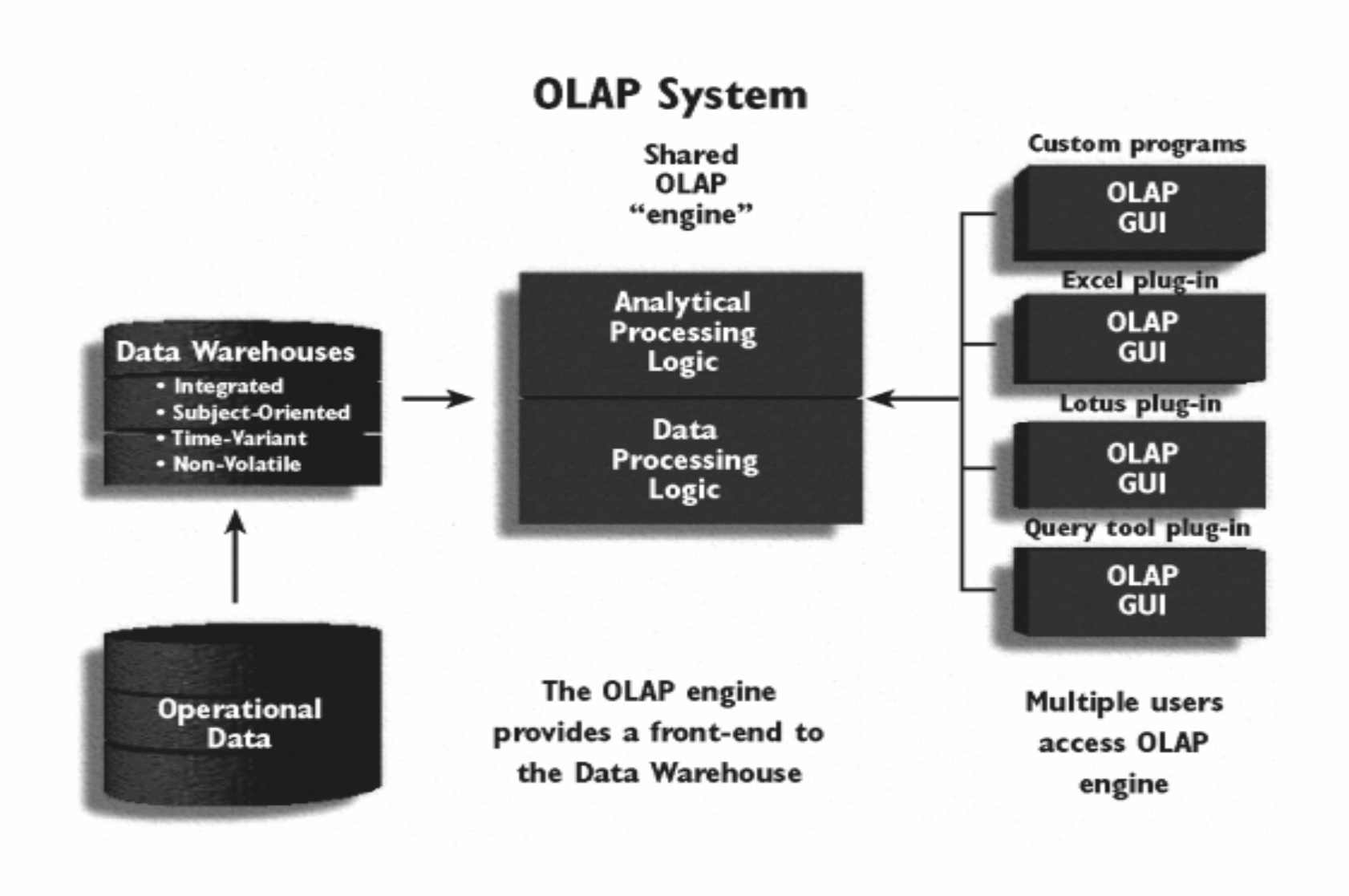


The OLAP system exhibits:

- *Client/Server Architecture*
- *Easy-to-use GUI*
  - *Dimensional Presentation*
  - *Dimensional Modeling*
  - *Dimensional Analysis*
- *Multidimensional Data*
  - *Analysis*
  - *Manipulation*
  - *Structure*
- *Database Support*
  - *Data Warehouse*
  - *Operational DB*
  - *Relational*
  - *Multidimensional*

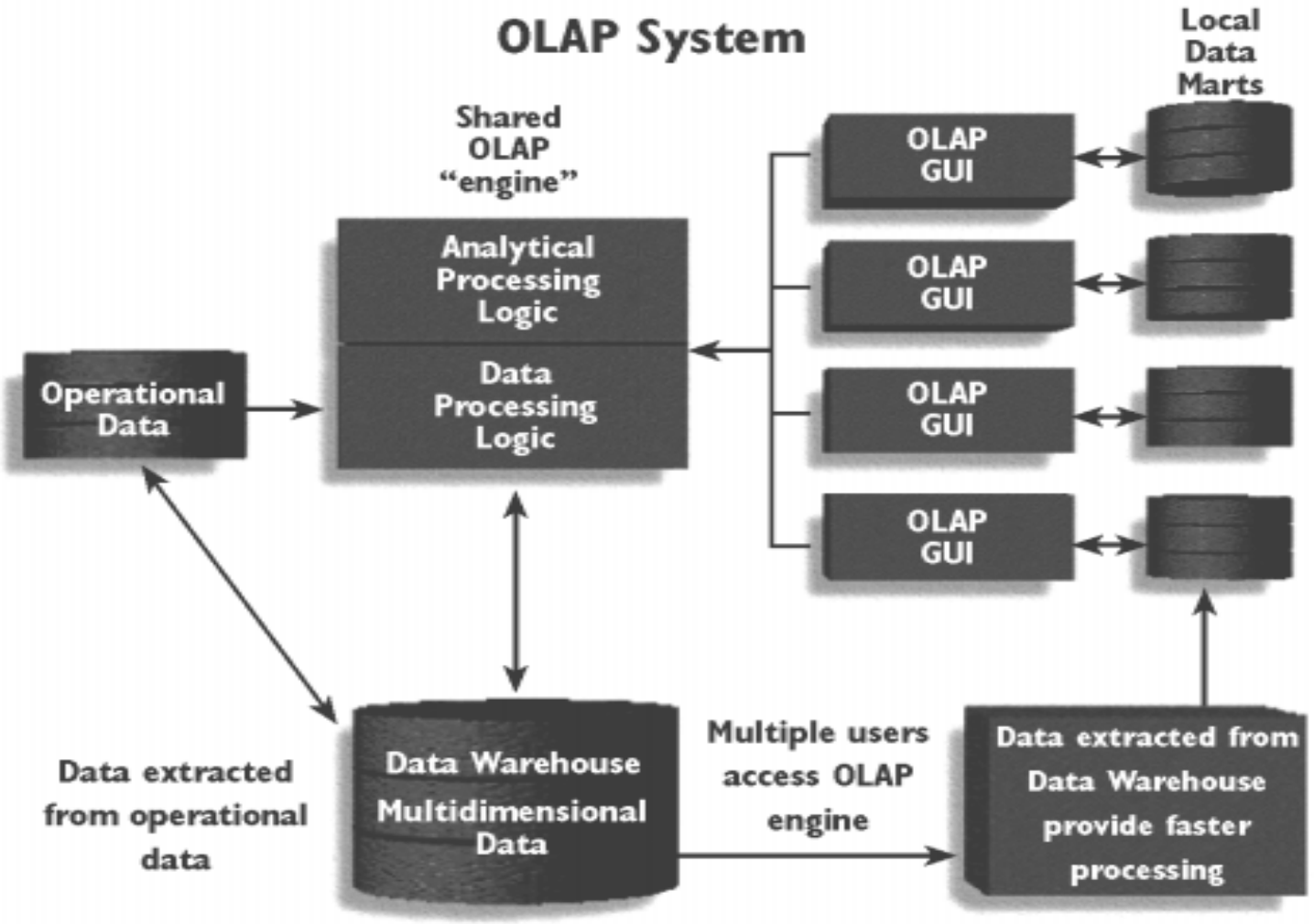
FIT 100

OLAP Server Arrangement



FIT 100

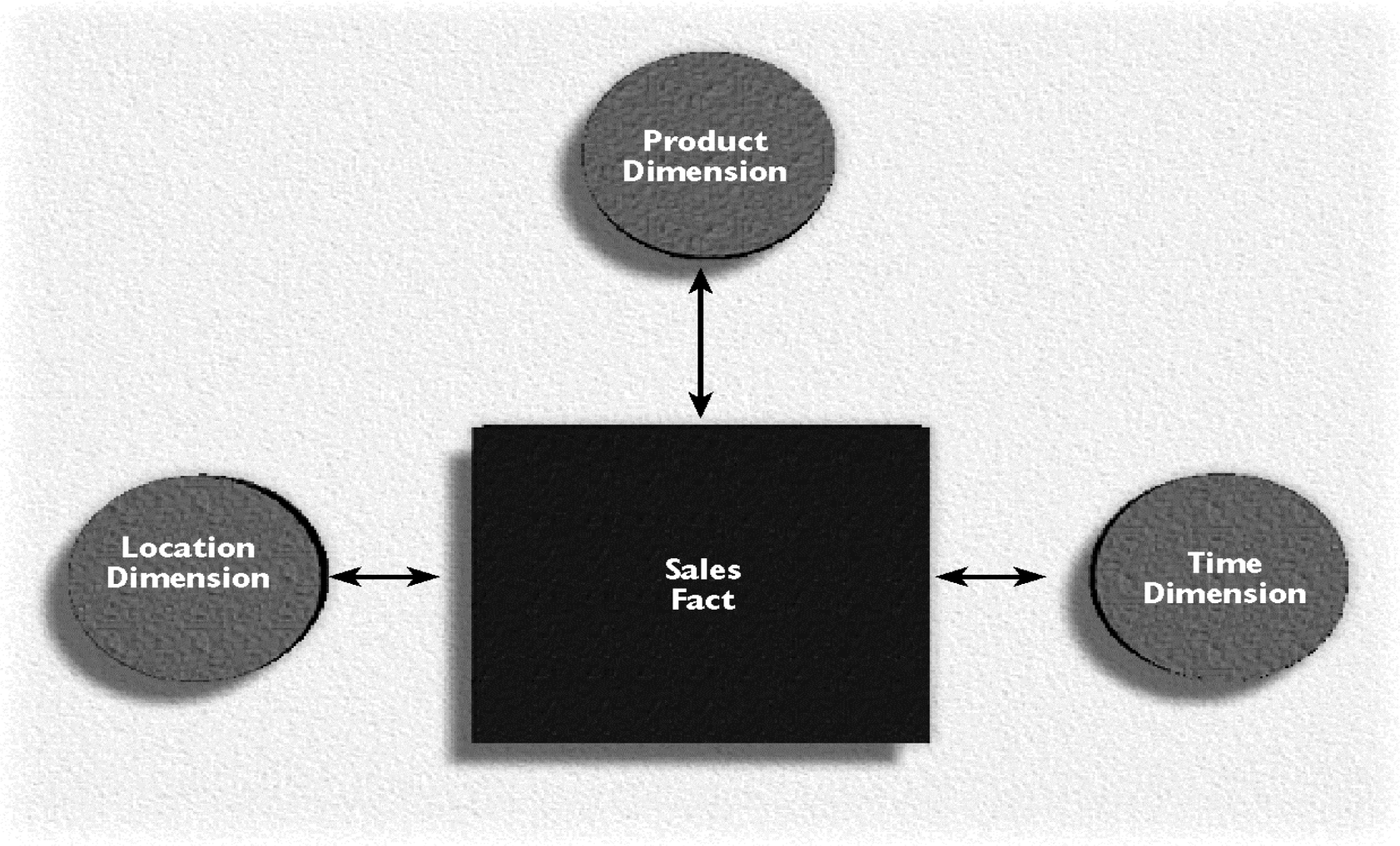
OLAP Server with Local Data Marts



FIT 100 Star Schema

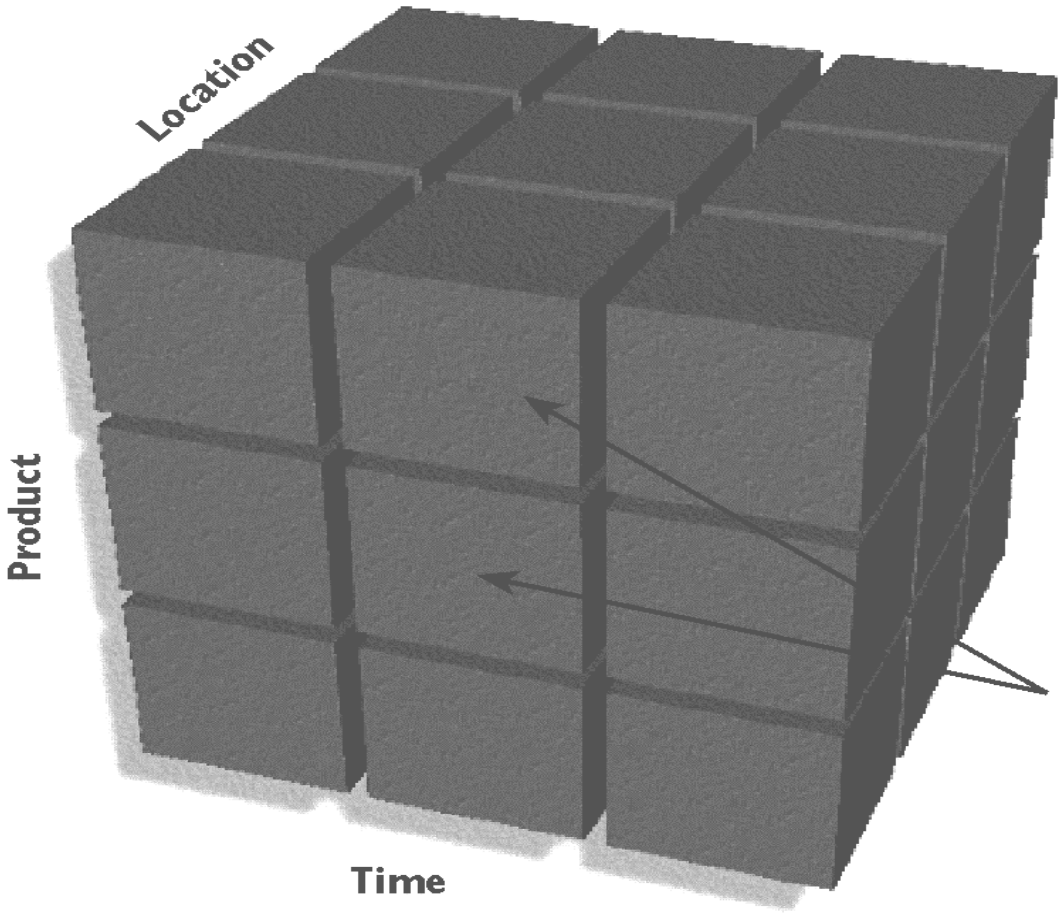
- Dimensions
  - **Dimensions** are qualifying characteristics that provide additional perspectives to a given fact.
  - Dimensions are stored in dimension tables.

FIT 100 A Simple Star Schema



FIT 100

A Three-Dimensional View of Sales

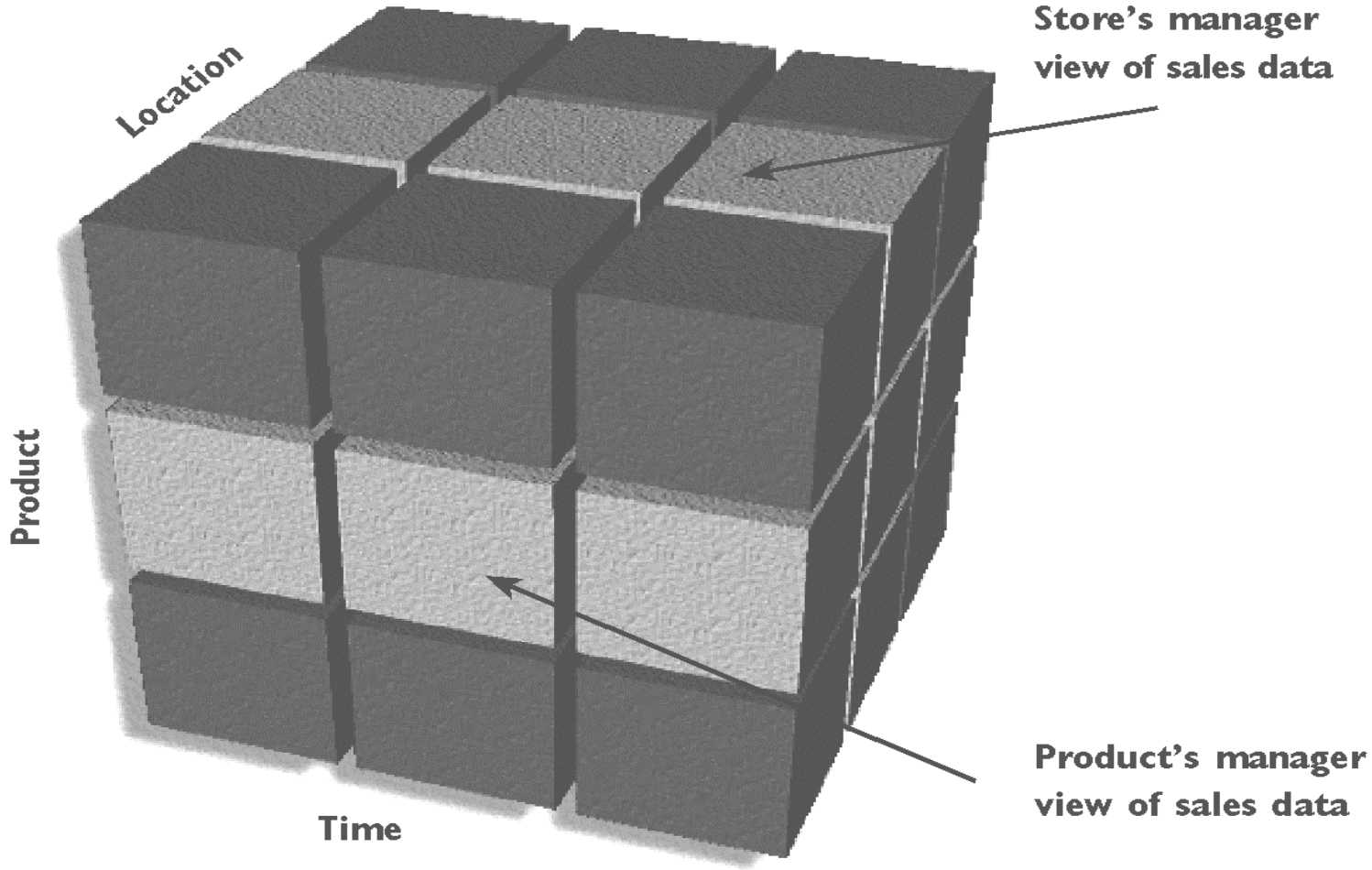


Conceptual three-dimensional cube of sales by product, location, and time.

Sales facts are stored in the cells at the intersection of each product, time, and location dimension.

FIT 100

Slice and Dice View of Sales



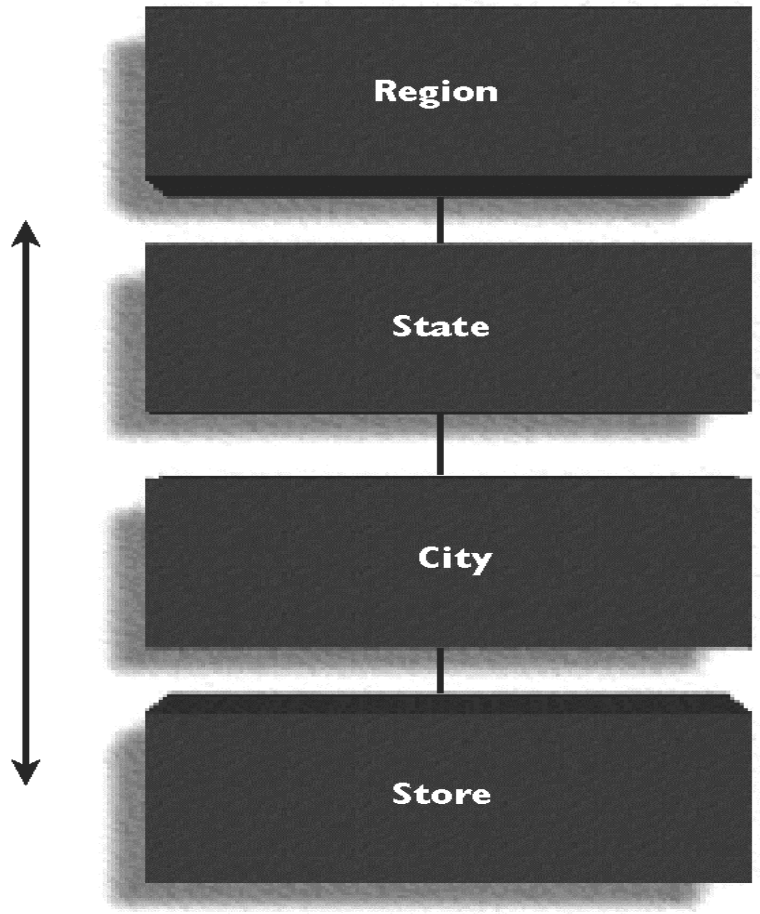
FIT 100 Star Schema

- Attribute Hierarchies
  - Attributes within dimensions can be ordered in a well-defined attribute hierarchy.
  - The attribute hierarchy provides a top-down data organization that is used for two main purposes:
    - Aggregation.
    - Drill-down/roll-up data analysis.

FIT 100

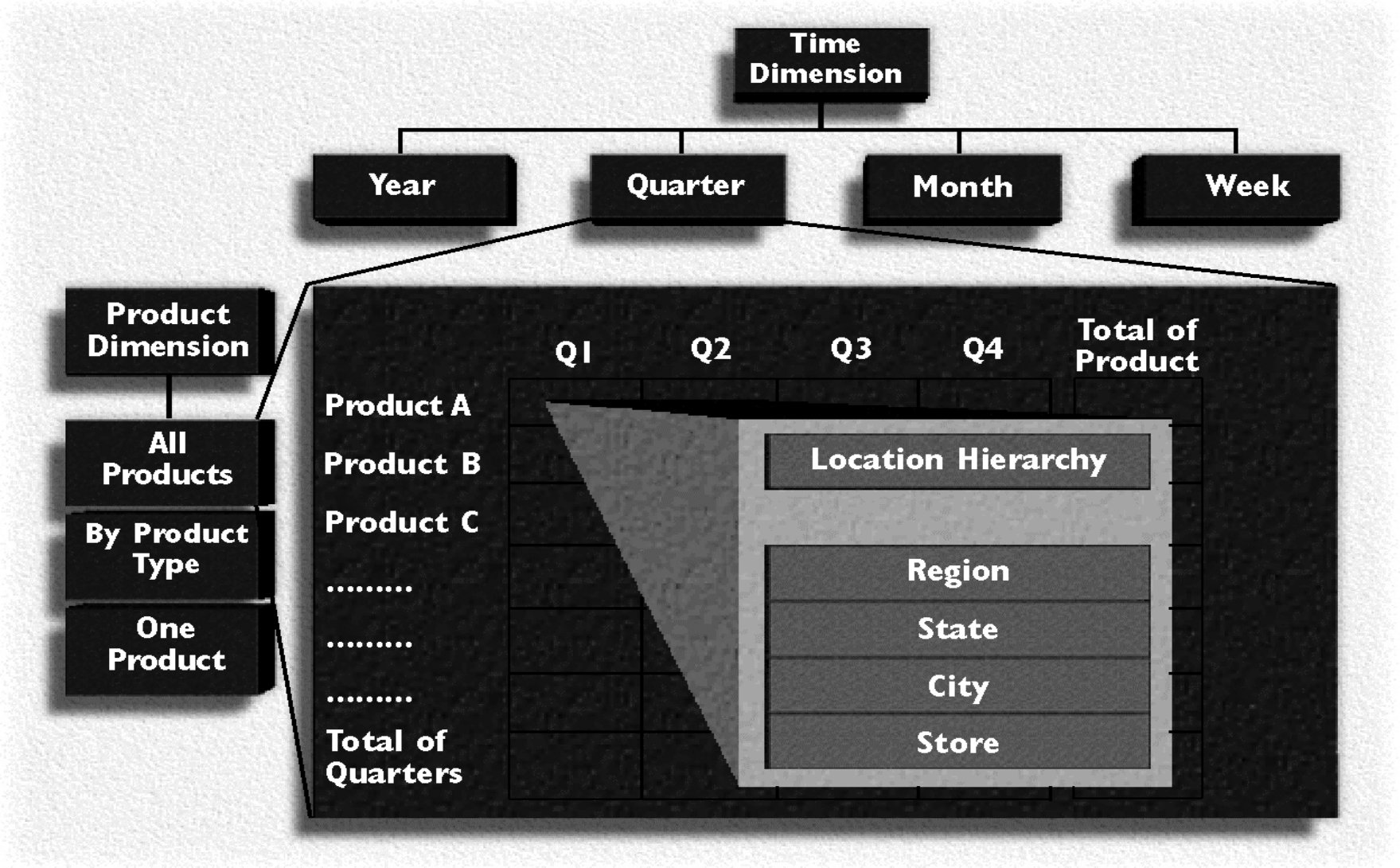
A Location Attribute Hierarchy

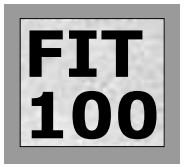
The attribute hierarchy allows the end user to perform drill-down and roll-up searches.



FIT 100

Attribute Hierarchies in Multidimensional Analysis





Star Schema

- Performance-Improving Techniques
  - Normalization of dimensional tables
  - Multiple fact tables representing different aggregation levels
  - Denormalization of fact tables
  - Table partitioning and replication

FIT 100 Data Warehouse Implementation

- The Data Warehouse as an Active Decision Support Network
  - A Data Warehouse is a dynamic support framework.
  - Implementation of Data Warehouse is part of a complete databased-system-development infrastructure for company-wide decision support.
  - Its design and implementation must be examined in the light of the entire infrastructure.

FIT 100 Data Mining

- **In contrast to the traditional (reactive) DSS tools, the data mining premise is proactive.**
- **Data mining tools automatically search the data for anomalies and possible relationships, thereby identifying problems that have not yet been identified by the end user.**
- **Data mining tools -- based on algorithms that form the building blocks for artificial intelligence, neural networks, inductive rules, and predicate logic -- initiate analyses to create knowledge.**

FIT 100 Data Mining

- Four Phases of Data Mining
  1. **Data Preparation**
     - Identify and cleanse data sets.
     - Data Warehouse is usually used for data mining operations.
  2. **Data Analysis and Classification**
     - Identify common data characteristics or patterns using
       - Data groupings, classifications, clusters, or sequences.
       - Data dependencies, links, or relationships.
       - Data patterns, trends, and deviations.

FIT 100 Data Mining

3. Knowledge Acquisition

- Select the appropriate modeling or knowledge acquisition algorithms.
- Examples: neural networks, decision trees, rules induction, genetic algorithms, classification and regression tree, memory-based reasoning, or nearest neighbor and data visualization).

4. Prognosis

- Predict future behavior and forecast business outcomes using the data mining findings.

FIT 100 Data Mining