Chapter 3

DATA WAREHOUSING

Learning Objectives

- Understand the basic definitions and concepts of data warehouses
- Understand data warehousing architectures
- 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

Operational vs. multidimensional view of sales

Database name: Ch13_Text

Table name: DW_INVOICE

INV_NUM	INV_DATE	CUS_NAME	INV_TOTAL
2034	15-May-10	Dartonik	1400.00
2035	15-May-10	Summer Lake	1200.00
2036	16-May-10	Dartonik	1350.00
2037	16-May-10	Summer lake	3100.00
2038	16-May-10	Trydon	400.00

Operational Data

Table name: DW_LINE

INV_NUM	LINE_NUM	PROD_DESCRIPTION	LINE_PRICE	LINE_QUANTITY	LINE_AMOUNT
2034	1	Optical Mouse	45.00	20	900.00
2034	2	Wireless RF remote and laser pointer	50.00	10	500.00
2035	1	Everlast Hard Drive, 60 GB	200.00	6	1200.00
2036	1	Optical Mouse	45.00	30	1350.00
2037	1	Optical Mouse	45.00	10	450.00
2037	2	Readster 56KB Ext. Modem	120.00	5	600.00
2037	3	Everlast Hard Drive, 60 GB	205.00	10	2050.00
2038	1	NoTech Speaker Set	50.00	8	400.00

Multidimensional View of Sales

	Time Dimension		
Customer Dimension	15-May-10	16-May-10	Totals
Dartonik	$1,400.00	$1,350.00	$2,750.00
Summer Lake	$1,200.00	$3,100.00	$4,300.00
Trydon		$400.00	$400.00
Totals	$2,600.00	$4,850.00	$7,450.00

Sales are located in the intersection of a customer row and time column.

Aggregations are provided for both dimensions.

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

Data Warehousing Definitions and Concepts

- **Data mart**

  A departmental data warehouse that stores only relevant data

- **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)**

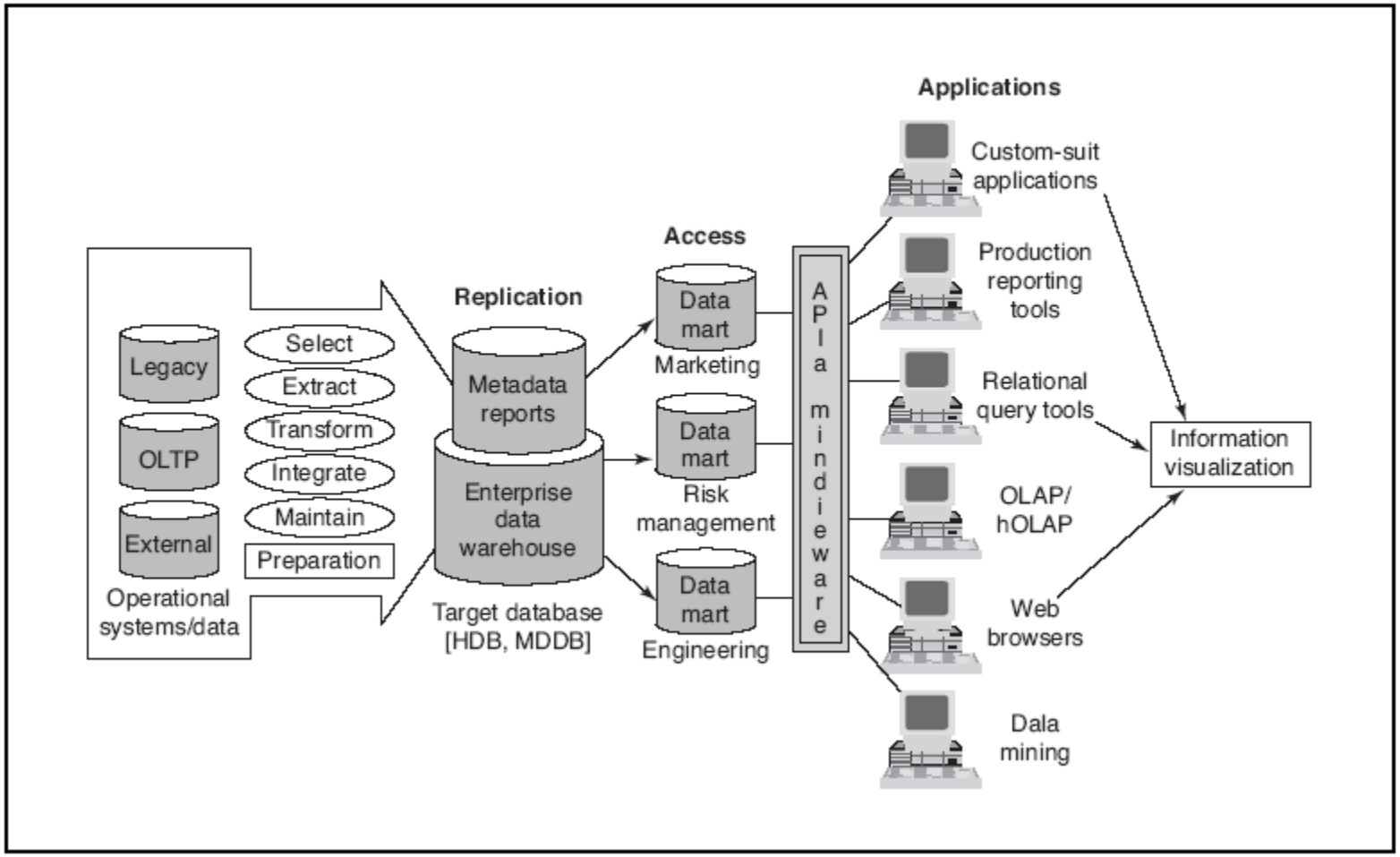
  A technology that provides a vehicle for pushing data from source systems into a data warehouse

- **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

FIGURE 5.1 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

- 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?*

Data Warehousing Process Overview

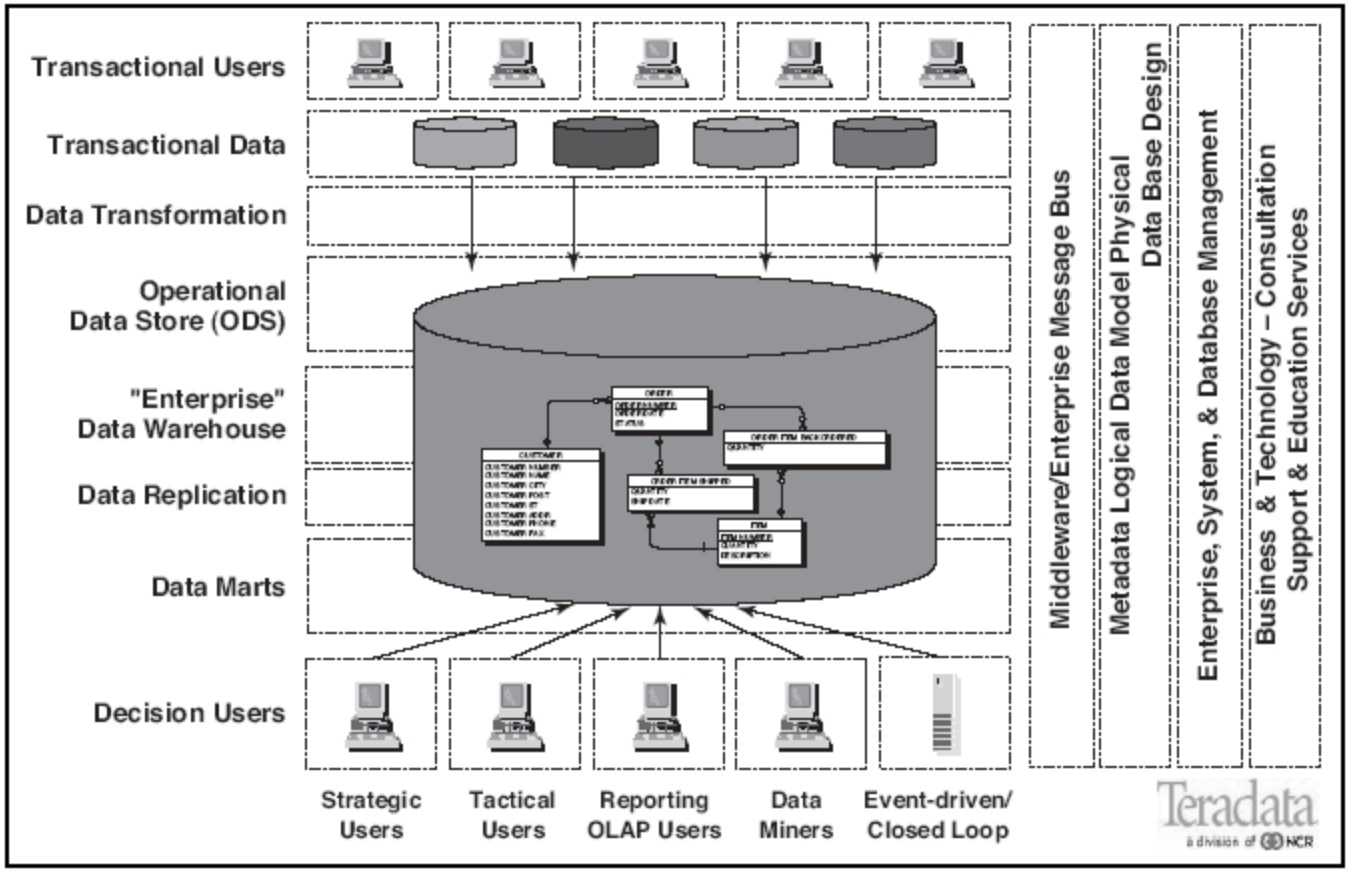


FIGURE 5.7 Teradata Corp.'s Enterprise Data Warehouse

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

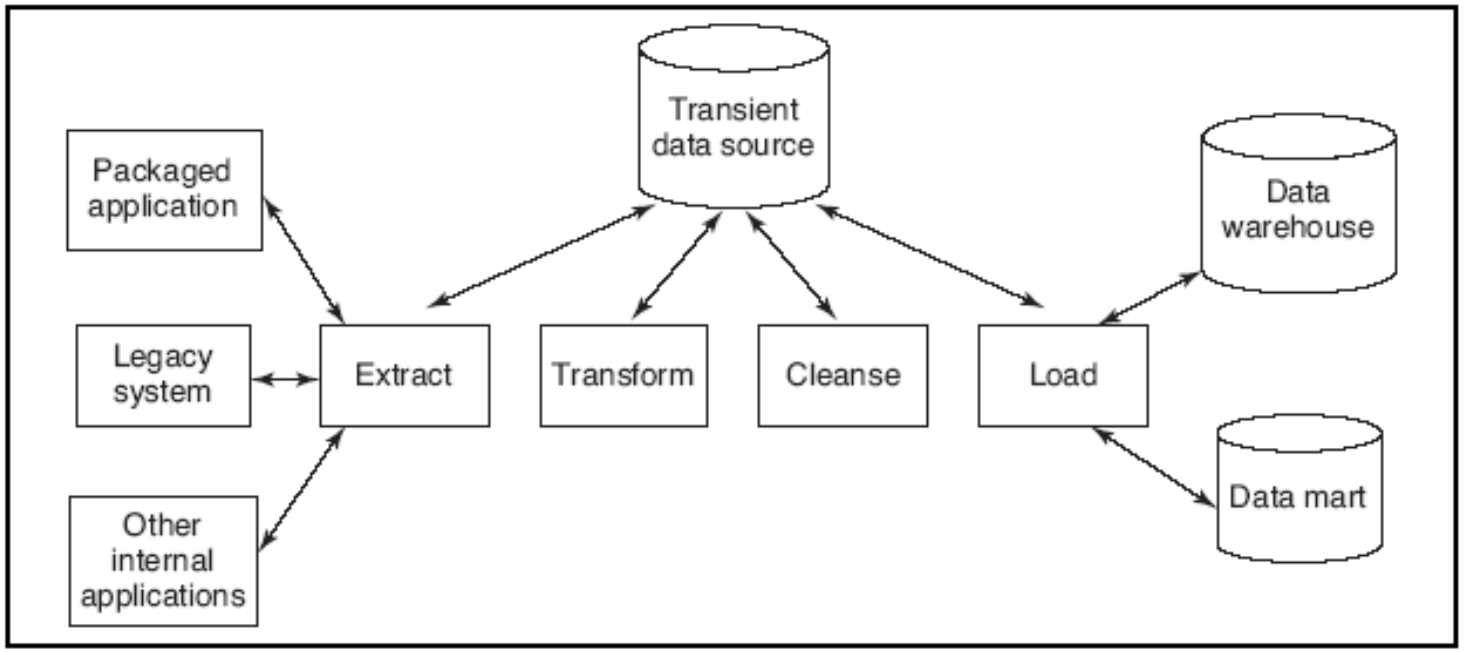
Data Integration and the Extraction, Transformation, and Load (ETL) Process

- **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 5.8 The ETL Process



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.
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.
8. The data warehouse contains data with several levels of detail: current detail data, old detail data, lightly summarized data, and highly summarized data.

Twelve Rules That Define a Data Warehouse

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 a time.
10. The data warehouse environment has a system that traces data sources, transformations, 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 chargeback mechanism for resource usage that enforces optimal use of the data by end users.

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 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

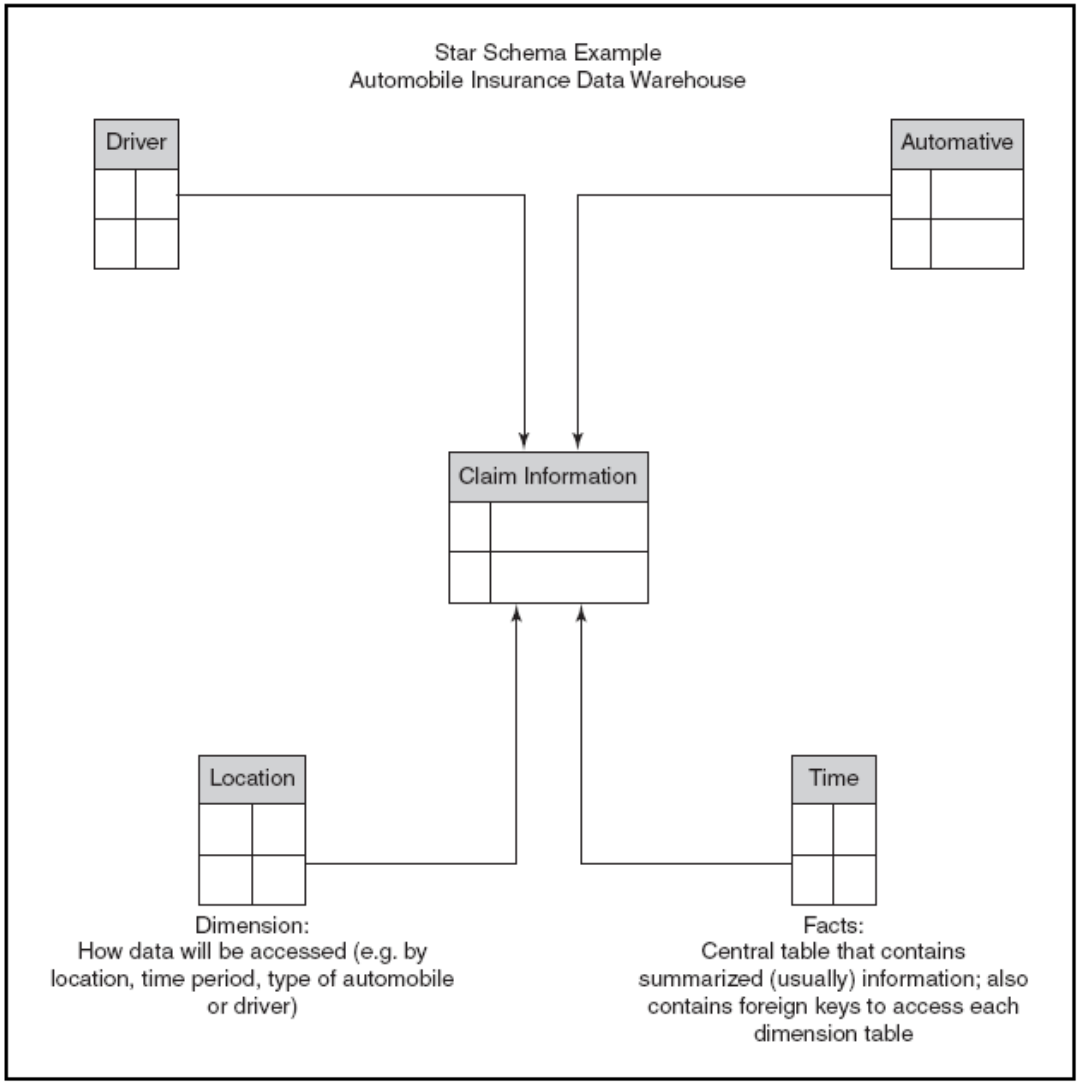
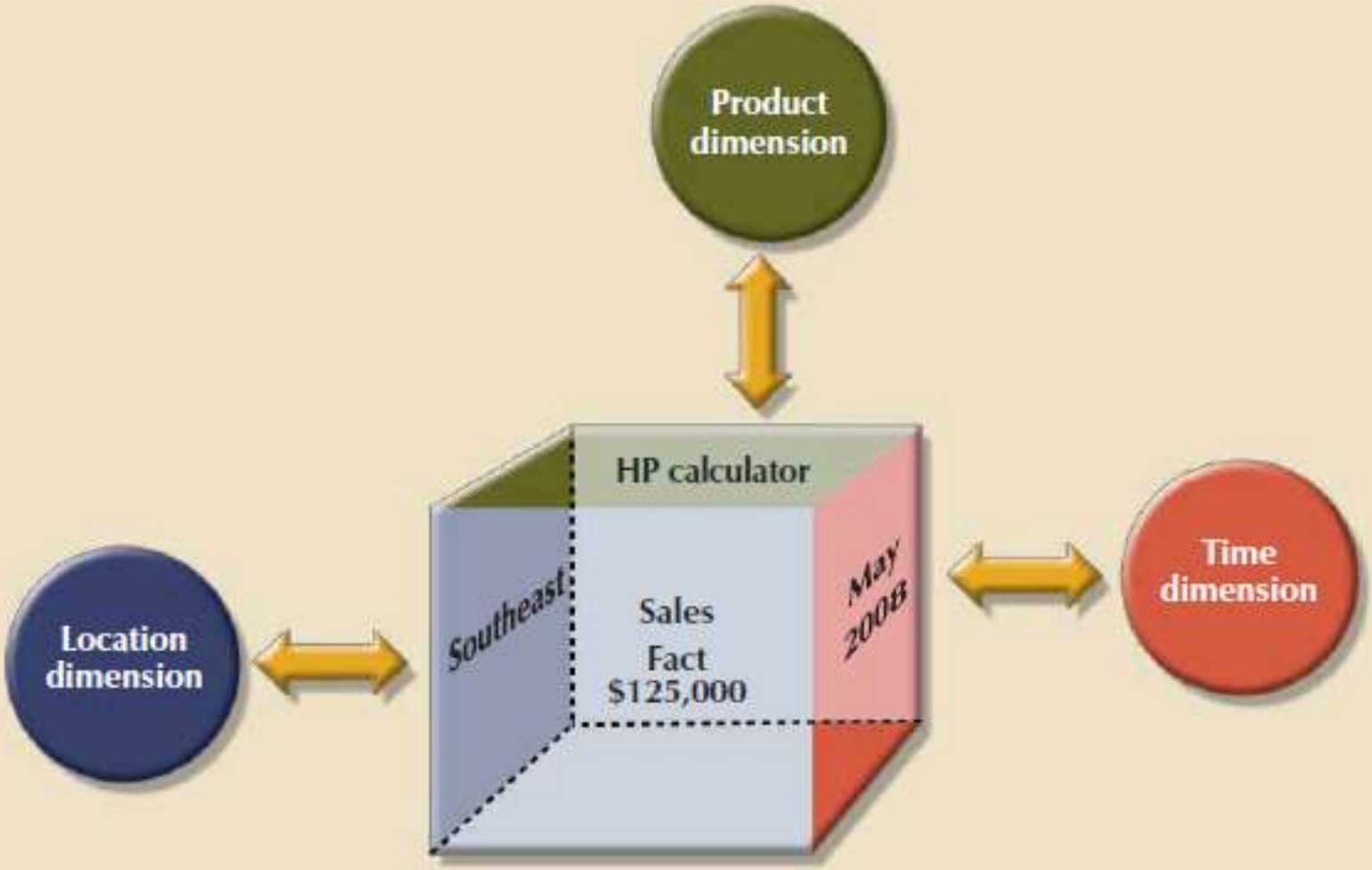
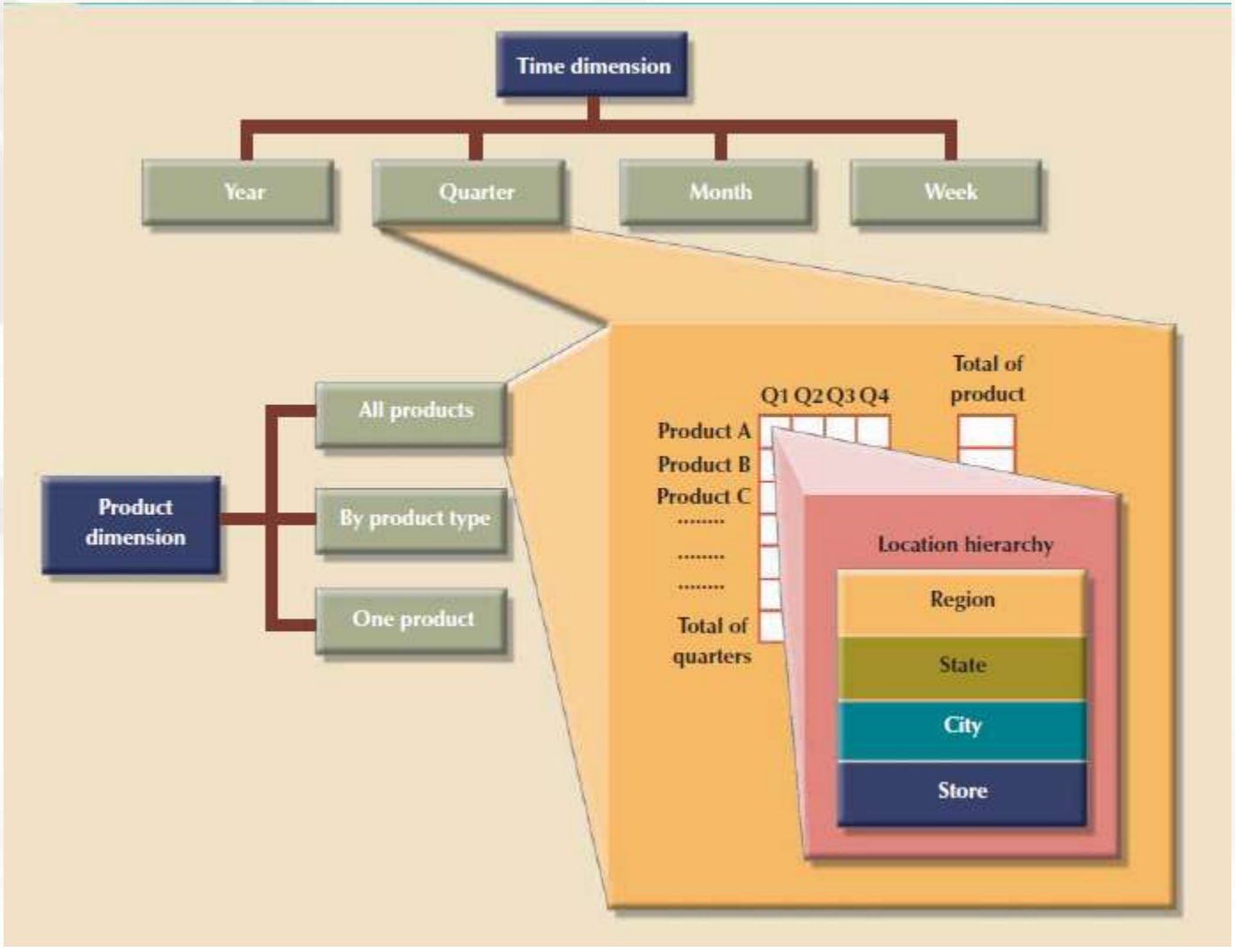


FIGURE 5.9 Star Schema

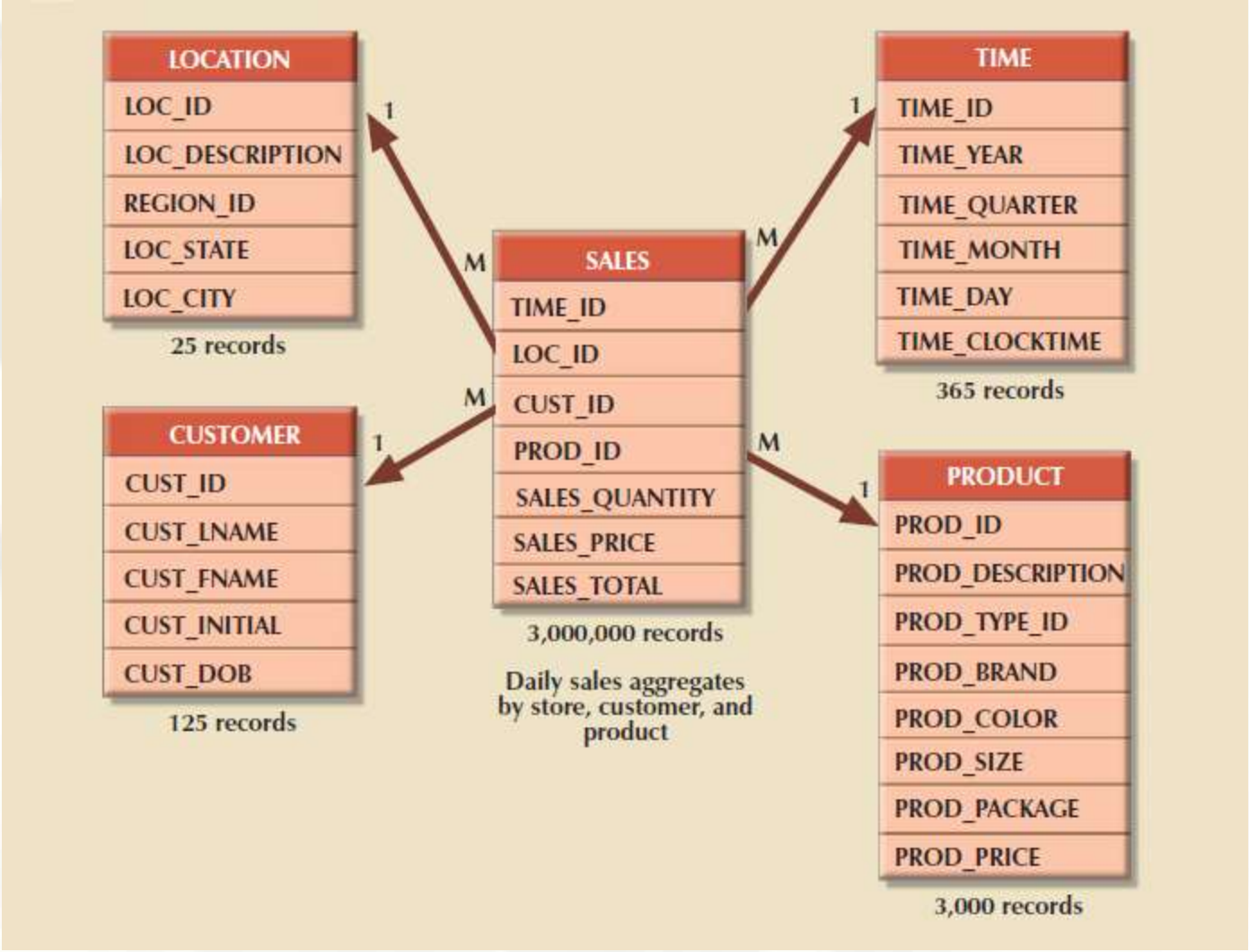
Simple star schema



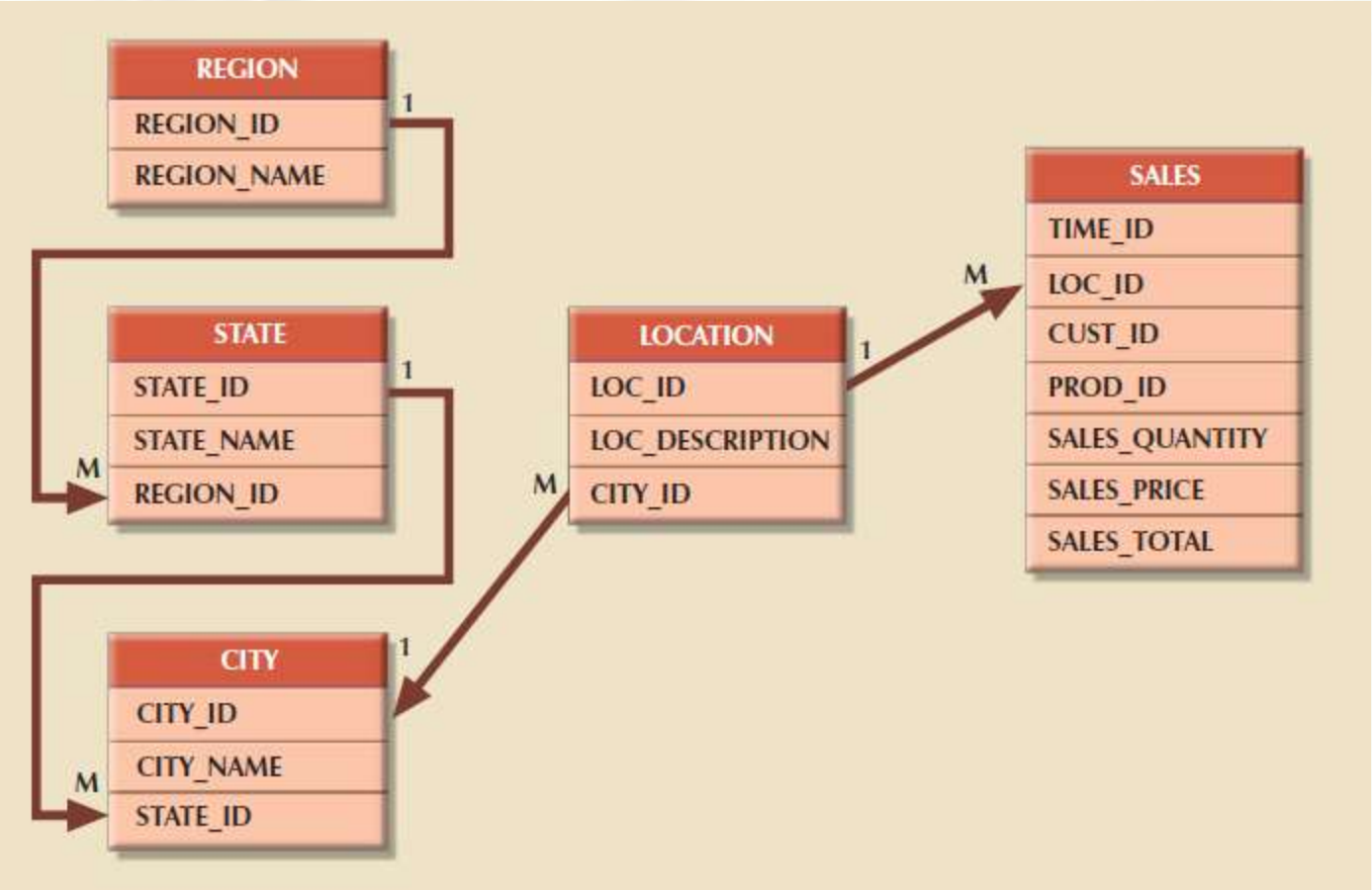
Attribute hierarchies in multidimensional analysis



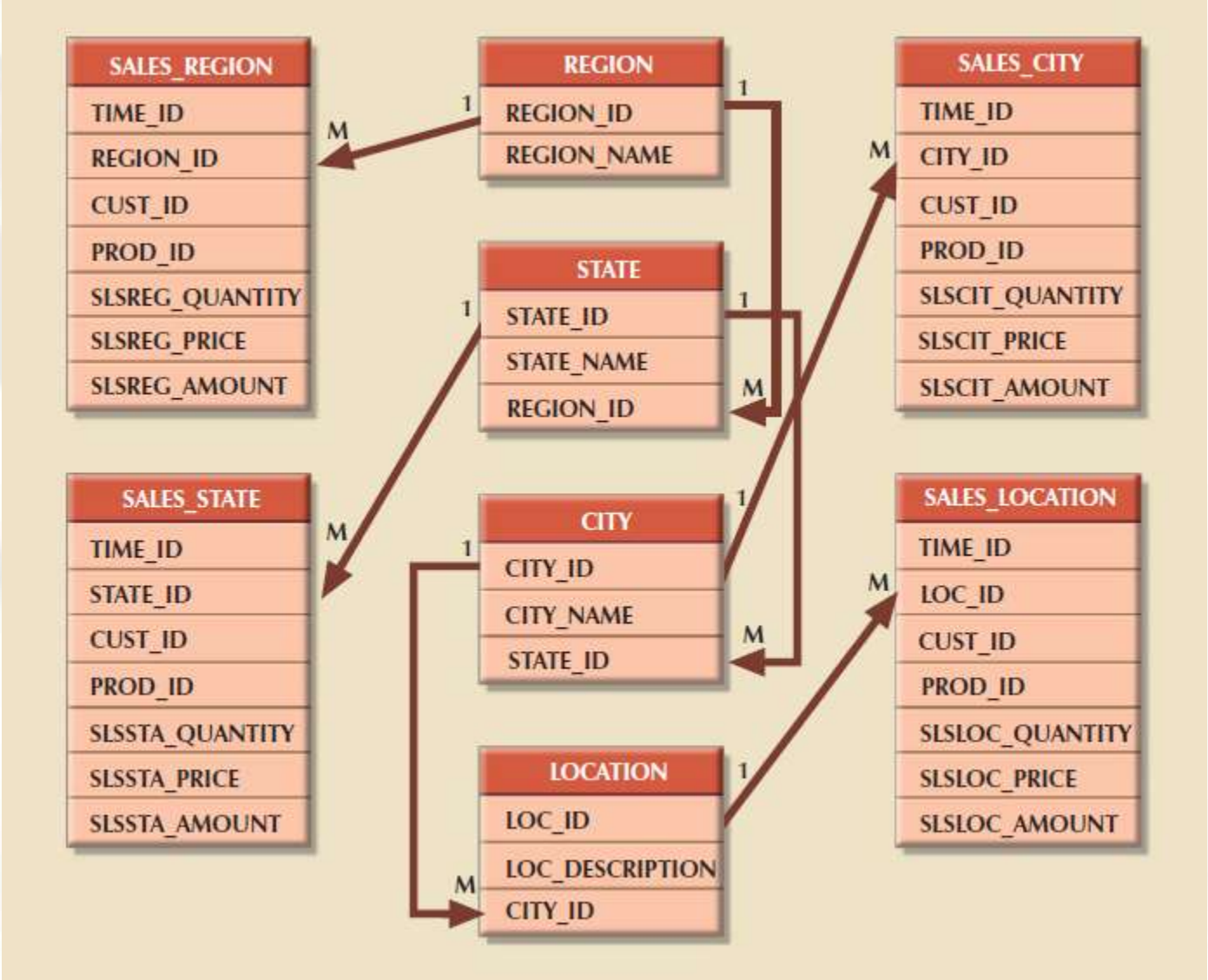
Star schema for SALES



Normalized dimension tables



Multiple fact tables



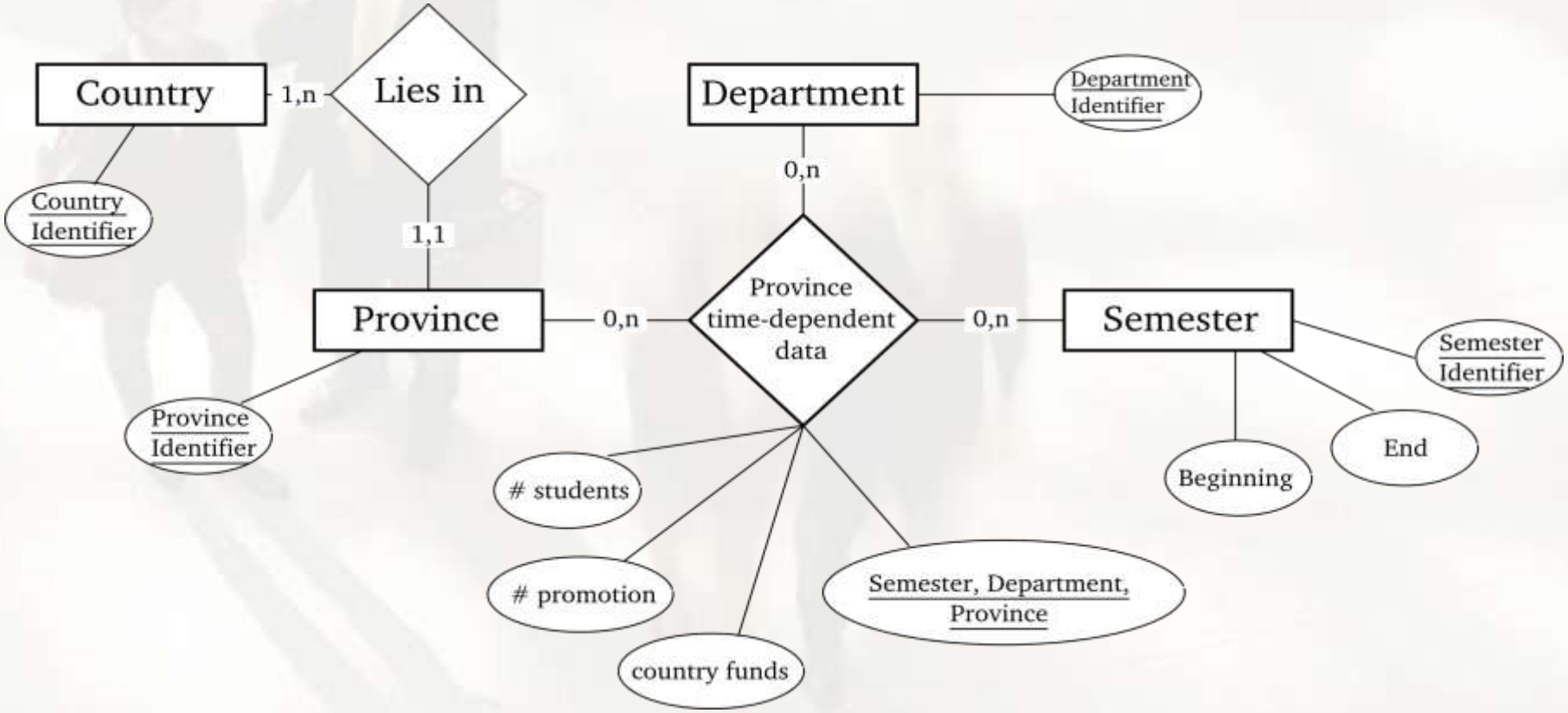
From E/R to ME/R
Multidimensional Entity/Relationship

Transactional structures in analytical map

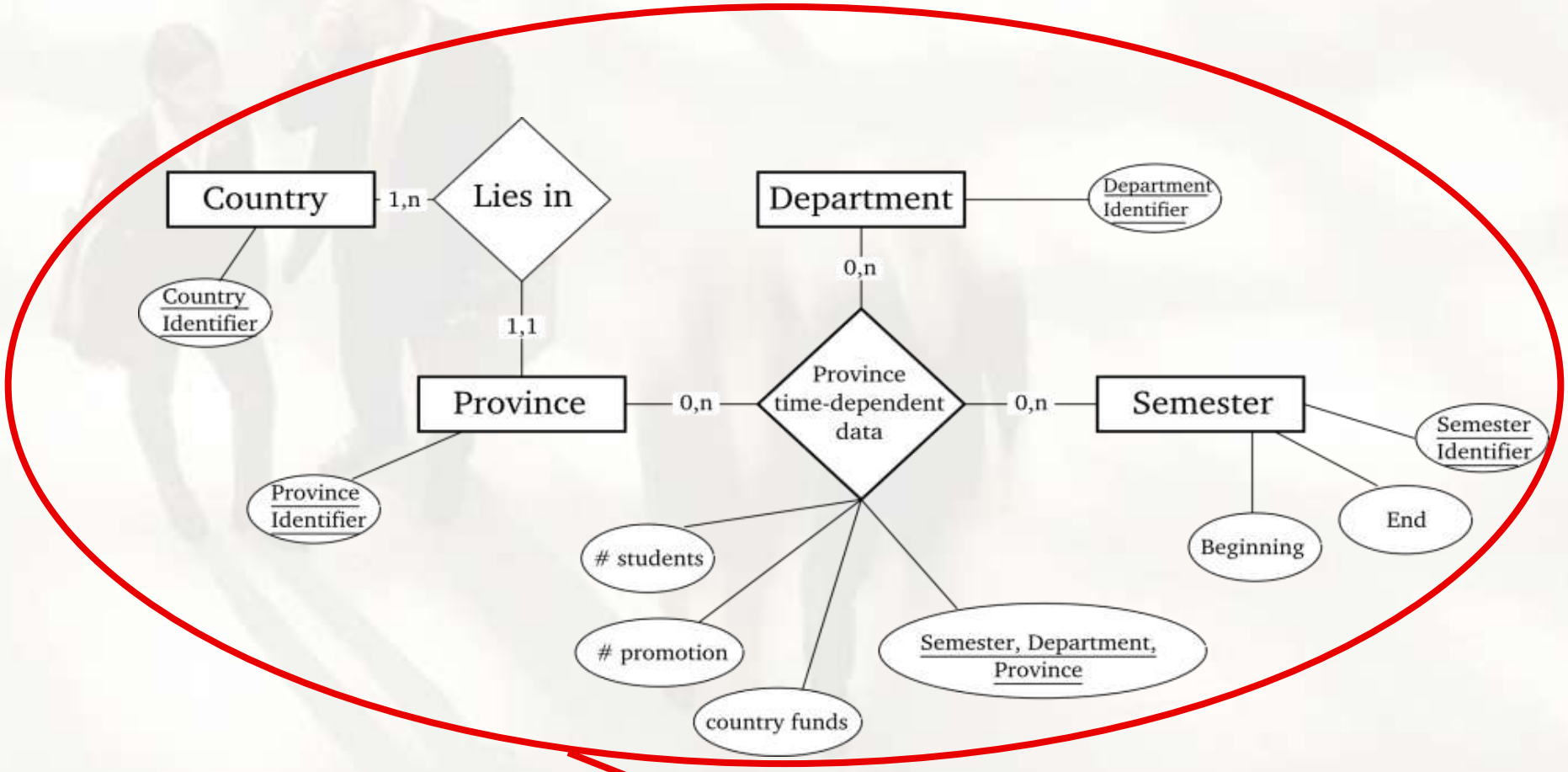
3 Steps Method

Step	Title	Description
1	Identifying business processes	Splitting an E/R into one or more business processes
2	Creating fact relationships	N-m relationships between strong entities revealed the facts relation, the numerical attributes are candidates for key figures
3	Forming dimensions	Content summarization of the remaining entities into groups that are dominated by powerful entities.

Example 1

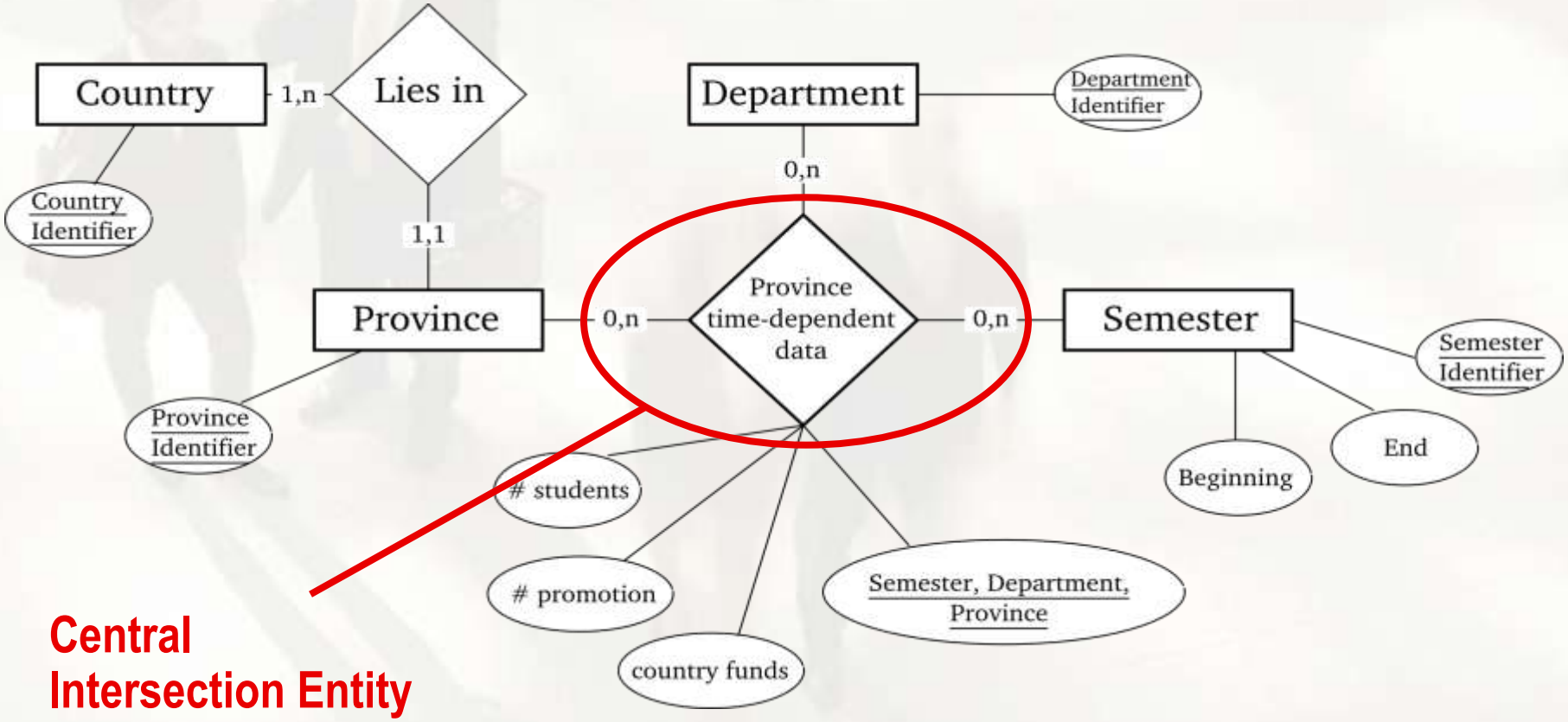

Country
1,n
Lies in
Country Identifier
1,1
Province
Province Identifier
0,n
Province time-dependent data
0,n
Department
Department Identifier
0,n
Semester
Semester Identifier
Beginning
End
students
promotion
country funds
Semester, Department, Province

1. Identifying Relevant Data



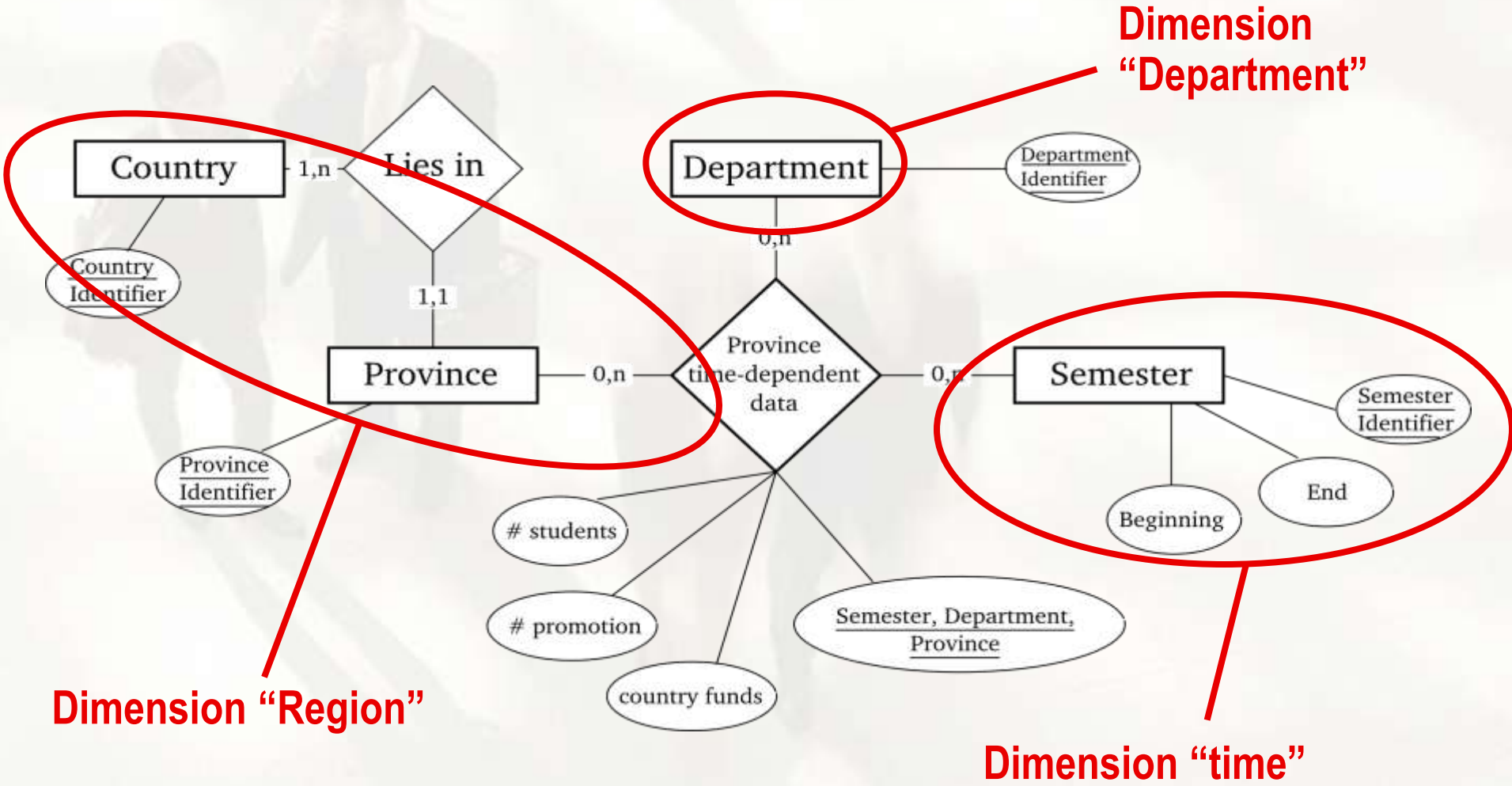
Data of interest

2. Finding Intersection Entities

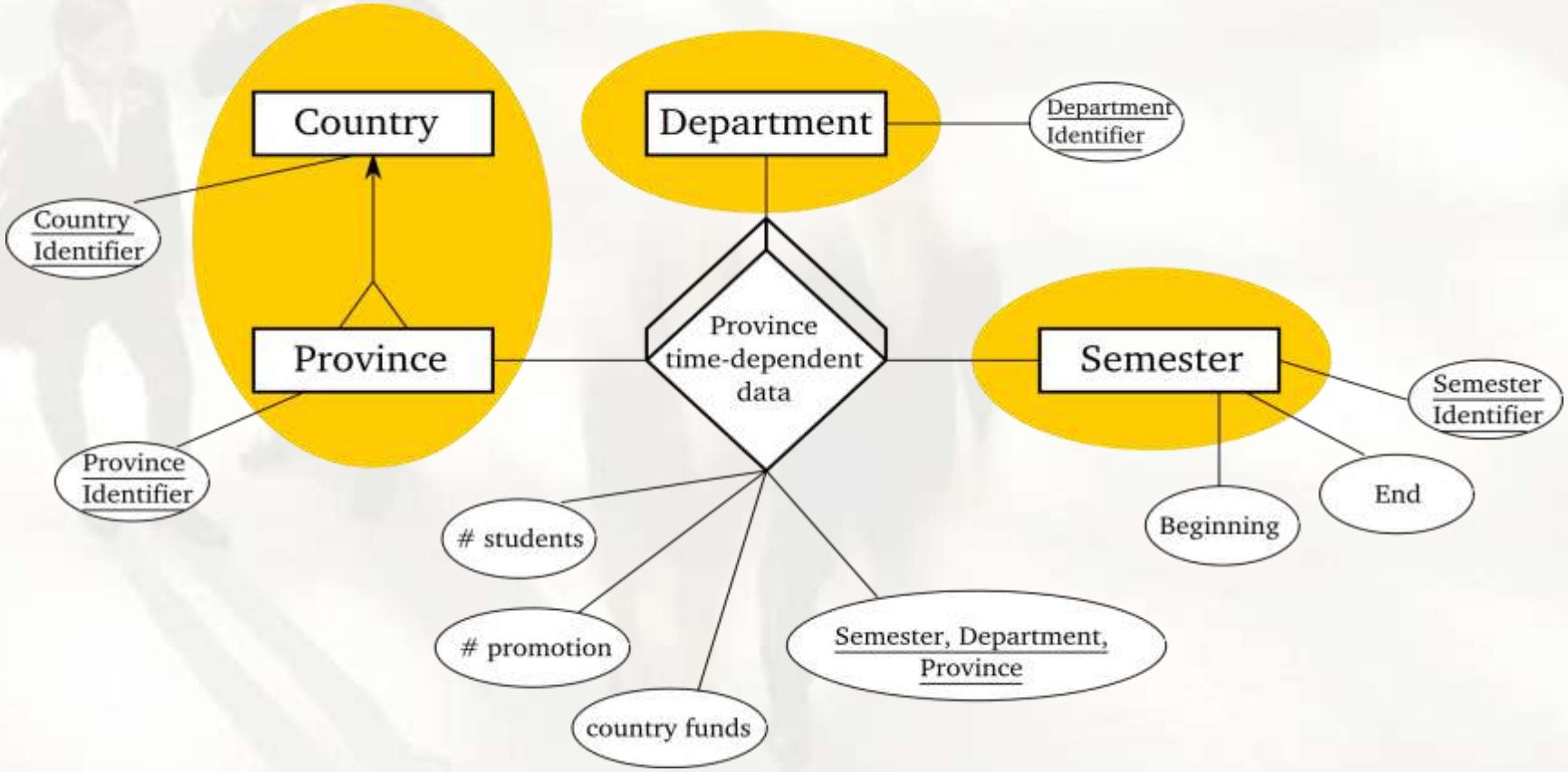


Central Intersection Entity

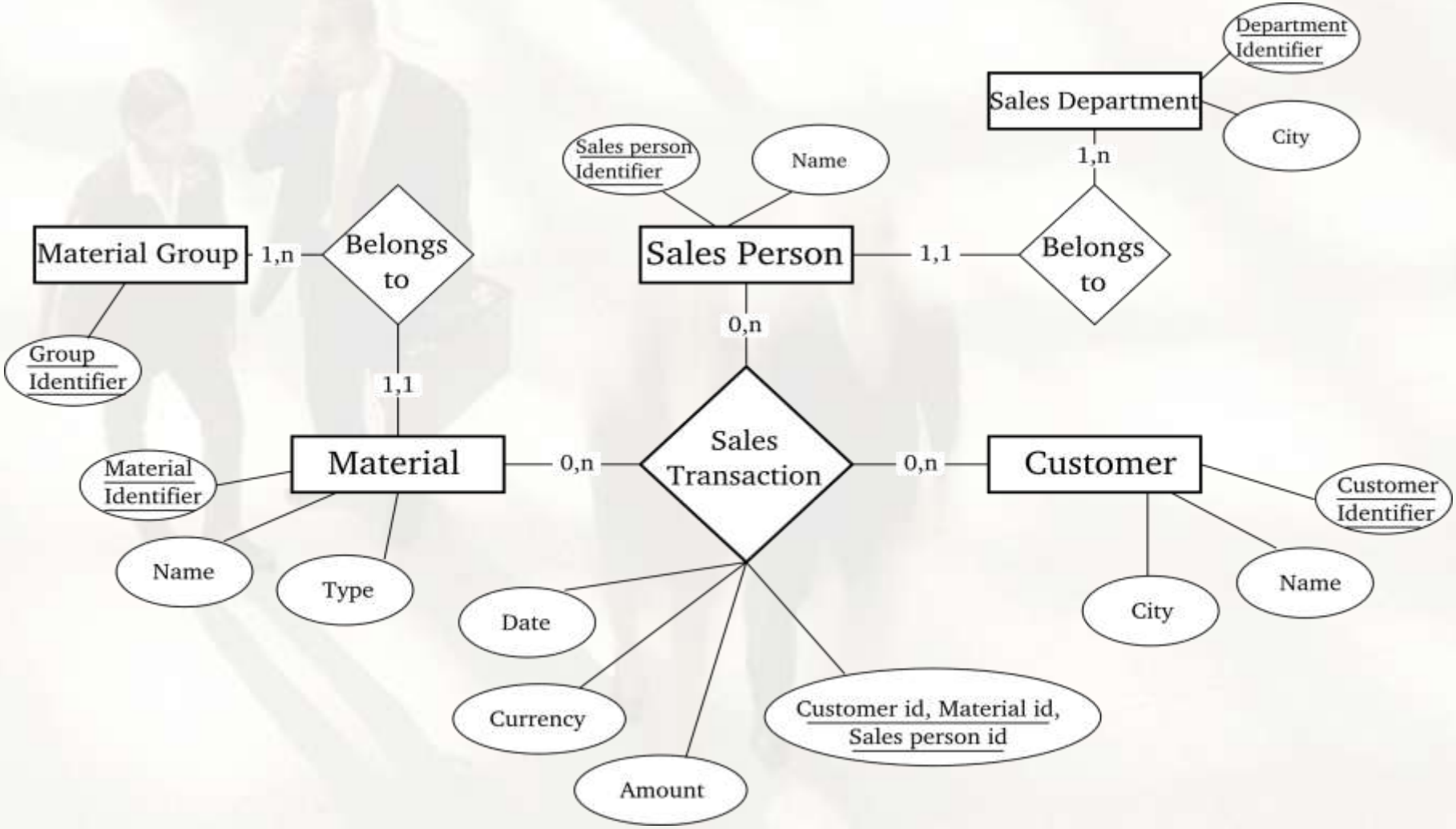
3. Forming Dimensions



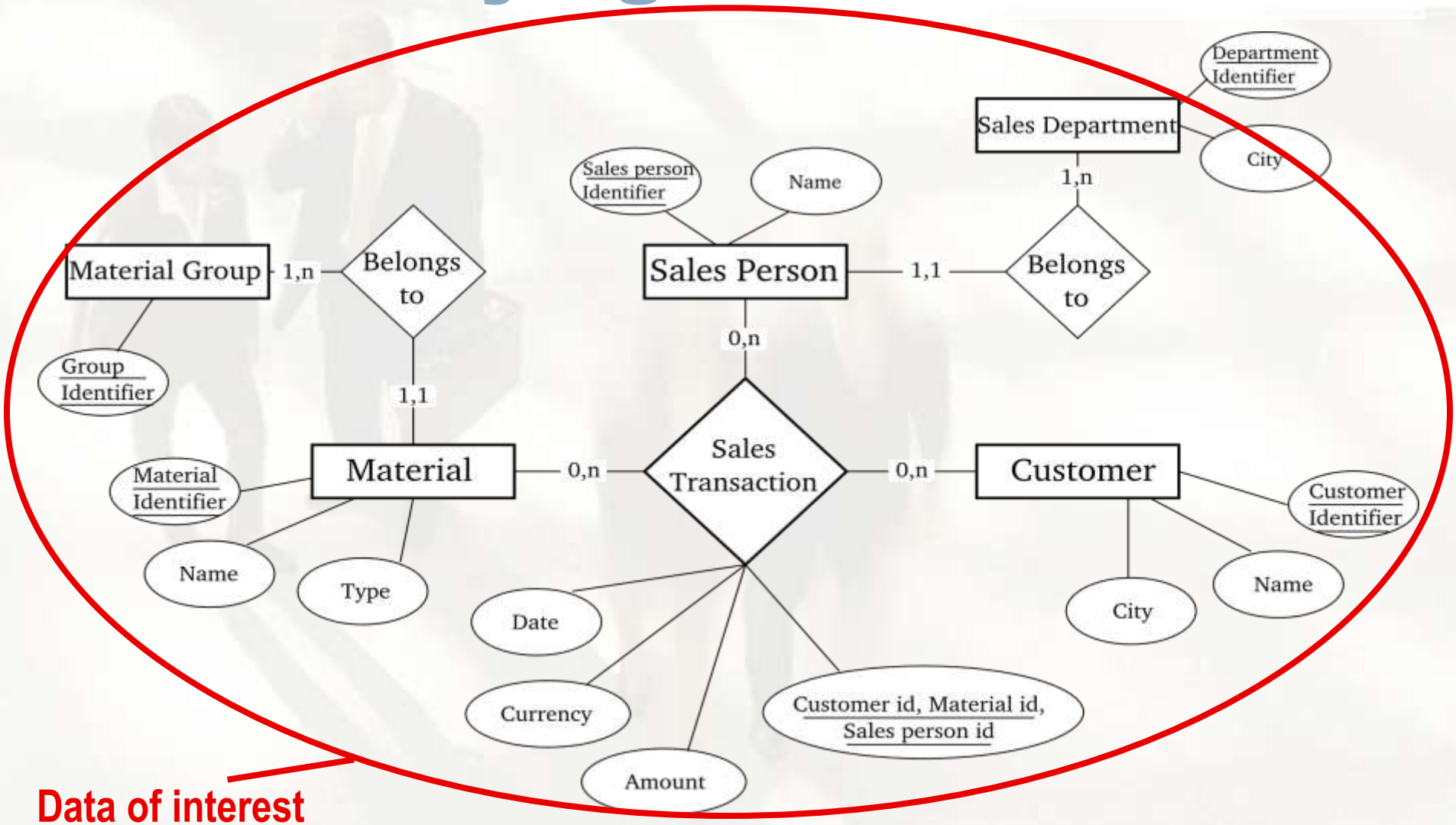
MERM


Country
Department
Department Identifier
Country Identifier
Province
Province time-dependent data
Semester
Semester Identifier
Province Identifier
End
Beginning
students
promotion
Semester, Department, Province
country funds

Example 2

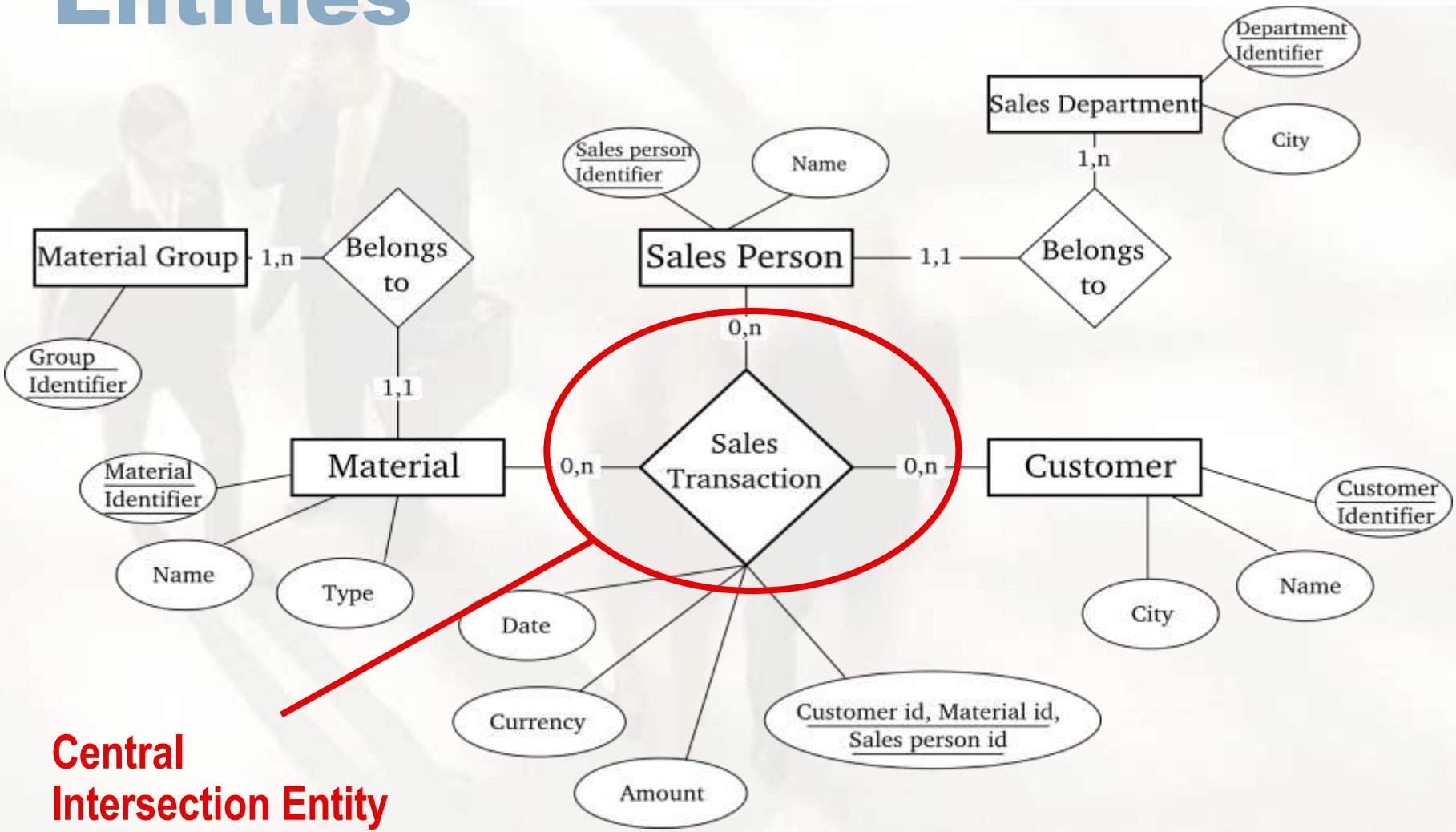

Department Identifier
Sales Department
City
1,n
Sales person Identifier
Name
Material Group
1,n
Belongs to
Sales Person
1,1
Belongs to
0,n
Group Identifier
1,1
Material Identifier
Material
0,n
Sales Transaction
0,n
Customer
Customer Identifier
Name
Type
Date
City
Name
Currency
Customer id, Material id, Sales person id
Amount

1. Identifying Relevant Data



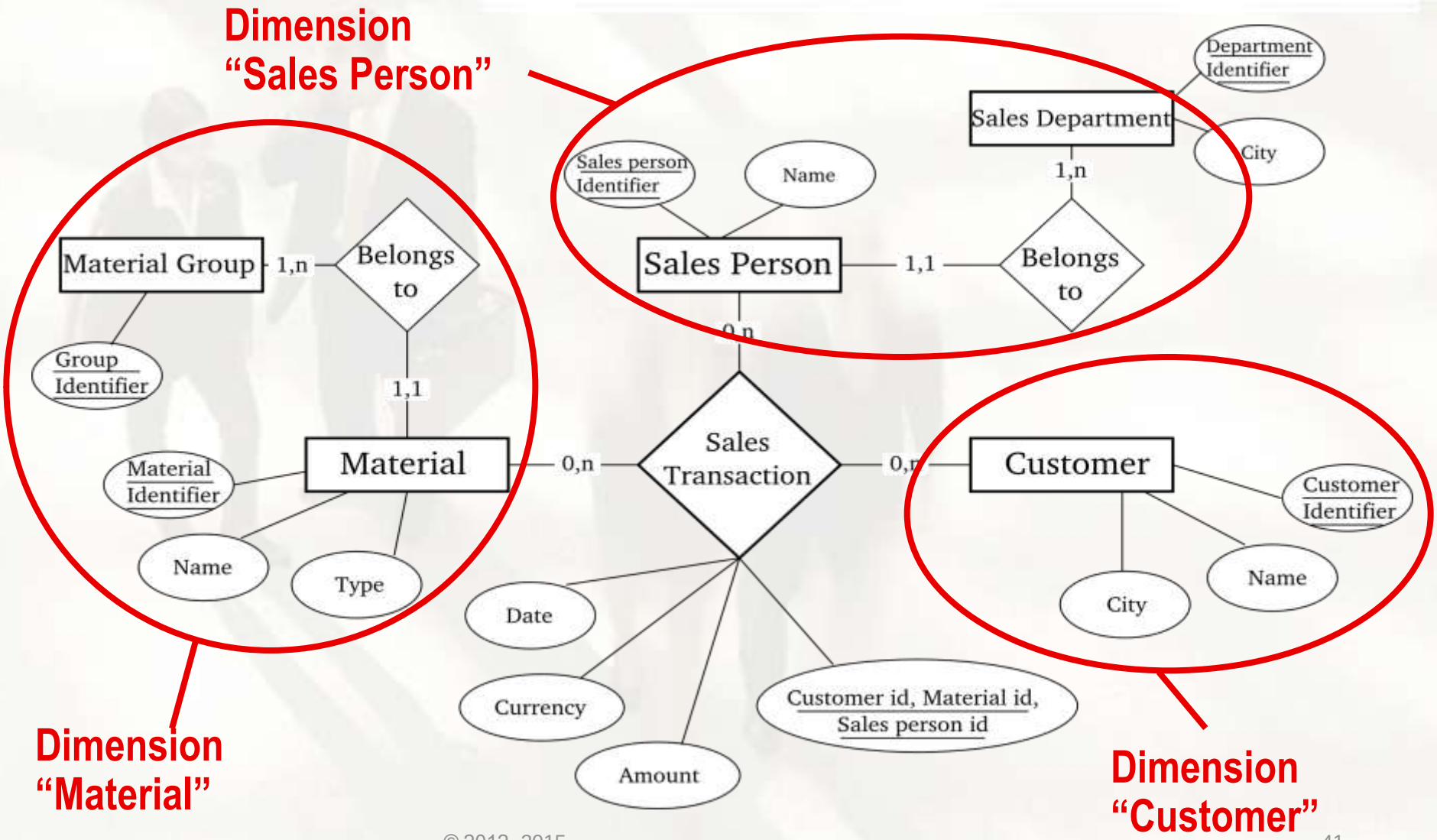
Data of interest

2. Finding Intersection Entities

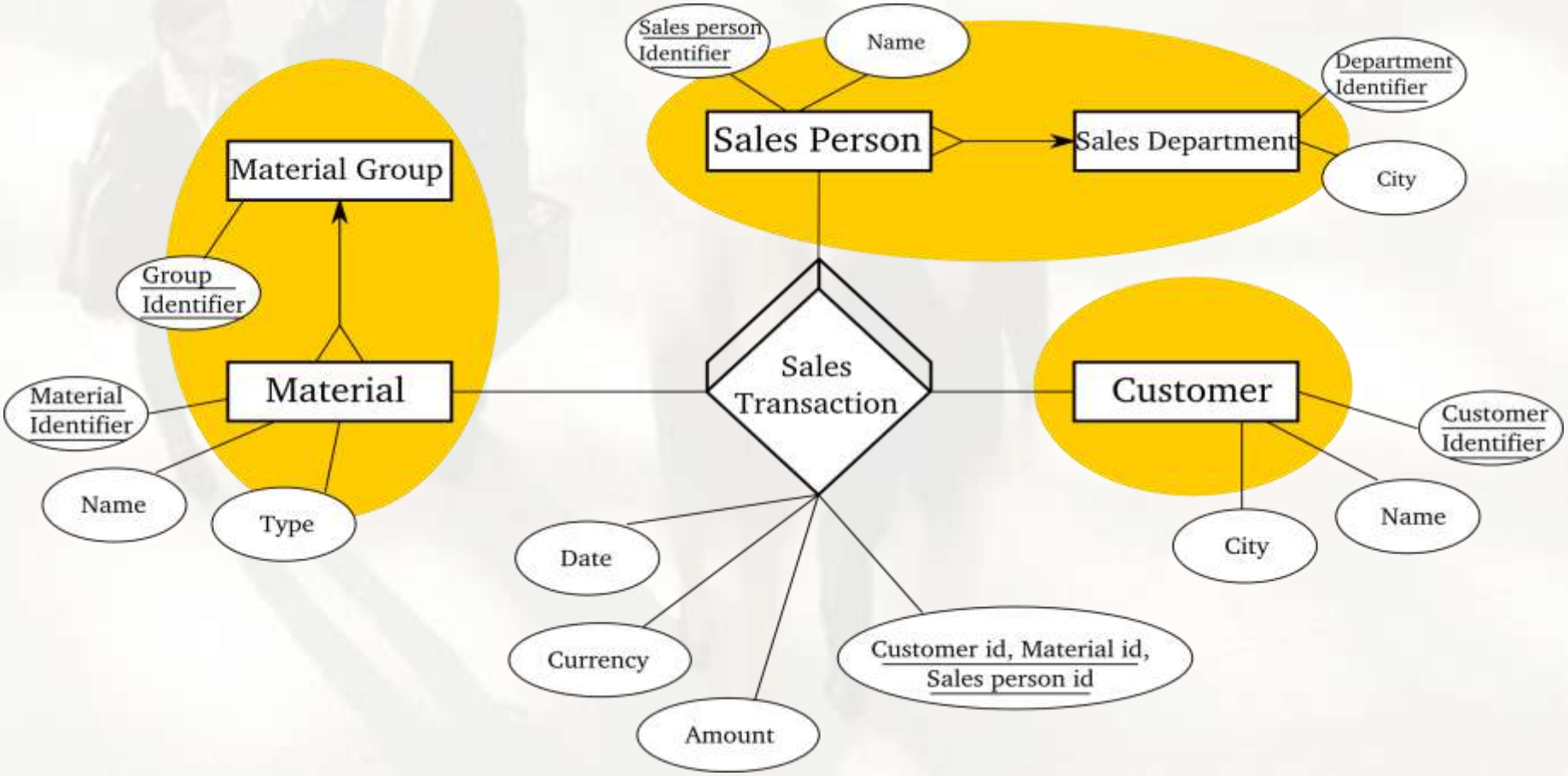


Central Intersection Entity

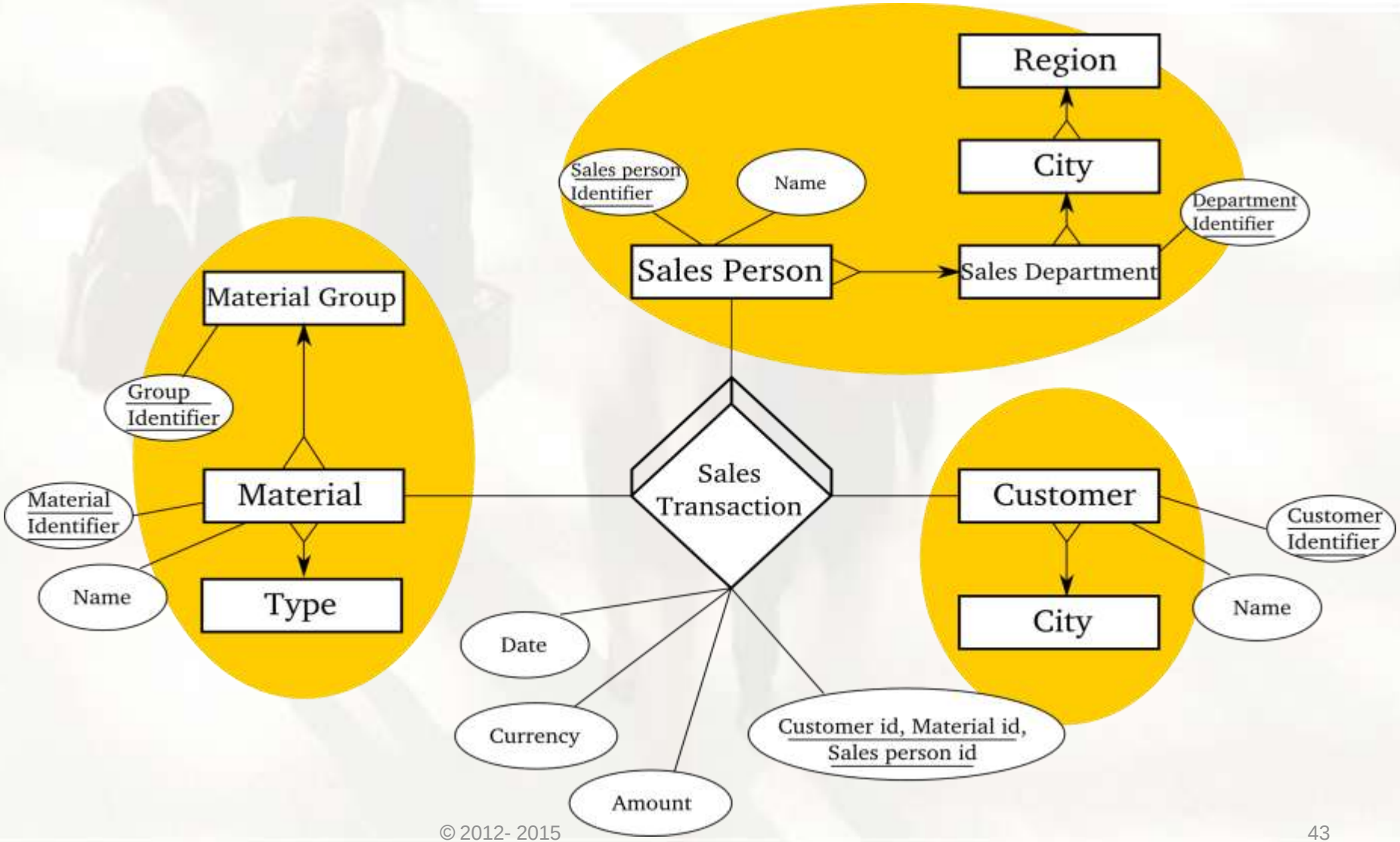
3. Forming Dimensions



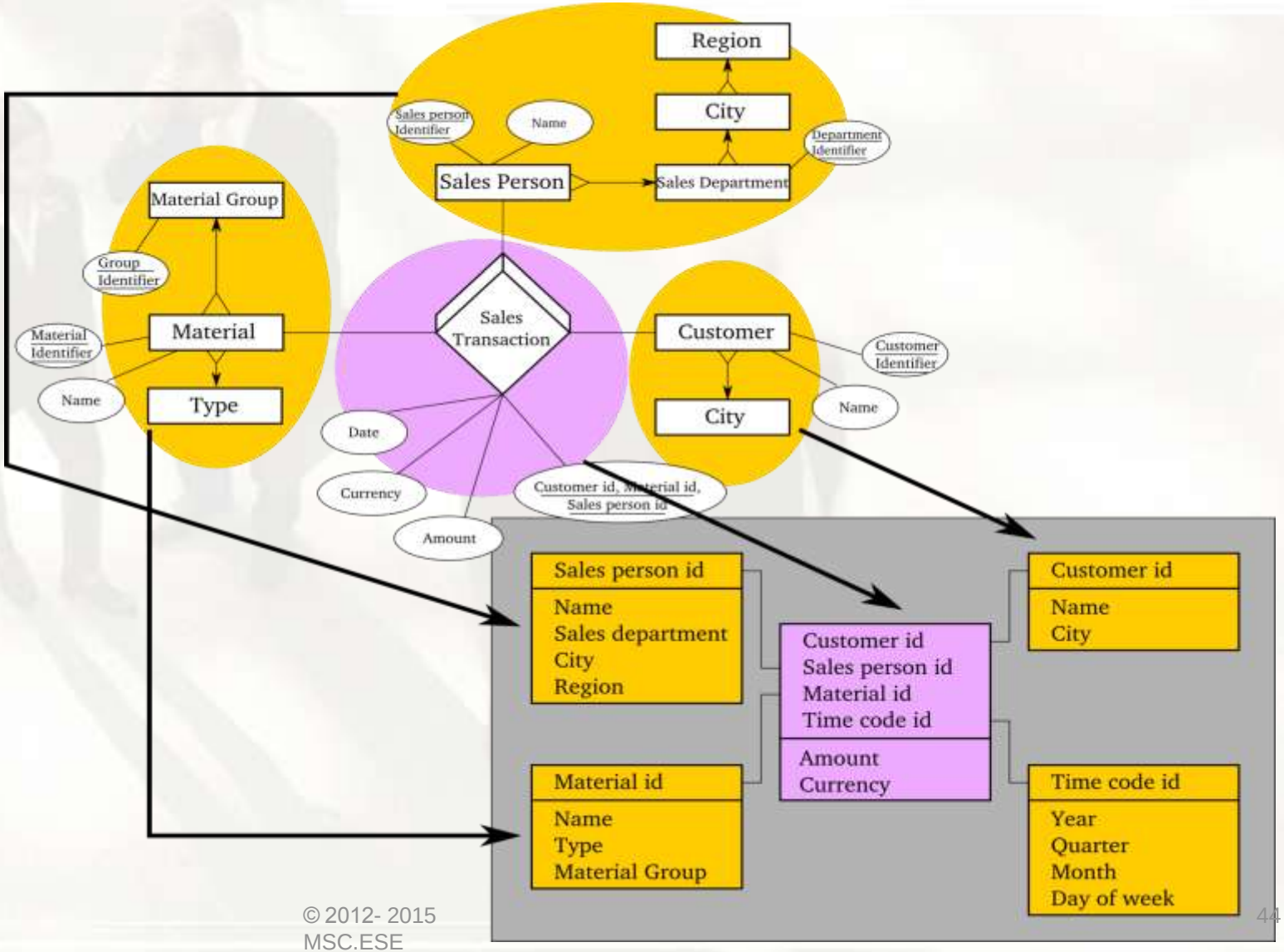
3. Forming Dimensions



4. Further refinement



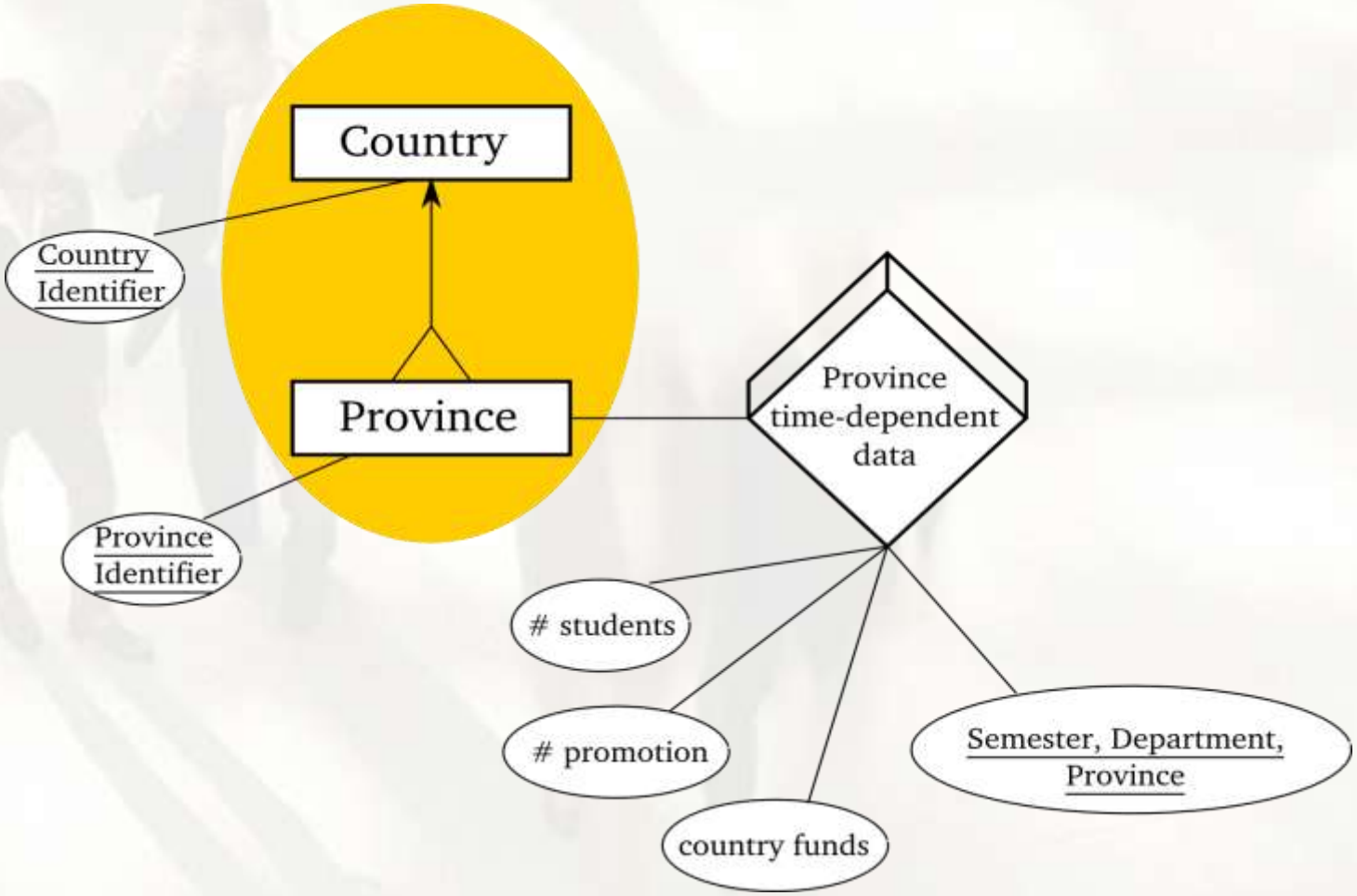
Into tables



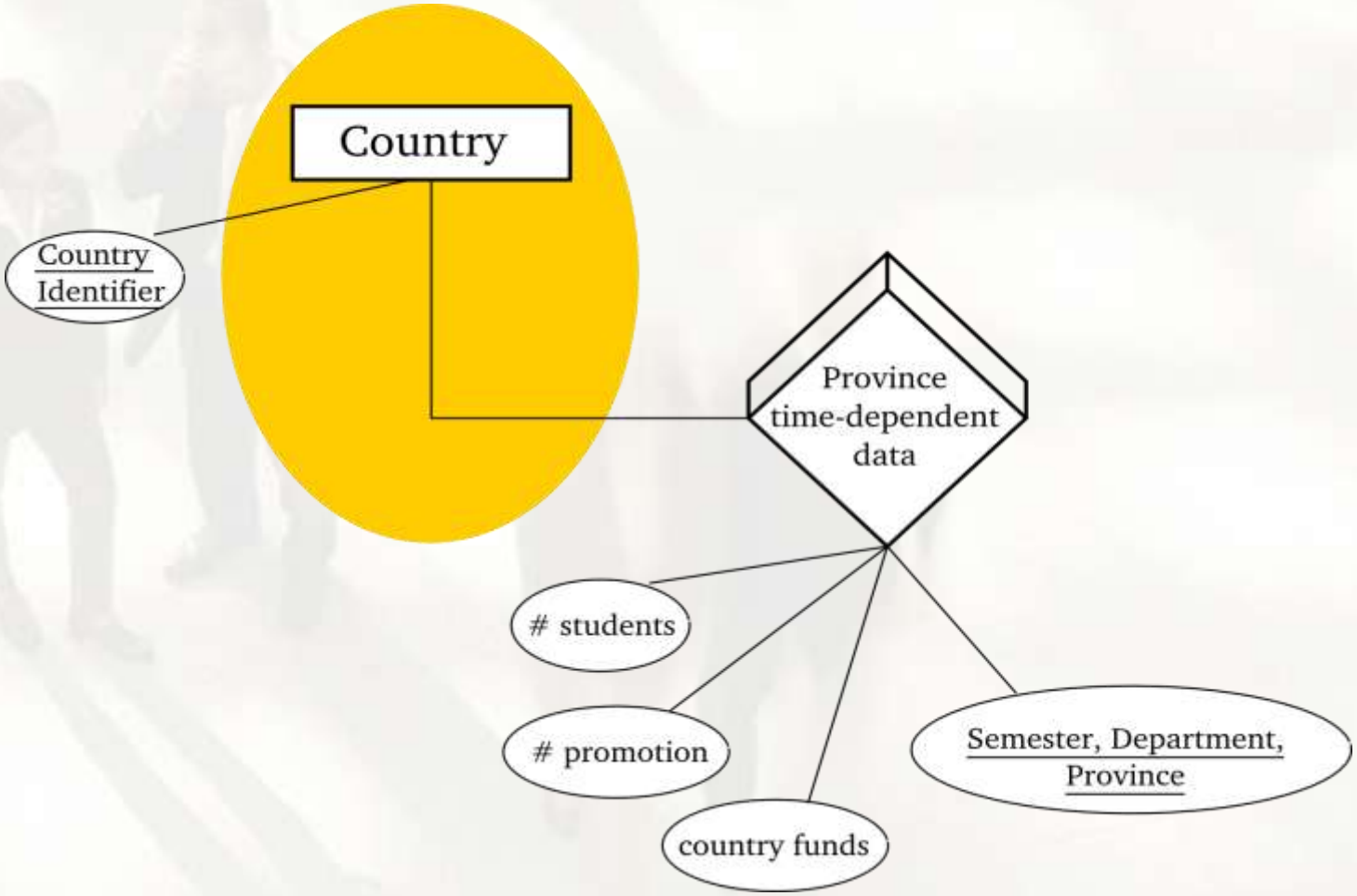
Granularity

- = "Detail" of a data structure
- High granularity: the data are described by many characteristics
- Low granularity: the data are described by fewer characteristics
- Positive impact on results of the query
- Negative effects on performance

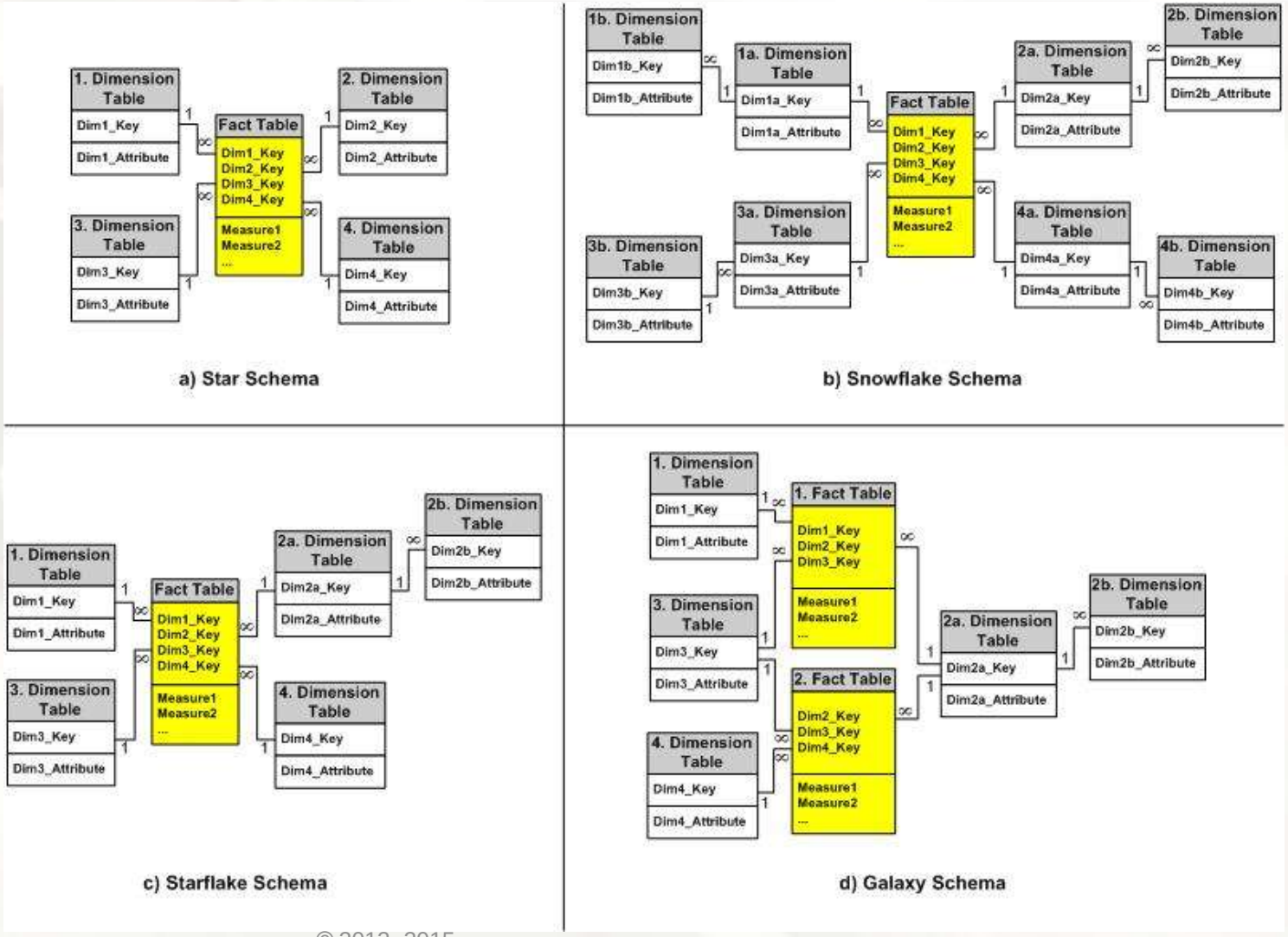
High Relative Granularity



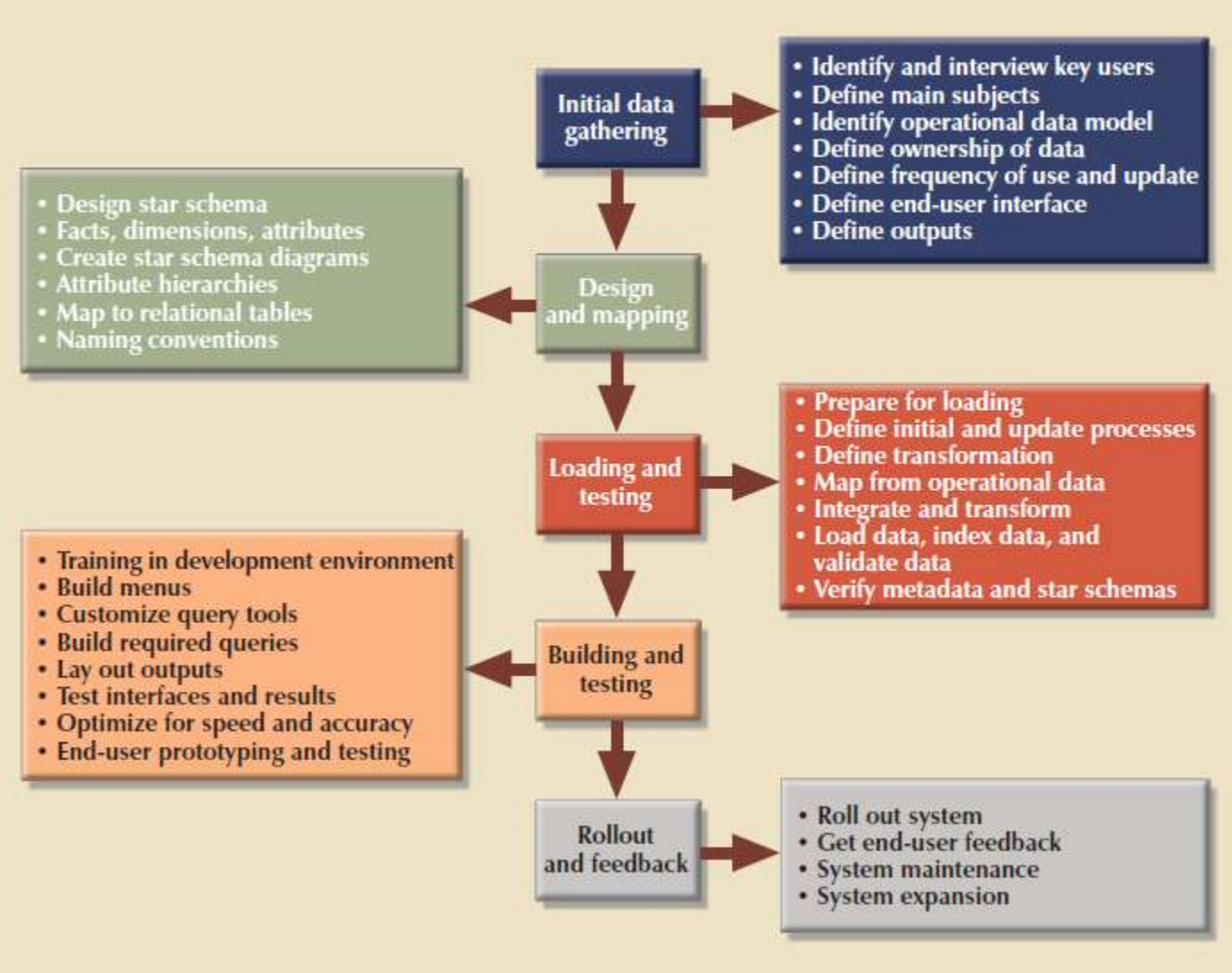
Low Relative Granularity



Common Schema Models for Data Warehousing



Data warehouse design and implementation road map



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

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

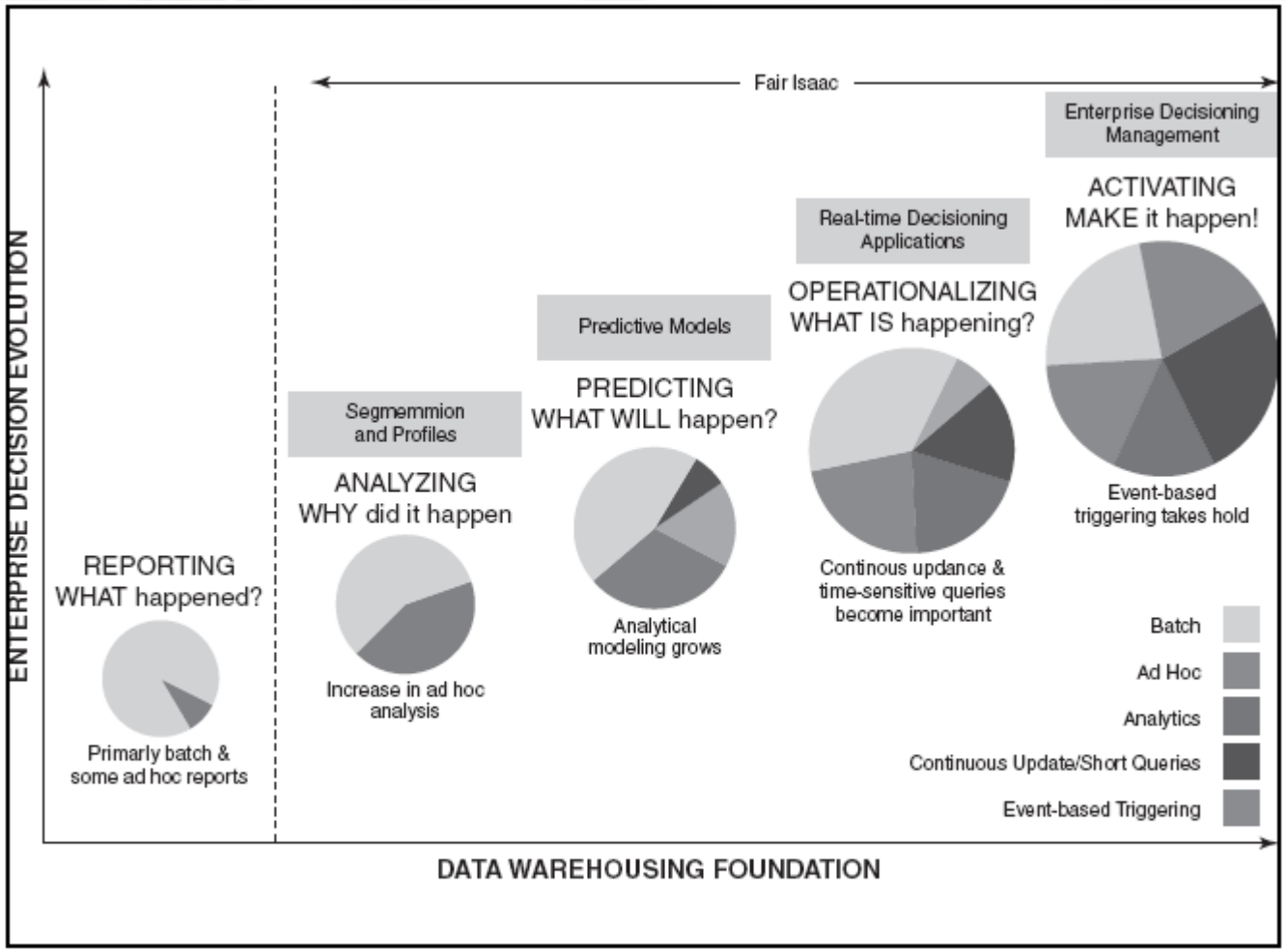
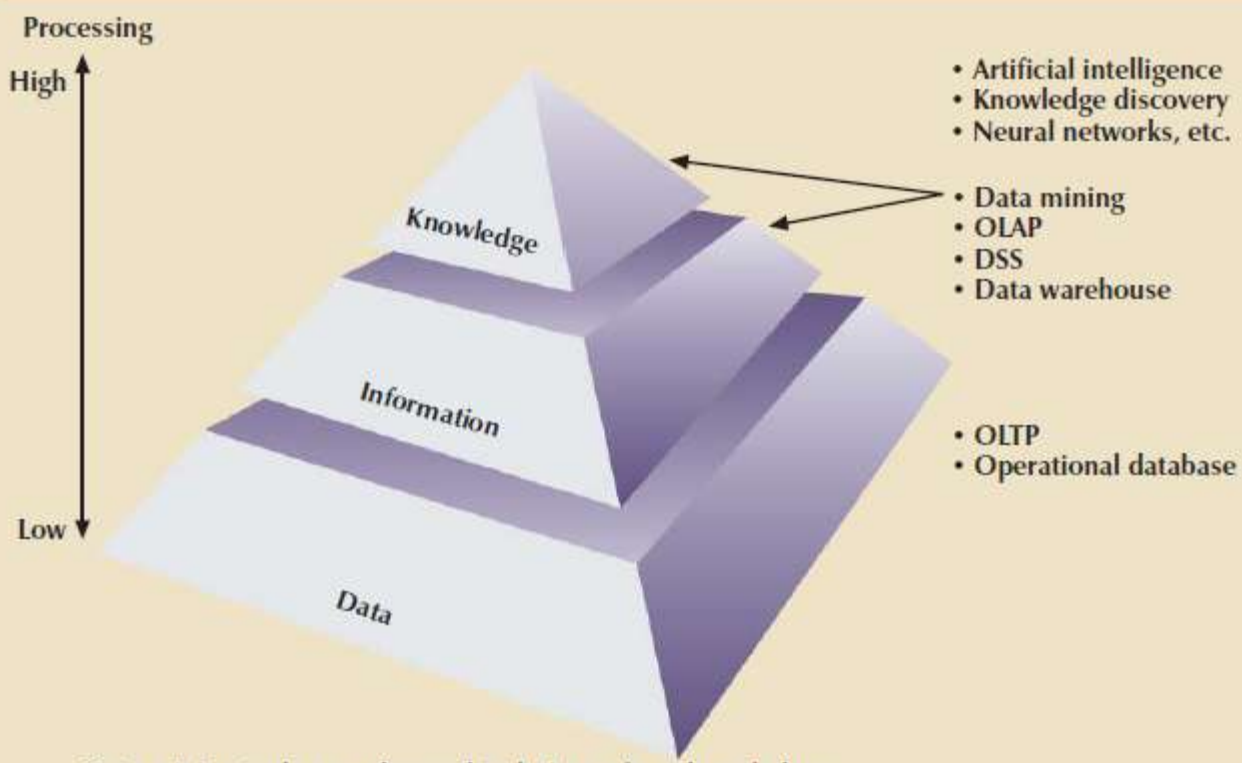


FIGURE 5.10 Enterprise Decision Evolution

Extracting knowledge from data



Data-mining tools use advanced techniques from knowledge discovery, artificial intelligence, and other fields to obtain "knowledge" and apply it to business needs. Knowledge is then used to make predictions of events or forecasts of values such as sales returns. Several OLAP tools have integrated at least some of these data-mining features in their products.