Previews of TDWI course books offer an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews cannot be printed.

TDWI strives to provide course books that are content-rich and that serve as useful reference documents after a class has ended.

This preview shows selected pages that are representative of the entire course book; pages are not consecutive. The page numbers shown at the bottom of each page indicate their actual position in the course book. All table-of-contents pages are included to illustrate all of the topics covered by the course.

This page intentionally left blank.

THE DATA WAREHOUSING INSTITUTE

TDWI Dimensional Data Modeling Primer

from Requirements to Business Analysis

The Data Warehousing Institute takes pride in the educational soundness and technical accuracy of all of our courses. Please give us your comments—we'd like to hear from you. Address your feedback to:

email: info@tdwi.org

Publication Date: February 2010

TABLE OF CONTENTS

THE DATA WAREHOUSING INSTITUTE

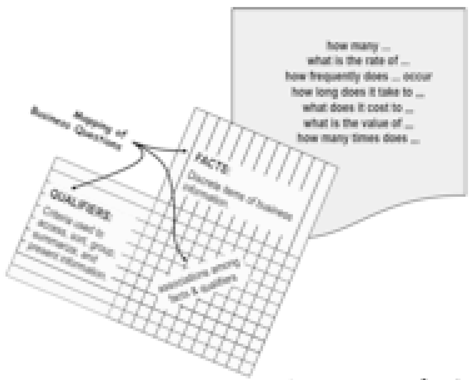
Module 1

Dimensional Modeling Concepts

Dimensional Modeling Basics

Dimensional Data Models

Conceptual Data Model



- an analysis process
- model of business requirements
- starting from vague and uncertain
- evolving to specific and certain
- business questions list & fact/qualifier matrix

Logical Data Model

- a business (non-technical) design process
- model of business solution
- starting from specific business requirements
- evolving to product specification
- logical dimensional data model

Physical Data Model

- an implementation (technical) design process
- model of technology solution
- starting from product specification
- evolving to database specification
- star schema

Dimensional Modeling Basics

Dimensional Data Models

LEVELS OF MODELING

Dimensional data design, like any other design process, involves a transition from abstract to highly specific. Abstract models are a means to understand requirements, while implementation models must specify the solution precisely. A three-level approach to data modeling works well where:

- **Conceptual modeling** describes the business needs.
- **Logical modeling** describes the business solution.
- **Physical modeling** specifies the technology solution.

CONCEPTUAL MODELS

Conceptual models are produced through analysis of business needs, and are intended to structure business requirements such that they can be verified and used as input to logical data modeling. When designing dimensional data, conceptual models include a representative list of business questions and analysis of those questions to understand them as a collection of data facts and data qualifiers.

LOGICAL MODELS

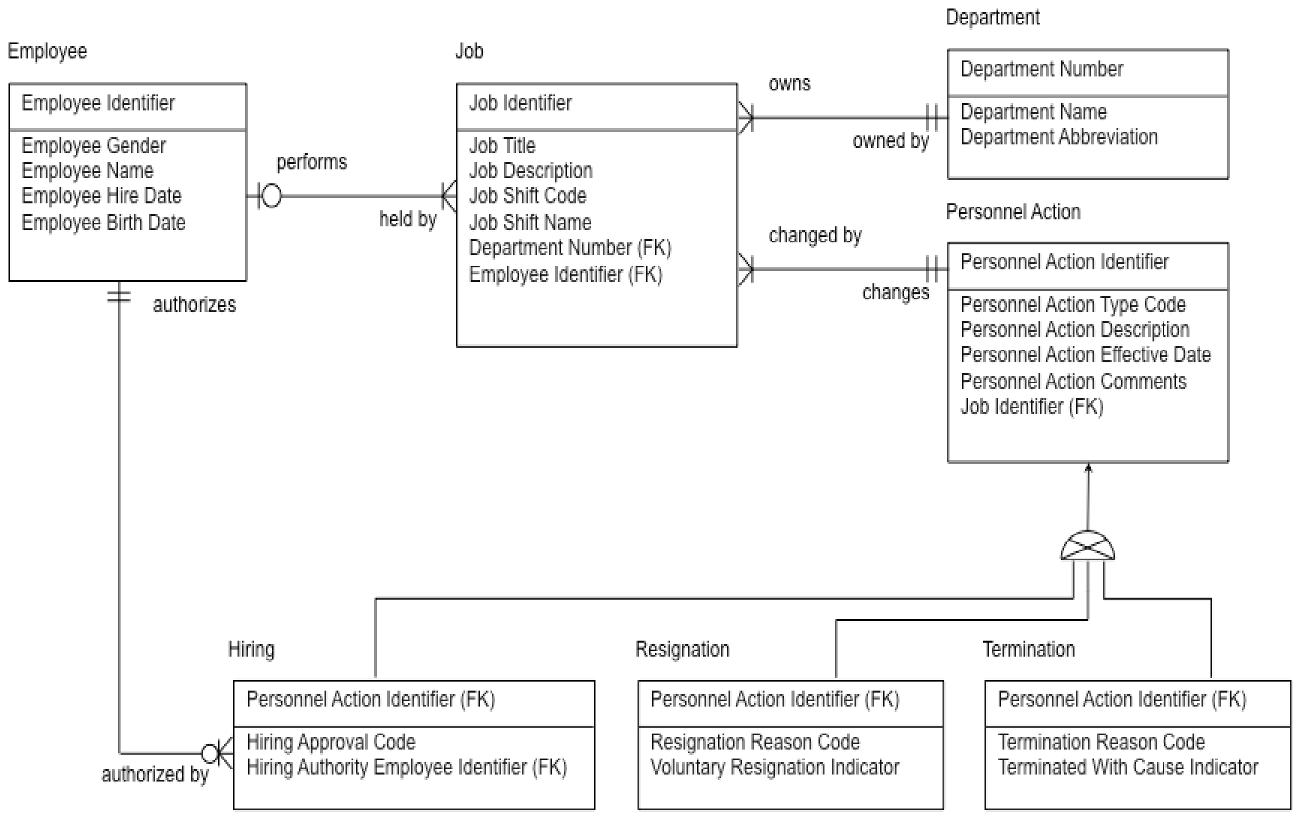
Logical modeling is a point of transition. The modeling activities change from analysis (understanding the needs) to design (understanding the solutions needed to satisfy those needs). A logical model is analogous with a product specification, describing *what* is to be produced but not detailing *how* it is to be assembled. A logical dimensional model meets this objective for design of dimensional data.

PHYSICAL MODELS

Physical models are produced by technical design processes. They describe and specify technology solutions. The physical design process transforms a logical model into a specification for implementation. Physical modeling adds the details necessary to describe *how* a product is structured and assembled. When designing dimensional data, the most common and widely accepted physical model is a star-schema.

Comparing E-R and Dimensional Models

A Quick Review of Entity-Relationship Modeling



Comparing E-R and Dimensional Models

A Quick Review of Entity-Relationship Modeling

ENTITY-RELATIONSHIP MODELS

This diagram illustrates a simple entity-relationship (ER) model. The primary components of an ER model are:

- **Entity**—An entity is a subject about which the business has the need, will, and means to collect data—a person, place, thing, concept, or event that is of business interest. Entities are represented as labeled boxes in the ER model. Examples in the diagram include EMPLOYEE, JOB, and DEPARTMENT.

- **Attribute**—An attribute is a property or characteristic of an entity that can be collected as data. Attributes are listed inside the box of the entities that they describe. *Job_shift_code* and *job_title*, for example, are attributes of the entity JOB.

- **Relationship**—A relationship is an important association between pairs of entities that is of business interest and may be collected as data. Examples of relationships in this diagram include EMPLOYEE PERFORMS JOB and DEPARTMENT OWNS JOB.

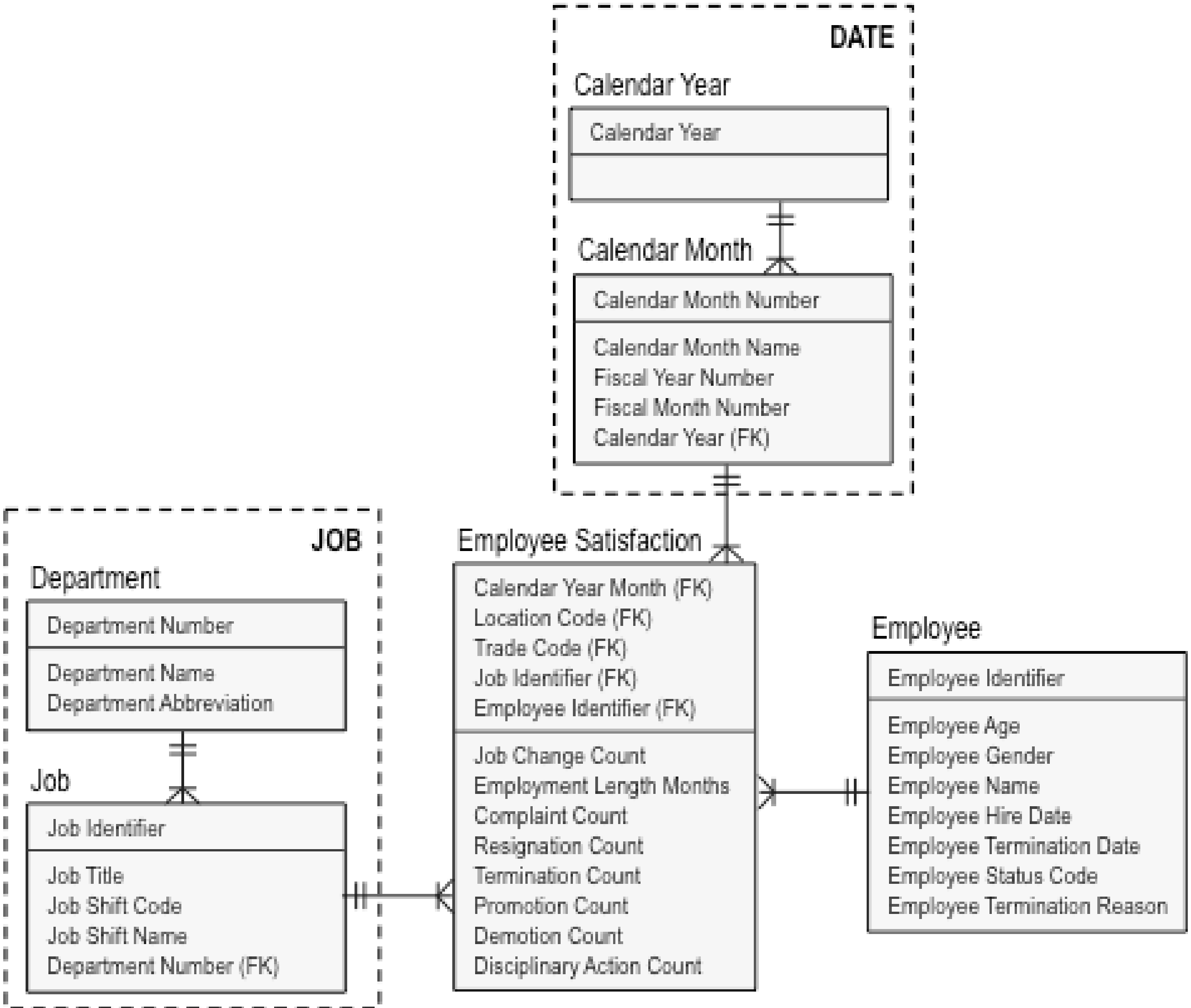
Other important ER concepts include:

- **Cardinality,** which describes the number of occurrences of each entity type that may participate in an occurrence of a relationship. Cardinality options include *zero or one*, *exactly one*, *one or more*, and *zero or more*. In this diagram some of the relationship cardinalities are: a PERSONNEL ACTION changes *exactly one* JOB; a JOB is changed by *one or more* PERSONNEL ACTIONs; a JOB is held by *zero or one* EMPLOYEEs. Sometimes cardinality as defined above is divided with "cardinality" indicating the maximum number and "optionality" indicating the minimum number.

- **Specialization** (also called subtyping) creates a hierarchy or parent/child relationship between an ENTITY super-type and its sub-types. Specialization makes sense when the entity sub-types have unique attributes or participate in relationships not common to all sub-types. In this diagram, PERSONNEL ACTION is an entity super-type with three sub-types.

Comparing E-R and Dimensional Models

Introduction to Dimensional Models



Comparing E-R and Dimensional Models

Introduction to Dimensional Models

COMPONENTS AND STRUCTURE OF THE DIMENSIONAL MODEL

This diagram illustrates a logical dimensional model. The components of the model include:

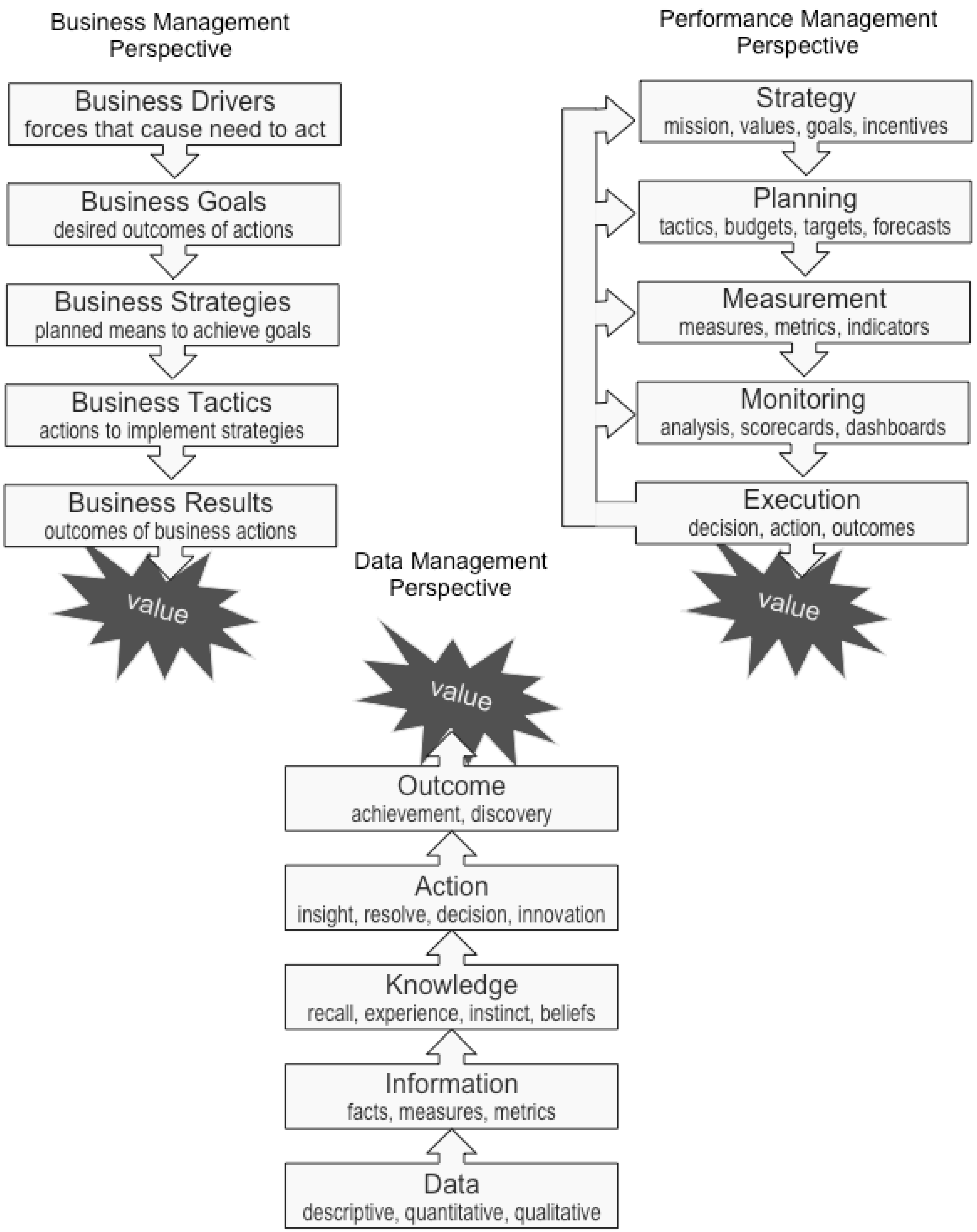
- A **meter** that contains related **measures** of business interest. Each logical dimensional model has only one meter. In this example, the meter is EMPLOYEE SATISFACTION. Related measures include *complaint_count*, *resignation_count*, *termination_count*, *promotion_count, demotion_count, disciplinary_action_count*, etc. Visually, meters in a logical dimensional model look much like entities in the ER model, and measures look much like attributes in the ER model.
- **Dimensions** are the business constraints that provide the means to select, sort, filter, and summarize business measures. In this diagram, the dimensions are EMPLOYEE, DATE, and JOB.
- **Dimension Levels** describe hierarchies that exist within dimensions. This diagram contains the multi-level dimension—JOB—which contains dimension levels of DEPARTMENT and JOB. The EMPLOYEE dimension contains dimension level EMPLOYEE, and the DATE dimension contains dimension levels YEAR and MONTH. Visually, dimension levels in a logical dimensional model look much like entities in an ER model.
- **Dimension Attributes** include identifiers and descriptive data about dimension levels. Some of the dimension attributes of JOB are *job_id*, *job_shift_code*, and *job_title*. Dimension attributes are diagrammed in the same way as attributes of entities in an ER model.
- **Associations within a dimension** describe the dimension hierarchy and are represented as one-to-many relationships from one dimension level to another—for example, one DEPARTMENT contains many JOBS.
- **Dimension to meter associations** are represented as one-to-many relationships from the <u>lowest level</u> of each dimension to the meter. In this diagram, the examples are MONTH to EMPLOYEE SATISFACTION, JOB to EMPLOYEE SATISFACTION, and EMPLOYEE to EMPLOYEE SATISFACTION. What this means is that each unique set of EMPLOYEE SATISFACTION measures is for one MONTH, one JOB, and one EMPLOYEE.

Module 2

Requirements Gathering for Dimensional Modeling

Business Context for Data Modeling

Business Value



Business Context for Data Modeling

Business Value

IMPORTANCE OF BUSINESS CONTEXT

Business context is essential to delivering business value. The context determines the nature of the business intelligence program—the people who will use information, the kinds of information that they need, the business processes to be affected, and the information services to be provided. Business context provides the means to align business intelligence results with business goals.

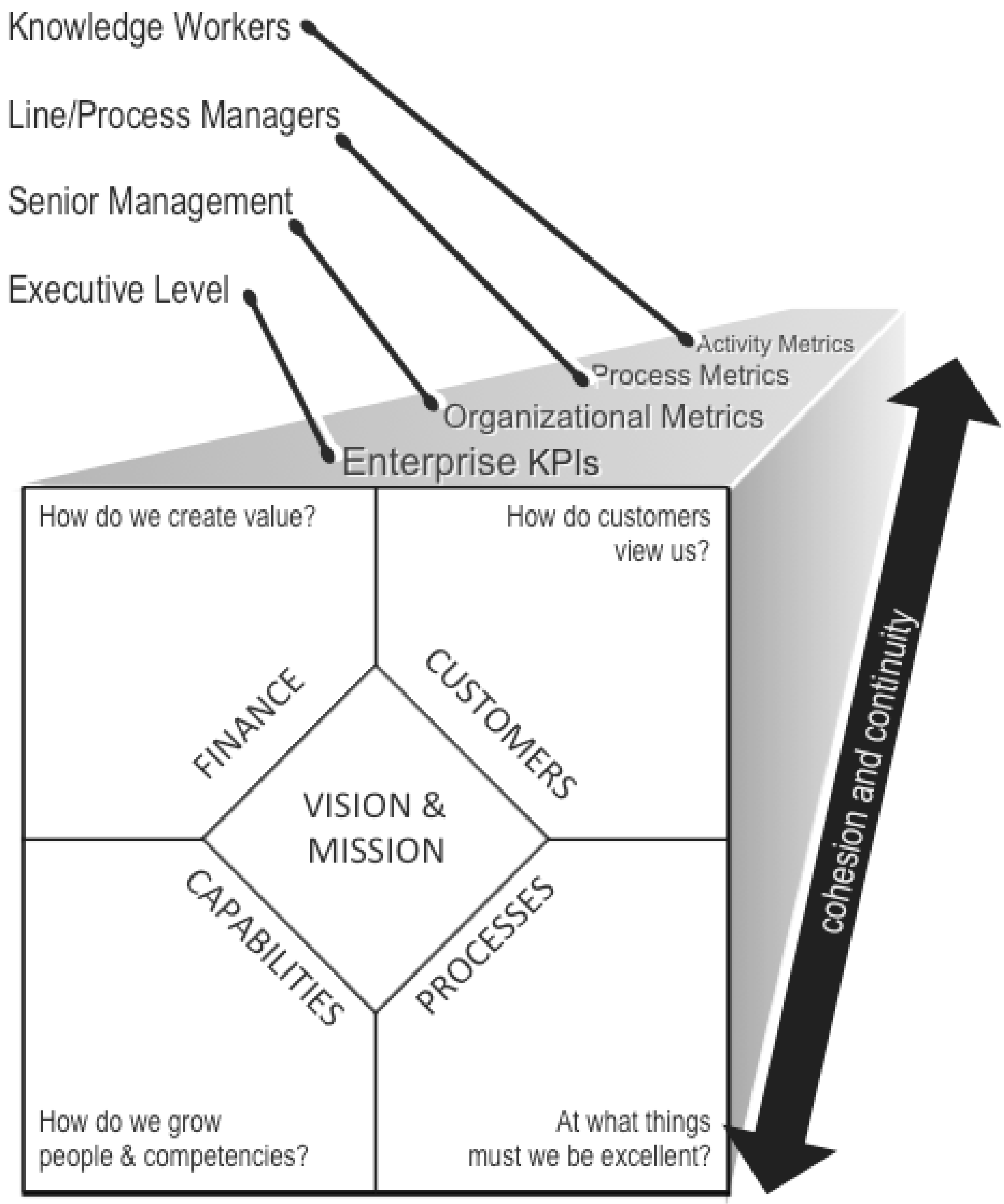
BUSINESS VALUE PERSPECTIVES

As previously described, dimensional modeling is really a business modeling process – modeling of business processes and activities. The “big picture” view of modeling context begins by understanding how we create business value. The facing page illustrates three perspectives of value creation:

- Business management views value as positive results of business activities that align with the tactics, strategies, goals, and drivers of the business.
- Data management views value as business outcomes that are enabled through useful data that enhances knowledge to inform business decisions and actions.
- Performance management views value as successful execution of continuously evolving strategy and planning where data and information are critical elements of a feedback system for continuous growth and improvement.

Business Context for Data Modeling

Business Alignment



Business Context for Data Modeling

Business Alignment

CONTEXT FOR BUSINESS METRICS

The need for business metrics can't be satisfied with a "one size fits all" approach. Understanding the many facets of metrics helps to align solutions with many and varied needs throughout an organization. The diagram on the facing page illustrates these facets as classes of metrics and levels of metrics.

CLASSES OF METRICS

The metrics classification shown here is based upon the Balanced Scorecard (BSC) approach (*The Balanced Scorecard: Translating Strategy into Action,* Kaplan and Norton, Harvard Business School Press, 1996). The BSC approach identifies four categories of metrics important to every business: (1) **Financial metrics** quantify economic factors and fiscal performance. Typical metrics focus on profit, loss, return on investment, etc. (2) **Process metrics** measure the internal processes that produce products, deliver services, and sustain the business. Common process metrics include resource utilization, cycle time, etc. (3) **Customer metrics** quantify common customer factors such as satisfaction and retention, and extend to include satisfaction with and value received from products and services. (4) **Growth metrics** provide insight into creation and innovation capabilities of employees, organizations, and internal systems. The employee satisfaction example that we're using is an example of growth metrics. Metrics in all four classes are associated with the vision and strategy of the organization.

LEVELS OF METRICS

Everyone involved in decision-making processes can get value from metrics. But not everyone needs the same level of detail or the same breadth of scope. Executives, for example, typically need highly summarized measures of a few enterprise-wide business indicators—the key performance indicators (KPIs) of the business. At the opposite extreme, knowledge workers need more detailed metrics with scope limited to the activities with which they are involved. The diagram on the facing page illustrates four common levels of metrics. Note that cohesion and consistency across the levels is an important consideration. The dimensional model supports retrieval of cohesive metrics at the desired level of detail.

Business Questions as Requirements Models

Examples

1. What is the job turnover rate by department, employee gender, employee age and trade? How has it changed historically over the past three years?

2. What is the average length of employment? Break down by age, gender, and shift. Show trends by month for the past five years.

3. How many employee complaints are filed each month? Count by labor union affiliation, shift, and department.

4. What is the rate of job turnover by department, shift, and location? How does it change from month to month?

5. What is the frequency of employee complaints by length of service? How does if differ between departments? How has it changed over time?

Business Questions as Requirements Models

Examples

SAMPLE QUESTIONS

Although we don't normally think of a list as a model, a list of business questions is a model of requirements for business information. It is not practical to develop an exhaustive list of all questions that might ever be asked in a specific domain. Fortunately an exhaustive list isn't needed. A representative and robust list serves as a model of the kinds of questions that need to be answered. A major focus should be on identifying the business constraints (i.e., dimensions) and the most significant measures of the business domain. When this is done well, the analysis and design processes of dimensional modeling extend, expand, and refine the collection of measures and dimensions such that the resulting data structure has the ability to answer many questions not found on the original list.

The list of questions on the facing page provides a set of examples that we will build upon throughout this course. By the end of this course, you'll see how extension, expansion, and refinement work to create a rich and adaptable data structure from these questions.

IMPORTANCE OF QUESTIONS

Often, business people want to simply tell the analyst what information they need. While getting a list of data elements is useful, it does not provide the context, and that context is the basis of the dimensional model design. The business questions provide that context.

Fact/Qualifier Analysis

Mapping Business Questions

2. What is the average length of employment? Break down by age, gender, and shift. Show trends by month for the past five years.

3. How many employee complaints are filed each month? Count by labor union affiliation, shift, and department.

4. What is the rate of job turnover by department, shift, and location? How does it change from month to month?

	job turnover rate	avg. length of employment	number of complaints									
department	1,4		3									
employee gender	1	2										
employee age	1	2										
trade	1											
year	1	2										
shift	4	2	3									
month	4	2	3									
labor union			3									
location	4											

Fact/Qualifier Analysis

Mapping Business Questions

EXAMPLE CONTINUED

This example extends the matrix to map three additional business questions from the EMPLOYEE SATISFACTION domain.

Notice that a single fact—*job turnover rate*—appears in both question 1 and question 4. The fact appears only once in the matrix; however, new qualifiers and new fact/qualifier associations are added as a result of question 4.

Also observe that many of the qualifiers apply to multiple business questions. Each qualifier appears only one time in the matrix, but participates in many associations with the set of facts.

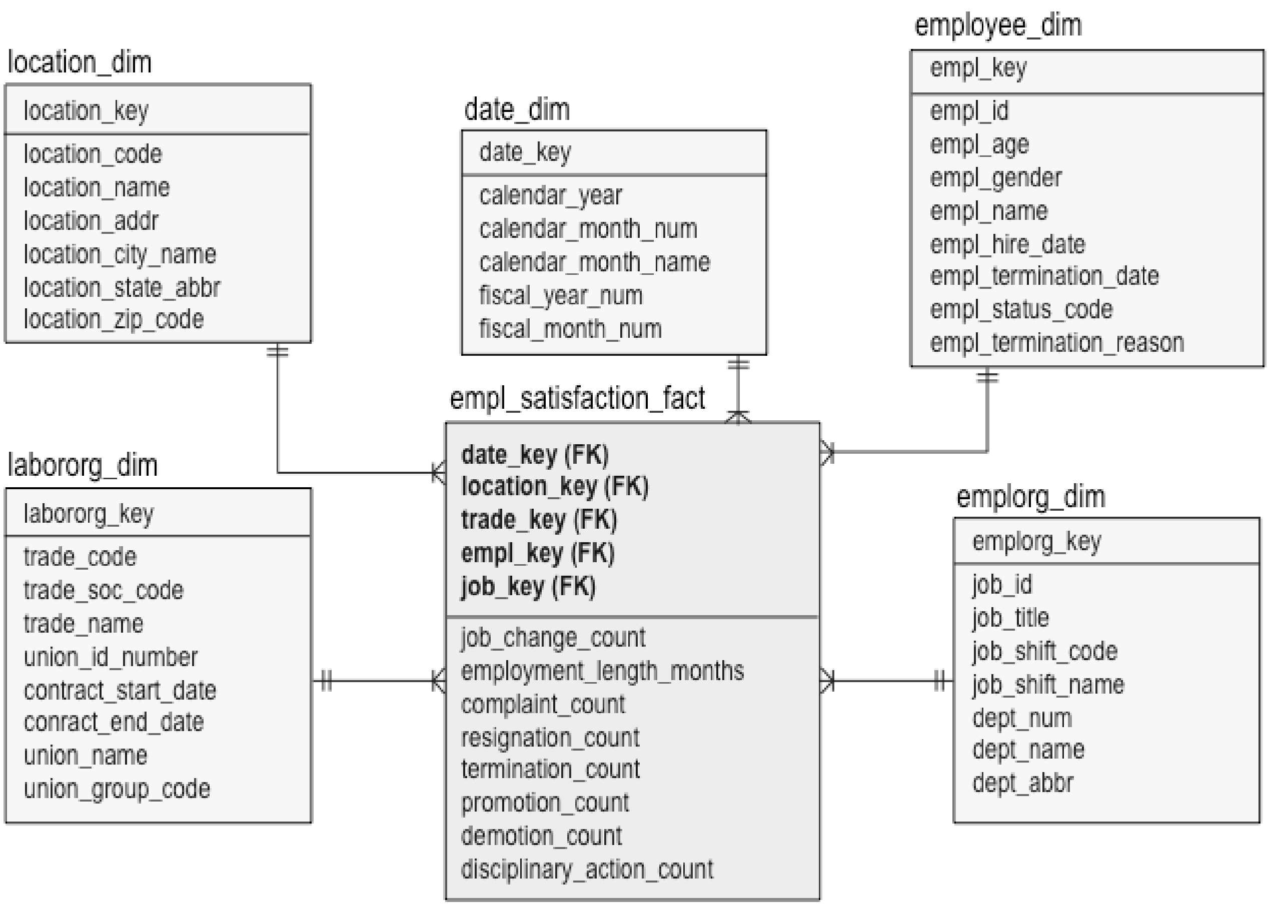
Further, note that a single association of fact and qualifier, such as j*ob turnover rate* with *department*, may occur in more than one business question. When this occurs, *all* questions are recorded in the association cell of the matrix.

Module 3

Logical Dimensional Modeling

Star Schema Fact Tables

Defining the Fact Table Key



Star Schema Fact Tables

Defining the Fact Table Key

DIMENSION TO FACT NAVIGATION

The foundation of OLAP processing is navigation from a set of dimensions to the facts associated with those dimensions. For example, the business question:

> 4. What is the rate of job turnover by department, shift, and location? How does it change from month to month?

asks that a specific measure (rate of job turnover) be reported by a group of dimensions (department, shift, location, and month). Each unique value of *job_turnover_rate* is determined from a unique combination of key values for all of the participating dimensions.

Note that *job_turnover_rate* does not exist in the star schema on the facing page. It is a measure that must be calculated based on the values of other data in the fact table. Derived measures are discussed on the next set of pages.

A COMPOSITE KEY

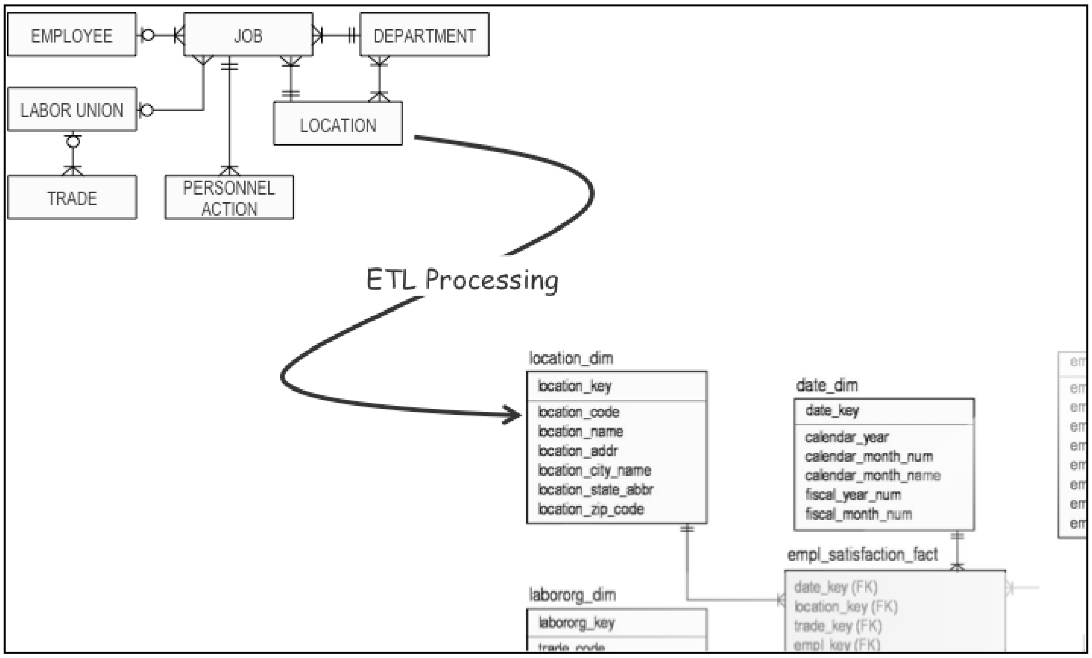
The fact table key is simply the composite of all dimension table keys—a concatenation of dimension surrogate keys. This works well because the keys are used only for navigation by the OLAP tool and are never exposed to a business user of the tool.

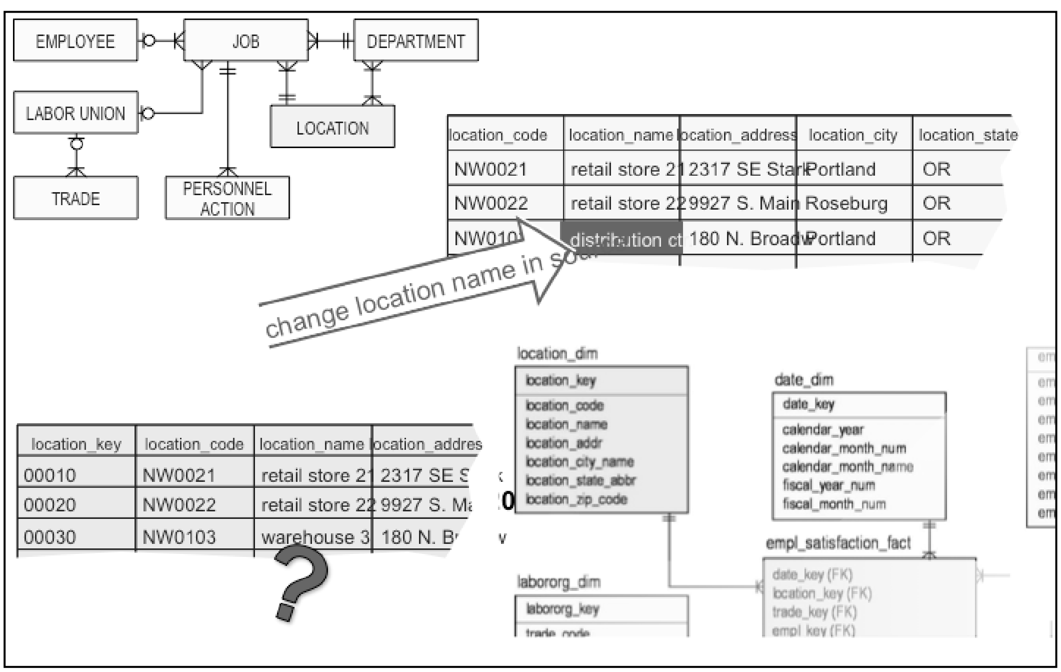
TABLELESS DIMENSIONS

Note that the dimensions tables for employee age and gender have been removed. Once the keys *emp_age* and *emp_gender* are migrated into the fact table, there is no need to maintain them redundantly as single column tables. It is also acceptable, but not essential, to remove those elements from the *employee* dimension table.

Star Schema Design Challenges

Slowly Changing Dimensions





Star Schema Design Challenges

Slowly Changing Dimensions

DESIGNING FOR DIMENSION VOLATILITY

One of the challenges of OLAP analysis is that of analyzing data over time when the dimensions don't remain constant across the time period of interest. Changing data values are common in dimensional data—a problem known as "slowly changing dimensions." The terminology doesn't mean to imply that some dimensions change more quickly than others; rather, that dimension values change less frequently than the values of measures in a fact table.

When the data values in a dimension change, three options are possible to implement the dimension. Using Ralph Kimball's widely accepted terminology, these options are:

- Type 1 dimensions that overwrite dimension rows without retaining a history of dimension changes.
- Type 2 dimensions that preserve dimension history by inserting new rows into the dimension table.
- Type 3 dimensions that keep versions of dimension values by creating new columns for changed data.

Star Schema Design Challenges

Slowly Changing Dimensions—Type 2 Example

Type 2 SCD – Keep Current and Row History

Source Data Change

location_code	location_name	location_address	location_city	location_state
NW0021	retail store 21	2317 SE Stark	Portland	OR
NW0022	retail store 22	9927 S. Main	Roseburg	OR
NW0103	distribution ctr.	180 N. Broadw	Portland	OR

Dimension Before Change

location_key	location_code	location_name	location_address	location_city	location_stat
00010	NW0021	retail store 21	2317 SE Stark	Portland	OR
00020	NW0022	retail store 22	9927 S. Main	Roseburg	OR
00030	NW0103	warehouse 3	180 N. Broadw	Portland	OR

Dimension After Change

location_key	location_code	location_name	location_address	location_city	location_stat
00010	NW0021	retail store 21	2317 SE Stark	Portland	OR
00020	NW0022	retail store 22	9927 S. Main	Roseburg	OR
00030	NW0103	warehouse 3	180 N. Broadw	Portland	OR
00040	NW0103	distribution ctr.	180 N. Broadw	Portland	OR

new row & new surrogate key →

Star Schema Design Challenges

Slowly Changing Dimensions—Type 2 Example

KEEPING ALL OF THE HISTORY

Type 2 dimensions insert a new row, with a new surrogate key value, into the dimension table whenever a change occurs to the value of a dimension element. This is the most common design choice for slowly changing dimensions.

The obvious advantage is that a type 2 dimension preserves all of the history of dimension data changes. For a structure that is also dimensioned by time, there is no need to time-stamp dimension records. Each row of the fact table is associated with only one row of each dimension table, so the time dimension serves as the time stamp. When the date or time dimension doesn't correspond with the timing of data changes, then it is common to add a column to record the effective date of each row.

The most significant disadvantage of type 2 dimensions is rapid growth. Type 2 dimensions may result in very large dimension tables, contributing to both database size and query performance concerns.

Modeling Process Summary

From Business Requirements to Star Schema

- dimensional modeling
  - business requirements
    - define the scope (metrics for business management)
    - develop a list of business questions
    - map questions into fact/qualifier matrix
  - logical dimensional modeling
    - model the meter
      - name the meter
      - select measures from F/Q matrix
    - model the dimensions
      - select qualifiers from F/Q matrix
      - find hierarchies in the qualifiers
      - add dimensions & associate with meter
      - refine the dimensions
      - add dimension attributes
    - refine the meter
      - declare the grain
      - refine the measures
  - physical dimensional modeling
    - name the dimensions
    - model the dimension tables
    - define the dimension table keys
    - model the fact table
    - define the fact table key

Modeling Process Summary

From Business Requirements to Star Schema

PROCESS OVERVIEW

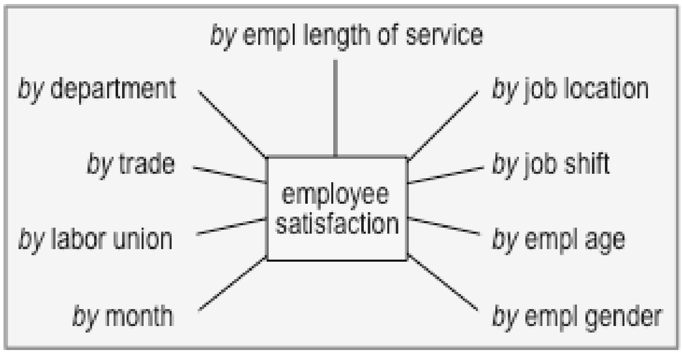
The facing page illustrates dimensional data design activities from requirements through physical design. This diagram provides a simple, easy-to-reference summary of the dimensional modeling process from beginning to end.

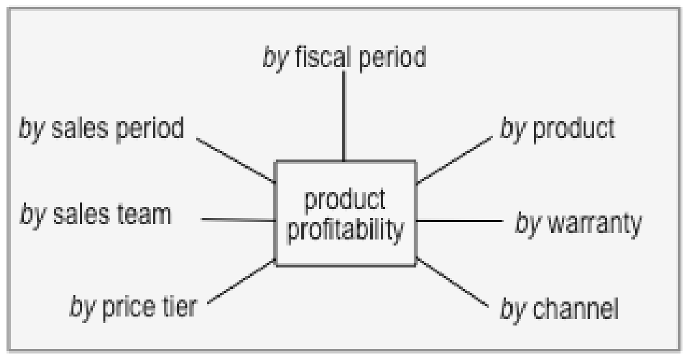
Module 5

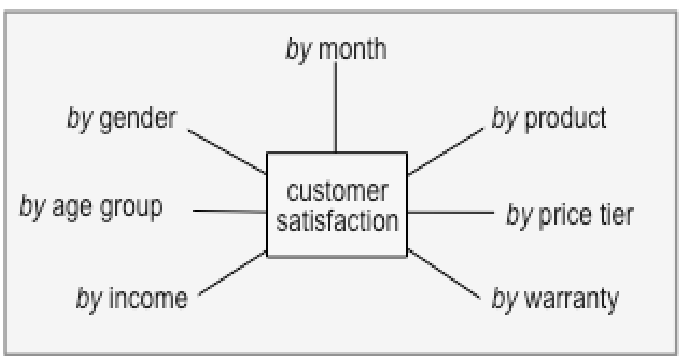
Dimensional Data and Business Analytics

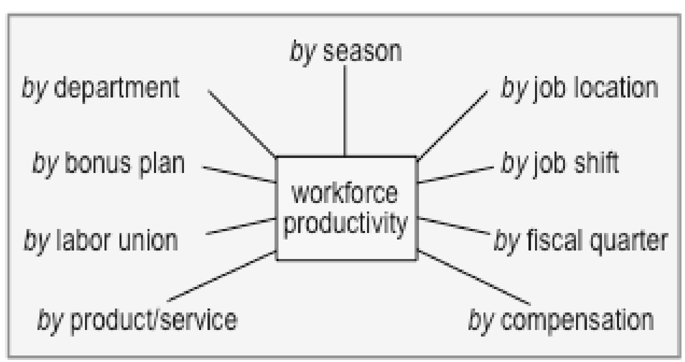
Delivering Business Value

Data Enabled Business Analysis









Delivering Business Value

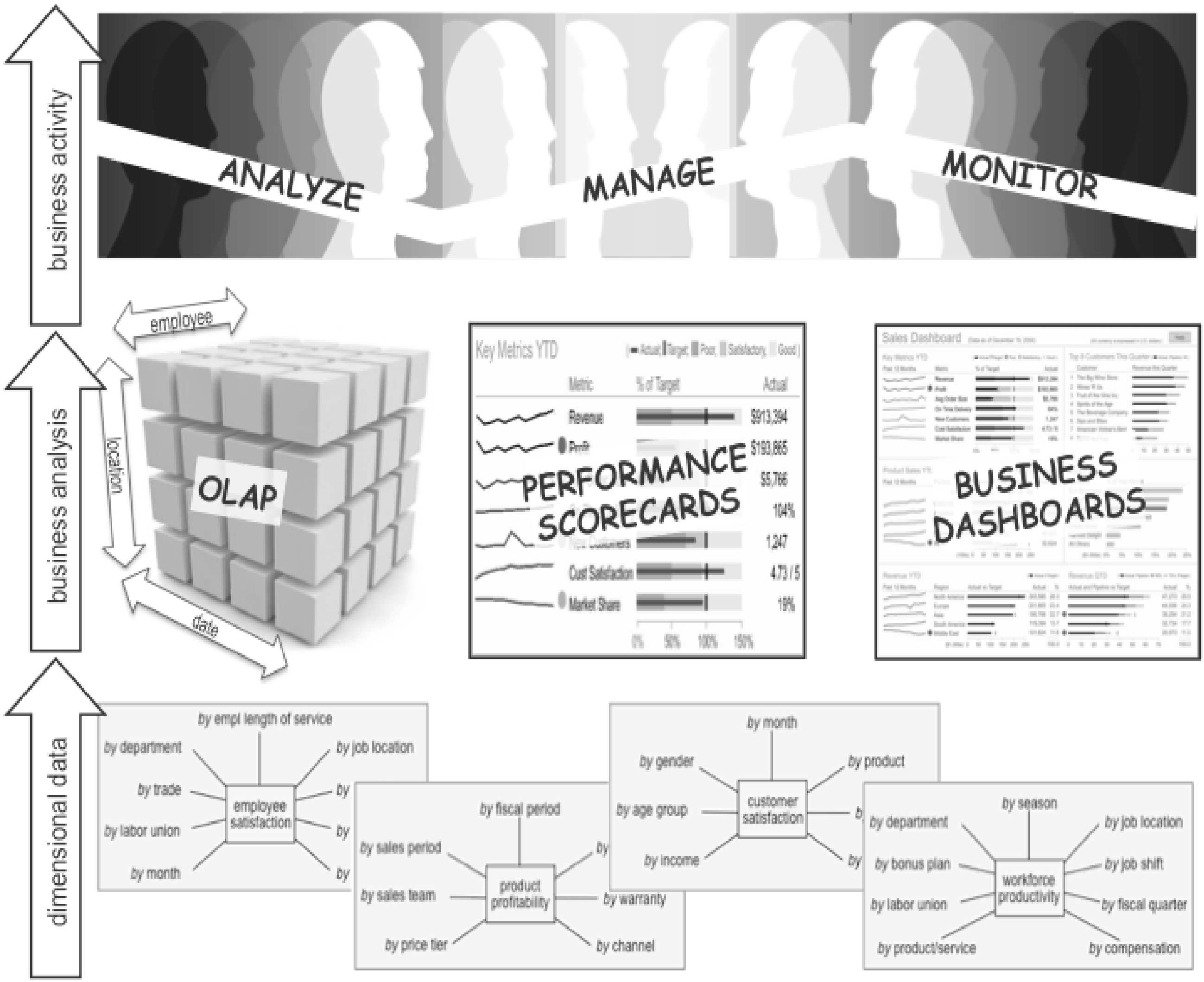
Data Enabled Business Analysis

MANY DATA MARTS

A typical business intelligence environment encompasses many different data marts at different levels of detail, with different focus and purpose, yet with substantial overlap of dimensions and some overlap of measures. The example that we've followed throughout the course – employee satisfaction – is but a single data model in a more complex environment. The employee satisfaction model is a summary, and it is likely that a data mart exists to contain more detailed, atomic level data. It is also likely that many other data marts exist with similar measures and similar dimensions. Consistency and conformity across data marts is essential to support cross-functional business analysis.

Delivering Business Value

Collecting, Analyzing, and Using Business Metrics



Delivering Business Value

Collecting, Analyzing, and Using Business Metrics

DATA IN SUPPORT OF BUSINESS ANALYSIS, MANAGEMENT, AND MONITORING

Dimensional data alone has no inherent value. Value is achieved only through the use of the data to understand trends and make decisions. Multi-dimensional data is the standard today to collect and store business metrics. Business analysis, however, continues to be a job for people.

OLAP is the most common category of business analysis tools currently in use. OLAP is powerful, the tools are relatively easy to learn and use, and they are a widely accepted and integral part of business analysis.

Beyond OLAP, dimensional data is used to populate scorecards that are used to manage business performance, and to populate business dashboards to monitor business results.

THE DATA WAREHOUSING INSTITUTE

Exercises

Exercise Instructions and Worksheets

Exercise 2: Business Questions

Instructions

STARTING TO UNDERSTAND REQUIREMENTS

Using the following statement as a basis, develop a list of questions to meet the need.

> Your instructor wants to track her/his performance as a teacher. The goal is continuous improvement as an instructor by understanding what is and is not effective and successful in the classroom. Items of interest include: instructor rating by student evaluations across time, which courses receive the highest ratings, etc.

The instructor is the business subject matter expert and the source of information for this exercise. You may ask questions or conduct an informal interview to gather the information that you need. Remember to address all the following aspects:

- **People**—Who are the stakeholders for monitoring and evaluating instructor performance? What metrics do they use today? What unanswered questions do they have? What levels of metrics do they need?

- **Performance**—Which classes of metrics are within the scope of modeling—Financial? Process? Customer? Growth? How do stakeholder questions relate to these four categories of performance management?

- **Process**—Which business processes are within the scope of modeling? What can be measured about those processes? What metrics are needed to effectively manage the processes? What metrics are needed for knowledge workers to perform the process activities? Which process components make good metrics subjects?

