TDWI Data Modeling

Data Analysis and Design for BI and Analytics Solutions

Transforming Data With Intelligence™

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.

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

info@tdwi.org

Publication Date: April 2021

TABLE OF CONTENTS

COURSE OBJECTIVES

To learn:

- ✓ ***Differences in modeling techniques for business transactions, business events, and business metrics***
- ✓ ***Different types of data and their implications***
- ✓ ***How modeling processes differ based on analytics objectives and data management technology***
- ✓ ***Application of business context to modeling activities***
- ✓ ***The role of business requirements in BI data modeling***
- ✓ ***The role of source data analysis in data modeling***
- ✓ ***Use of normalized modeling techniques for data warehouse analysis and design***
- ✓ ***Use of dimensional modeling techniques for BI and data mart analysis and design***
- ✓ ***The roles of generalization and abstraction in data warehouse design***
- ✓ ***The roles of identity and hierarchy management in data model design***
- ✓ ***How time-variant data is represented in data models***
- ✓ ***Implementation and optimization considerations for data stores***

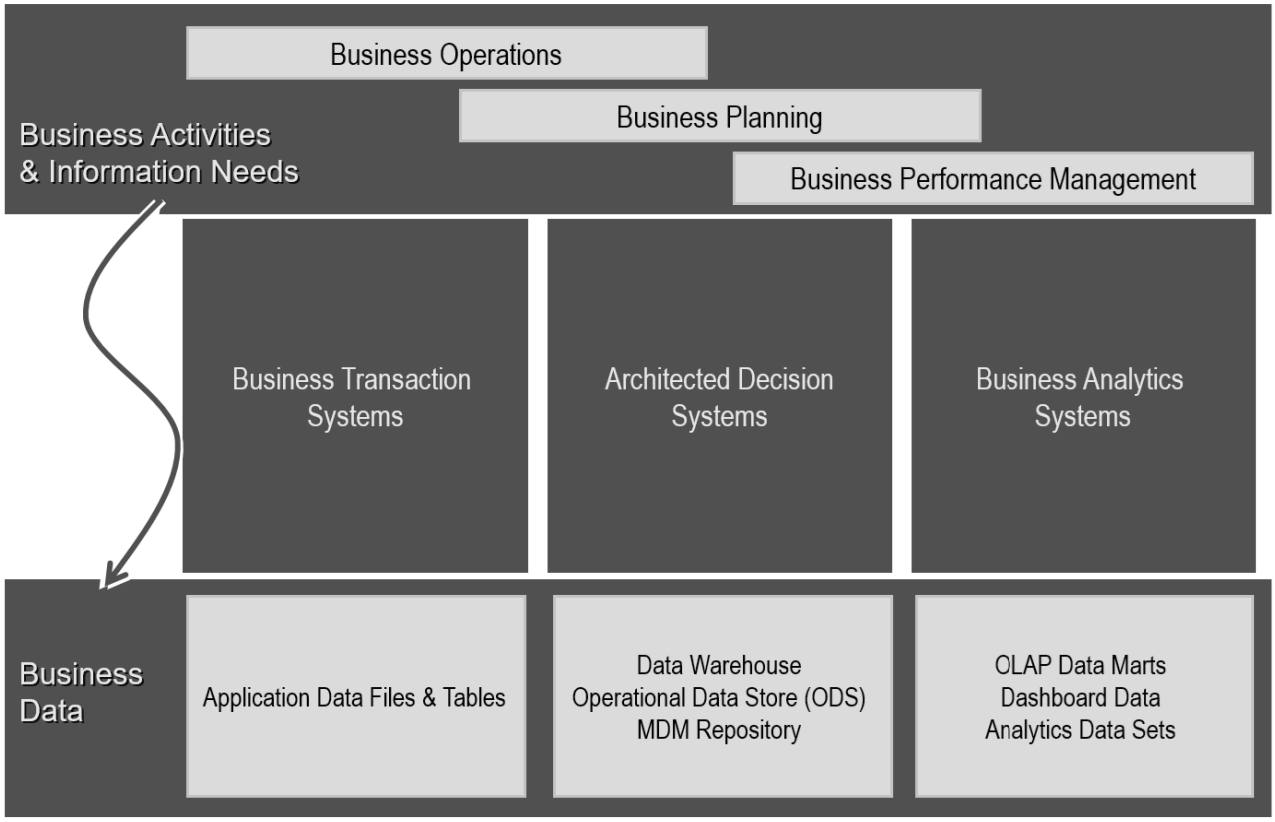
tdwi
Transforming Data With Intelligence™

Module 1

Data Modeling Concepts

The Data Modeling Life Cycle

Where Data Modeling Begins and Ends



The Data Modeling Life Cycle

Where Data Modeling Begins and Ends

THE BEGINNING

A *data model* is an abstract representation of the data in an enterprise or of the information that is derived from that data. Data and information can (and should) be represented at multiple levels of abstraction, each providing a different perspective and understanding of the data. The highest level of abstraction is a business context view with both external (outside looking in) and internal (looking from within) perspectives. Thus data modeling begins with business activities and the information needs of those activities. This view describes the scope of and the context for business information requirements—a sensible start to *modeling the right data*.

THE END

The natural conclusion of data modeling is implemented data—data files and database tables. Far from the top level of abstraction, implemented data is beyond the bottom tier; it is no longer abstract but real and physical. At this level attention turns to *the right implementation for the data.*

THE MIDDLE

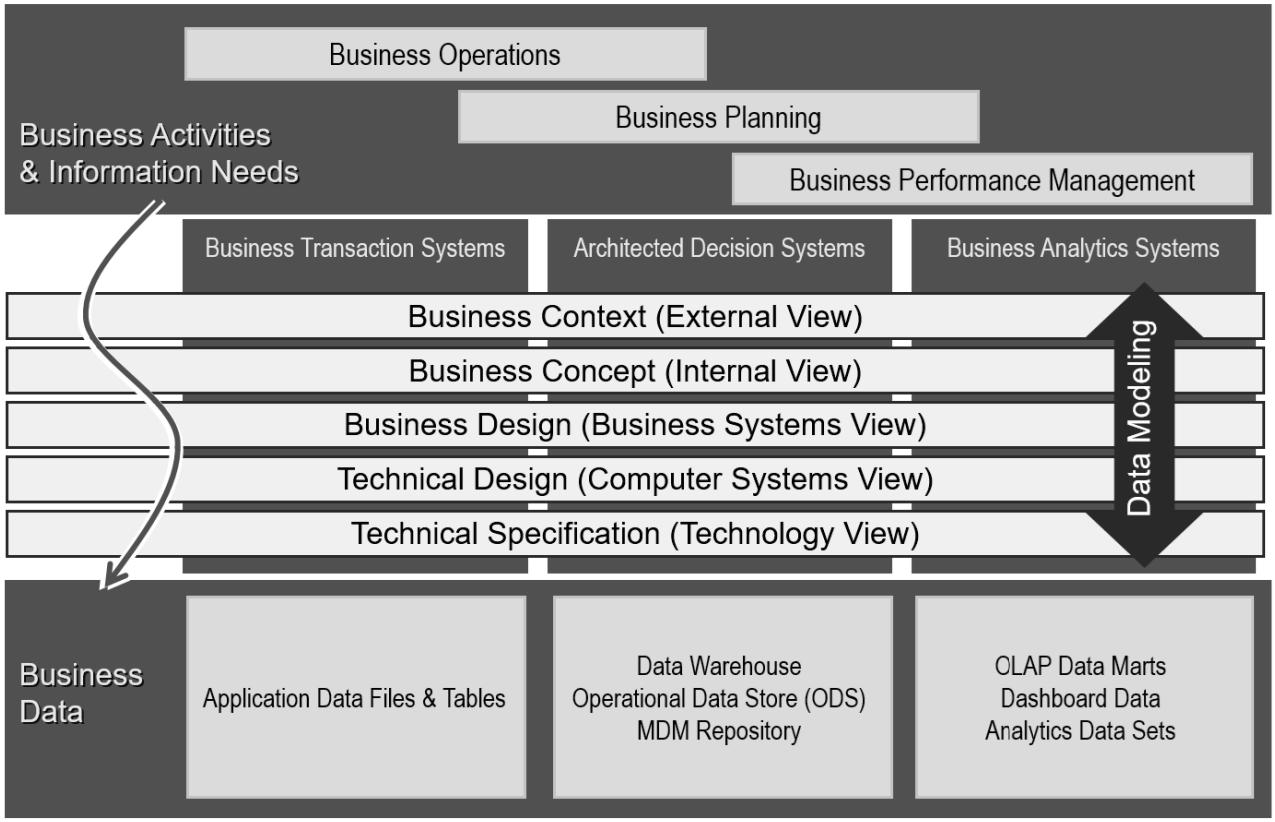
The complexities and challenges of data modeling lie between the top layer of context and the bottom tier of implementation. Getting from the right data to the right implementation involves an understanding of the business and many information systems, ranging from those that capture data in the course of business activities to those that turn data into information and supply that information to the business.

IN REVERSE

In business analytics, it is often necessary to explore data sets that are not fully understood. This search for value reverses the process described above. Analysis starts with a physical data structure that is the subject of exploration; the process ends with an understanding of its content and applicability in the enterprise.

The Data Modeling Life Cycle

Between Business Needs and Implemented Data



The Data Modeling Life Cycle

Between Business Needs and Implemented Data

BUSINESS AND SYSTEMS VIEWS

Business context alone is insufficient to describe business requirements for data and information. Similarly, implementation is not adequate to design and deploy information systems and databases. Multiple levels of abstraction are needed to manage complexity, understand and document from multiple perspectives, communicate effectively, and provide a natural progression from need to solution. The six layers of modeling abstraction are based on the Zachman Framework (*A Framework for Information Systems Architecture, IBM Los Angeles Scientific Center Report*, John A. Zachman, 1986). Zachman's framework addresses many aspects of information systems architecture. This course is concerned solely with the data. To learn more about the Zachman Framework, visit www.zachman.com. The six levels of data modeling are:

CONTEXTUAL

Context modeling provides a view of the scope of the planned data warehousing program. Context models communicate understanding of the business requirements and establish a context for analysis. This level corresponds to the *Executive Perspective* of the Zachman Framework.

CONCEPTUAL

Models at this level are about understanding the required set of data stores. Conceptual models describe data requirements from a business point of view without the burden of technical details. This corresponds to the *Business Management Perspective* of the framework.

LOGICAL

Models at this level refine conceptual models by documenting entities, their attributes, and their relationships. These models are technology-oriented designs, although they are platform-independent. This level corresponds to the *Architect Perspective* of Zachman's framework.

STRUCTURAL

Structural models move a step closer to implementation. They specify the design necessary for the warehouse/marts to maintain history, to distribute data, and to provide for ease of use. Structural modeling corresponds to the *Engineer Perspective* of the Zachman framework.

PHYSICAL

Physical models represent the detailed specification of what is physically implemented using specific technology. This level corresponds to the *Technician Perspective* of Zachman's framework.

FUNCTIONAL

This level does not describe models, but implemented and instantiated data stores. This is the tangible result of modeling activity and the product of the development process. This level corresponds to the *Enterprise Perspective* of the Zachman framework.

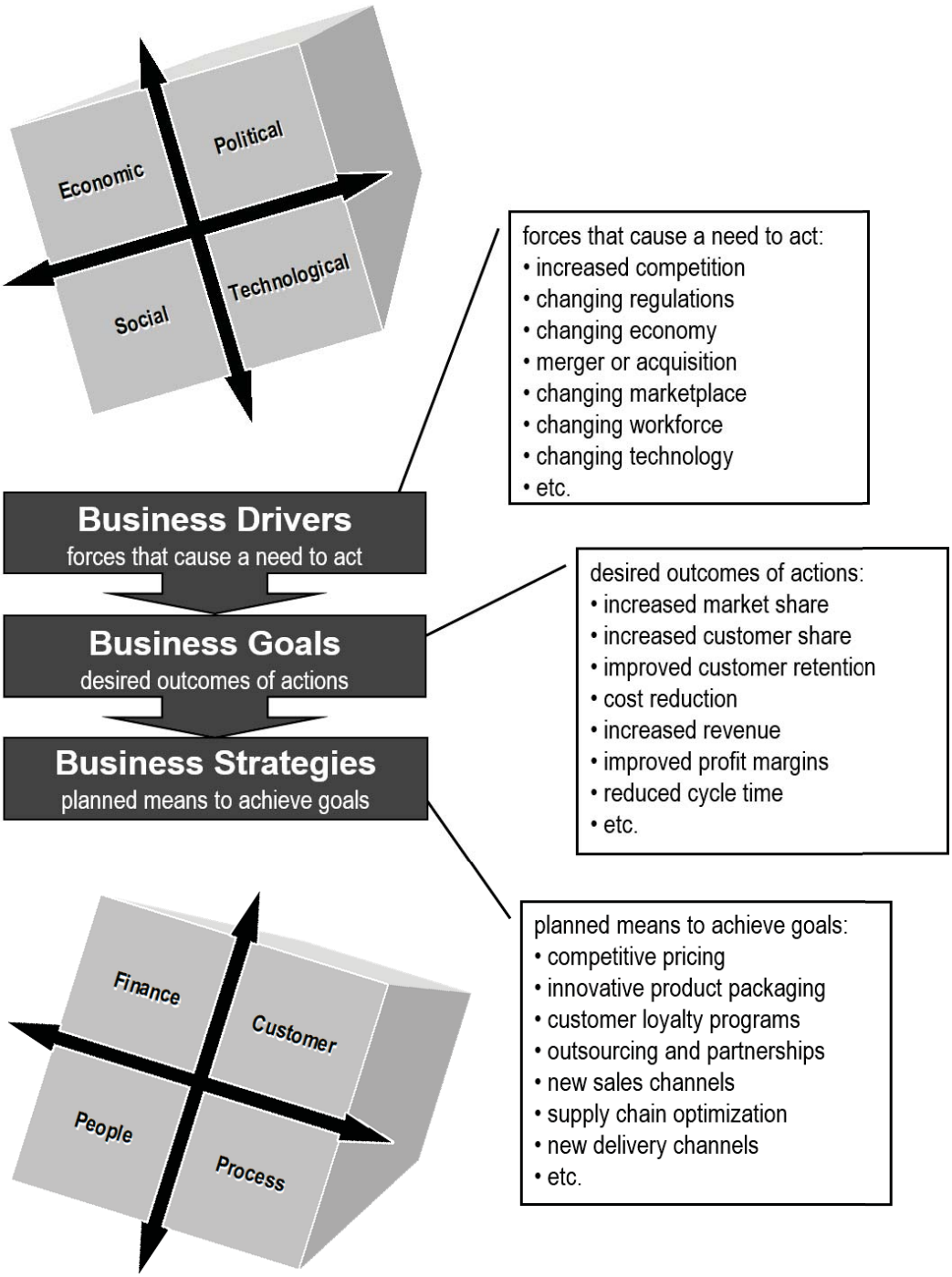
Transforming Data With Intelligence™

Module 2

Business Data Models

Business Context

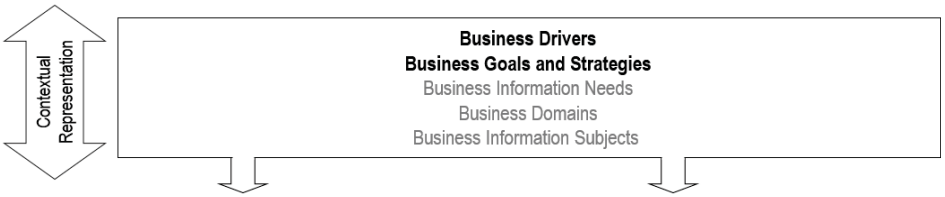
Business Drivers, Goals, and Strategies



Business Context

Business Drivers, Goals, and Strategies

THE MODELING FRAMEWORK



WHY MODEL

Business context determines the nature of data and information services: the business processes to be affected, the kinds of applications to be implemented, and the information services to provide. Business context provides the means to align data with business goals. The contextual representation is generally determined prior to initiating a project. Project team members should be aware of and use the context to establish a more complete and relevant set of requirements.

BUSINESS DRIVERS

Business drivers are those things that are strategically important in positioning the business to achieve its short- and long-term goals. They are the external forces that have significant influence on the operation and performance of a business. Drivers create the need to take action, but they don't dictate the actions to be taken. Common business drivers include changing economic, political, social, and technological factors.

BUSINESS GOALS

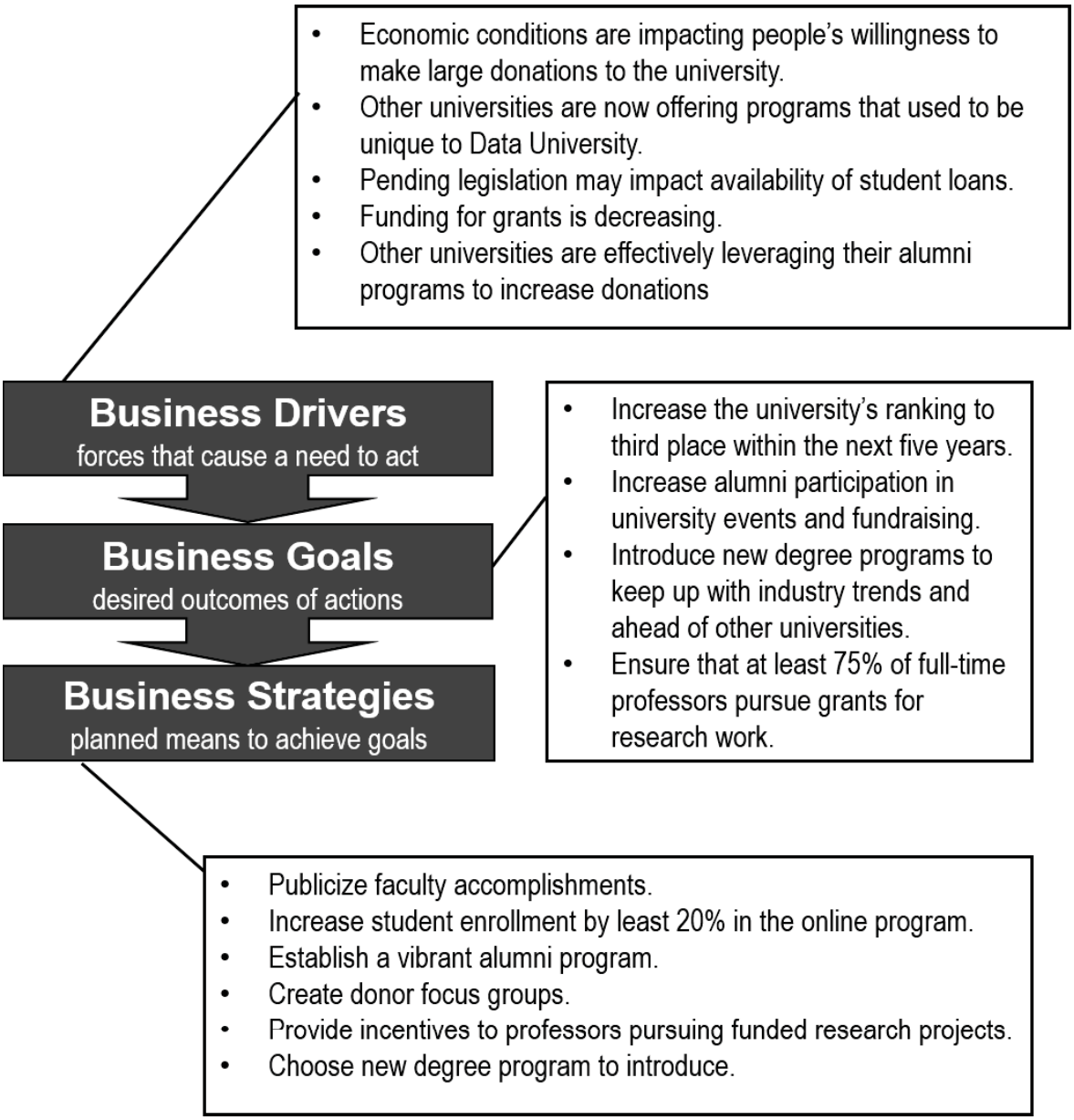
Business goals are the things that the business wants to accomplish to respond to business drivers. Drivers create the need to act. Goals describe the desired outcomes of taking action. Goals are commonly related to financial or operational performance (i.e., cost reduction, generation of revenue, increased market share, etc.). Goals are most effective in setting data management priorities and directions when they are: (1) described by clear, concise, understandable statements, (2) specific enough that the level of achievement can be measured, and (3) of high business priority.

BUSINESS STRATEGIES

Business strategies are plans to turn the defined business goals into reality. They include the things that the business will do to shape its future and achieve the business goals. The range of strategies is broad—introducing new products, exploiting new sales channels, pricing competitively, optimizing business processes, etc. Strategies help to determine which business processes and organizations most need to be information-enabled.

Business Context

Business Drivers, Goals, and Strategies



Business Context

Business Drivers, Goals, and Strategies

BUSINESS DRIVERS

Throughout this course, examples will be presented based on a hypothetical university that provides post-secondary education to students and has a large faculty which engages in research and publication.

Sample business drivers for the Data University could include:

- Tuition costs are having an increasing impact on university selection
- Donors are demanding innovative programs
- Many universities are offering online programs

BUSINESS GOALS

Sample business goals for the Data University could include:

- Become recognized as a leading university in our fields
- Create and publicize innovative programs
- Increase donor involvement

BUSINESS STRATEGIES

Sample business strategies for the Data University could include:

- Engage field luminaries as advisors or professors
- Create donor focus groups
- Implement alumni relationship organization
- Expand unique online offerings
- Increase the proportion of students enrolled in online programs by 20 percent

- What are some of the business drivers for your company?
- What goals has your company created as a result?
- What are some of your business strategies to achieve those goals?

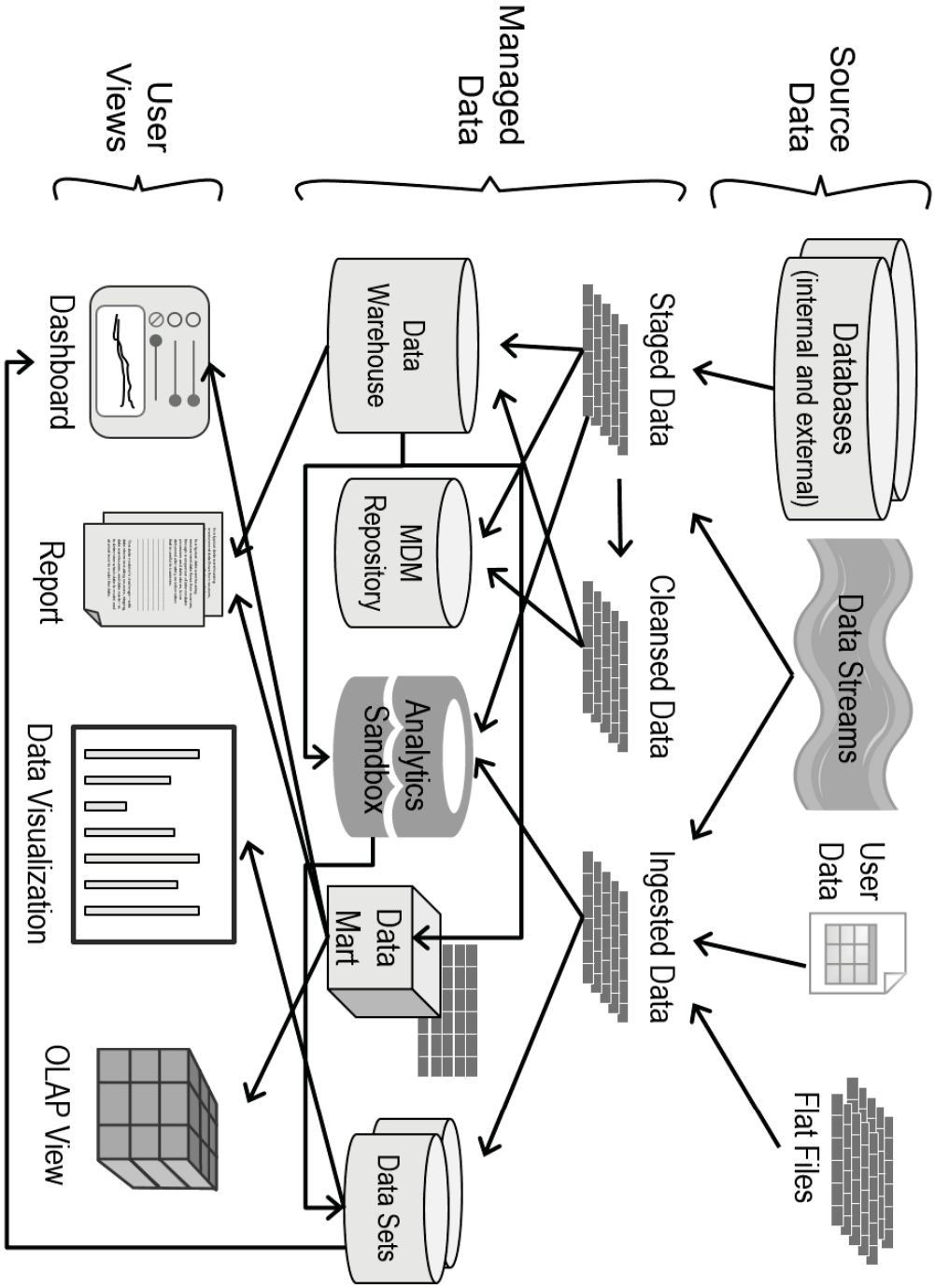
tdwi
Transforming Data With Intelligence™

Module 3

Logical Data Models

What to Model

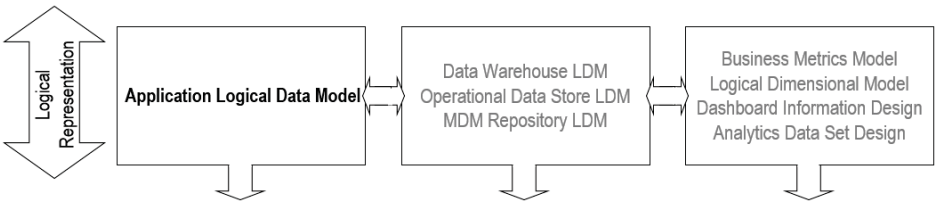
The Data and Information Pipeline



What to Model

The Data and Information Pipeline

THE MODELING FRAMEWORK



FLOW OF DATA

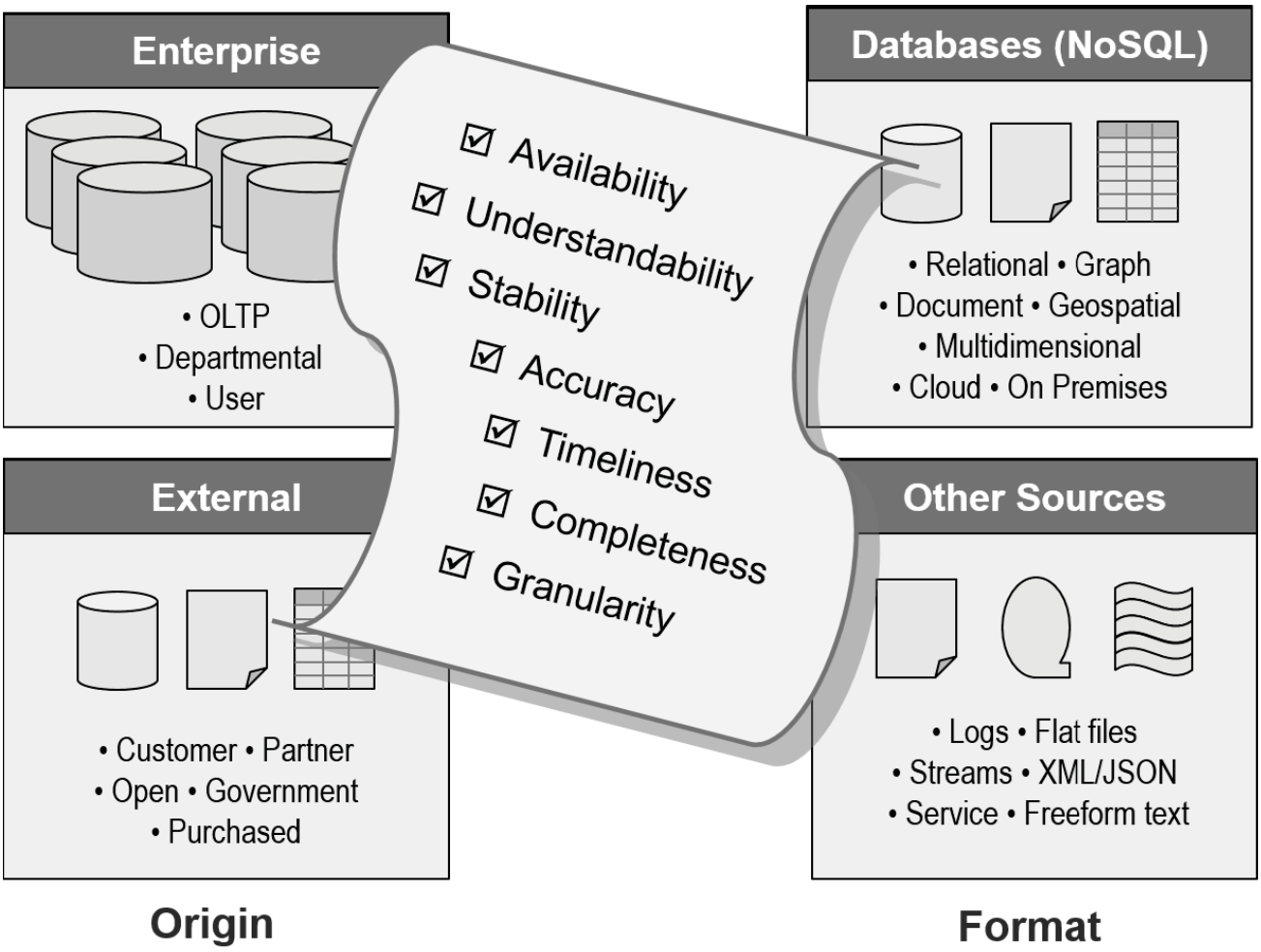
In a typical data warehousing environment data flows from sources, through a sequence of intermediate processes and data stores, to be delivered ultimately as information that is useful to business. The data modeler's challenge—with data stores including sources, staging, data warehouses, and data marts—is to determine which data to model and at what level to model the data.

Every project must determine the right amount of data modeling to be done. Compromises are often necessary to achieve appropriate balance between data complexity and the real project constraints of time and resources. A practical minimum standard includes:

- User view modeling that is sufficient to understand end-user requirements for data.
- Modeling of managed data that is detailed enough to meet the data delivery requirements for user products, including data flows, cleansing, and integration.
- Source data modeling as needed to understand the content and structure of data sources.

Understanding Data Sources

Why Sources Matter



Understanding Data Sources

Why Sources Matter

CHOOSING THE RIGHT DATA

Choosing the right data sources is a critical step to getting the right data in the warehouse and delivering the right information to the business. Making the right choices depends on real understanding of the contents of each data source, followed by careful qualification of sources based on the criteria that matter to your project and your data warehouse. Some of the common qualification criteria are listed below.

Qualifying Criteria	Assessment Questions
Availability	How available and accessible is the data? Are there technical obstacles to access? Or ownership and access authority issues?
Understandability	How easily understood is the data? Is it well documented? Does someone in the organization have depth of knowledge? Who works regularly with this data?
Stability	How frequently do data structures change? What is the history of change for the data? What is the expected life span of the potential data source?
Accuracy	How reliable is the data? Do the business people who work with the data trust it?
Timeliness	When and how often is the data updated? How current is the data? How much history is available? How available is it for extraction?
Completeness	Does the scope of data correspond to the scope of the data warehouse? Is any data missing?
Granularity	Is the source the lowest available grain (most detailed level) for this data?

BEYOND ENTERPRISE DATA

The advent of big data analytics has pushed the boundary for what is considered an acceptable data source. Although past solutions were limited to enterprise data extracted from relational databases, some analytics applications may leverage data that originates outside the enterprise and information that is managed in a nonrelational format.

In this environment, the concept of data quality takes on increased nuance. Rather than asking, “is the data high quality,” it becomes necessary to ask, “is the quality of this data suited to this business need.” An enterprise dashboard likely requires high-quality data, while less reliable data may be acceptable for an analytic model that forecasts market demand.

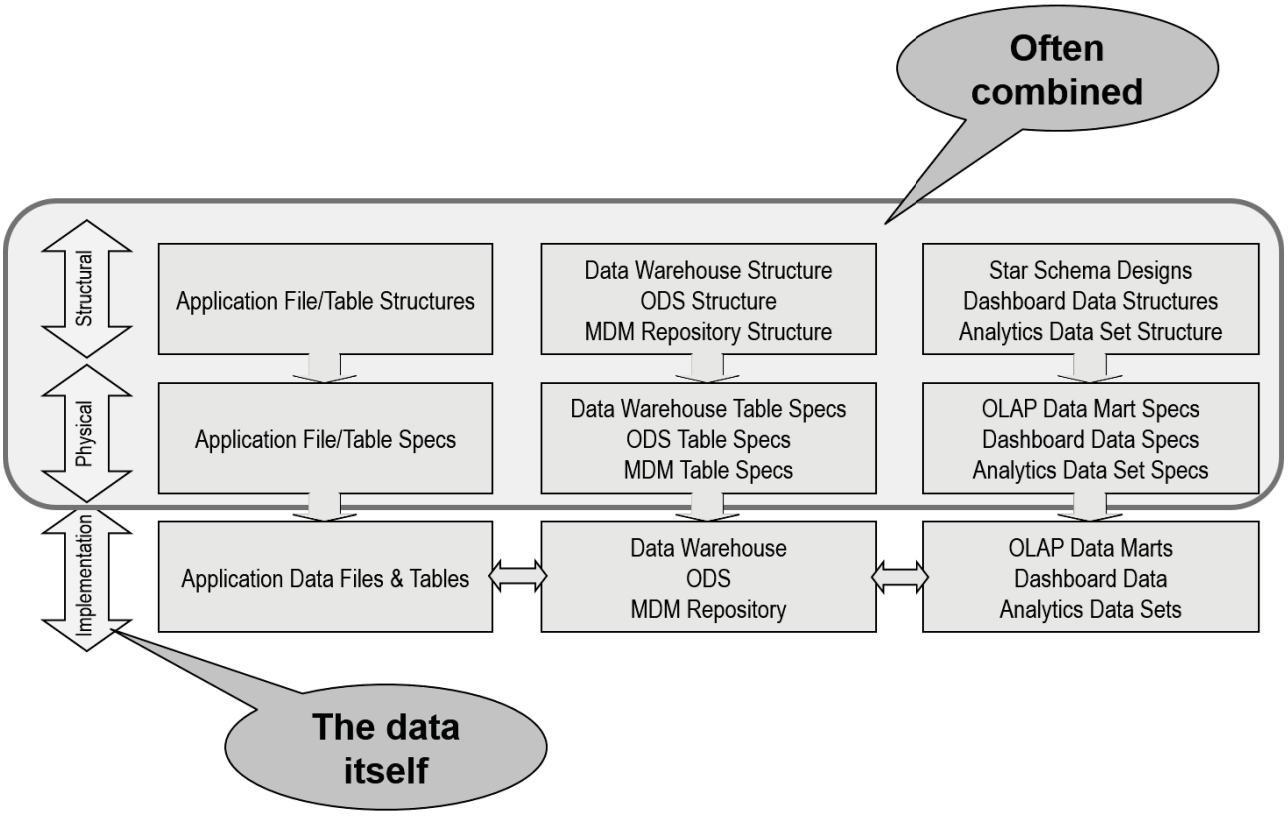
tdwi
Transforming Data With Intelligence™

Module 4

Physical Data Models

Defining Physical Modeling

Both Structural and Physical



Defining Physical Modeling

Both Structural and Physical

PHYSICAL AND STRUCTURAL MODELS

The enterprise architecture framework introduced in Module 1 divided modeling activities across six levels of abstraction. These layers ranged from a contextual representation of the business situation to the implementation of actual data.

Once data modelers have completed logical modeling activities, they often move to a stage called "physical modeling" that actually combines the physical and structural layers of this framework.

This simplification is both acceptable and commonplace. In fact, most modeling tools combine these categories into a single modeling activity. In this module, physical and structural modeling will sometimes be collapsed under the "physical" moniker, in acknowledgement of industry practices.

It is useful to note that many teams divide responsibilities at this stage, even though the result is a single model. For example, a data modeler may define the structural features of the solution, such as the tables, columns, and keys, while a database administrator takes care of the detailed specifications, such as data types, indexing, and so forth.

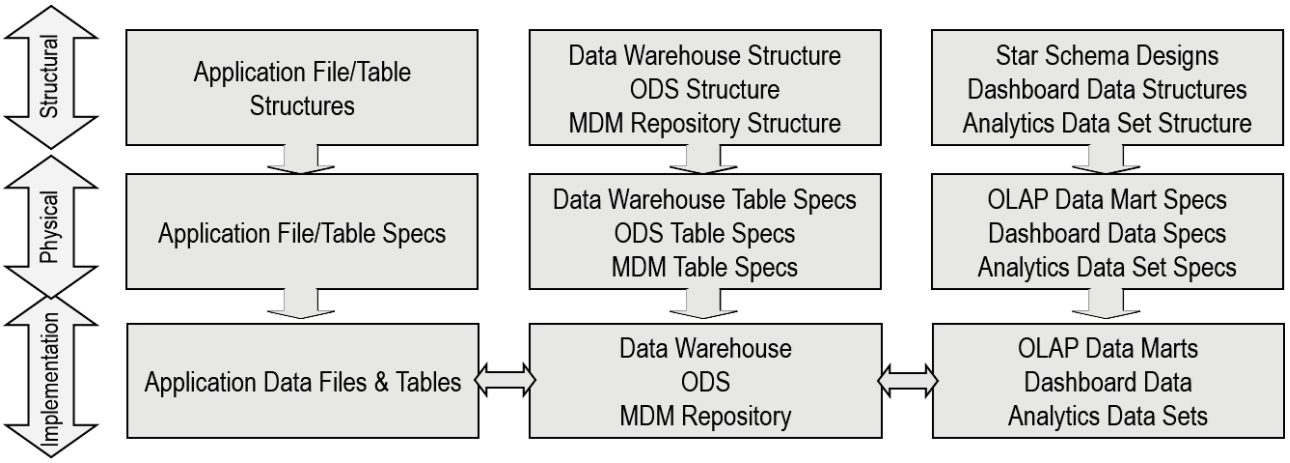
IMPLEMENTATION

The bottom-most layer of the enterprise architecture framework is the implementation layer. This component represents the data itself. An additional model is not produced for this layer.

Data Structure in Transaction Systems

Extracting the Structure of Existing Data

Table	Column	Description	ID	Entity
Grades	Student ID	Unique identifier	Yes	
	Student Name	Name of student		Student
	Student Number	Assigned student number		Student
	Course Number	Catalog number of course		Course
	Grade	Letter grade for the course		
	Credit Hours	Credit hours for the course		Course
	Program Name	Program student is enrolled		Program
Student	Student ID	Unique identifier	Yes	
	Student Name	Name of student		
	Student Birth Date	Birth date of student		
	Student Marital Status	Marital status of student		



Data Structure in Transaction Systems

Extracting the Structure of Existing Data

LEVELS OF TRANSACTION SYSTEM MODELS

Four levels of transaction systems data and models are of interest for the value that they contribute to data warehouse modeling:

- *Implemented Data* already exists as application data files and tables.
- *Physical models* usually exist in part as descriptions of data structures (DDL, COBOL copy code, etc.). Multiple potentially confusing descriptions may exist for a single file or table. This is especially true of older legacy systems.
- *Structural Models* may need to be created to identify and resolve conflict where multiple physical descriptions exist for a file or table.
- *Logical Models* exist infrequently. Where logical models are present, they are often incomplete or out-of-date, and are almost certainly limited to a single application's view of the data.

BOTTOM-UP DATA SOURCE MODELING

It is valuable to build a source data model in reverse—working from implemented data backward to an E-R model. The process is one of examining every data element (column or field) in every data store (file or table) to answer the questions:

- What business fact (attribute) does this data element contain?
- What thing (entity) is it a fact about?
- Does the data element identify the entity that it describes?
- Does the data element indicate a relationship to another entity?

The answers can be represented in a table such as that on the opposite page, which is readily translated into an E-R diagram if necessary.

This bottom-up understanding of source data is referred to as *data profiling*, and it can be supported by automated tools that analyze data and identify patterns, domains of values, apparent key relationships, and more. A data profiling process reveals the truth about the data, which often contradicts existing artifacts such as schema designs or E-R documentation which has not been kept up-to-date.

This page intentionally left blank.