Previews of TDWI course books are provided as an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews can not 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. The pages shown are not consecutive. The page numbers as they appear in the actual course material are shown at the bottom of each page. All table-of-contents pages are included to illustrate all of the topics covered by a course.

TDWI Data Warehousing Concepts and Principles

An Introduction to the Field of Data Warehousing

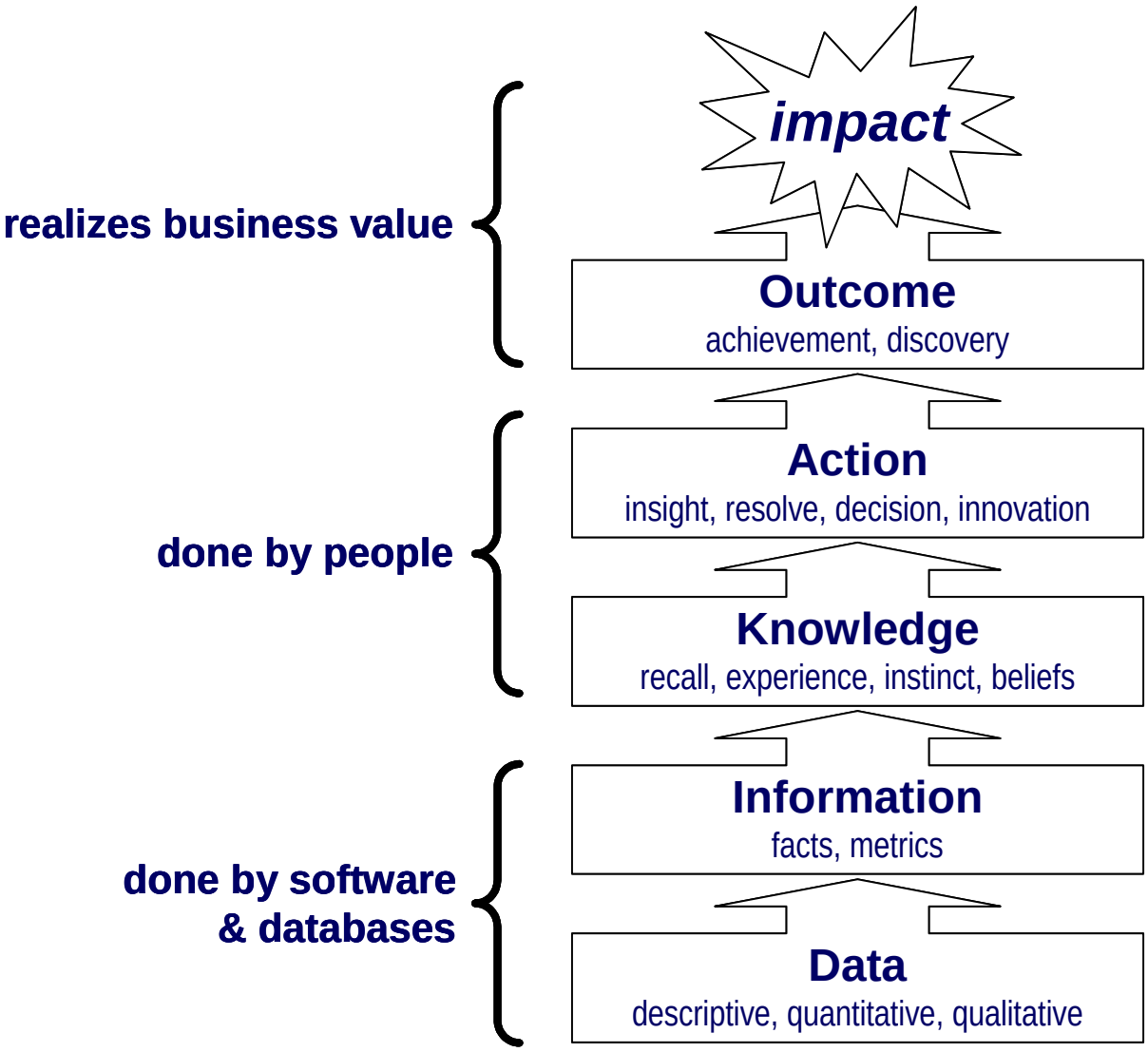
TABLE OF CONTENTS

Module 1

Data Warehousing Concepts

Data Warehousing Basics

Understanding Data, Information, and Knowledge



Data Warehousing Basics

Understanding Data, Information, and Knowledge

DATA

Data is composed of individual and discrete facts that collect descriptive, quantitative, and qualitative values of business interest. Data warehousing involves two types of data – operational data which describe the day-to-day events and transactions of the business, and informational data that are reconciled, integrated, and cleansed to constitute the raw material from which information is constructed.

INFORMATION

Information is an organized collection of data presented in a specific and meaningful context. The purpose of business information is to *inform* people and processes – to provide facts and metrics vital to the processes and useful to the people who carry out those processes. Information adds to the collection of knowledge that is available to business people and business processes.

KNOWLEDGE

Knowledge is a personal and individual thing. Here we leave the realm of what computers and software do, and enter the domain of what people do. Knowledge encompasses the familiarity, awareness, understanding, and perceptions of a person about a given subject. Knowledge is gained through many channels including study, recall, experience, instinct, and beliefs. These factors are different for each person, thus the knowledge of every individual is unique

ACTIONS AND OUTCOMES

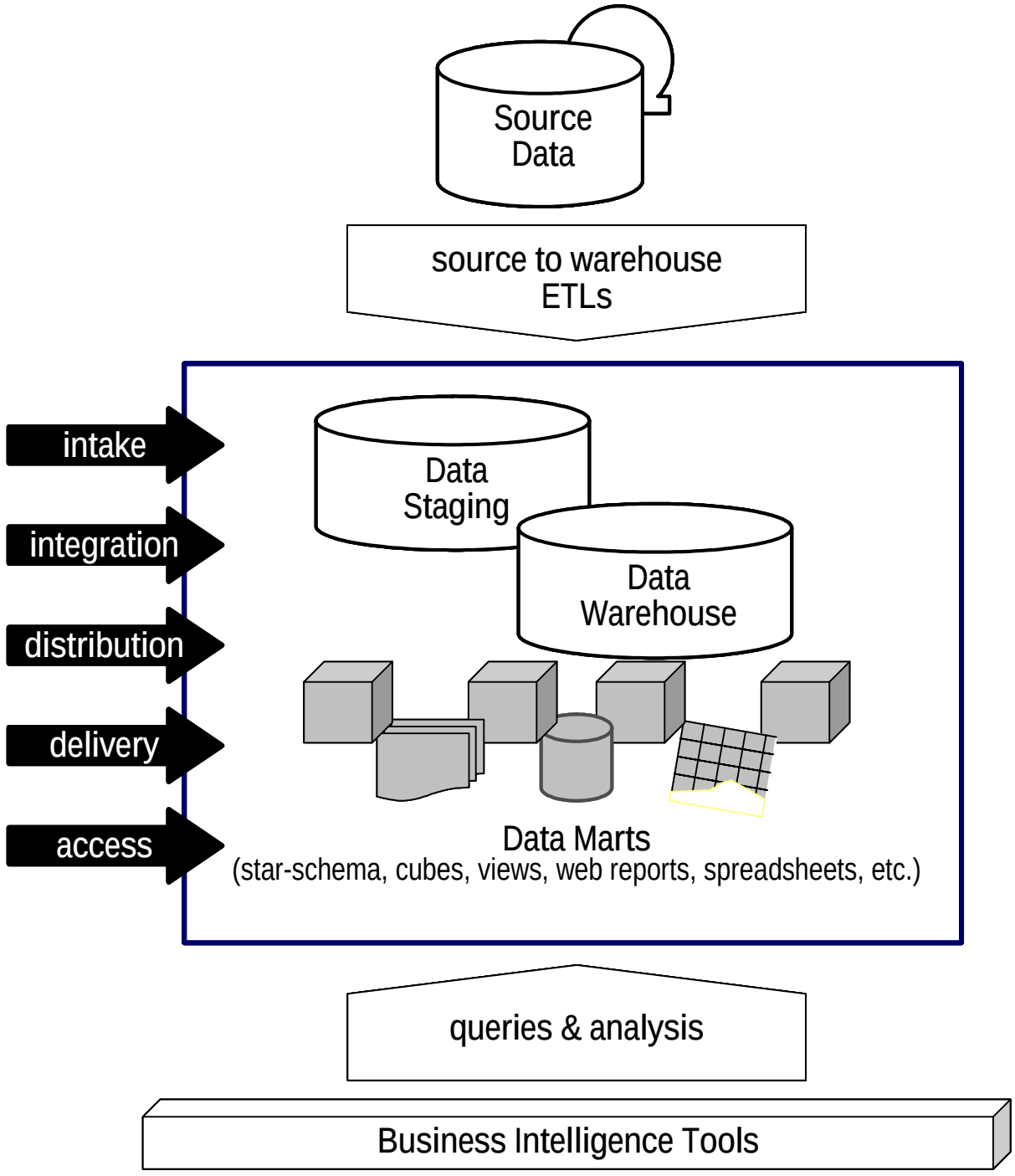
Action is a process of doing something. Effective action is the process of doing the right thing. It is described as a process because we need to look beyond the event of doing and consider the activities and behaviors that lead to that event. Any combination of insight, resolve, decision, and innovation may drive a person to act – the “doing” part of action. Outcomes are the results of actions. Favorable business outcomes are generally those that reduce cost, save time, optimize resources, increase revenue, satisfy customers, or otherwise help to fulfill the business mission and goals.

IMPACT AND VALUE

Value is realized at the bottom line of the business – when outcomes reduce cost or increase revenue either directly or indirectly. The value of an action is determined by the outcomes produced. The value of information is derived through contribution to valued action – providing support for insight, resolve, decision, and innovation. The value of the data warehouse depends entirely on the value of the information services that it delivers.

Warehousing Data Stores

Data Store Responsibilities



Warehousing Data Stores

Data Store Responsibilities

THE ROLES Every data warehousing environment, regardless of architecture and flow of data, must provide for five roles to be complete. Different architectures assign these roles to data stores in various ways.

INTAKE Data stores with intake responsibility receive data into warehousing environment. Data is acquired from multiple source systems, of varying technologies, at different frequencies, and into numerous warehousing files and/or tables. Further, the data typically requires many and diverse transformations. Most data is extracted from operational systems whose data is most certainly not all clean, error-free and complete. Data cleansing is commonly performed as part of the intake process to ensure completeness and correctness of data.

INTEGRATION Integration describes how the data fits together. The challenge for the warehousing architect is to design and implement consistent and interconnected data that provides readily accessible, meaningful business information. Integration occurs at many levels – "the key level, the attribute level, the definition level, the structural level, and so forth …" (*Data Warehouse Types*, www.billinmon.com) Additional data cleansing processes, beyond those performed at intake, may be required to achieve desired levels of data integration.

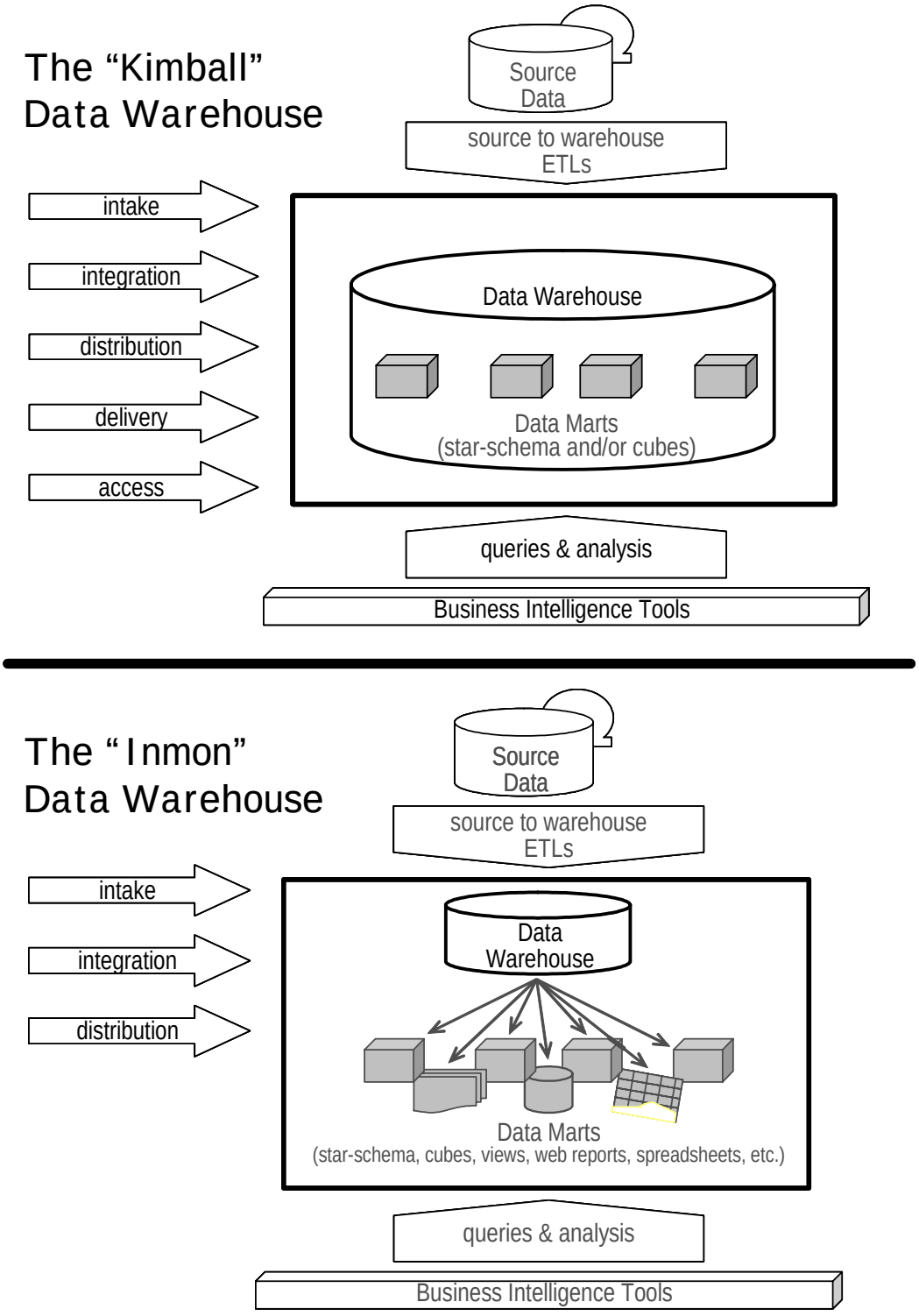
DISTRIBUTION Data stores with distribution responsibility serve as long-term information assets with broad scope. Distribution is the progression of consistent data from such a data store to those data stores designed to address specific business needs for decision support and analysis.

DELIVERY Data stores with delivery responsibility combine data as "in business context" information structures to present to business units who need it. Delivery is facilitated by a host of technologies and related tools - data marts, data views, multidimensional cubes, web reports, spreadsheets, queries, etc.

ACCESS Data stores with access responsibility are those that provide business retrieval of integrated data – typically the targets of a distribution process. Access-optimized data stores are biased toward easy of understanding and navigation by business users.

Warehousing Data Stores

The Data Warehouse



Warehousing Data Stores

The Data Warehouse

CENTRAL DATA WAREHOUSE (HUB)

As previously discussed, Inmon defines a data warehouse "a subject-oriented, integrated, non-volatile, time-variant, collection of data organized to support management needs." (W. H. Inmon, Database Newsletter, July/August 1992) The intent of this definition is that the data warehouse serves as a single-source hub of integrated data upon which all downstream data stores are dependent. The Inmon data warehouse has roles of intake, integration, and distribution.

KIMBALL'S DEFINITION (BUS)

Kimball defines the warehouse as "nothing more than the union of all the constituent data marts." (Ralph Kimball, et. al, The Data Warehouse Life Cycle Toolkit, Wiley Computer Publishing, 1998) This definition contradicts the concept of the data warehouse as a single-source hub. The Kimball data warehouse assumes all data store roles -- intake, integration, distribution, access, and delivery

DIFFERENCES IN PRACTICE

Given these two predominant definitions of the data warehouse - Inmon's (hub-and-spoke architecture) and Kimball's (bus architecture), what are the implications with regard to the five roles of a data store – intake, integration, distribution, access and delivery?

	Inmon Warehouse	Kimball Warehouse
intake	fills the intake role, but may be downstream from staging area	Fills the intake role – downstream from "backroom" transient staging
integration	Primary integrated data store with data at the atomic level	Integration through standards and conformity of data marts
distribution	Designed and optimized for distribution to data marts	Distribution is insignificant because data marts are a subset of the data warehouse
access	May provide limited data access to some "power" users	Specifically designed for business access and analysis
delivery	Not designed or intended for delivery	Supports delivery of information to the business

Data Warehousing Deliverables

Results of Architecture, Implementation & Operation Activities

Architecture
- data warehousing program charter
- data warehousing readiness assessment
- defined business architecture
- defined data architecture
- defined technology architecture
- defined project architecture
- defined organizational architecture

Implementation
- project plans
- target data models
- data warehousing process models
- deployed technology
- warehousing databases
- data acquisition processes
- data transformation processes
- data transport & load processes
- populated warehousing databases
- business analysis applications
- delivered data warehousing capabilities

Operation
- business services
- data refresh
- managed platforms
- managed environment
- customer service
- managed quality
- managed infrastructure

Data Warehousing Deliverables

Results of Architecture, Implementation & Operation Activities

ARCHITECTURE RESULTS

Architectural activities establish the standards, conventions, and guidelines that ensure consistency and integration among results of multiple implementation projects. Architectural work begins by defining a warehousing program and assessing organizational readiness. Architecture is broad in scope and focused on analysis and design in the following areas:

- Business Architecture – Understanding of business goals, drivers, and information needs.
- Data Architecture – Understanding of source data. Requirements and standards for warehousing data and warehouse metadata.
- Technology Architecture – Identification of standards for hardware, software, and communications technology. Specification of the data warehousing toolset.
- Project Architecture – Incremental development plan for the data warehouse. Defined scope of each increment. Sequence and dependencies among increments.
- Organizational Architecture – Identification of training, support, and communications responsibilities.

IMPLEMENTATION RESULTS

Where architecture is broad in scope, implementation narrows the scope to that of a single increment. Each increment is defined as a project that focuses on design, construction, and deployment of warehousing products including:

- Warehousing Databases – Data models and implemented databases for staging data, data warehouse, and data marts.
- Warehousing Processes – Source –to-target mapping, specification of data transformation rules, and development of processes to move data through the warehousing environment.
- Business Analysis Applications – Standard queries, decision support systems (DSS), warehouse published reports, and other standard means of receiving information from the data warehouse.

OPERATION RESULTS

Operation is the phase where data warehousing delivers value. That value is realized through business services that provide data and information and enable confident decisions and positive actions. Training, support, and administration are also key elements of data warehouse operation.

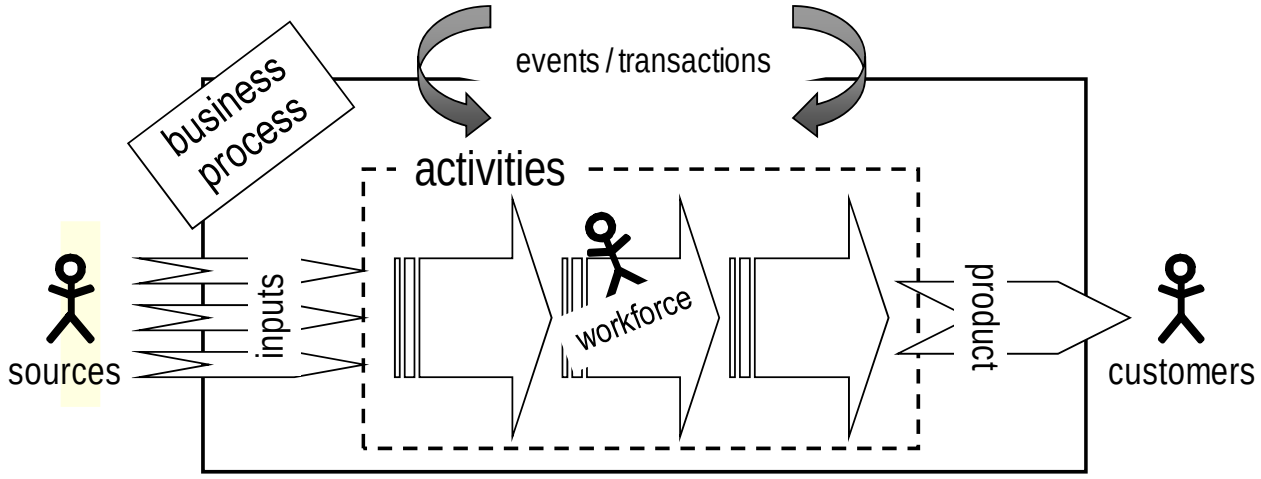
Module 2

Data Warehousing Architecture

This page intentionally left blank.

Business Architecture

Business Processes



- which processes are in scope of the warehousing program?
- who (customer, source, workforce) needs information?
- which business process components are information subjects?

- how can inputs be optimized?
- how can activities be streamlined?
- how can the workforce contribute?
- how can suppliers contribute?
- how can events be managed?
- how can product value be enhanced?

Business Architecture

Business Processes

UNDERSTANDING BUSINESS PROCESSES

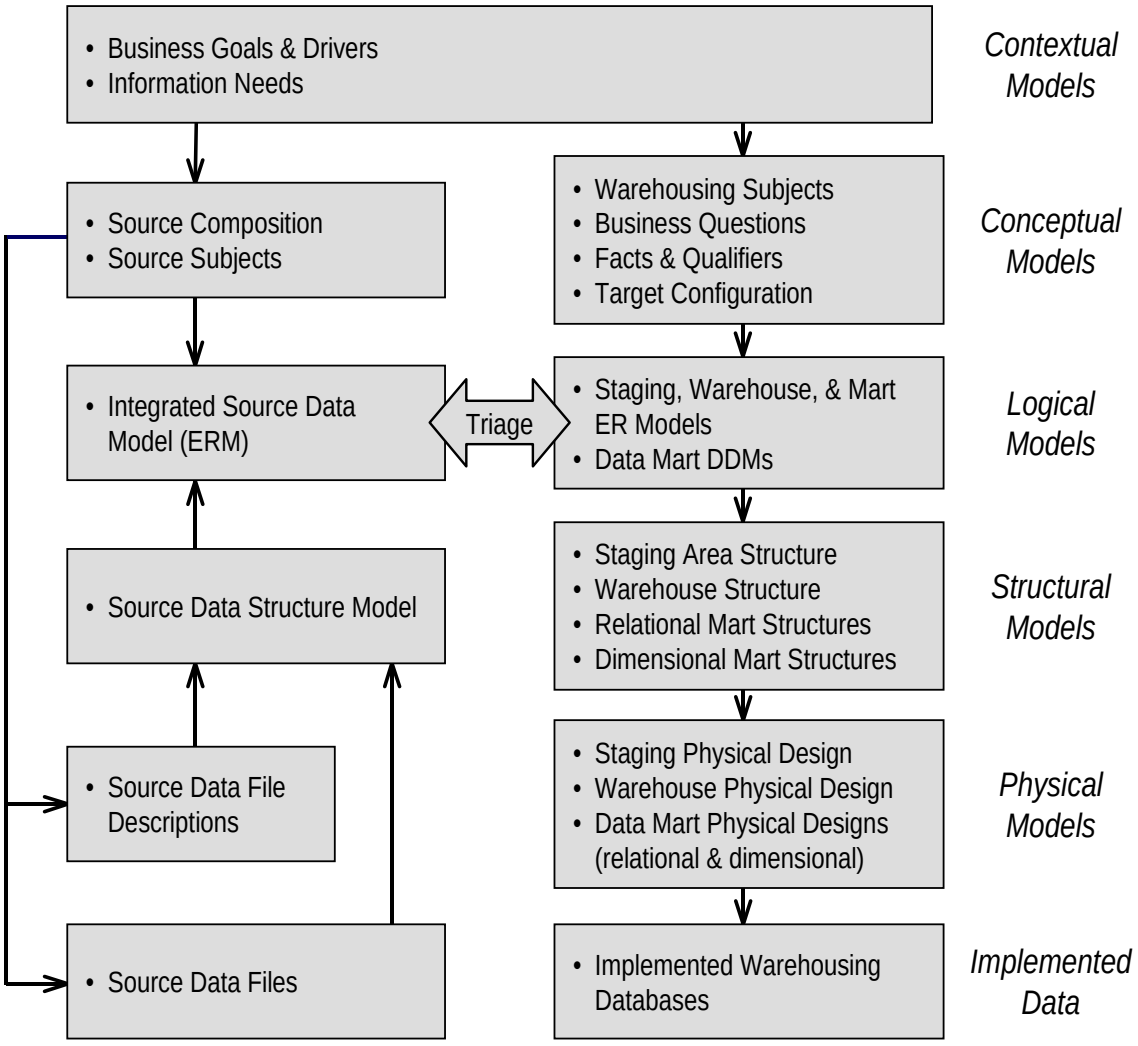
Business processes are the things that a business does to produce its products, deliver its services, manage its infrastructure, etc. Every business process can be understood in terms of the components of that process:

- the **product** that the process produces,
- the **customer** who uses the product,
- the **inputs** that are needed to produce the product,
- the **sources/suppliers** that provide the inputs,
- the **activities** that comprise the process,
- the **actors** who perform the activities,
- the **events** that drive the activities.

Recognizing which processes will be information-enabled through data warehousing, and which process components will become subjects of warehousing data, offers valuable input to all phases of data warehouse planning, development, and operation.

Data Architecture

Data Modeling Concepts



Data Architecture

Data Modeling Concepts

FAMILIAR DATA MODELING PRINCIPLES

Like application data modeling, warehouse modeling works well when practiced at multiple levels of abstraction. Modeling either application or warehouse data may develop any or all of:

- Contextual Models describing the scope of requirements, establishing a context for analysis.
- Conceptual Models describing requirements without consideration for computer implementation.
- Logical Models describing data from a computer system perspective, yet free of any implementation platform specifics.
- Structural Models specifying data structures that account for variables of access, navigation, security, distribution, and time-variance.
- Physical Models providing detailed design and specification of data structures to be implemented using a particular technology.

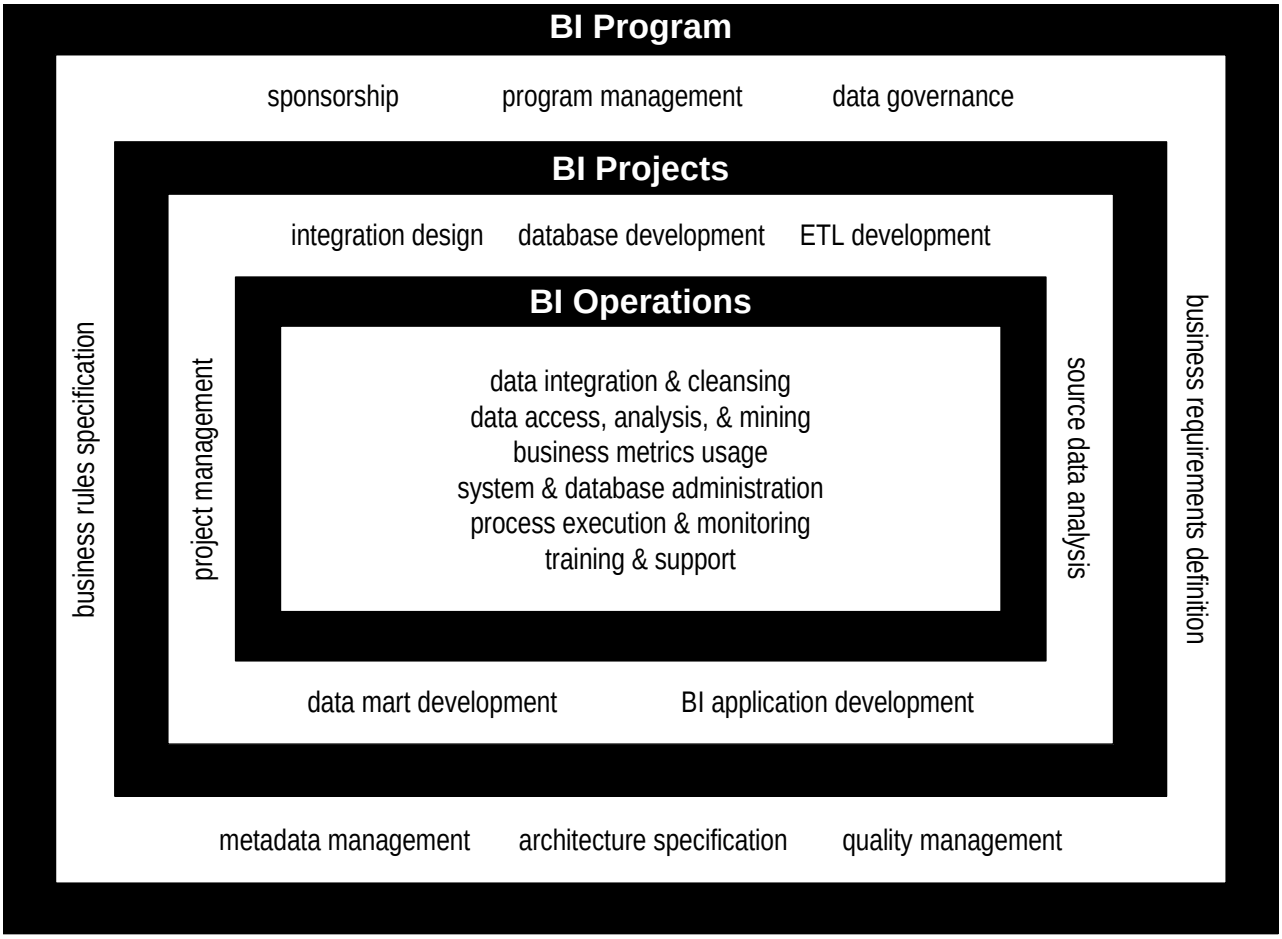
WAREHOUSE MODELING DIFFERENCES

Even the most experienced application data modelers are challenged by early warehouse modeling experiences. New issues, terminology, and techniques combine to make warehouse data modeling more complex than application data modeling. The primary differences include:

This Facet of Warehouse Modeling ...	Differs from Application Modeling in This Way ...
Multiple Data Types	Both source data and warehousing data need to be modeled. Each is modeled separately, and they are associated through a technique called “triage.”
Multiple Ways to Use Warehouse Data	Warehouse data uses range from publishing and managed query to complex OLAP applications and data mining. The ideal data structure depends on planned uses of the data.
Multiple Ways to Organize the Data	Warehouse databases may be organized relationally, dimensionally, or with a combination of the two techniques. The ideal organization depends on both the planned uses of the data and the characteristics of the data.
Multiple Modeling Techniques	The complexities of warehouse data modeling require that many modeling techniques be used. Matrix models, E/R models, subject models, dimensional models, star-schema, and snowflake-schema are all used to meet various data modeling needs.
Planned and Managed Redundancy	Redundancy, typically avoided in application databases, is an asset to warehouse databases. Planning and managing redundancy is a key skill for warehouse data modelers.
Large Data Volumes	Redundancy and time-variance combine to make a very large database (VLDB) a common warehouse consideration. Optimizing for data volumes and database size is a common requirement of warehouse modeling.

Organizational Architecture

Program, Project & Operations Roles



Organizational Architecture

Program, Project & Operations Roles

ROLES AND RESPONSIBILITIES

The program, project, and operation activities of data warehousing are different from those of developing and supporting operational systems. The work is different; therefore the roles and responsibilities are different. Data warehousing has different goals and challenges. It demands different kinds of organizations and teams. Common data warehousing roles and responsibilities include:

BI Program Roles & Responsibilities	
Program Management	Managing business/IT relationship, multiple dependent projects, issue resolution, etc.
Sponsorship	Advocacy, political will, resource acquisition, issue resolution, expectation setting, etc.
Data Governance	Data definitions, business rules alignment, data quality management, access authorization, etc.
Business Rules Specification	Business basis for data rules about content, relationships, correctness, integrity, etc.
Business Requirements Definition	Requirements for data & information, service levels, quality & reliability, etc.
Architecture Specification	Frameworks & standards for business alignment, data, technology, projects, etc.
Quality Management	Beyond data quality – quality of information, delivery, interface, reporting, services, etc.
Meta Data Management	Meta data strategy, meta data implementation, meta data content, etc.
BI Project Roles & Responsibilities	
Project Management	Work breakdown, scheduling, resource allocation, deliverables, deployment, etc.
Integration Design	Data source selection, source/target mapping, transformation rules, populating databases
Database Development	Logical and physical database design, database specification and creation
ETL Development	Analysis, design, construction, and deployment of data movement processes
Source Data Analysis	Data profiling, source content analysis, source data modeling
Data Mart Development	Analysis, design, construction, and deployment of data marts
BI Application Development	Analysis, design, construction, and deployment of information services & analytic applications
BI Operations Roles	
Data Integration & Cleansing	Maintenance and support of data migration processes; Continuous data quality management
Data Access, Analysis, & Mining	Access and application of data to make business decisions
Business Metrics Usage	Application of business measures to drive business actions
System & Database Administration	Installation, configuration, and management of BI operating platforms
Process Execution & Monitoring	Scheduling, execution, verification, and support of data warehousing processes
Training & Support	Customer care activities for BI customers

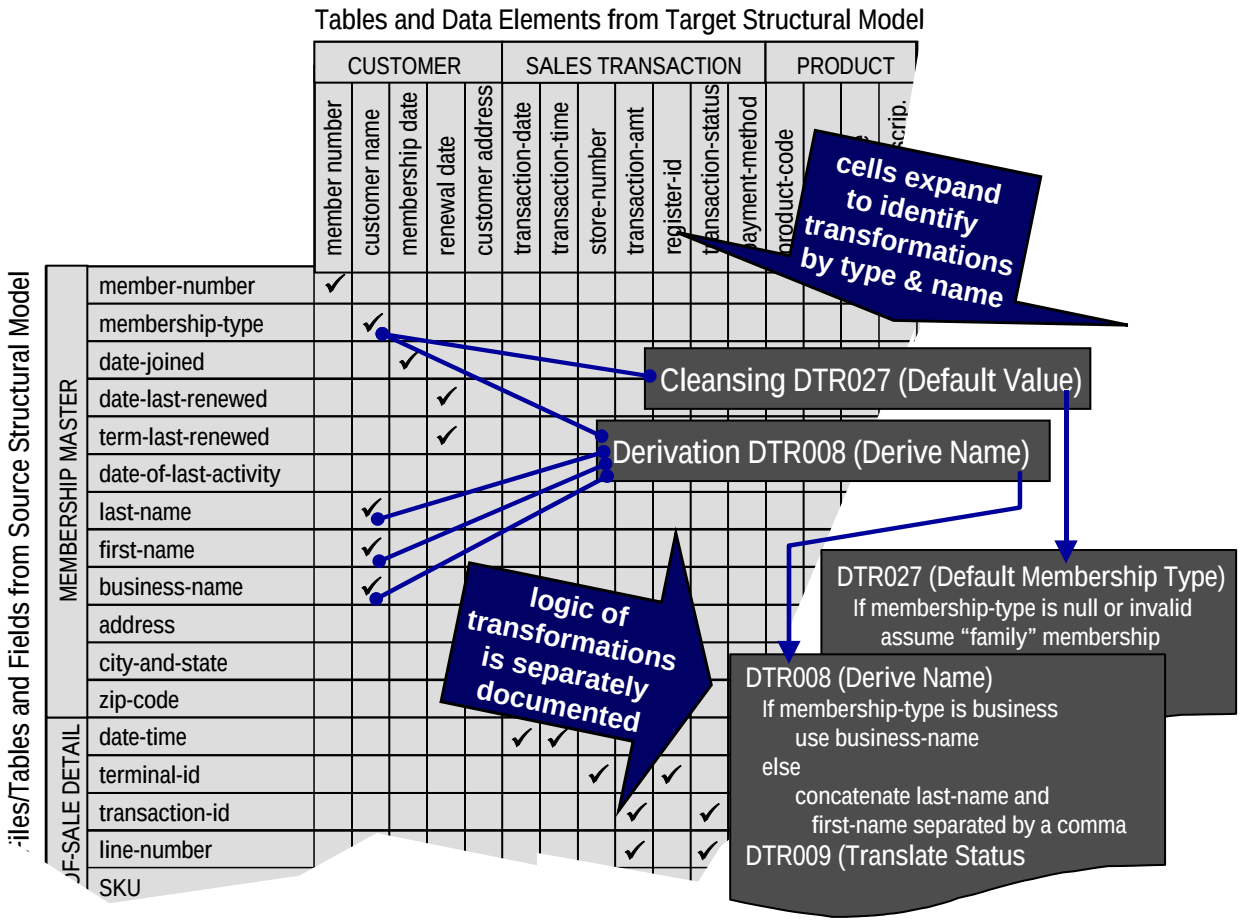
Module 3

Data Warehouse Implementation

This page intentionally left blank.

The Warehouse Process Model

Data Transformation Rules



The Warehouse Process Model

Data Transformation Rules

DETAILED SPECIFICATION

Specification of data transformations develops a large set of details about how source data is to be processed prior to loading of a warehousing database. Documenting data transformation must address both the *identification* of what transformations are needed, and the *logic* of the transformation process.

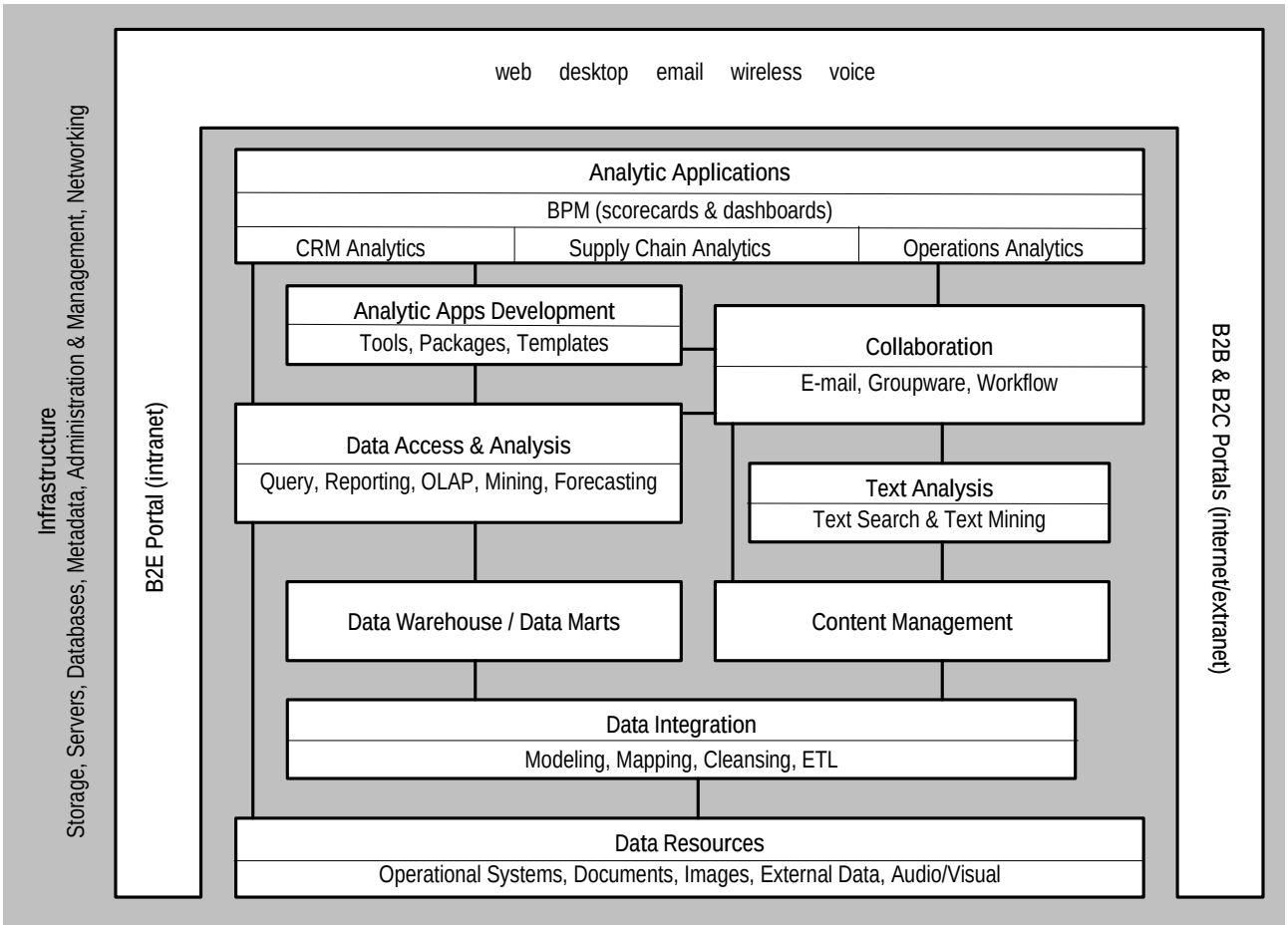
Documenting which transformations occur can readily be achieved by extending the source/target maps. View the set of logic for each transformation as a unique rule, and develop a convention for naming these rules. As each transformation need is identified, assign a name and place that name in the appropriate cell of a source/target map. Then document the logic of each transformation rule. For each source/target association consider possible rules for each of the transformation types.

In addition, consider need for data cleansing. Although data clean-up is not a unique transformation rule type, it is a common reason for filtering, conversion, and derivation.

Transform Need	Description
Specify Selection Requirements	Identifies and describes the selection processes needed to choose among multiple sources. The objective is to select the best data to be used for warehouse population. Selection requirements may exist at both data store and date element levels.
Specify Filtering Requirements	Identifies and describes the filtering processes needed to choose records from a source file (or rows from a source table) to be used for data warehouse population.
Specify Conversion and Translation Requirements	Identifies and describes the conversion and translation processes which change the formats and values of data elements. Conversion processing achieves consistency of formats and value sets among data extracted from multiple sources. Translation processes change data formats and values from encoded and cryptic to descriptive and meaningful.
Specify Derivation and Summarization Requirements	Identifies and describes needed derivation processes used to develop a value for a single data element by applying logic to the values of some other data elements. It also identifies and describes the processes through which summary data values are created.
Specify Clean-up Requirements	Identifies and describes the clean-up processing needed to ensure quality and integrity of the data that is placed into the data warehouse. Clean-up needs may exist at both data record and data element levels. Among the issues of clean-up processing are intra- and inter-record consistency checking, and decisions regarding elements with null values or invalid values.

Deployed Technology

Range and Roles of Technology



Deployed Technology

Range and Roles of Technology

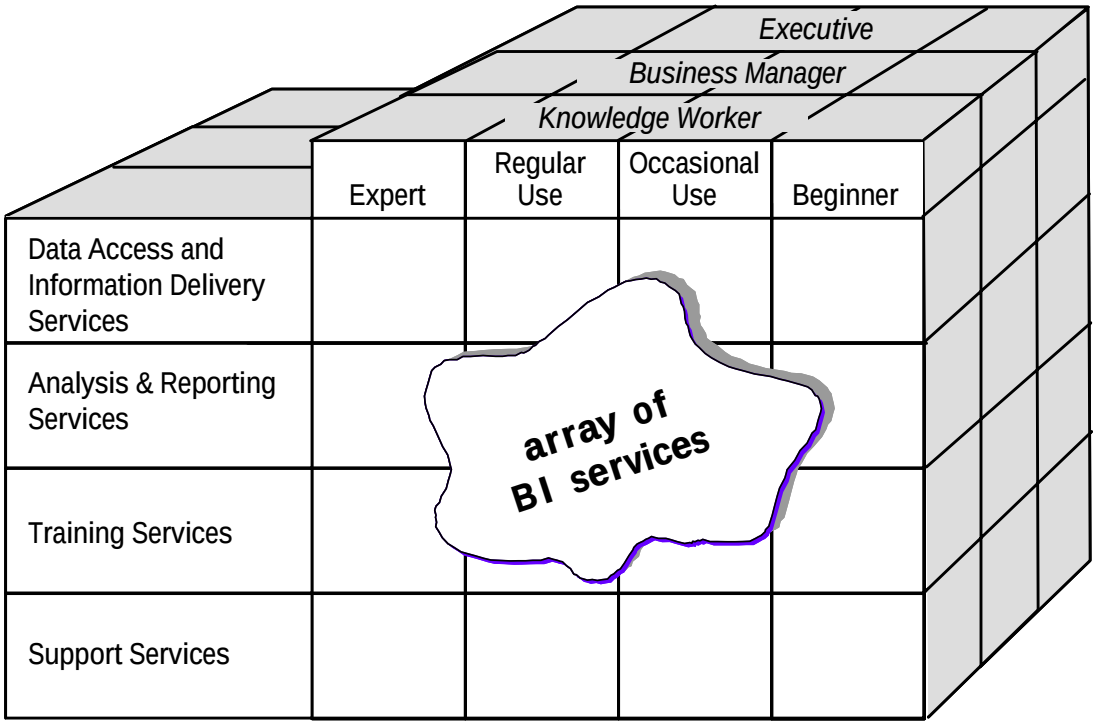
TECHNOLOGY ROLES AND RELATIONSHIPS

The technology framework illustrates classes of tools and technology from infrastructure through information delivery. This framework includes established and mainstream technologies as well as emerging technologies (content management, text analysis, text mining, collaboration, etc.) that are gaining significance in data warehousing. The major technology classes are:

Delivery	Delivery media includes web portals, desktop clients, email, wireless, voice print, pager and fax. Delivery technology sets include (1) B2E Portal – intranet business-to-enterprise delivery to the workforce, (2) B2B Portal– internet business-to-delivery to vendors, customers, partners and anyone with internet access, (3) B2C Portal – extranet business-to-customer delivery.
Analytic Applications	Analytic applications are the technology components of business applications, ranging from static reporting to dashboards and scorecards. They place information into business function context, i.e. Customer Relationship Management (CRM), Supply Chain Management (SCM), Business Performance Management (BPM), etc.
Analytic Application Development Tools	Tools, templates, and packaged applications to quickly build views, reports, dashboards, scorecards, and other applications to deliver information in context of a business function or business process.
Collaboration	Web applications to support employees, partners, customers, vendors and others to collaborate on documents, share business metrics, manage content, and work collectively. While reporting is still dominant today, collaboration capabilities will grow as the technology and market place mature.
Data Access & Analysis	Data access and analysis tools are today's most common delivery technologies. Unlike analytic applications, these tools focus on data before information, and they provide less business context than analytic applications. The most widely-used tools include managed reporting, query, and OLAP.
Text Analysis	Text analysis tools use semantics and statistical techniques to identify, tag, and select relevant content from text documents. Parsing, pattern recognition, natural language processing and other advanced techniques are used to transform unstructured text into data and/or information structures.
Data Warehouse / Marts	Data warehouses and data marts integrate and reconcile data from multiple data sources. Their purpose is to prepare data to serve as the raw material from which information is created. Regardless of the multiple definitions of *data warehouse* and *data mart* that are used in the industry, all warehouses and marts exist primarily to serve this purpose.
Content Management	Content management technology first emerged as an internet technology – to support management of content-rich web sites. Uses of the technology in BI are emerging as the industry evolves from data warehousing to business intelligence, and from integration of structured data integrating all types of business information resources. Basic content management functions include indexing, searching, and retrieval.
Data Resources	This class includes all sources from which data can be acquired. When both internal and external data are considered, and when both structured and unstructured data are included, the range of possible source technologies becomes exceptionally broad.
Infrastructure	This technology class describes the underlying hardware, software, networking, administration and support structures upon which systems and data sources are constructed and operated.

Delivery Results

Data and Information Services



The right kinds of services matched to the customer's roles, responsibilities, and experience level

Delivery Results

Data and Information Services

MEETING CUSTOMER NEEDS

A mature data warehousing environment includes a robust set of services that support the goal of delivering the right services to the right people at the right time. A three-dimensional view of the services array is useful to classify services and to assess customer needs and match them with available services. The services dimensions are:

- Classification of customers as
  - *Knowledge workers* who carry out the day-to-day activities of the business
  - *Managers* responsible for performance of individual business processes
  - *Executives* responsible for business performance across many business processes

- Classification of customer experience as
  - *Experts* who use the data warehouse regularly and have a high level of computer and analytic skills combined with an intimate knowledge of data warehouse content
  - *Regular users* of the data warehouse with moderate computer and analytic skills combined with a working knowledge of data warehouse content
  - *Occasional users* of the data warehouse who may have necessary computer and analytic skills, but have limited knowledge of data warehouse content
  - *Beginners* with little or no knowledge of data warehouse content, and who may have limited computer or analytic skills

- Classification of services as
  - *Data access and information delivery* services that make data and information available to the business.
  - *Analysis and reporting* services that deliver analytic applications of greater complexity than simple data access and information delivery.
  - *Training services* that develop customer skill and ability to use the data warehouse, with a goal of making each customer self-sufficient.
  - *Support services* that augment the services culture, enhance communications with customers, and ensure rapid resolution of problems.

Module 4

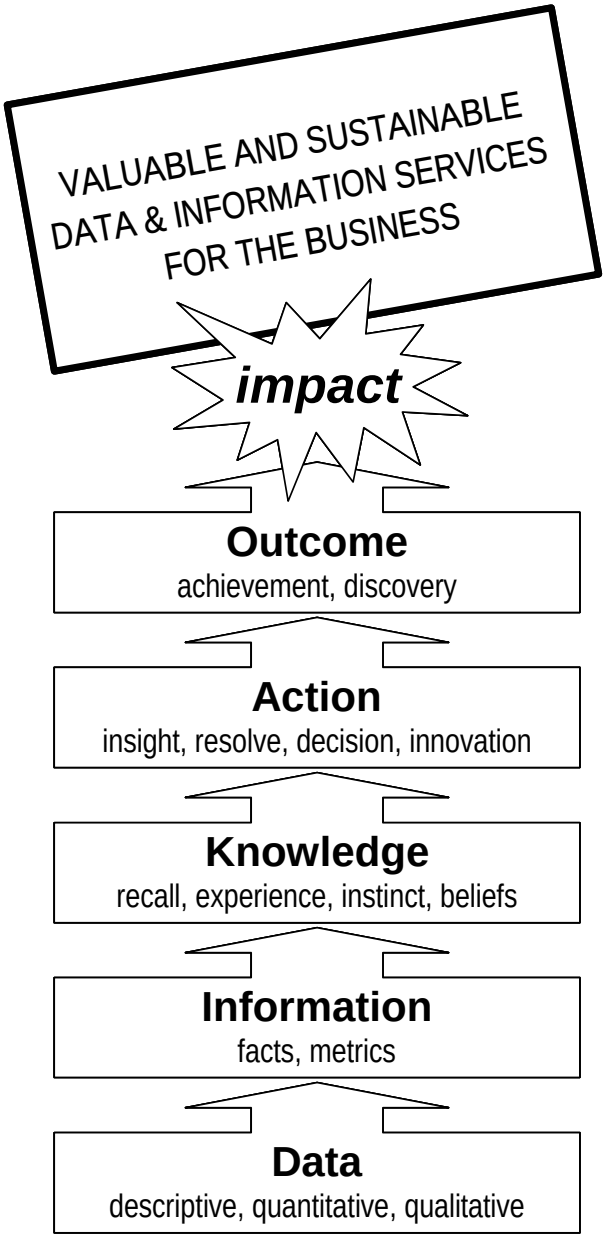
Data Warehouse Operation

This page intentionally left blank.

Business Services

Valuable and Sustainable Services

Operation
- business services
- data refresh
- managed platforms
- managed environment
- customer service
- managed quality
- managed infrastructure



Business Services

Valuable and Sustainable Services

WAREHOUSING FOR THE LONG TERM

Sustaining the data warehouse demands a commitment to delivering reliable and valuable business services in an environment of high-frequency change. Value is sustained by ensuring continuous alignment with changing business needs and with a changing customer base.

Reliability is sustained by attention to all of the "under the hood" components upon which the services depend including:

- Regular, routine, and dependable data refresh despite changing data sources and systems.
- Effectively managed technology platforms from data acquisition to information delivery in a climate of rapid technological change.
- Managed environment including security, growth, capacity planning, and configuration management.
- Customer service including support, help desk, and training services.
- Continuous quality management for all aspects of quality – business quality, data and information quality, and technical quality.
- Actively managed infrastructure that ensures continued alignment of people, processes, and technology for optimum business value.

Managed Quality

Dimensions of Quality

Managed Quality

Dimensions of Quality

QUALITY IMPROVEMENT

Quality, as with any other aspect of business, is effectively managed with measures and metrics. A metrics foundation for quality management includes both measures of product quality and measures of the process that produces the product. In the case of business intelligence, the products are BI results – information delivered to the business, analytics used by the business, actions and outcomes enabled through BI, etc. The processes are those necessary to execute the entire chain of events from data warehousing to business action, and to sustain a BI program over time. Product measures are used to detect defects in BI products and to improve those products. Process measures help to identify causes of defects and prevent reoccurrence through process improvement. A mature quality process regularly adjusts quality targets to achieve continuous improvement.

DIMENSIONS OF QUALITY

Business intelligence quality is much more than simple data quality. Data quality is, in fact, a relatively small and easy piece of the overall quality domain. BI quality is measured and managed in three major categories:

- **Business Quality** directly affects the business value derived from BI, and the economic success of the BI program.
- **Information Quality** is related to acceptance and use of BI products – the extent to which BI customers value those products. Information quality is a significant factor in political success of BI.
- **Technical Quality** involves choosing the right technologies, configuring multiple technologies to work well together, and using the right tools for the right job. High-quality implementation of technology is typically unnoticed by the business. Low-quality, however, is highly visible and directly affects overall acceptance, usage, trust, value realization, and sustainability of a BI program.

Managed Infrastructure

Processes, Technology, and People

Managed Infrastructure

Processes, Technology, and People

COMPLEX INFRASTRUCTURE

Infrastructure is the foundation upon which BI operates and grows. While the infrastructure supports development, it's more critical role is in operating and sustaining BI solutions. Operation and sustenance are both more demanding and of longer duration than development. An effective BI infrastructure is one in which processes, technology, and people work seamlessly to support a BI culture and to realize business value from BI solutions.

PROCESS

This course has already discussed the analytics processes of BI. When successful, BI becomes a key component in decision making processes. It depends, however, on many other processes to achieve this level of success. The process components of BI infrastructure are program management, change management, data governance, development methodology, project management, data warehouse administration, and metadata management.

TECHNOLOGY

While technology can't create BI, neither can BI be created without use of technology. Blending the right technologies with the process and people components of BI is a key to success. Technology infrastructure includes data warehousing tools, BI tools, and enabling/infrastructure hardware and software.

PEOPLE

People are integral to effective BI. Neither processes nor technology can deliver value independently of the knowledge, decisions, and actions of people. Human infrastructure is arguably the single most important of all BI infrastructure categories. Identifying the right set of roles and responsibilities, assigning them to people with the right skills, and constructing the right kinds of organizations and relationships are all critical to BI success.

Module 5

Summary and Conclusions

Common Mistakes

From TDWI's *10 Mistakes* Series

	An effective project manager will not …
1	Accept an unrealistic schedule.
2	Take on a failing project.
3	Launch a project with a dysfunctional team.
4	Choose the wrong sponsor.
5	Accept unrealistic expectations.
6	Expand the project scope.
7	Skip the project plan.
8	Fail to put the project agreement in writing.
9	Let IT drive the project.
10	Give others authority to select software.
11	Market the project alone.

	Effective team-builders will avoid …
1	Hiring yourself.
2	Squelching disagreement.
3	Confusing titles with roles and responsibilities.
4	"Talking the walk."
5	Thinking one size fits all.
6	Pointing fingers.
7	Interviewing only for technical skills.
8	Limiting leadership.
9	Becoming too task focused.
10	Believing that all decisions are created equal.

	An effective data modeler will avoid …
1	Not gathering business requirements.
2	Saving time by not creating a subject area model.
3	Delivering normalized tables to drive data mart design.
4	Designing the staging process for ease of developers at end-user expense.
5	De-normalizing without starting from a fully normalized data model.
6	Allowing users to drive the level of detail.
7	Not modeling all levels of a multi-tiered warehousing environment.
8	Developing a data model from a list of required data elements.
9	Believing you must choose between relational and dimensional models.
10	Jumping straight into data mart design.

