CP5806
WEEK 2

Presented by Trevor (trevor.andersen@jcu.edu.au)

DUE DATES

- Assignment 1 due Sunday 26 July, 11:59 pm

JULY							
	M	T	W	T	F	S	S
O Week			1	2	3	4	5
wk 1	6	7	8	9	10	11	12
wk 2	13	14	15	16	17	18	19
wk 3	20	21	22	23	24	25	26
wk 4	27	28	29	30	31		

AUGUST							
	M	T	W	T	F	S	S
wk 4						1	2
wk 5	3	4	5	6	7	8	9
wk 6	10	11	12	13	14	15	16
wk 7	17	18	19	20	21	22	23
O Week	24	25	26	27	28	29	30
wk 1	31						

JAMES COOK UNIVERSITY AUSTRALIA

WEEK 2 LEARNING OUTCOMES

- Understand the need for strategic information and DW
- Distinguish between database and DW
- Distinguish between DW and data marts
- Examine each building block of DW
- Examine DW architecture
- Understand the significant role of metadata in DW

JAMES COOK UNIVERSITY AUSTRALIA

TOPICS FOR WEEK 2

- Topic 1: The background to DW
- Topic 2: DW building blocks
- Topic 3: DW planning and project management
- Topic 4: DW architecture and infrastructure
- Topic 5: Metadata in DW

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 1: THE BACKGROUND TO DW

- Data warehousing is about **analysing business needs** and building an **integrated, subject-oriented, time-variant and non-volatile** system for the query and analysis for business decisions
- **A database** is an organised collection of data.
- **A data warehouse** is a special type of database designed for analysis of data.
- A typical **database** is used for OnLine Transactional Processing (OLTP) while a general **data warehouse** is used for OnLine Analytical Processing (OLAP).

DW PROVIDES STRATEGIC INFORMATION

INTEGRATED	Must have a single, enterprise-wide view.
DATA INTEGRITY	Information must be accurate and must conform to business rules.
ACCESSIBLE	Easily accessible with intuitive access paths, and responsive for analysis.
CREDIBLE	Every business factor must have one and only one value.
TIMELY	Information must be available within the stipulated time frame.

Figure 1-2 Characteristics of strategic information.

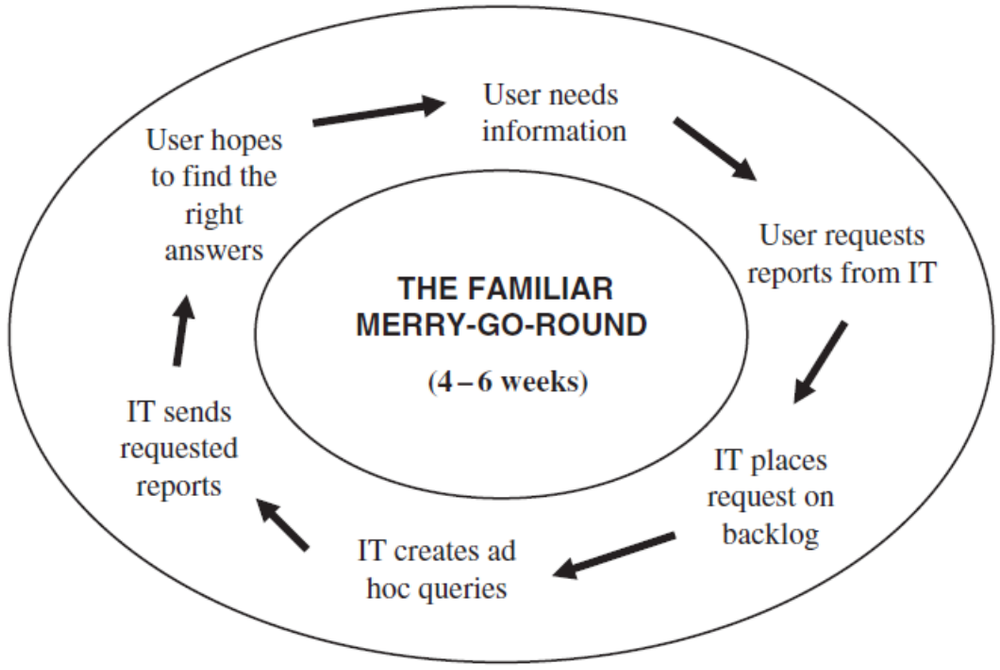


Figure 1-4 Inadequate attempts by IT to provide strategic information.

DW PROVIDES STRATEGIC INFORMATION

Get the data in

Making the wheels of business turn

- Take an order
- Process a claim
- Make a shipment
- Generate an invoice
- Receive cash
- Reserve an airline seat

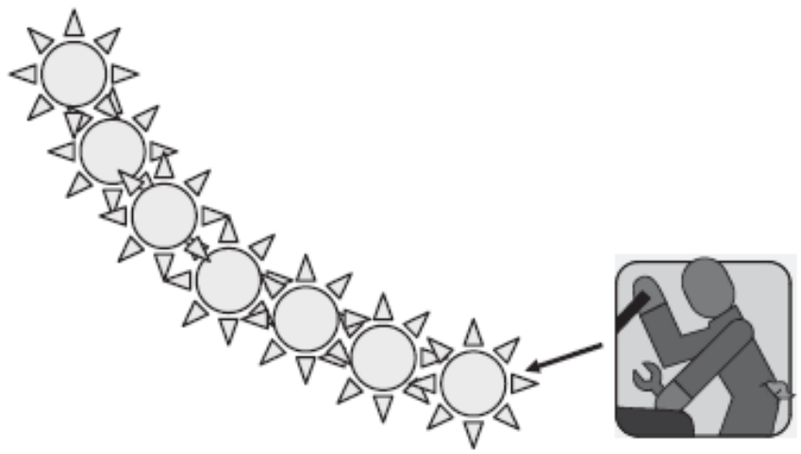


Figure 1-5 Operational systems.

Get the information out

Watching the wheels of business turn

- Show me the top-selling products
- Show me the problem regions
- Tell me why (drill down)
- Let me see other data (drill across)
- Show the highest margins
- Alert me when a district sells below target

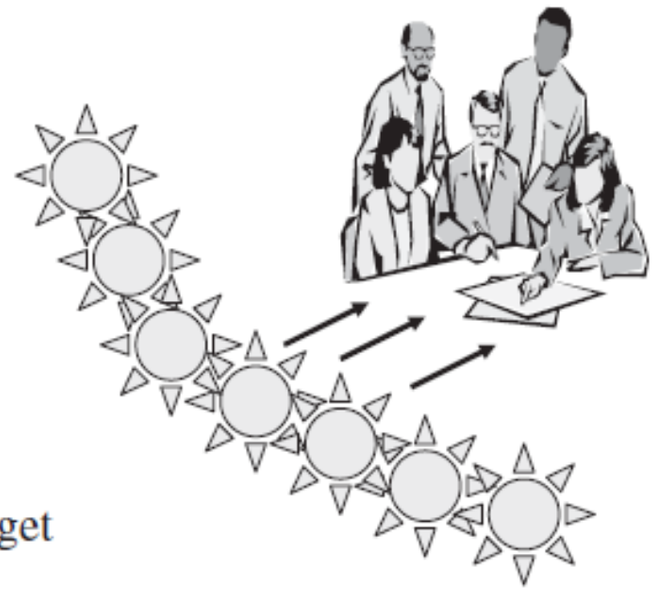


Figure 1-6 Decision-support systems.

DATABASE VS DATA WAREHOUSE

Database	Data warehouse
relational or object oriented data	heterogenous data types
two-dimensional data	multi-dimensional data
tables normalised to achieve efficient storage	tables denormalised to achieve faster querying, summarised
OLTP	OLAP

DBMS VS DATA WAREHOUSING

- Advantages of **DBMS**:
  - compactness of data representation
  - currency of data
  - concurrency control
  - write capability

- Advantages of **data warehousing**
  - effectiveness
  - independent processing
  - robustness
  - efficient query processing

TOPIC 2: DW BUILDING BLOCKS

- Defining features of DW
  - Subject-oriented data
  - Integrated data:
    - Requires standardisation:
      - Naming conventions
      - Codes
      - Data attributes
      - Measurements
  - Time-variant data
      - Allows for analysis of the past
      - Relates information to the present
      - Enables forecasts for the future
  - Nonvolatile data
  - Data granularity
    - refers to the level of detail. Many data warehouses have at least dual levels of granularity.

TOPIC 2: DW BUILDING BLOCKS

- Development of DW:
  - Difference between data warehouse and data marts
  - Top-down approach
  - Bottom-up approach
  - Architectural types:
    - Centralised data warehouse
    - Independent data marts
    - Federated
    - Hub-and-Spoke
    - Data-Mart bus

JAMES COOK UNIVERSITY AUSTRALIA

DATA WAREHOUSE VS DATA MARTS

DATA WAREHOUSE	DATA MART
◆ Corporate/Enterprise-wide	◆ Departmental
◆ Union of all data marts	◆ A single business process
◆ Data received from staging area	◆ STARjoin (facts & dimensions)
◆ Queries on presentation resource	◆ Technology optimal for data access and analysis
◆ Structure for corporate view of data	◆ Structure to suit the departmental view of data
◆ Organized on E-R model	

Figure 2-5 Data warehouse versus data mart.

DESIGN STRATEGY

	Advantages	Disadvantages
Top-down	➤ A truly corporate effort, an enterprise view of data ➤ Inherently architected, not a union of disparate data marts ➤ Single, central storage of data about the content ➤ Centralized rules and control ➤ May see quick results if implemented with iterations	➤ Takes longer to build even with an iterative method ➤ High exposure to risk of failure ➤ Needs high level of cross-functional skills ➤ High outlay without proof of concept
Bottom-up	➤ Faster and easier implementation of manageable pieces ➤ Favorable return on investment and proof of concept ➤ Less risk of failure ➤ Inherently incremental; can schedule important data marts first ➤ Allows project team to learn and grow	➤ Each data mart has its own narrow view of data ➤ Permeates redundant data in every data mart ➤ Perpetuates inconsistent and irreconcilable data ➤ Proliferates unmanageable interfaces

ARCHITECTURAL TYPES

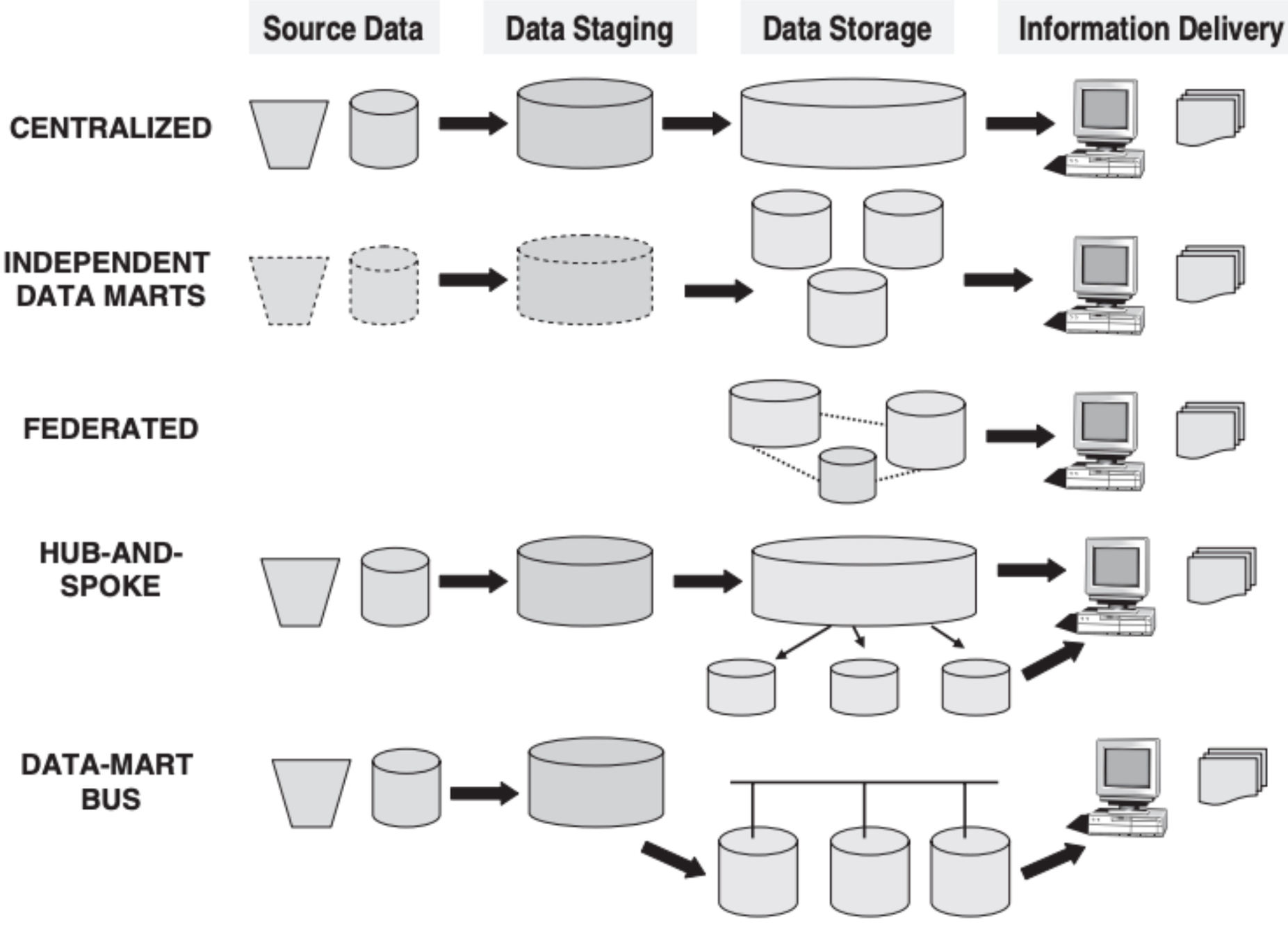


Figure 2-6 Data warehouse architectural types.

TOPIC 2: DW BUILDING BLOCKS

- Overview of building blocks:
  - Source data
    - Production data
    - Internal data
    - Archived data
    - External data
  - Data staging component
    - Data extraction
    - Data transformation
    - Data loading
  - Data storage component
  - Information delivery component
  - Metadata component
  - Management and control component

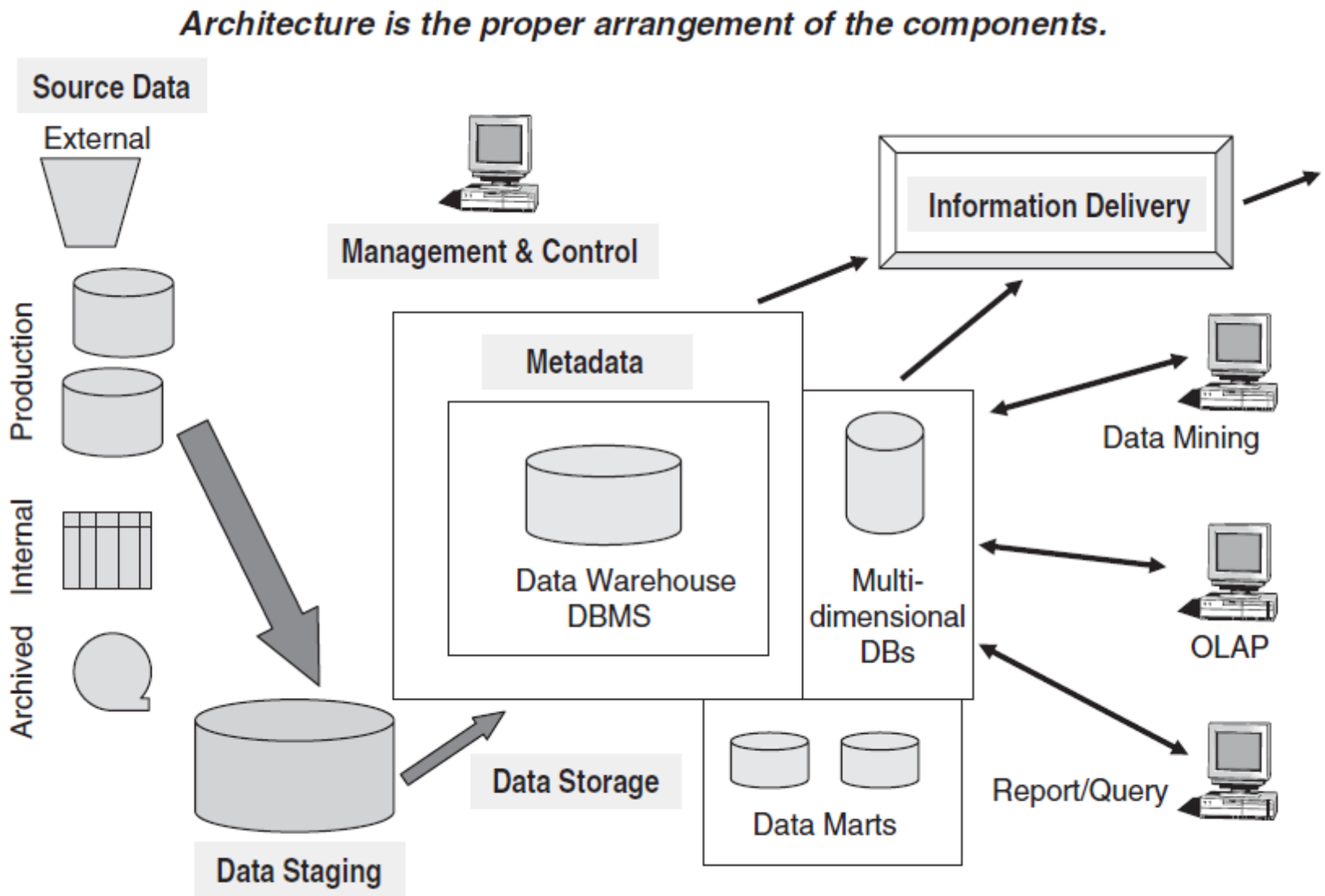


Figure 2-7 Data warehouse: building blocks or components.

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 2: DW BUILDING BLOCKS

Start-up task

Time: Set aside 20 minutes to complete this task.

Scenario: You are a data scientist and analyst on a project team building a data warehouse for CTB Travel, a tourism company. The company has three branches: Cairns, Townsville and Brisbane. The Cairns branch focusses on adventure activities such as bungee jumping, the Townsville branch focusses on water activities such as snorkelling, and the Brisbane branch focusses on city tour activities. The company does not need to have a single version of the truth but specific need.

Task: In the *Start-up task: CBT Travel* discussion forum, discuss your ideas regarding the following:

- What approach will you use in this scenario: top-down or bottom-up? Justify your answer.
- What type of architectural type will you adopt? Why?
- List possible data sources from which you will bring the data into your data warehouse. State your assumptions.

TOPIC 3: DW PLANNING AND PROJECT MANAGEMENT

- Data warehousing has become mainstream with the **spread of high-volume data warehouses** and the **rapid increase in the number of vendor products**.
- To be effective, modern data warehouses need to store multiple types of data: **structured and unstructured**, including documents, images, audio, and video.
- **Data visualisation** deals with **displaying information** in several types of visual forms: text, numerical arrays, spreadsheets, charts, graphs, and so on.
- Data warehouse performance may be improved by using **parallel processing** with appropriate hardware and software options.

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 3: DW PLANNING AND PROJECT MANAGEMENT

- Key issues for planning DW include
  - setting proper expectations
  - assessing risks
  - deciding between top-down or bottom-up approaches
  - choosing from vendor solutions
- A data warehouse project is much different from a typical OLTP system project. Use of **agile development methodology** is to be considered.
- **Participation of the users** is mandatory for success of the data warehouse project.
- Consider the **warning signs and success factors**: in the final analysis, adopt a practical approach to build a successful data warehouse.

TOPIC 3: DW PLANNING AND PROJECT MANAGEMENT

- Business data is **dimensional** in nature and the users of the data warehouse think in terms of business dimensions, such as product, geography, time, and promotion.
- **Information packages**—a new and useful concept—are the backbone of the requirements definition.
- Interviews, group sessions, and questionnaires are standard methods for collecting requirements.
- Scope and content of the requirements definition document include data sources, data transformation, data storage, information delivery, and information package diagrams.

TOPIC 3: DW PLANNING AND PROJECT MANAGEMENT

- Business data is **dimensional** in nature and the users of the data warehouse think in terms of business dimensions, such as product, geography, time, and promotion.

Marketing Vice President

How much did my new product generate
month by month, in the southern division, by customer demographic, by sales office, relative to the previous version, and compared to plan?

Marketing Manager

Give me sales statistics
by products, summarized by product categories, daily, weekly, and monthly, by sale districts, by distribution channels.

Financial Controller

Show me expenses
listing actual vs budget, by months, quarters, and annual, by budget line items, by district, division, summarized for the whole company.

Figure 5-1 Managers think in business dimensions.

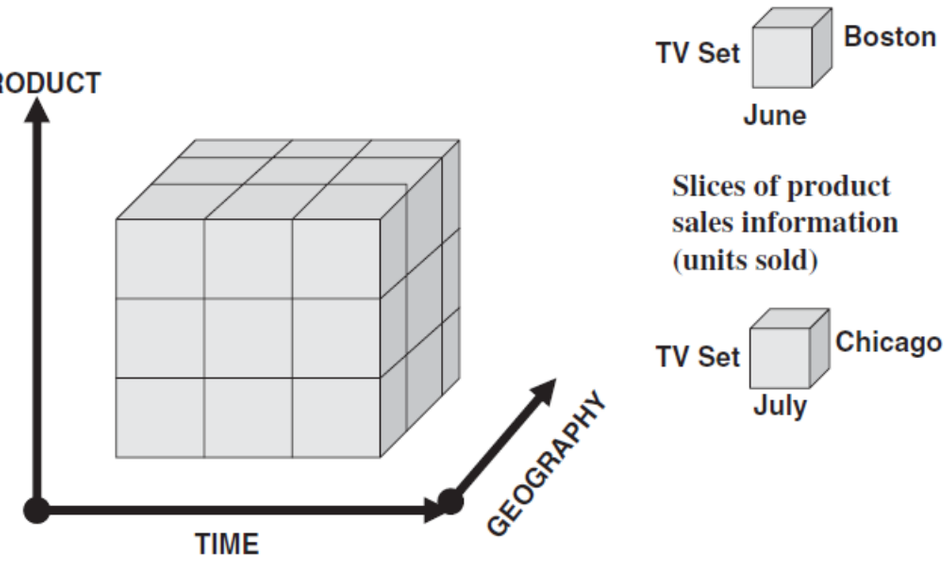


Figure 5-2 Dimensional nature of business data.

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 3: DW PLANNING AND PROJECT MANAGEMENT

- **Information packages**—a new and useful concept—are the backbone of the requirements definition.

Information Subject: Automaker Sales

Dimensions

Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		

Hierarchies/Categories

Facts: Actual Sale Price, MSRP, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance

Figure 5-5 Information package: automaker sales.

TOPIC 3: DW PLANNING AND PROJECT MANAGEMENT

- **Accurate requirements definition** in a data warehouse project is many times more important than in other types of projects.
- **Business requirements** condition the outcome of the data design phase.
- Every component of the data warehouse architecture is strongly influenced by the business requirements.
- In order to provide data quality, identify the **data pollution sources, the prevalent types of quality problems, and the means to eliminate data corruption** early in the requirements definition phase itself.
- Data storage specifications, especially the selection of the DBMS, are determined by business requirements.
- Business requirements strongly influence the information delivery mechanism. Requirements define how, when, and where the users will receive information from the data warehouse.

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

- Data warehouse architecture consists of distinct components with the **read-only data repository** as the centerpiece.
- A few typical data warehouse architectural types are in use at various organizations: **centrally as the single** enterprise data warehouse database or **as a collection cohesive data marts**.
- The architectural components support the functioning of the data warehouse in the three major areas of data acquisition, data storage, and information delivery.
- The architectural framework enables the flow of data from the data sources at one end to the user's desktop at the other.
- The technical architecture of a data warehouse is the complete set of functions and services provided within its components. It includes the procedures and rules needed to perform the functions and to provide the services. It encompasses the data stores needed for each component to provide the services.
- The flow of data from the source systems to end-users as business intelligence depends on the architectural type.

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

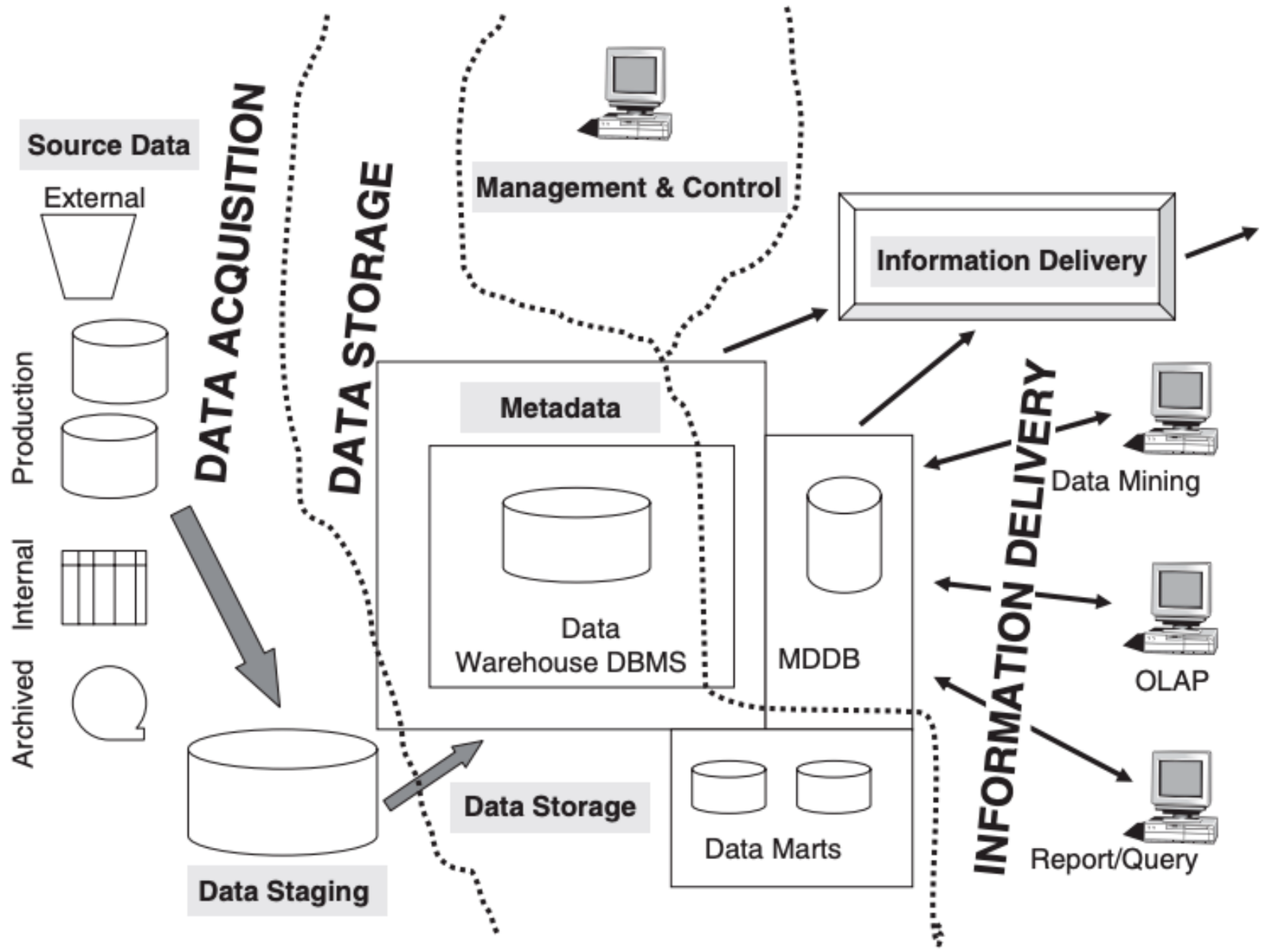


Figure 7-1 Architectural components in the three major areas.

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

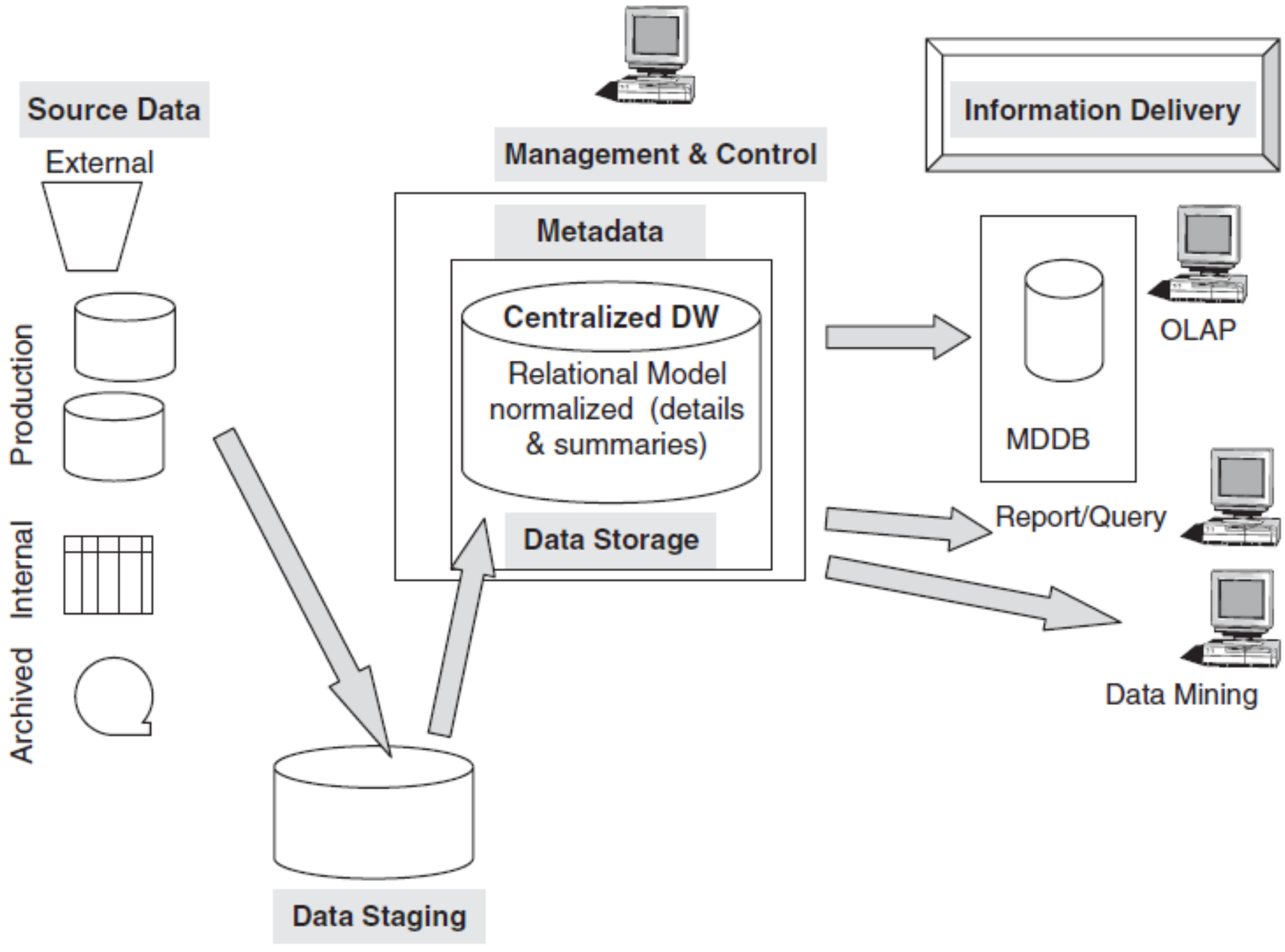


Figure 7-7 Overview of the components of a centralized data warehouse.

Inmon's approach

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

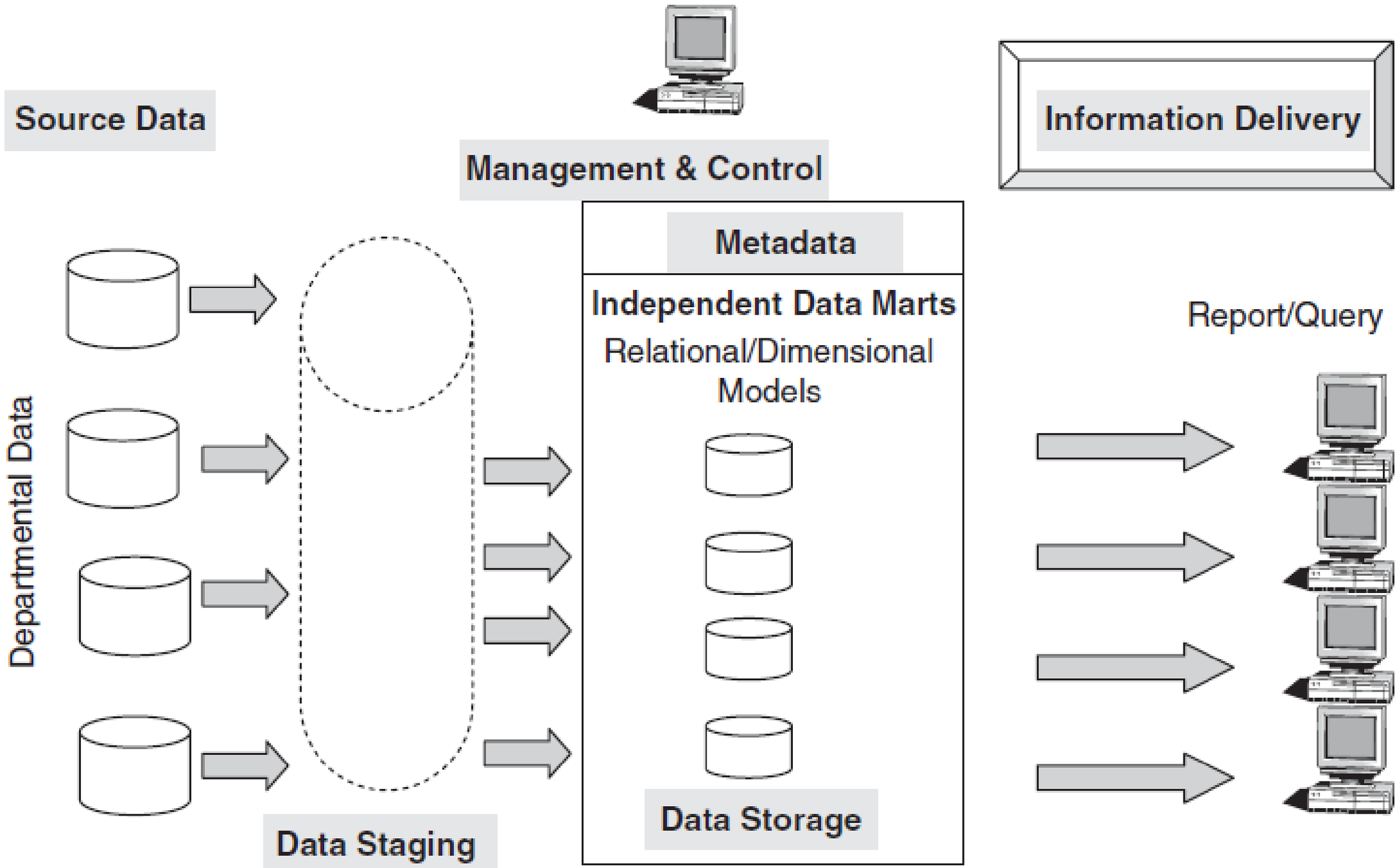


Figure 7-8 Overview of the components of independent data marts.

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

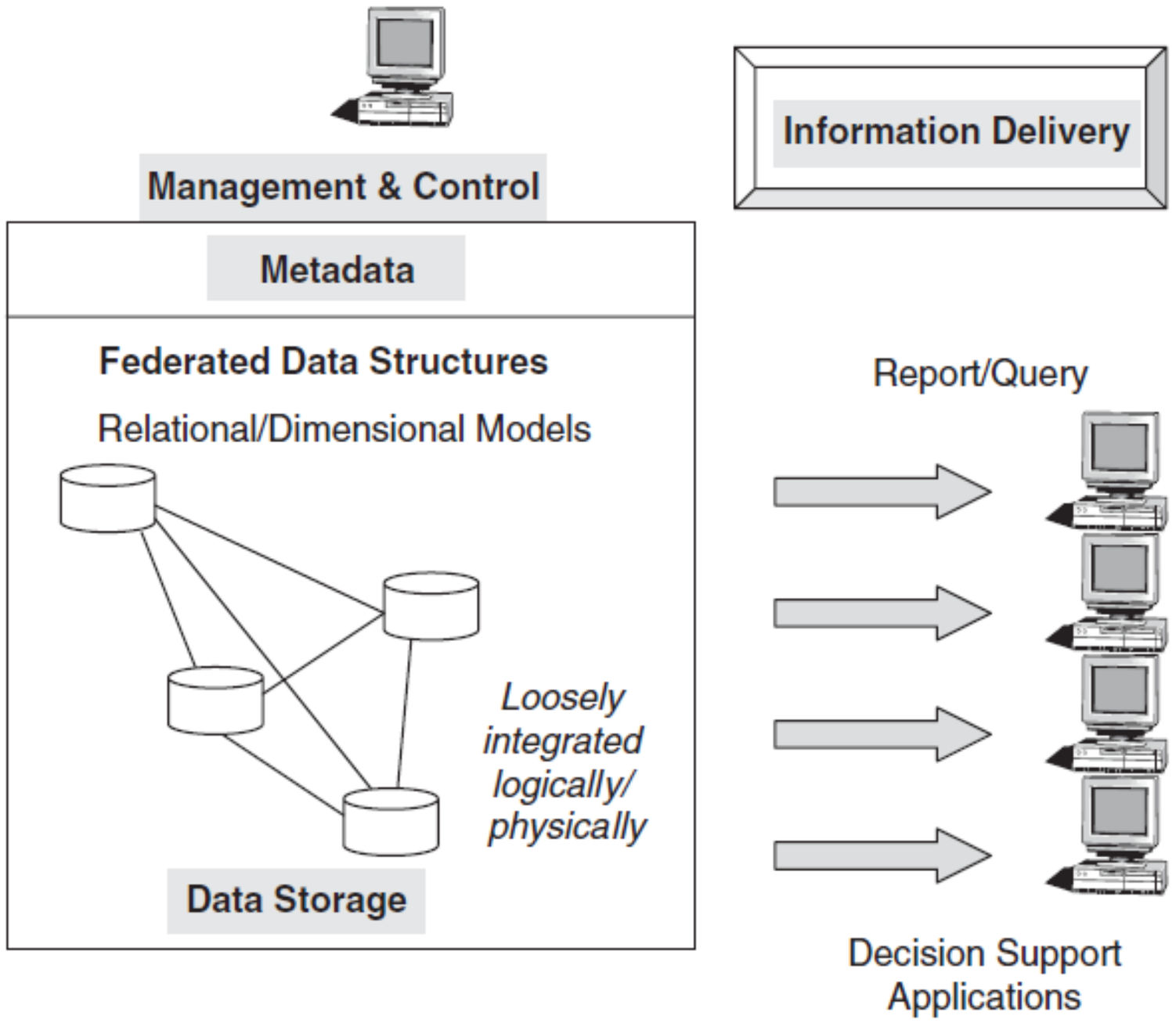


Figure 7-9 Overview of the components of a federated data warehouse.

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

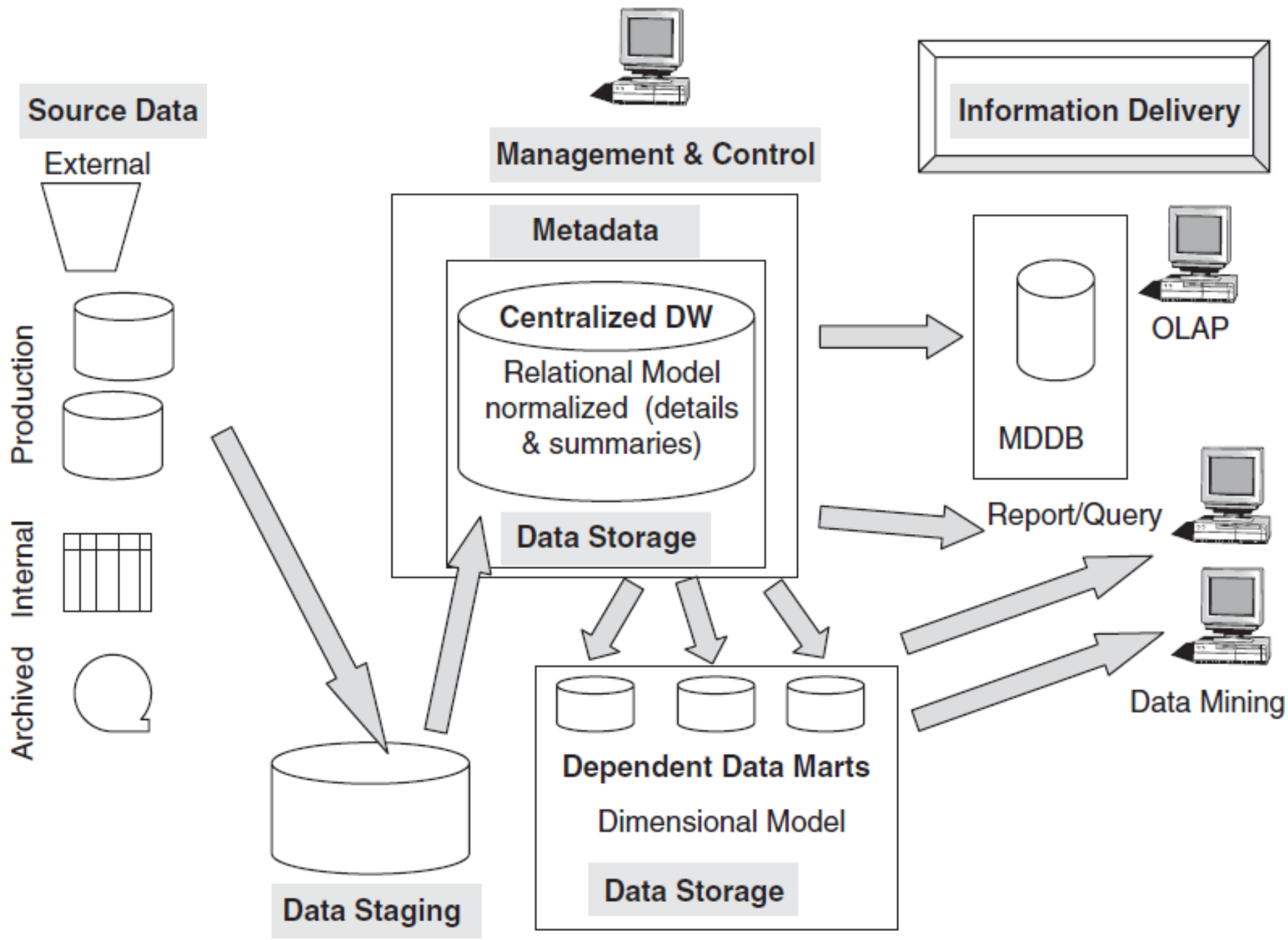


Figure 7-10 Overview of the components of a hub-and-spoke type of data warehouse.

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

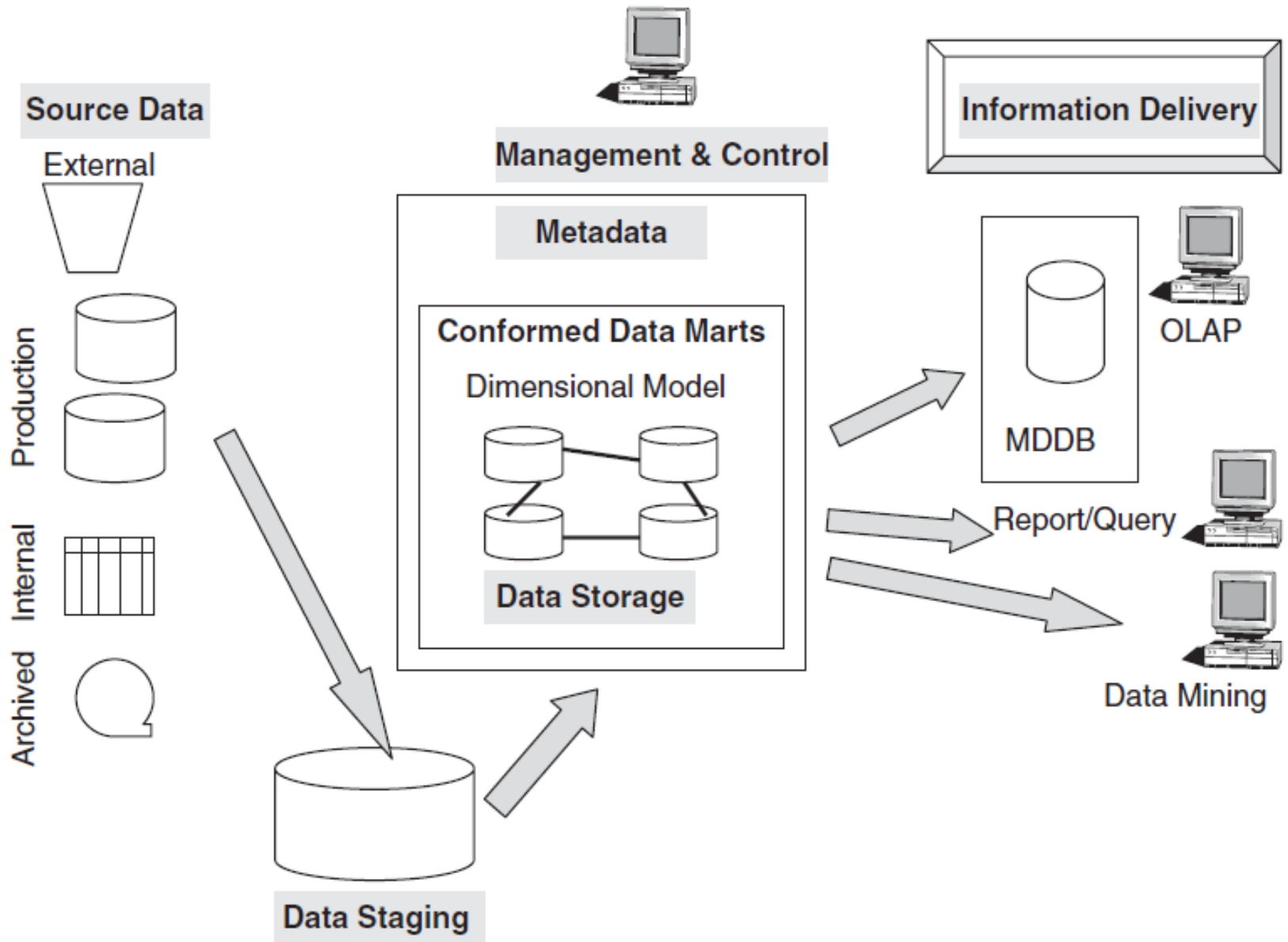


Figure 7-11 Overview of the components of a data-mart bus.

Kimball's approach

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

- Infrastructure acts as the foundation supporting the data warehouse architecture.
- The data warehouse infrastructure consists of **operational infrastructure and physical infrastructure**.
- **Hardware and operating systems** make up the computing environment for the data warehouse.
- Selecting the server hardware is a key decision. Invariably, the choice is one of **four parallel server architectures**.
- Parallel processing options are critical in the DBMS.
- Software tools are used in the data warehouse for data modeling, data extraction, data transformation, data loading, data quality assurance, queries and reports, and online analytical processing (OLAP). Tools are also used as middleware, alert systems, and for data warehouse administration.
- A data warehouse appliance is an architected device consisting of integrated hardware processors, storage, operating systems, and DBMS, specifically preoptimized for data warehousing.

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

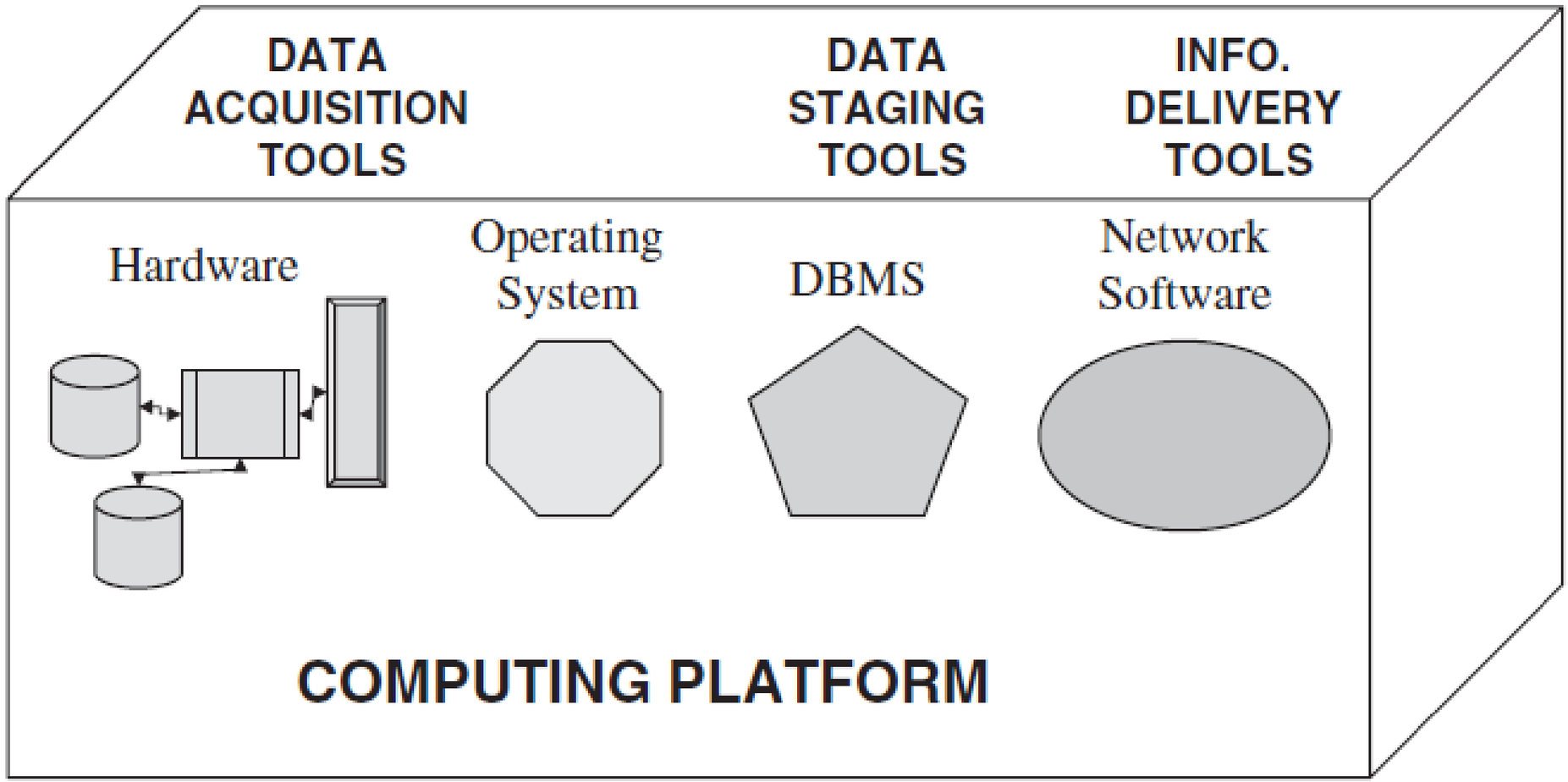


Figure 8-2 Physical infrastructure.

TOPIC 4: DW ARCHITECTURE AND INFRASTRUCTURE

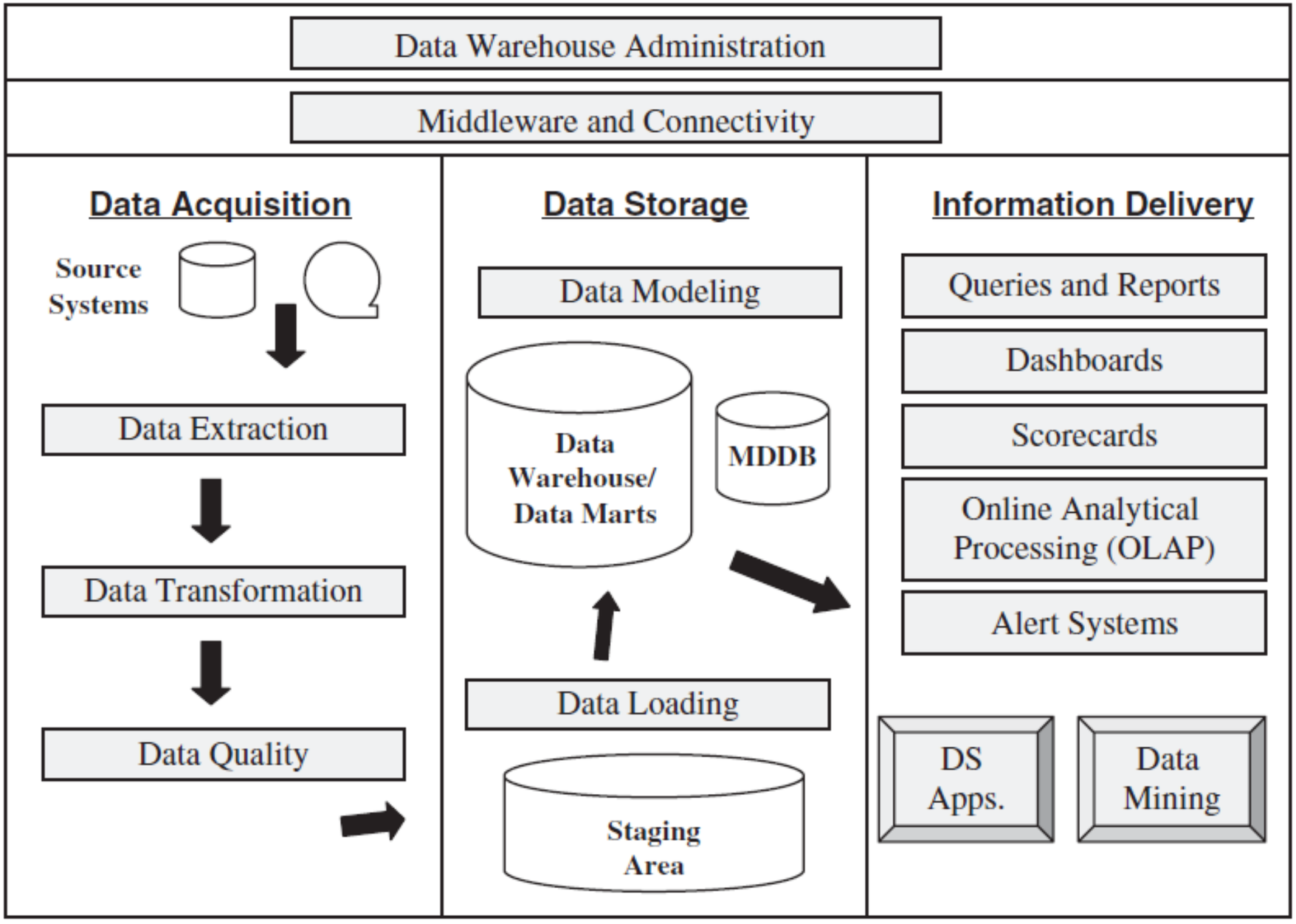


Figure 8-16 Tools for your data warehouse.

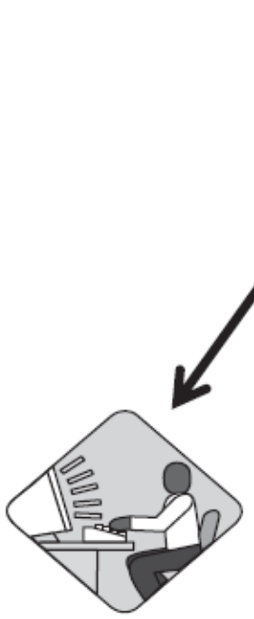
JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 5: METADATA IN DW

- Metadata is a critical need for **using, building, and administering** the data warehouse.
- For end-users, metadata is like a **roadmap to the data warehouse contents**.
- For IT professionals, metadata supports **development and administration functions**.
- Metadata types may be classified by the three functional areas of the data warehouse, namely, **data acquisition, data storage, and information delivery**.
- Business metadata connects the business users to the data warehouse.
- Technical metadata is meant for the IT staff responsible for development and administration.

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 5: METADATA IN DW



END-USERS

METADATA VITAL FOR END-USERS	METADATA ESSENTIAL FOR IT
Data content	
Summary data	
Business dimensions	
Business metrics	
Navigation paths	
Source systems	
External data	
Data transformation rules	
Last update dates	
Data load/update cycles	
Query templates	
Report formats	
Predefined queries/reports	
OLAP data	

Figure 9-5 Metadata vital for end-users.

METADATA ESSENTIAL FOR IT	
	Source data structures
	Source platforms
	Data extraction methods
	External data
	Data transformation rules
METADATA VITAL FOR END-USERS	Data cleansing rules
	Staging area structures
	Dimensional models
	Initial loads
	Incremental loads
	Data summarization
	OLAP system
	Web-enabling
	Query/report design

IT PROFESSIONALS

Figure 9-6 Metadata essential for IT.

JAMES COOK UNIVERSITY AUSTRALIA

TOPIC 5: METADATA IN DW

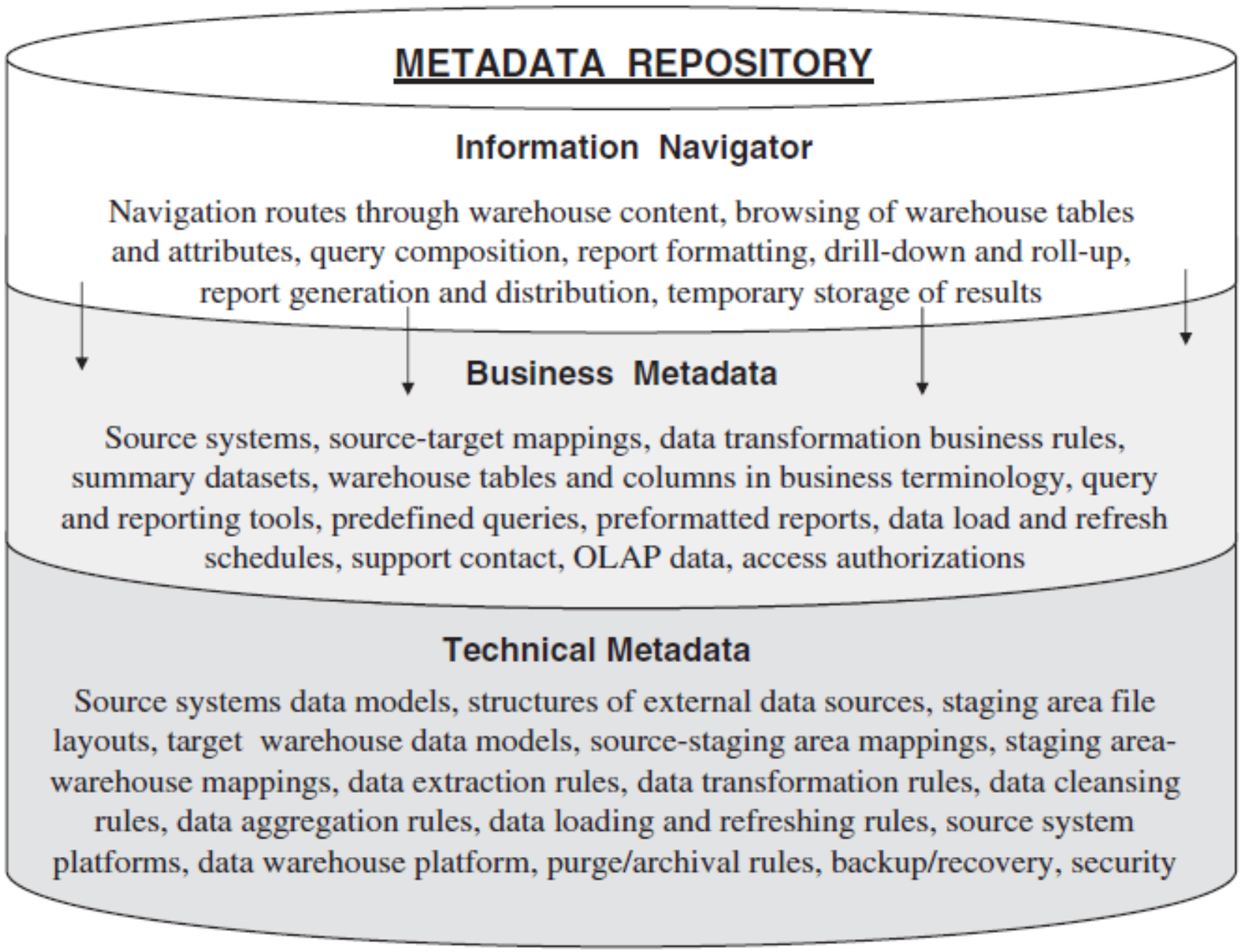


Figure 9-11 Metadata repository.

JAMES COOK UNIVERSITY AUSTRALIA

ASSESSMENT 1 THE DIKW HIERARCHY

- Specifications:
  - weight: 25%
  - Due: Sunday 26 July 11:59PM (Week 3)
  - 1500 words limitation, less than 5 pages and in 12pt Arial
    - Each task should be one page, around 300 words
  - APA 6th referencing style

Criteria	HD 85-100%
Readability **1%**	- Excellent progression of topics with a convincing logical flow - A highly conventional academic writing style, including the use of appropriate terminology and unbiased language - Excellent writing without typographical or grammatical errors
Citation/ Referencing **1%**	- All citations correctly adhere to APA 6th ed. referencing conventions - Excellent use of in-text citations for all claims and statements - Excellent list of references with no errors
Justification 3%	- Excellent development of decision, supported by logical, valid and comprehensive justifications