Architectural Components of DWH

Introduction

Manoj Kumar

What is a Data Warehouse ?

A data warehouse is a subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions.

- WH Inmon

Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Data Warehouse - Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse -Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”.

Data Warehouse - Non Updatable

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms.
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

Alternate Definitions

A collection of integrated, subject oriented databases designed to support the DSS function, where each unit of data is relevant to some moment of time - Imhoff

Data Warehouse is a repository of data summarized or aggregated in simplified form from operational systems. End user orientated data access and reporting tools let user get at the data for decision support - Babcock

Do we need a separate database ?

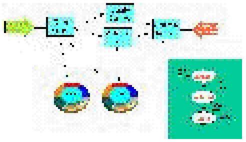
- OLTP and data warehousing require two very differently configured systems
- Isolation of Production System from Business Intelligence System
- Significant and highly variable resource demands of the data warehouse
- Cost of disk space no longer a concern
- Production systems not designed for query processing

OLTP Systems Vs Data Warehouse

Remember
Between OLTP and Data Warehouse systems

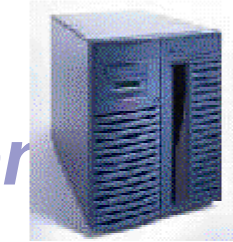
users are different

data content is different,



data structures are different

hardware is differen



Understanding The Differences Is The Key

OLTP Vs Warehouse

Operational System	Data Warehouse
Transaction Processing	Query Processing
Predictable CPU Usage	Random CPU Usage
Time Sensitive	History Oriented
Operator View	Managerial View
Normalized Efficient Design for TP	Denormalized Design for Query Processing

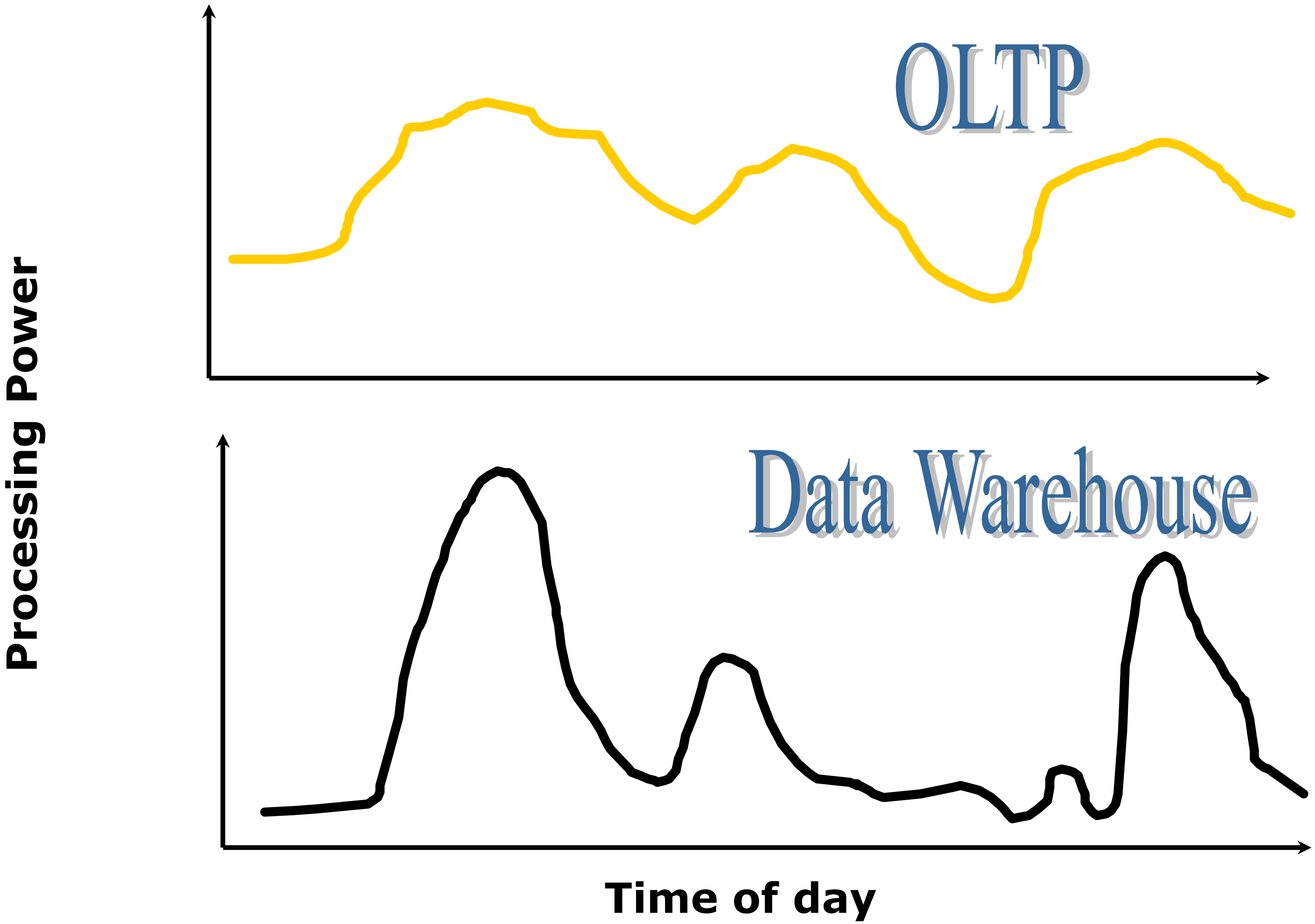
OLTP Vs Warehouse

Operational System	Data Warehouse
Designed for Atmocity, Consistency, Isolation and Durability	Designed for quite or static database
Organized by transactions (Order, Input, Inventory)	Organized by subject (Customer, Product)
Relatively smaller database	Large database size
Many concurrent users	Relatively few concurrent users
Volatile Data	Non Volatile Data

OLTP Vs Warehouse

Operational System	Data Warehouse
Stores all data	Stores relevant data
Performance Sensitive	Less Sensitive to performance
Not Flexible	Flexible
Efficiency	Effectiveness

Capacity Planning



Processing Load Peaks During the Beginning and End of Day

Data Marts

- Enterprise wide data warehousing projects have a very large cycle time
- Getting consensus between multiple parties may also be difficult
- Departments may not be satisfied with priority accorded to them
- Sometimes individual departmental needs may be strong enough to warrant a local implementation
- Application/database distribution is also an important factor

Data Marts

Subject or Application Oriented Business View of Warehouse

- Finance, Manufacturing, Sales etc.
- Smaller amount of data used for Analytic Processing
- Address a single business process

A Logical Subset of The Complete Data Warehouse

Data Warehouse and Data Mart

	Data Warehouse	Data Marts
Scope	• Application Neutral • Centralized, Shared • Cross LOB/enterprise	• Specific Application Requirement • LOB, department • Business Process Oriented
Data Perspective	• Historical Detailed data • Some summary	• Detailed (some history) • Summarized
Subjects	• Multiple subject areas	• Single Partial subject • Multiple partial subjects • OLTP snapshots

Data Warehouse and Data Mart

	Data Warehouse	Data Marts
Data Sources	• Many • Operational/ External Data	• Few • Operational, external data • OLTP snapshots
Implement Time Frame	• 9-18 months for first stage • Multiple stage implementation	• 4-12 months
Characteristics	• Flexible, extensible • Durable/Strategic • Data orientation	• Restrictive, non extensible • Short life/tactical • Project Orientation

Warehouse or Mart First ?

Data Warehouse First	Data Mart first
Expensive	Relatively cheap
Large development cycle	Delivered in < 6 months
Change management is difficult	Easy to manage change
Difficult to obtain continuous corporate support	Can lead to independent and incompatible marts
Technical challenges in building large databases	Cleansing, transformation, modeling techniques may be incompatible

Different kinds of Information Needs

Current	**Is this medicine available in stock**	OLTP
Recent	**What are the tests this patient has completed so far**	ODS
Historical	**Has the incidence of Tuberculosis increased in last 5 years in Southern region**	Data Warehouse

OLTP Vs ODS Vs DWH

Characteristic	OLTP	ODS	Data Warehouse
Audience	Operating Personnel	Analysts	Managers and analysts
Data access	Individual records, transaction driven	Individual records, transaction or analysis driven	Set of records, analysis driven
Data content	Current, real-time	Current and near-current	Historical
Data granularity	Detailed	Detailed and lightly summarized	Summarized and derived
Data organization	Functional	Subject-oriented	Subject-oriented
Data quality	All application specific detailed data needed to support a business activity	All integrated data needed to support a business activity	Data relevant to management information needs

OLTP Vs ODS Vs DWH

Characteristic	OLTP	ODS	Data Warehouse
Data redundancy	Non-redundant within system; Unmanaged redundancy among systems	Somewhat redundant with operational databases	Managed redundancy
Data stability	Dynamic	Somewhat dynamic	Static
Data update	Field by field	Field by field	Controlled batch
Data usage	Highly structured, repetitive	Somewhat structured, some analytical	Highly unstructured, heuristic or analytical
Database size	Moderate	Moderate	Large to very large
Database structure stability	Stable	Somewhat stable	Dynamic

OLTP Vs ODS Vs DWH

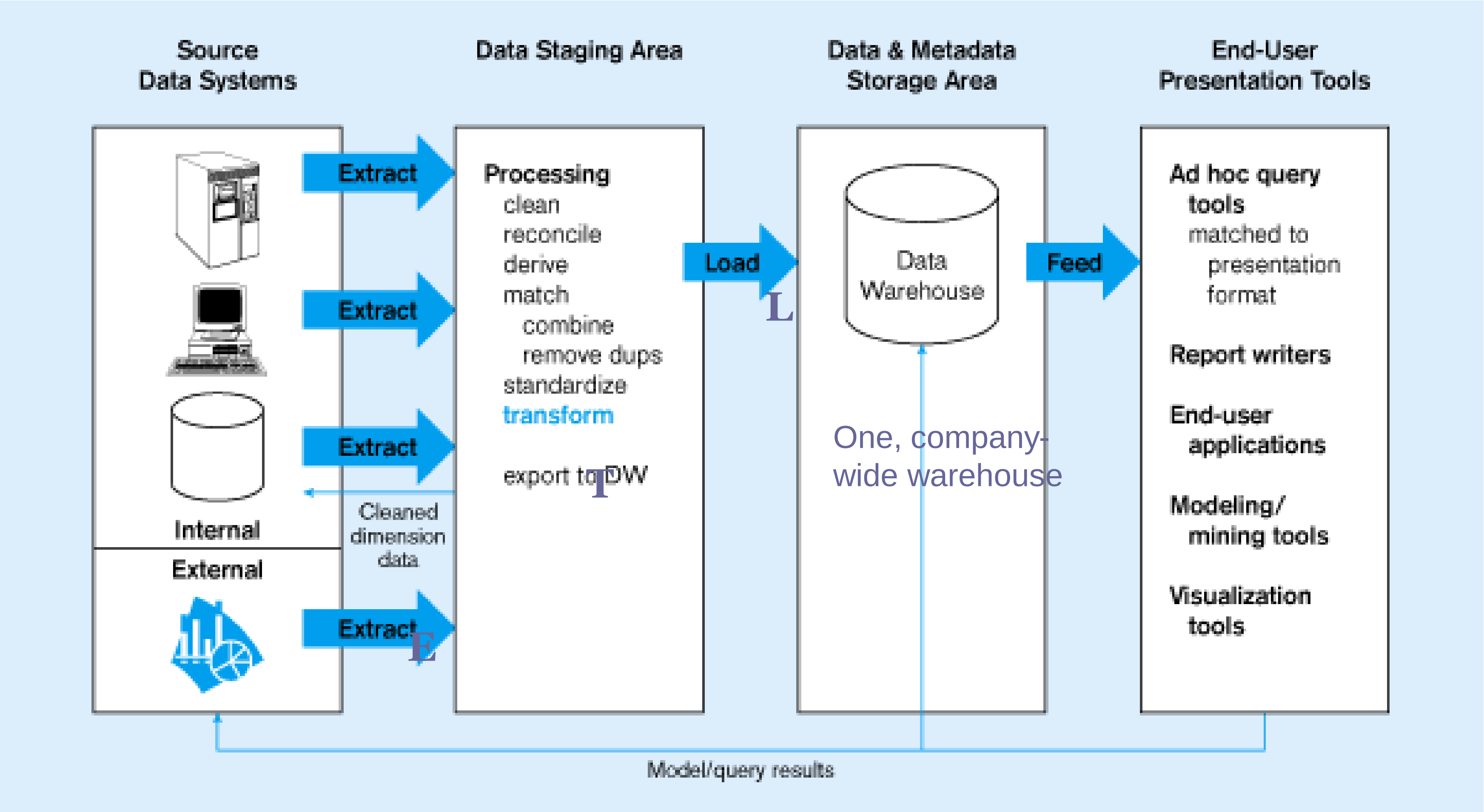
Characteristic	OLTP	ODS	Data Warehouse
Development methodology	Requirements driven, structured	Data driven, somewhat evolutionary	Data driven, evolutionary
Operational priorities	Performance and availability	Availability	Access flexibility and end user autonomy
Philosophy	Support day-to-day operation	Support day-to-day decisions & operational activities	Support managing the enterprise
Predictability	Stable	Mostly stable, some unpredictability	Unpredictable
Response time	Sub-second	Seconds to minutes	Seconds to minutes
Return set	Small amount of data	Small to medium amount of data	Small to large amount of data

Data Warehouse Architectures

- 1.Generic Two-Level Architecture
- 2.Independent Data Mart
- 3.Dependent Data Mart and Operational Data Store
- 4.Logical Data Mart and Active Warehouse
- 5.Three-Layer architecture

All involve some form of ***extract***, ***transform*** and ***load*** (**ETL**)

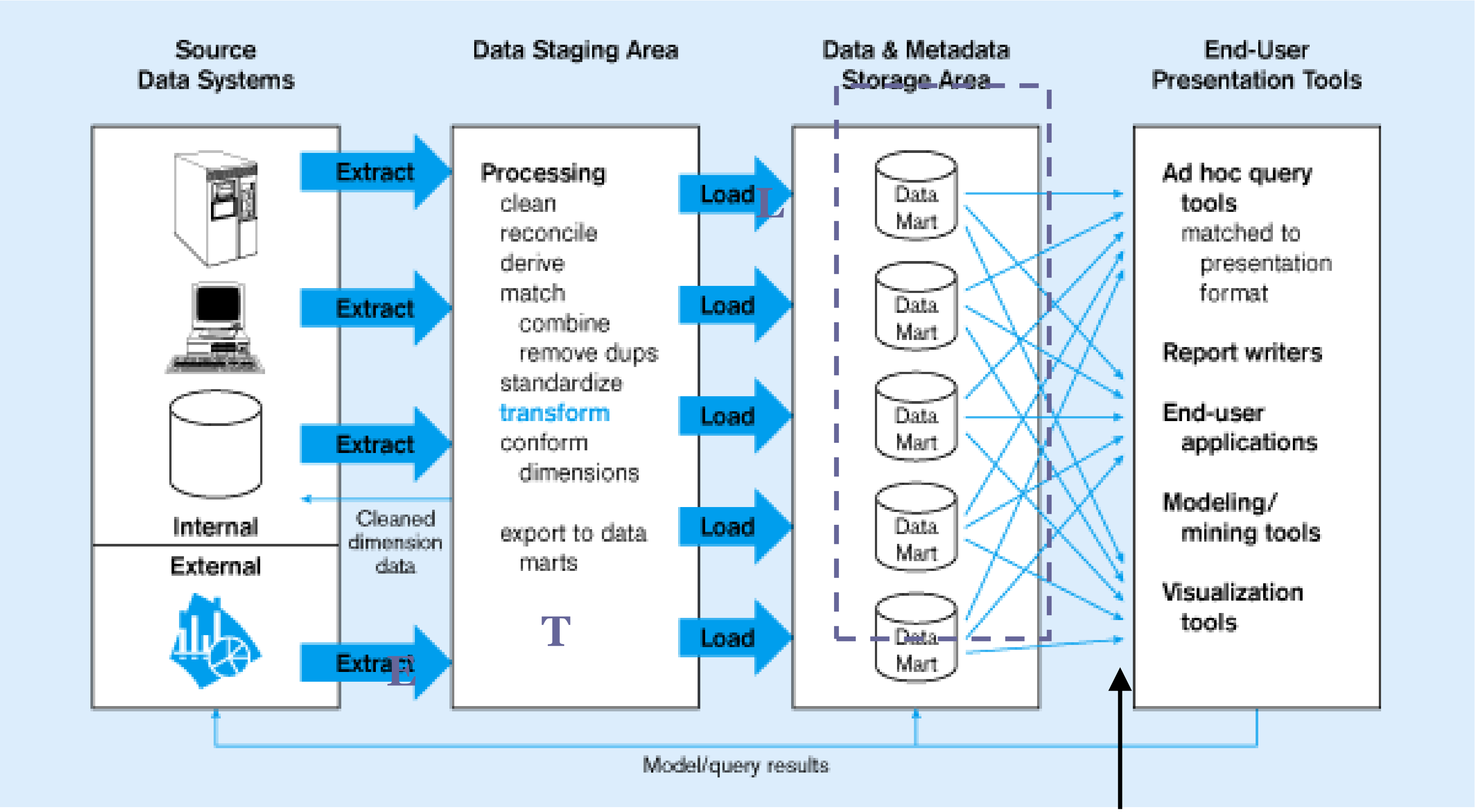
Figure 11-2: Generic two-level architecture



Periodic extraction ➔ data is not completely current in warehouse

Figure 11-3: Independent Data Mart

Data marts:
Mini-warehouses, limited in scope



Separate ETL for each ***independent*** data mart

Data access complexity due to ***multiple*** data marts

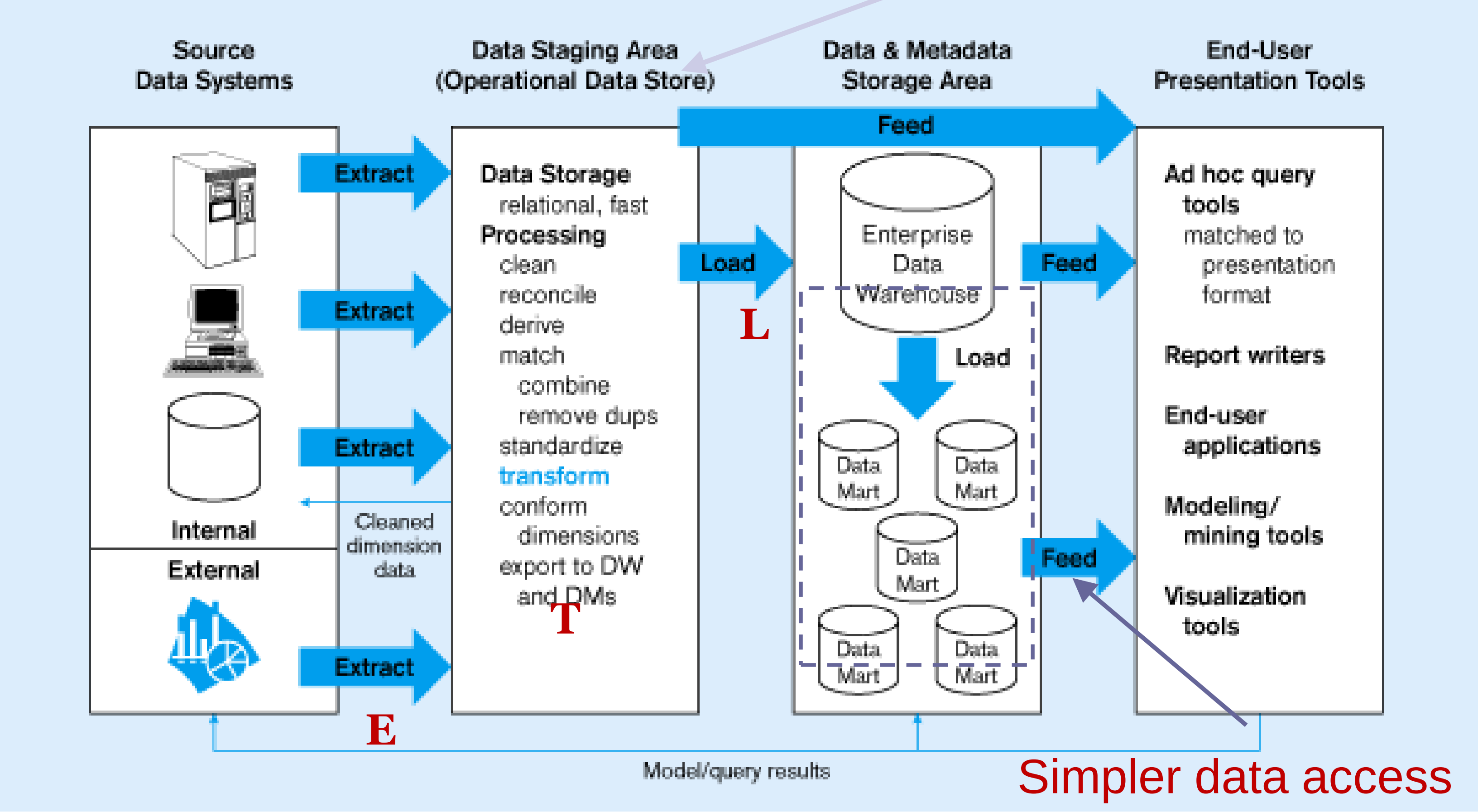
Independent Data mart

- **Independent data mart:** a data mart filled with data extracted from the operational environment without benefits of a data warehouse.

Figure 11-4:

Dependent data mart with *operational data store*

ODS provides option for obtaining ***current*** data



Simpler data access

Single ETL for ***enterprise data warehouse (EDW)***

Dependent data marts loaded from EDW

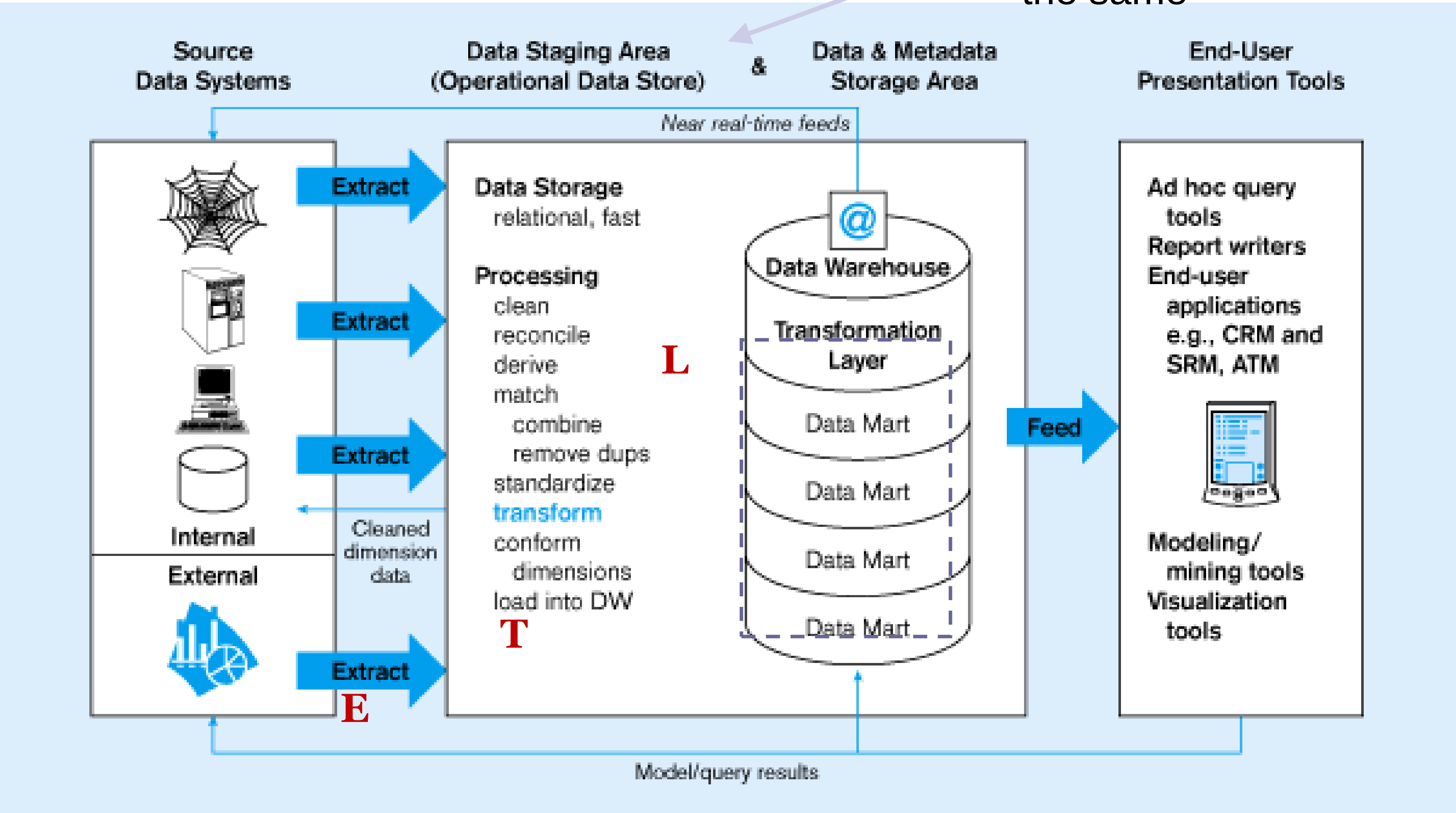
Dependent data mart - Operational data store

- **Dependent data mart**: A data mart filled exclusively from the enterprise data warehouse and its reconciled data.
- **Operational data store (ODS)**: An integrated, subject-oriented, updatable, current-valued, enterprise-wise, detailed database designed to serve operational users as they do decision support processing.

Figure 11-5:

Logical data mart and @ctive data warehouse

ODS and data warehouse are one and the same



Near real-time ETL for
@active Data Warehouse

Data marts are NOT separate databases, but logical ***views*** of the data warehouse

➔ Easier to create new data marts

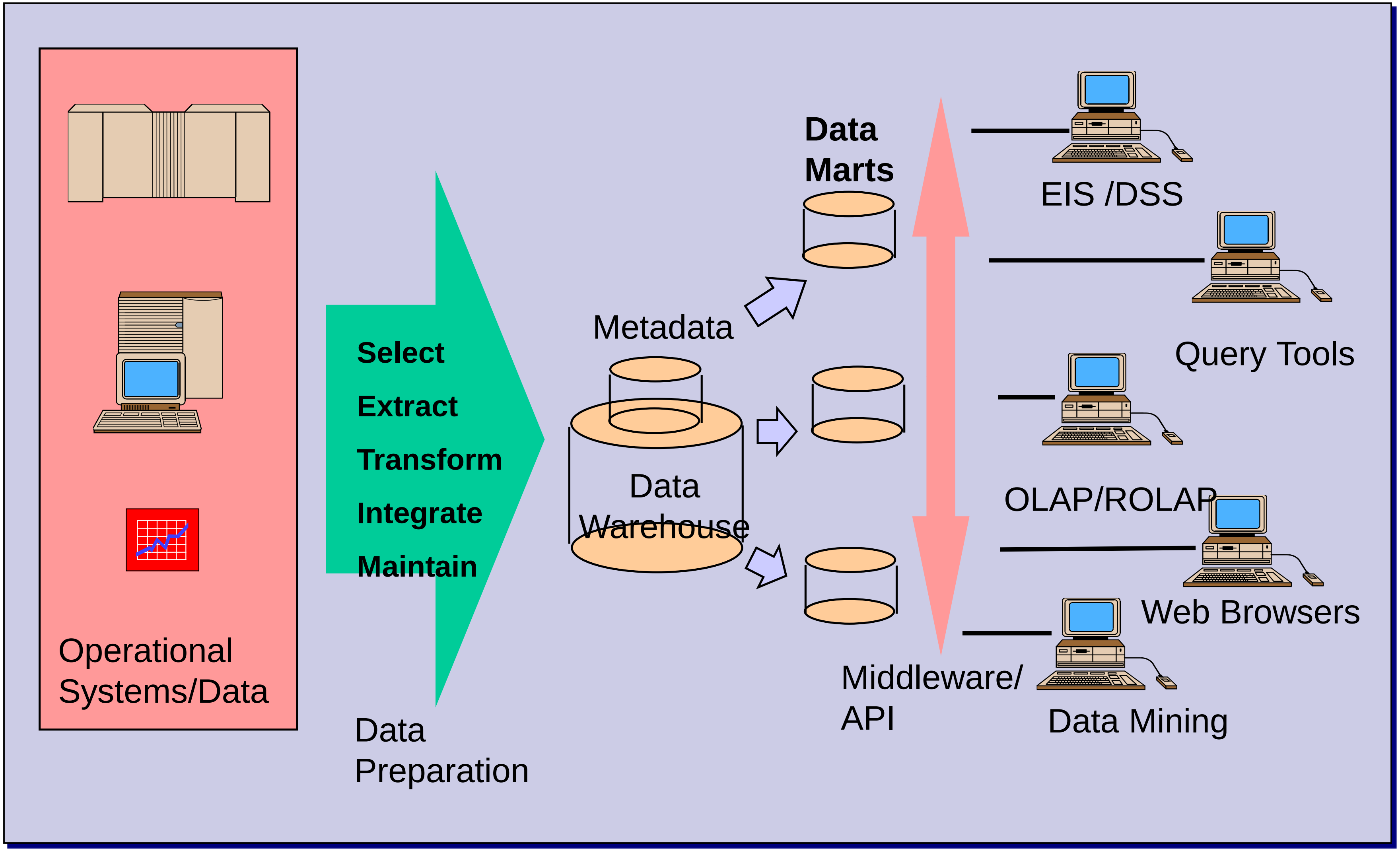
Three-layer architecture
Reconciled and derived data

- *Reconciled data*: detailed, current data intended to be the single, authoritative source for all decision support.
- *Derived data*: Data that have been selected, formatted, and aggregated for end-user decision support application.
- *Metadata*: technical and business data that describe the properties or characteristics of other data.

Other data warehouse changes

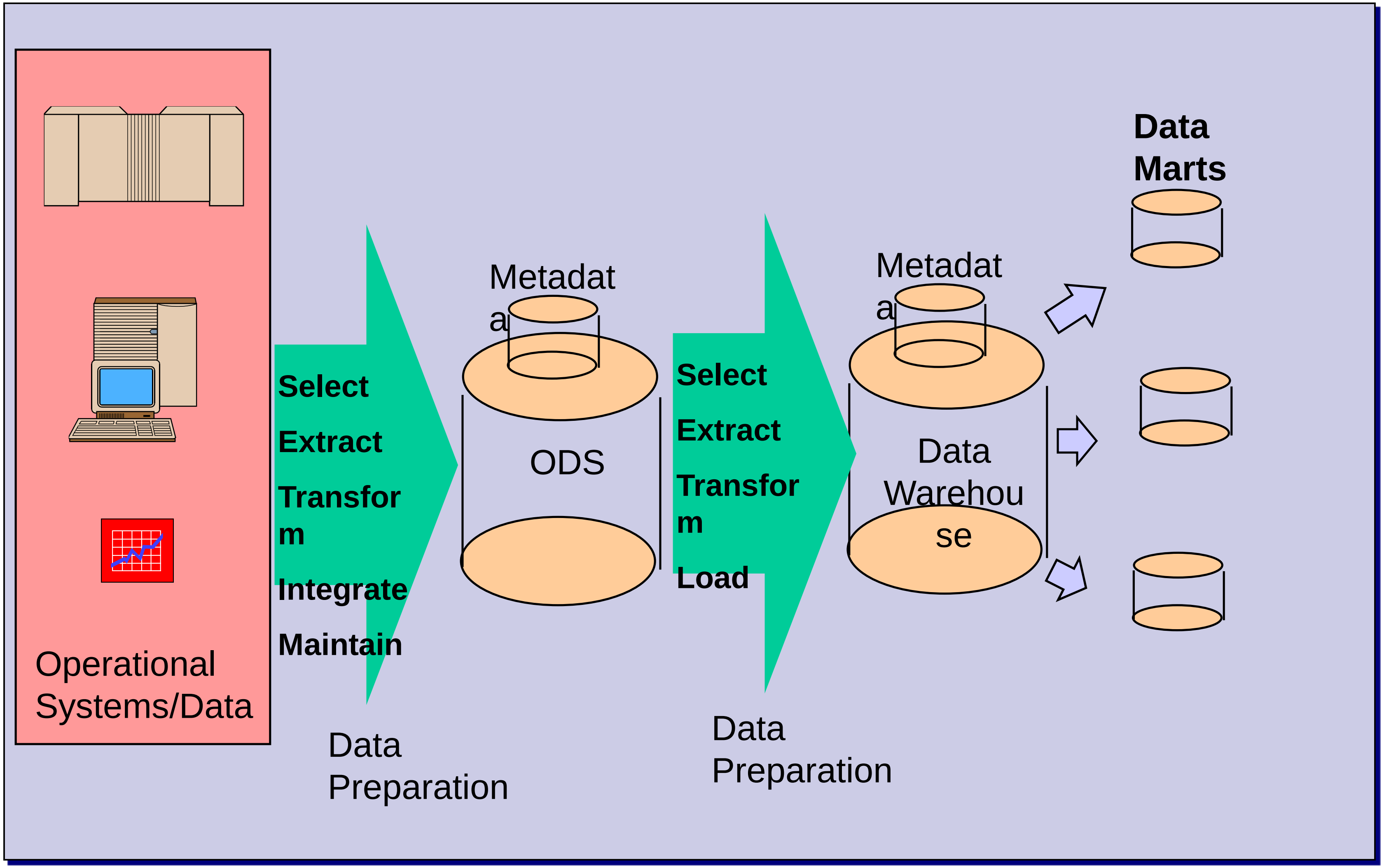
- New descriptive attributes
- New business activity attributes
- New classes of descriptive attributes
- Descriptive attributes become more refined
- Descriptive data are related to one another
- New source of data

Typical (Graphical) DWH Architecture



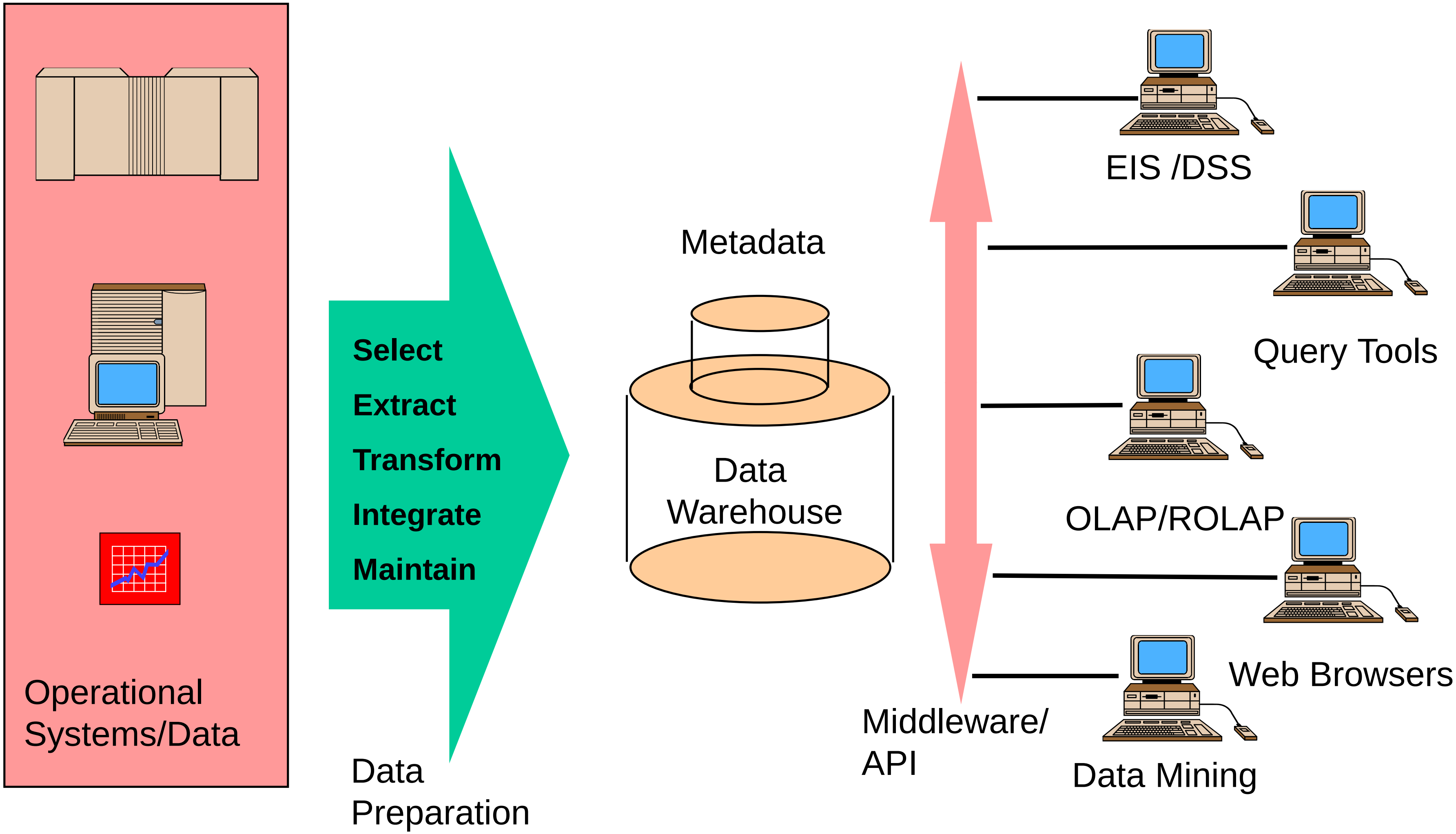
Multi-tiered Data Warehouse without ODS

Typical (Graphical) DWH Architecture



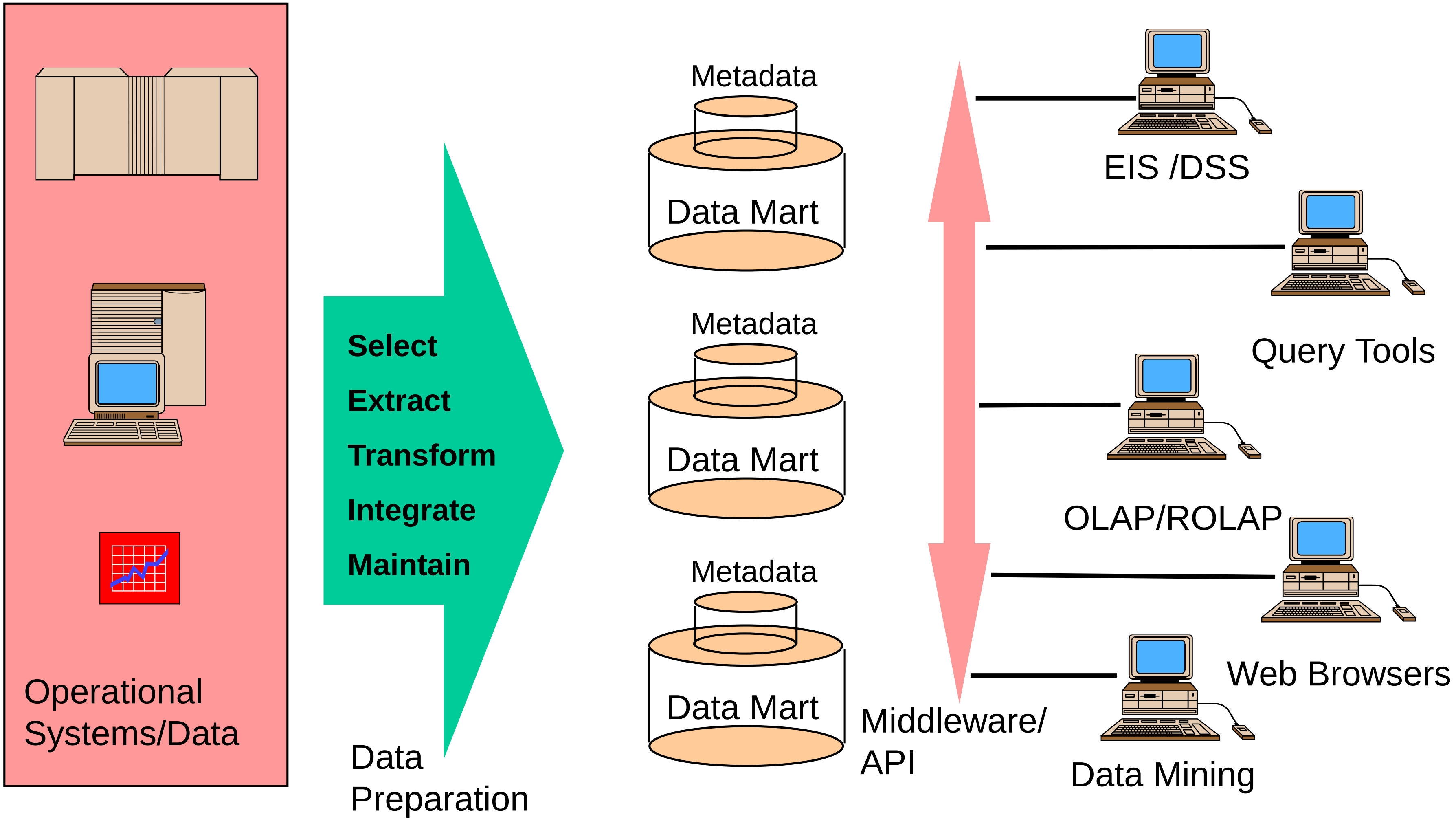
Multi-tiered Data Warehouse with ODS

Warehouse Architecture - 1



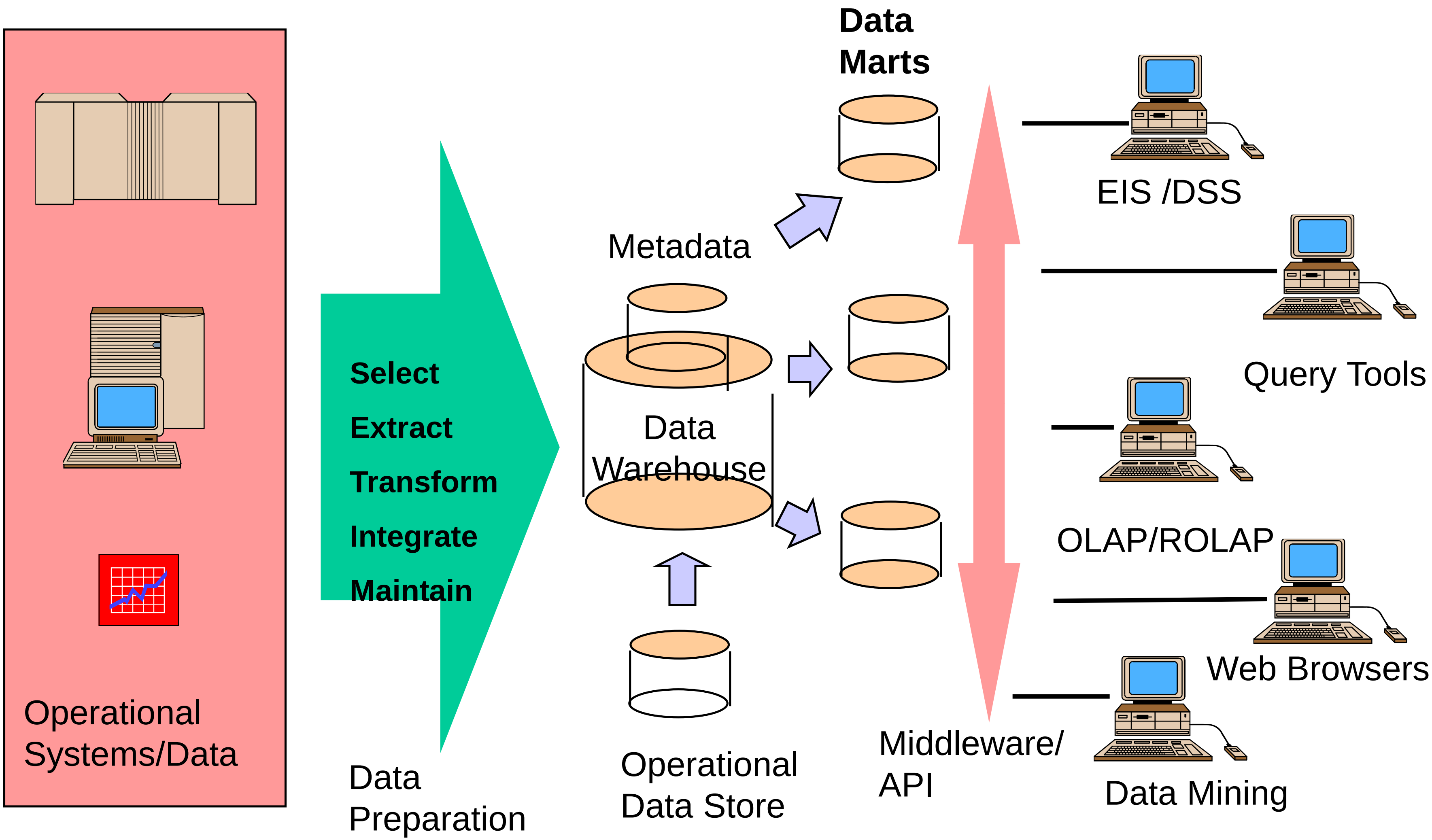
Enterprise Data Warehouse

Warehouse Architecture - 2



Single Department Data Mart

Warehouse Architecture - 3



Multi-tiered Data Warehouse

Kimball’s View

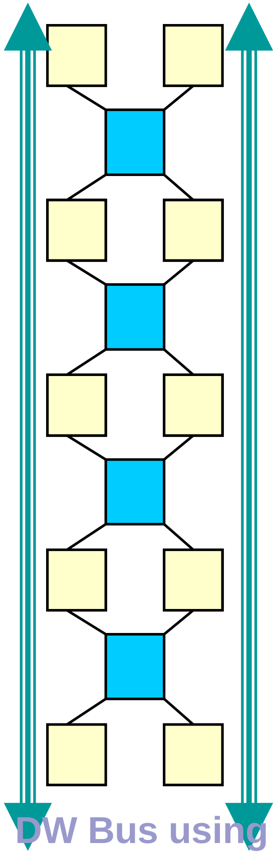
Operational Systems

Staging Area

Data Warehouse
Server
Processes

- Extract
- Scrubbing
- Transformation
- Load Jobs
- Aggregation Jobs
- Replication
- Monitoring
- Management
- Meta Data Repository
- Meta Data Population
- Meta Data Maintenance

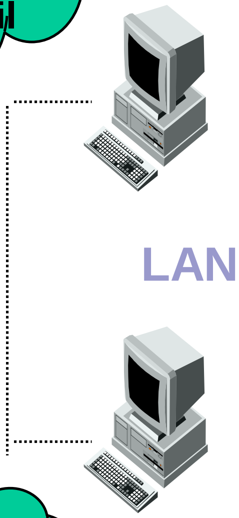
Presentation Server



DW Bus using
Conformed Dimensions

Each Star
is
a Data
Mart and
has both
summary
and detail
data

DW is
sum
total of all
Data
Marts



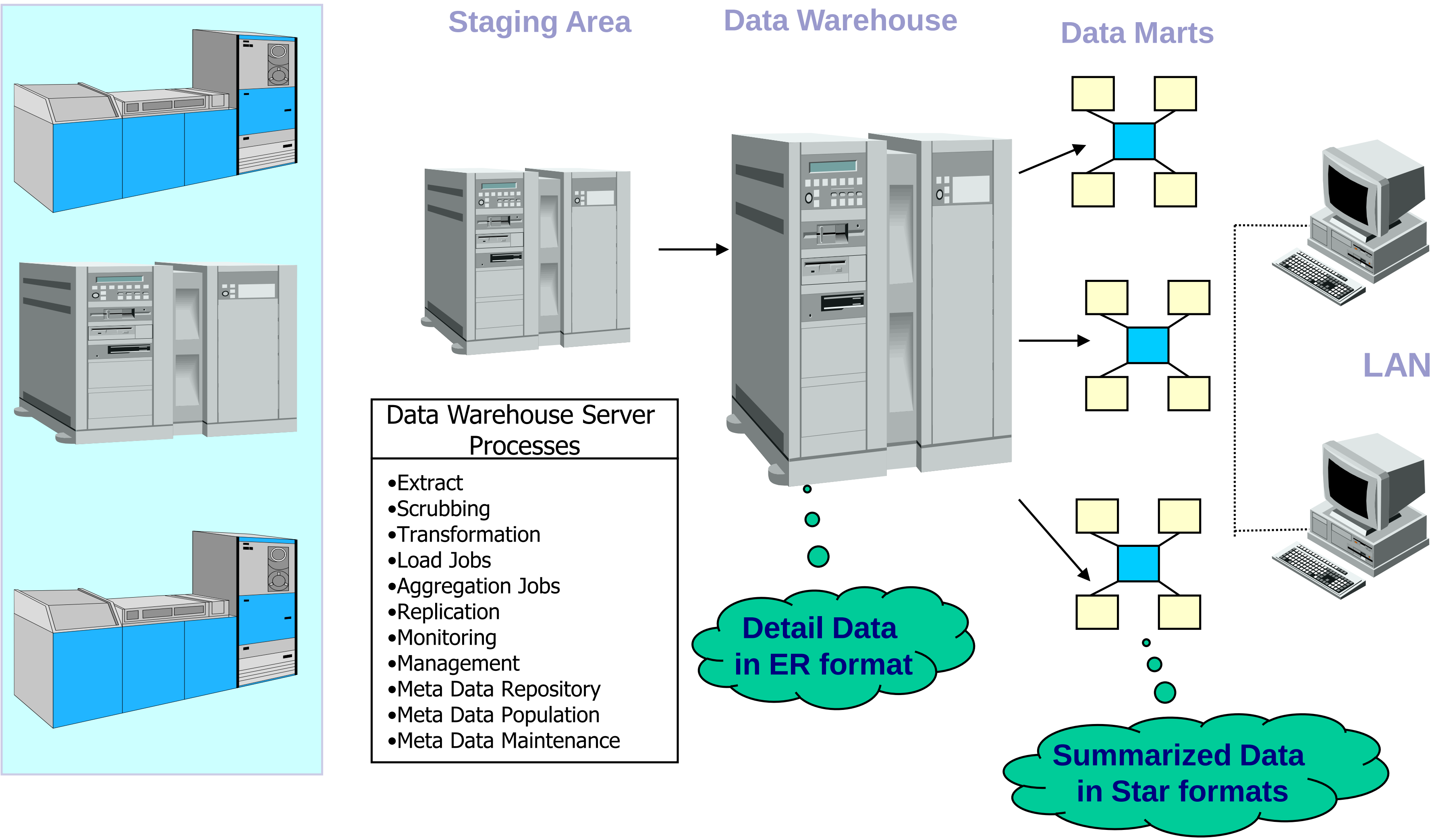
LAN

Multiple Data Marts With Conformed Dimensions

Ralph Kimball: Bottom-Up Approach

- Recommends, start with **small** mission critical **Data Marts** that serve analytic needs of departments….
- Then integrate these data marts for data consistency through a so called information **bus**
- Uses **star schemas** or snowflakes to organize the data in *dimensional Modeling data warehouse*
- Kimball gives his opinion of **Independent** data **marts**
- More Simpler, Cost effective & Quicker to Deliver
- DWH Bus Architecture consist of both the **Atomic** and **Aggregated** Data Marts data, stored in a **star** schema

Inmon’s View



Data Warehouse Server Processes

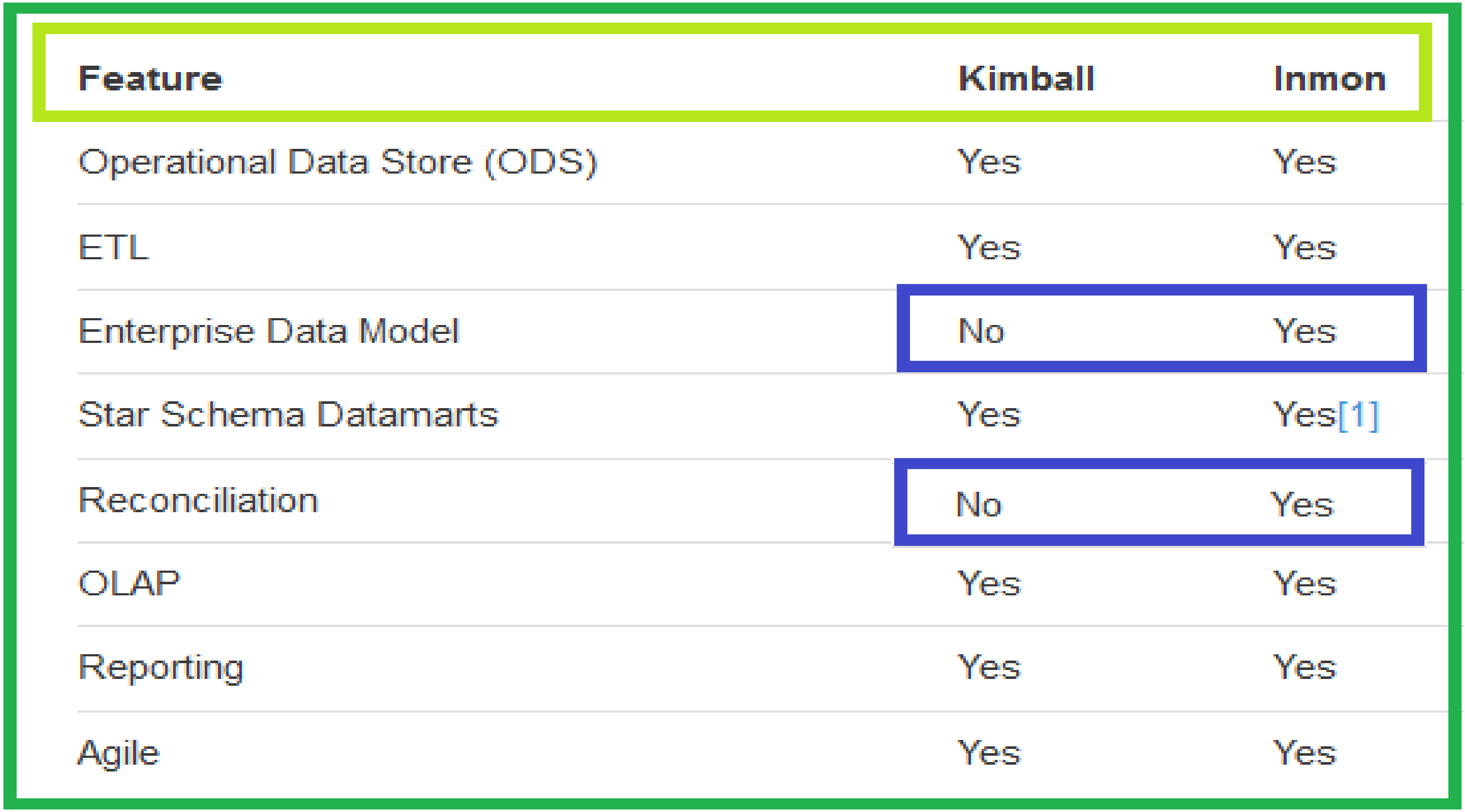
- Extract
- Scrubbing
- Transformation
- Load Jobs
- Aggregation Jobs
- Replication
- Monitoring
- Management
- Meta Data Repository
- Meta Data Population
- Meta Data Maintenance

Data Warehouse (ER) Feeding Multiple Data Marts (Star Schema)

Bill Inmon: Top-Down Approach

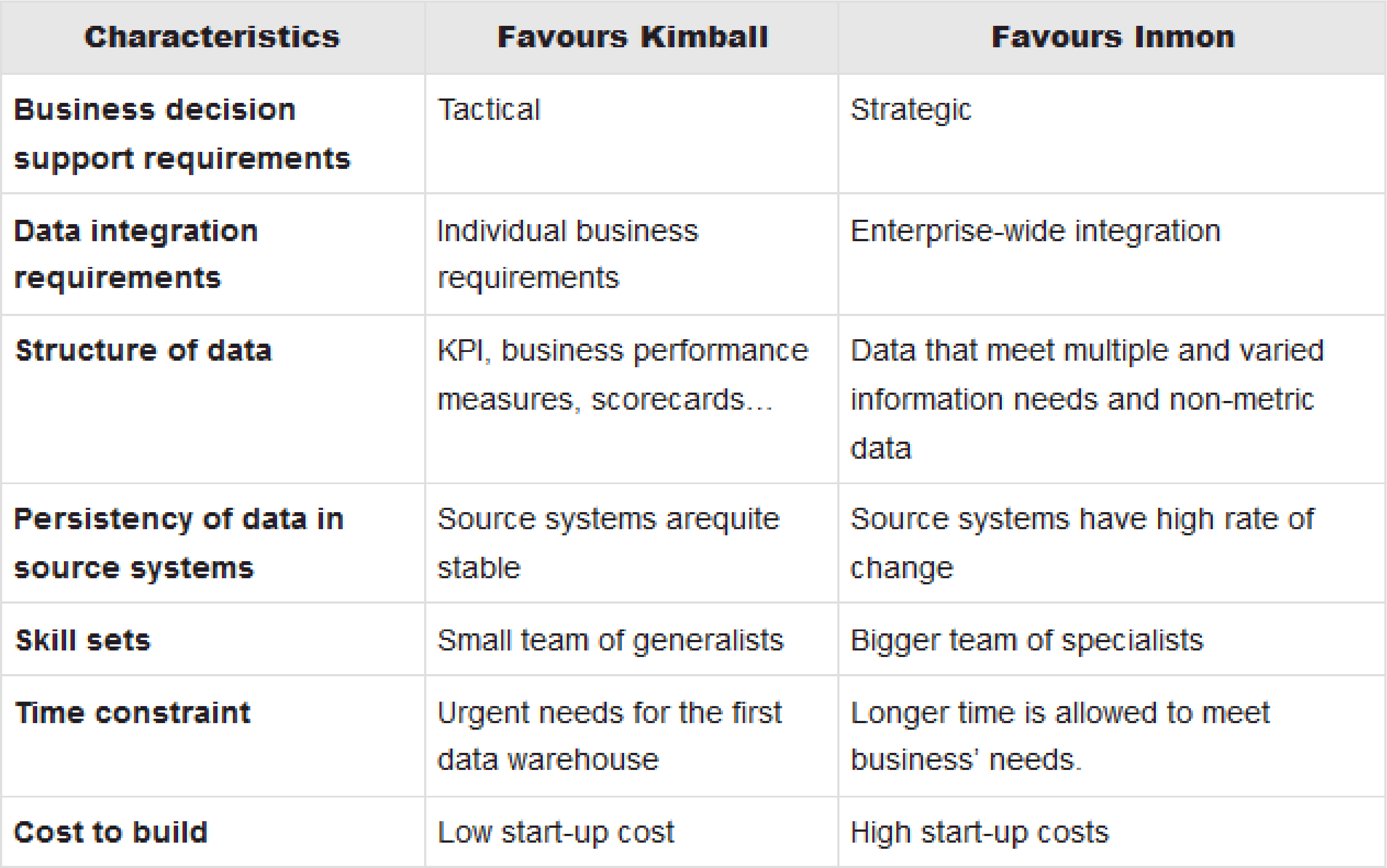
- Recommends big **Centralized enterprise** data warehouse where all available data from transaction systems are consolidated into....
- a subject-oriented, integrated, time-variant and non-volatile collection of data into DSS... then data marts are built for analytic needs of depts
- Uses **ER model** to organize the data in enterprise *data warehouse*
- Inmon gives his opinion of **Dependent** data **marts**
- More Complex, Expensive & Longer to Deliver

Kimball Vs Inmon



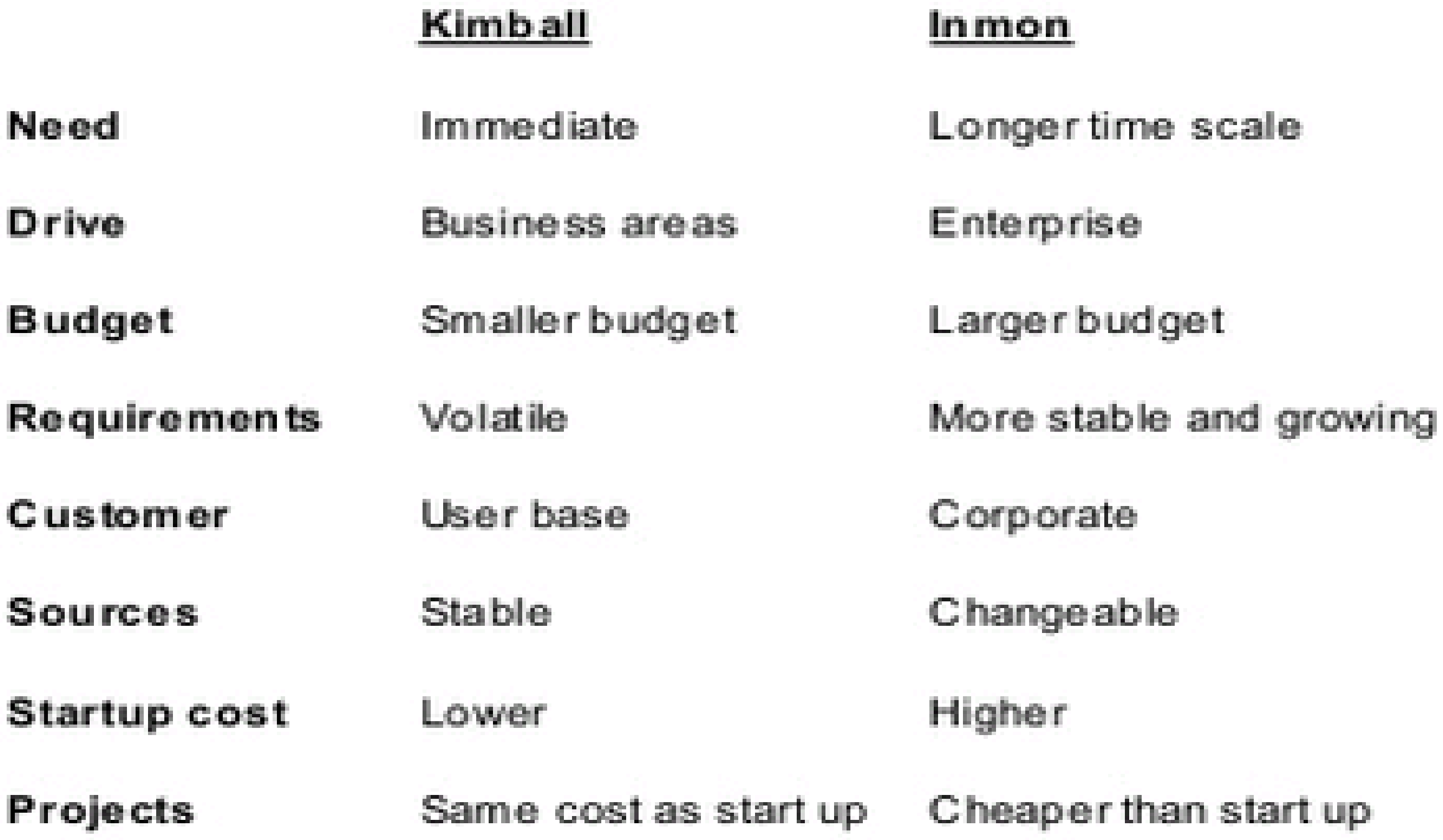
Feature	Kimball	Inmon
Operational Data Store (ODS)	Yes	Yes
ETL	Yes	Yes
Enterprise Data Model	No	Yes
Star Schema Datamarts	Yes	Yes[1]
Reconciliation	No	Yes
OLAP	Yes	Yes
Reporting	Yes	Yes
Agile	Yes	Yes

Kimball vs. Inmon approach



Characteristics	Favours Kimball	Favours Inmon
Business decision support requirements	Tactical	Strategic
Data integration requirements	Individual business requirements	Enterprise-wide integration
Structure of data	KPI, business performance measures, scorecards…	Data that meet multiple and varied information needs and non-metric data
Persistency of data in source systems	Source systems arequite stable	Source systems have high rate of change
Skill sets	Small team of generalists	Bigger team of specialists
Time constraint	Urgent needs for the first data warehouse	Longer time is allowed to meet business' needs.
Cost to build	Low start-up cost	High start-up costs

Kimball vs. Inmon approach



	Kimball	Inmon
Need	Immediate	Longer time scale
Drive	Business areas	Enterprise
Budget	Smaller budget	Larger budget
Requirements	Volatile	More stable and growing
Customer	User base	Corporate
Sources	Stable	Changeable
Startup cost	Lower	Higher
Projects	Same cost as start up	Cheaper than start up

Complementary Approach

Common elements: in Both Approaches

- There is **no right or wrong approach** and it totally depends on *kind of requirement, project nature* to decide the approach.
- Both Kimball and Inmon’s architectures share a same common feature that each has a **single integrated repository** of atomic data.
- Both architectures have an enterprise focus that supports information **analysis** across the organization.
- Both enables to address the business requirements not only within a **subject area** but also across subject areas.

Complementary Approach

Common elements: in Both Approaches

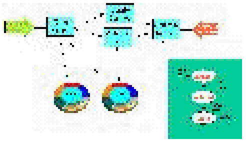
- When it comes to data **modelling**, depends on specific requirements, sometimes makes sense to take a **hybrid** approach**.**
- Both these models have their own strengths and weakness.
- All enterprises require a means to store, analyze and interpret the data they generate and accumulate in order to implement critical decisions that range from “continuing to exist” to maximizing prosperity.

OLTP Systems Vs Data Warehouse *Application*

Between OLTP and DWH systems

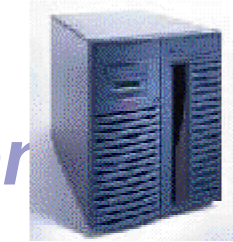
users are different

data content is different,



data structures are different

hardware is differen



Understanding The Differences Is The Key

Examples Of Some Applications

- Target Marketing
- Market Segmentation
- Budgeting
- Credit Rating Agencies
- Financial Reporting and Consolidation

- Market Basket Analysis - POS Analysis
- Fraud Management
- Profitability Management
- Event tracking