BITS Pilani
Pilani Campus

SS ZG515 - Data Warehousing

Dr. Yashvardhan Sharma
CSIS Dept., BITS-Pilani

Need for Data Warehousing

- Companies, over the years, gathered huge volumes of data
- “Hidden Treasure”
- Can this data be used in any way?
- Can we analyze this data to get any competitive advantage?
- If yes, what kind of advantage?

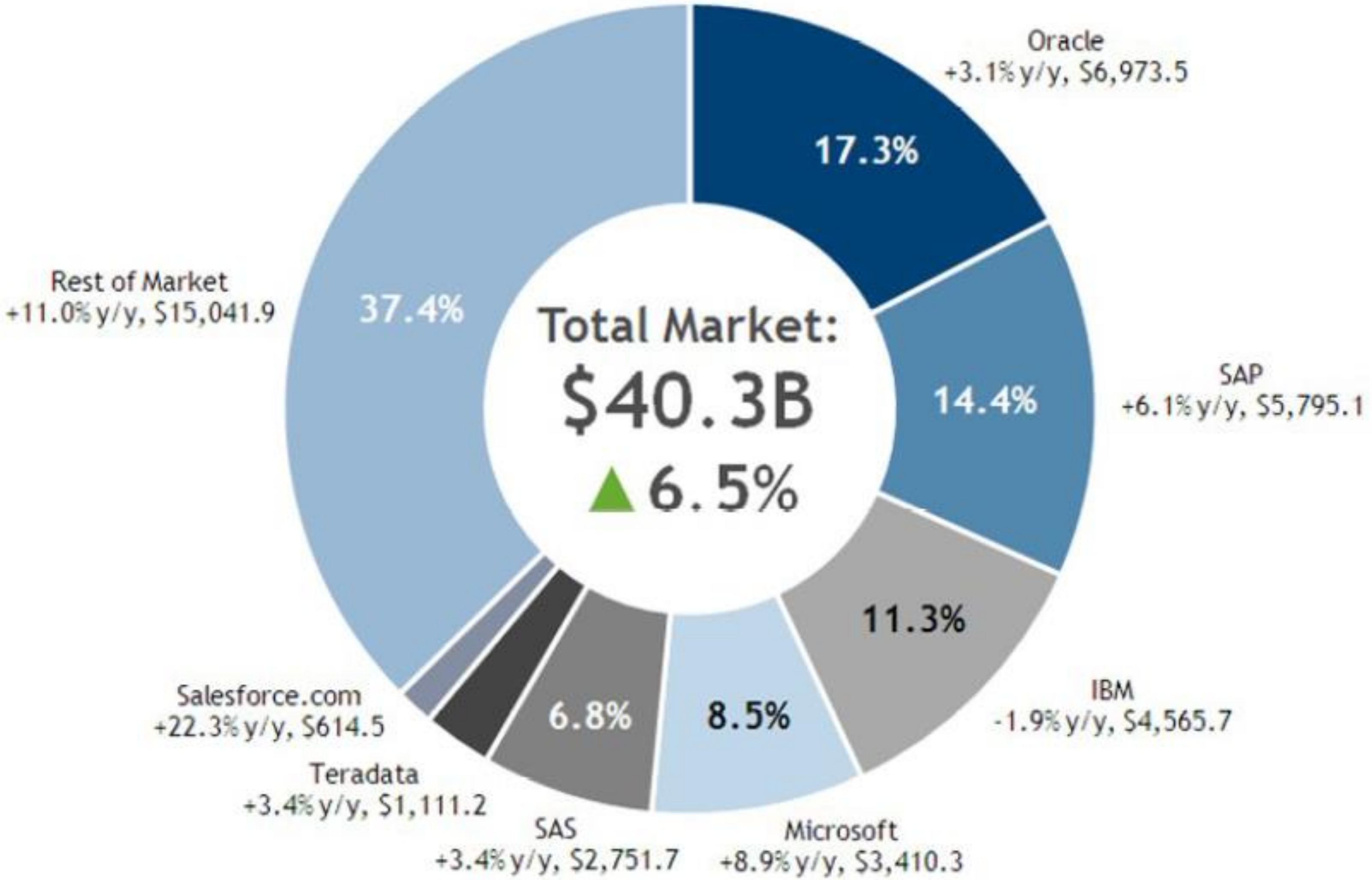
Benefits of Data Warehousing

- Allows “efficient” analysis of data
- Competitive Advantage
- Analysis aids strategic decision making
- Increased productivity of decision makers
- Potential high ROI
- Classic example: Diaper and Beer

Decision Support Systems, DW, & OLAP

- Information technology to help the knowledge worker (executive, manager, analyst) make faster and better decisions.
- Data Warehouse is a DSS
- A data warehouse is an architectural construct of an information system that provides users with current and historical decision support information that is hard to access or present in traditional operational systems.
- Data Warehouse is not an Intelligent system
- On-Line Analytical Processing (OLAP) is an element of DSS

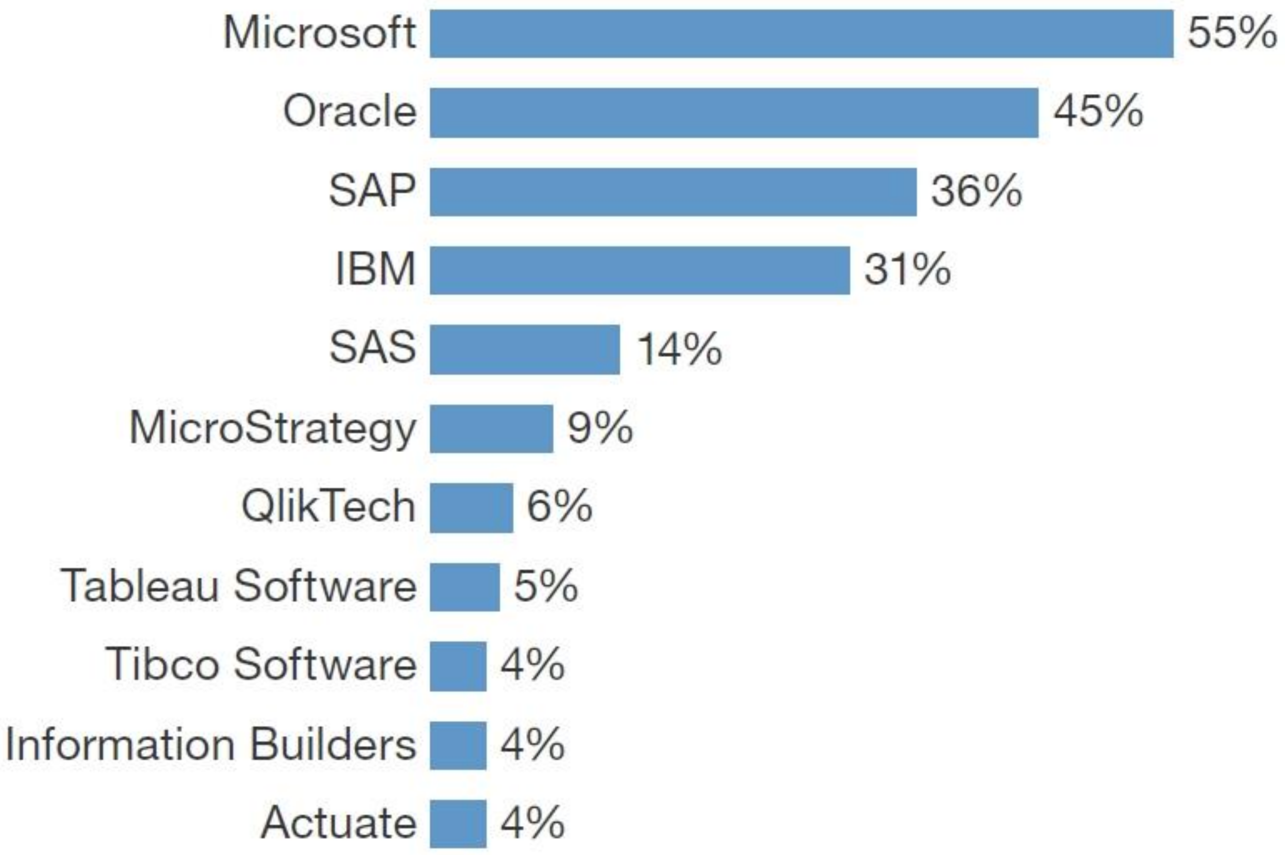
Worldwide Business Analytics Software 2014 Share Snapshot



Note: 2014 Share (%), Growth (%), and Revenue ($M)

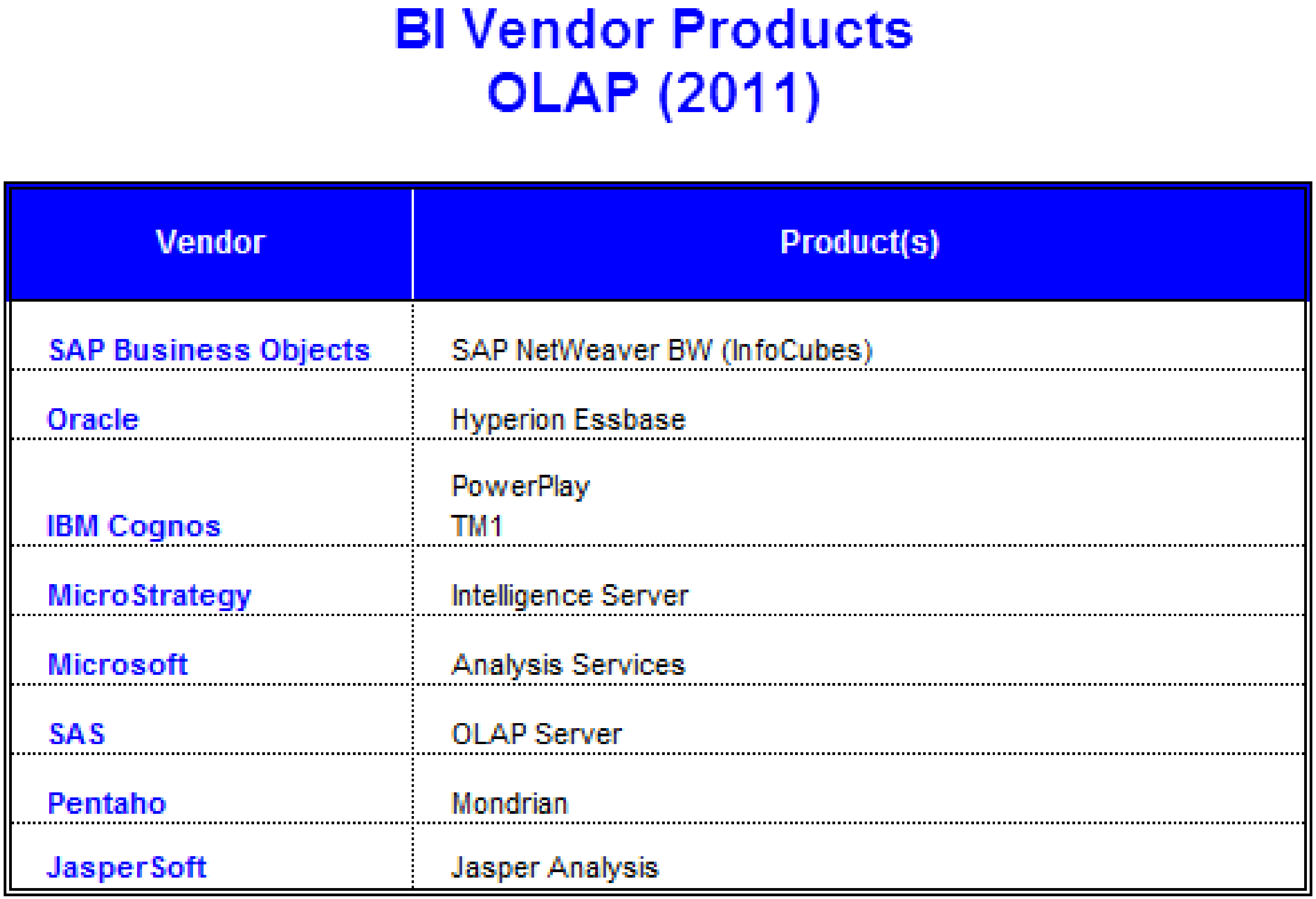
Source: IDC, 2015

“Which vendors’ BI tools do you currently use?”*



Base: 634 IT executives and technology decision-makers
(multiple responses accepted)

Data Warehouse: Major Players

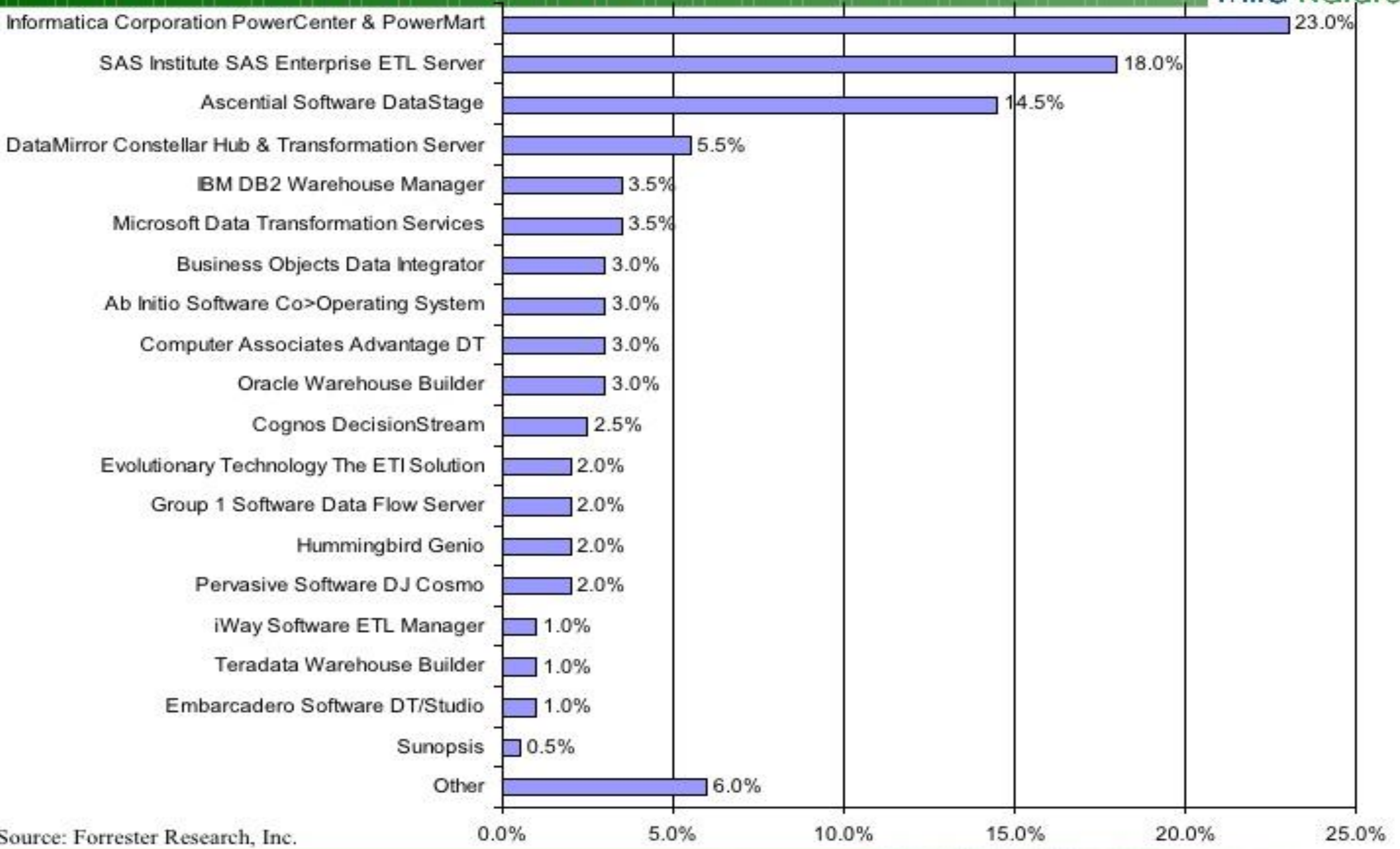


BI Vendor Products OLAP (2011)

Vendor	Product(s)
SAP Business Objects	SAP NetWeaver BW (InfoCubes)
Oracle	Hyperion Essbase
IBM Cognos	PowerPlay TM1
MicroStrategy	Intelligence Server
Microsoft	Analysis Services
SAS	OLAP Server
Pentaho	Mondrian
JasperSoft	Jasper Analysis

Vendor Market Share

Third Nature



Source: Forrester Research, Inc.

TDWI May 2007 Mark Madsen Slide 13

Data Warehouse: Characteristics

- Analysis driven
- Ad-hoc queries
- Complex queries
- Used by top managers
- Based on Dimensional Modeling
- Denormalized structures

Data Warehouse

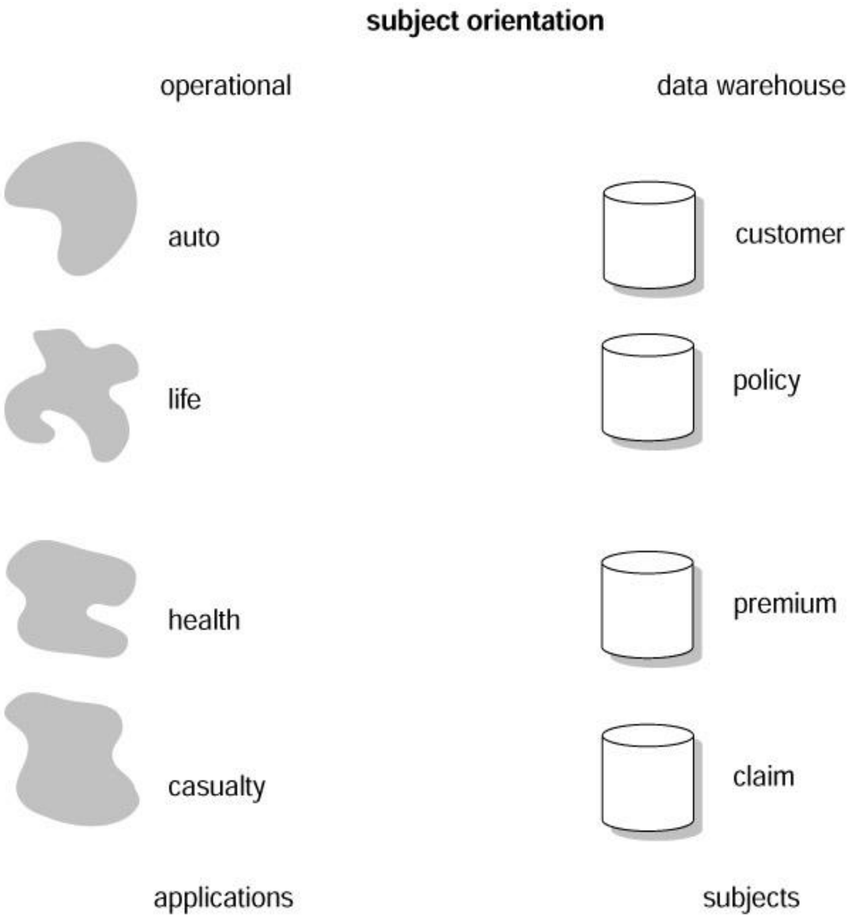
- A decision support database that is maintained separately from the organization’s operational databases.
- ”Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions.”
- A data warehouse is a
  - subject-oriented,
  - integrated,
  - time-varying,
  - non-volatile

“collection of data that is used primarily in organizational decision making” -Bill Inmon

R. Kimball’s definition of a DW

- A **data warehouse** is a copy of transactional data specifically structured for querying and analysis.
- According to this definition:
  - The form of the stored data (RDBMS, flat file) has nothing to do with whether something is a data warehouse.
  - Data warehousing is not necessarily for the needs of "decision makers" or used in the process of decision making.

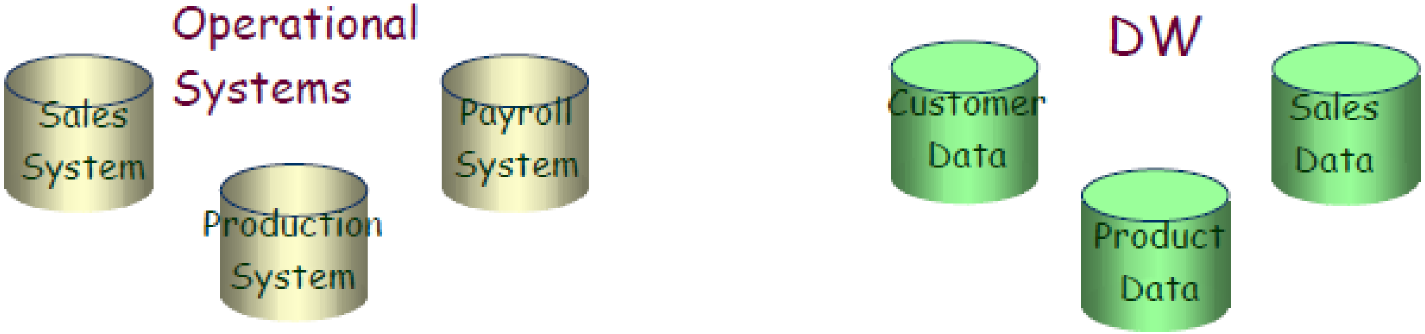
Subject-Oriented Data Collections



- ➢Classical operations systems are organized around the functional applications of the company.
- ➢For an insurance company, the applications may be auto, health, life, and casualty. The major subject areas of the insurance corporation might be customer, policy, premium, and claim.
- ➢For a manufacturer, the major subject areas might be product, order, vendor, bill of material, and raw goods.
- ➢For a retailer, the major subject areas may be product, SKU, sale, vendor, and so forth.
- ➢Each type of company has its own unique set of subjects

Data Warehouse — Subject-Oriented

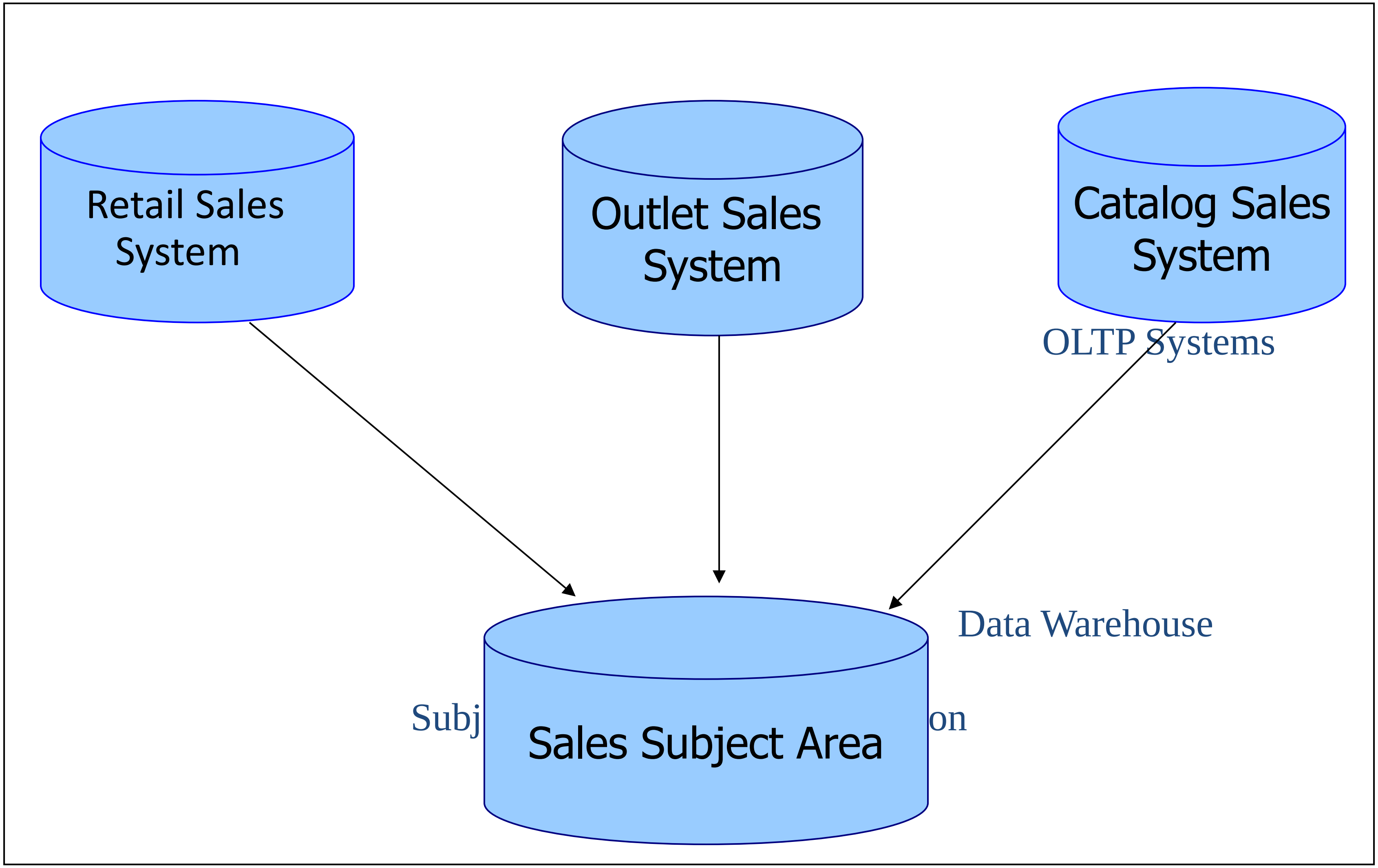
- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the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



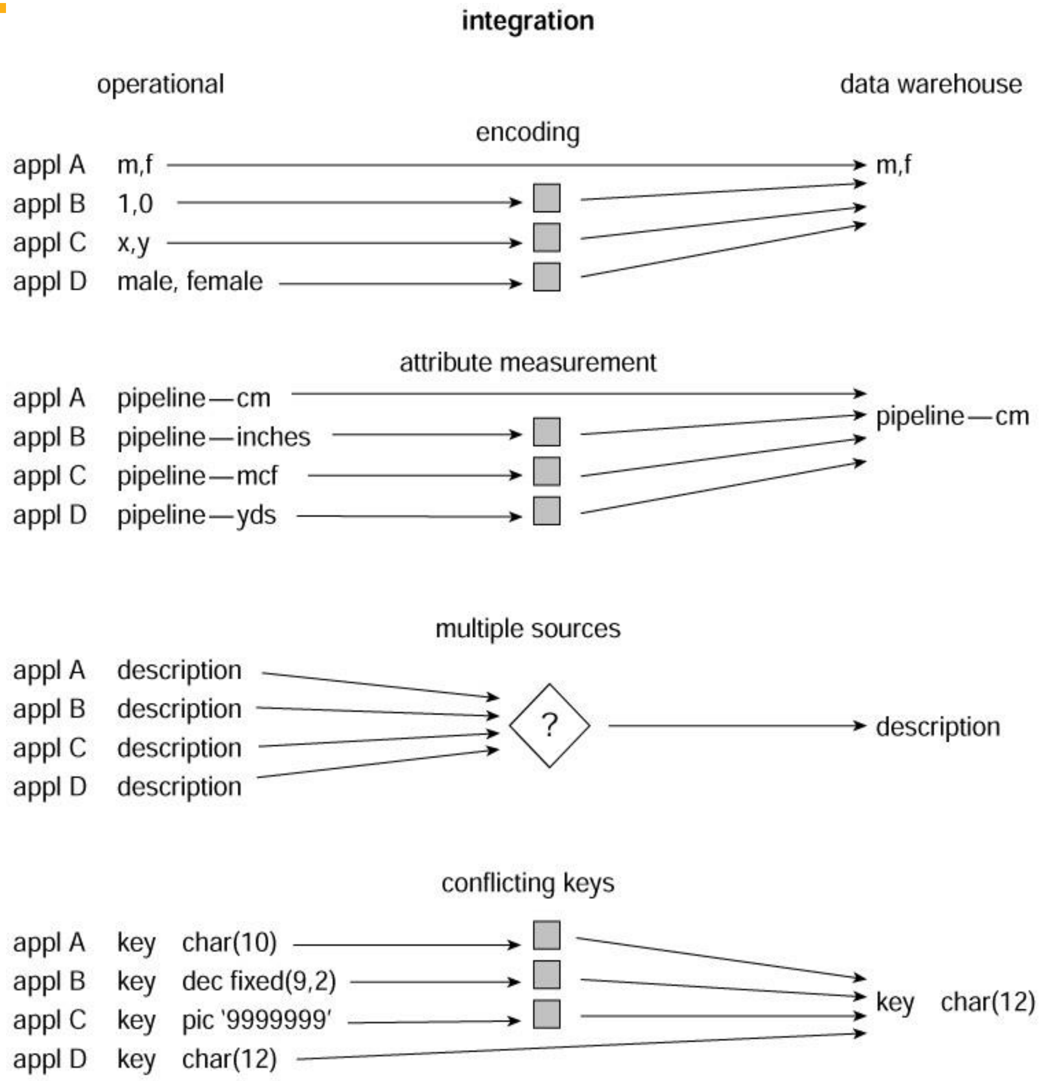
Subject Oriented

- Data Warehouse is designed around "subjects" rather than processes
- A company may have
  - Retail Sales System
  - Outlet Sales System
  - Catalog Sales System
- DW will have a Sales Subject Area

Subject Oriented



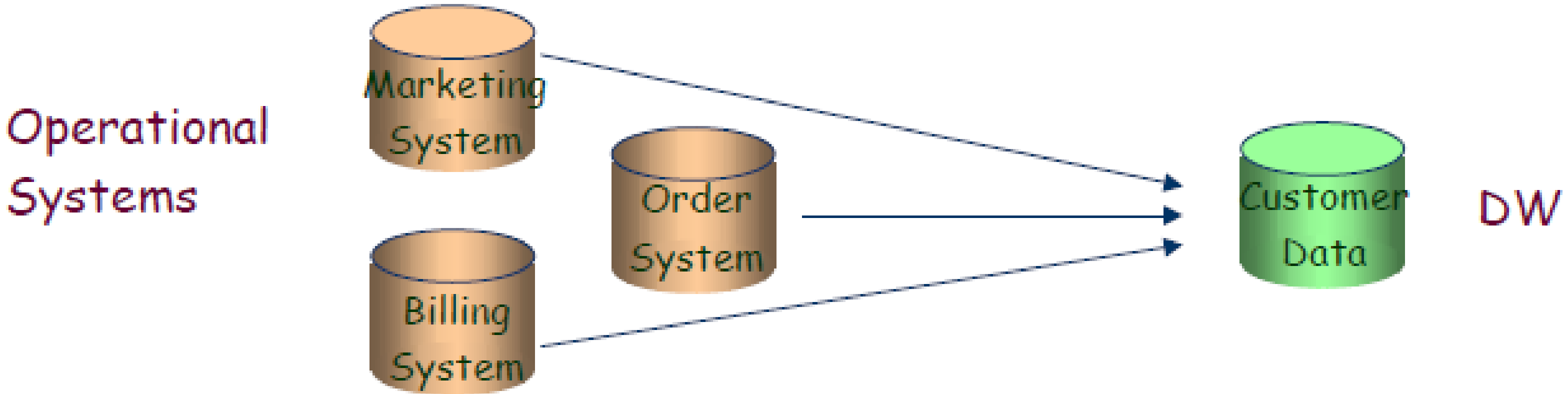
Integrated Data Collections



Of all the aspects of a data warehouse, integration is the most important. Data is fed from multiple disparate sources into the data warehouse. As the data is fed it is converted, reformatted,resequenced, summarized, and so forth. The result is that data—once it resides in the data warehouse—has a single physical corporate image.

Data Warehouse — Integrated

- Constructed by integrating multiple, heterogeneous data sources —— relational or other databases, flat files, external data
- Data cleaning and data integration techniques are applied. —— Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources —— When data is moved to the warehouse, it is converted.

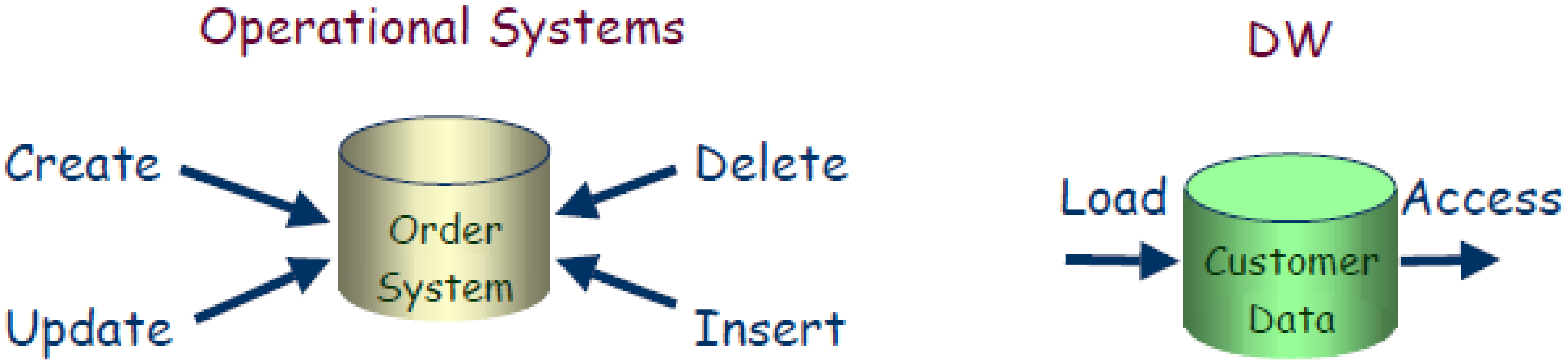


Integrated

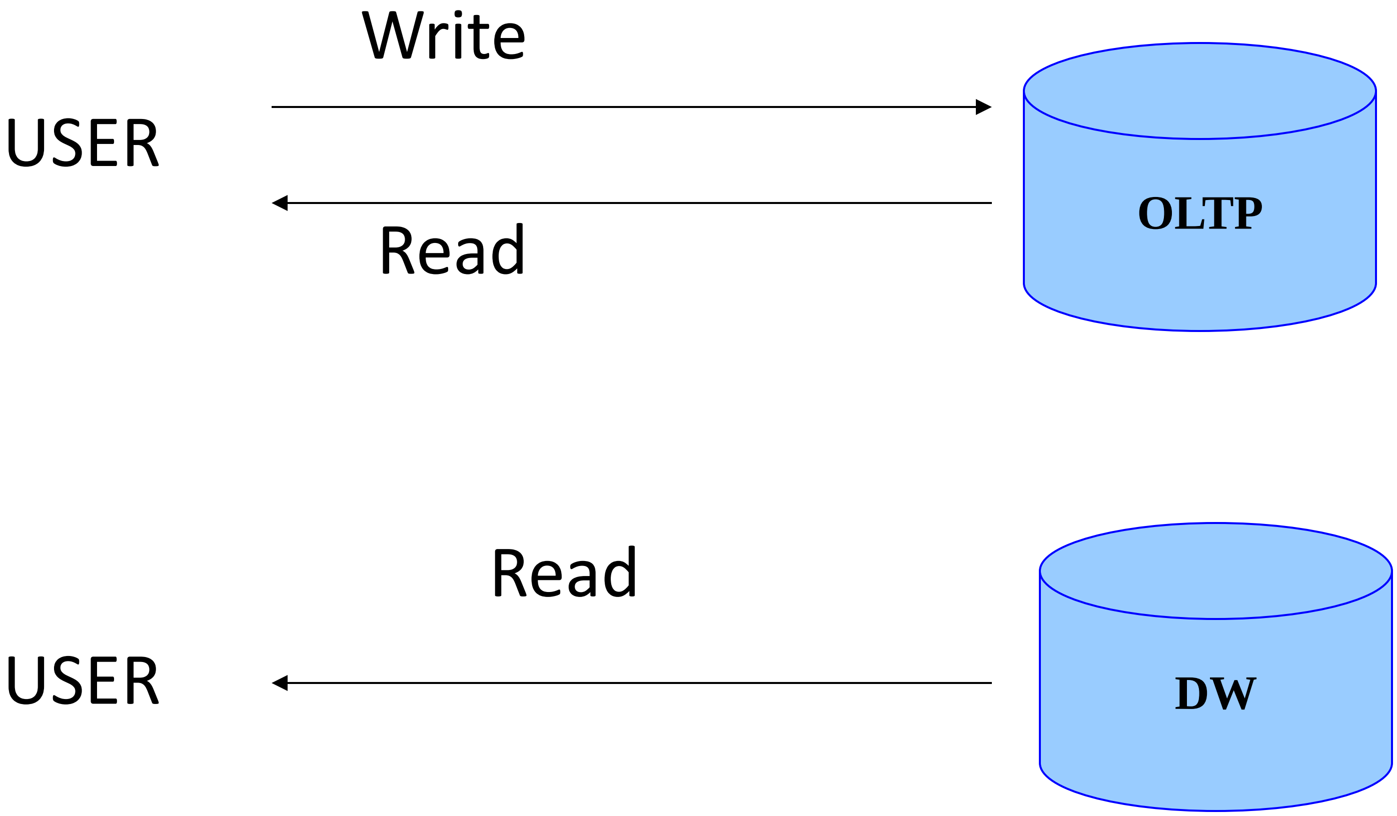
- Heterogeneous Source Systems
- Little or no control
- Need to Integrate source data
- For Example: Product codes could be different in different systems
- Arrive at common code in DW

Data Warehouse — Non-Volatile

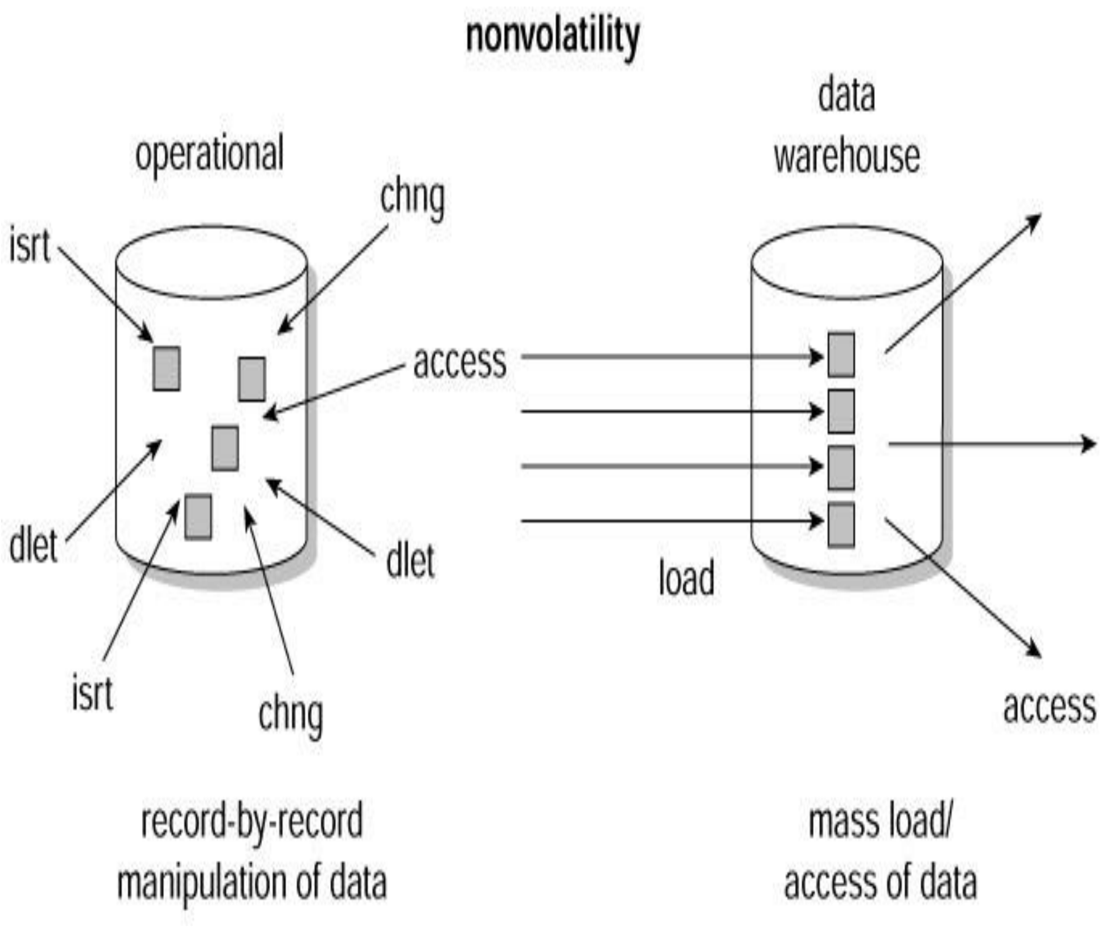
- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms



Non-Volatile (Read-Mostly)



Non-volatile Data Collections



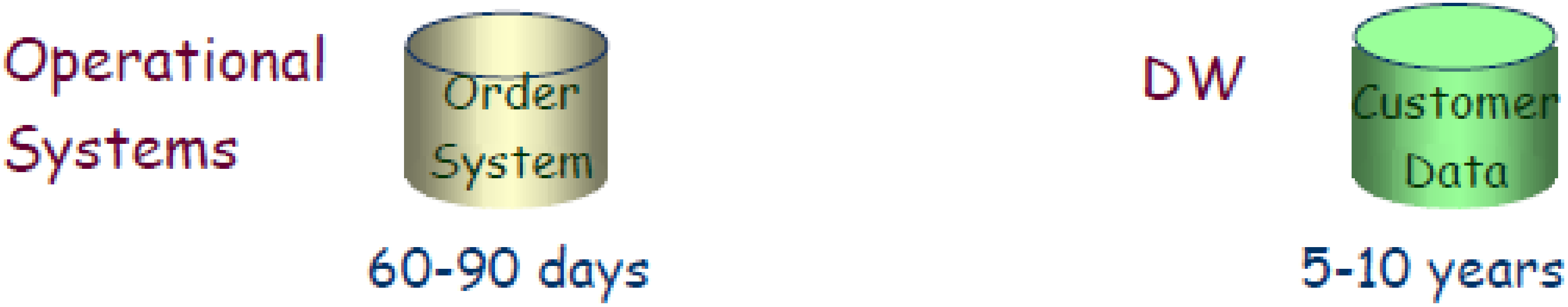
Data is updated in the operational environment as a regular matter of course, but warehouse data exhibits a very different set of characteristics. Data warehouse data is loaded and accessed, but it is not updated (in the general sense). Instead, when data in the data warehouse is loaded, it is loaded in a snapshot, static format. When subsequent changes occur, a new snapshot record is written. In doing so a history of data is kept in the data warehouse.

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

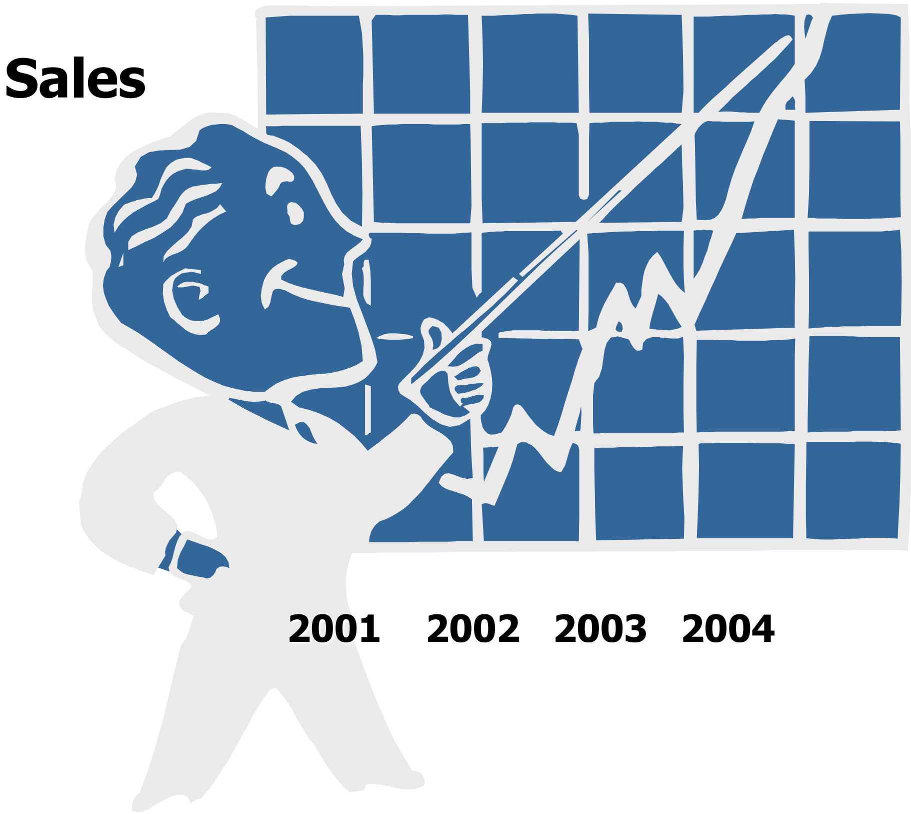
– But the key of operational data may or may not contain “time element”.



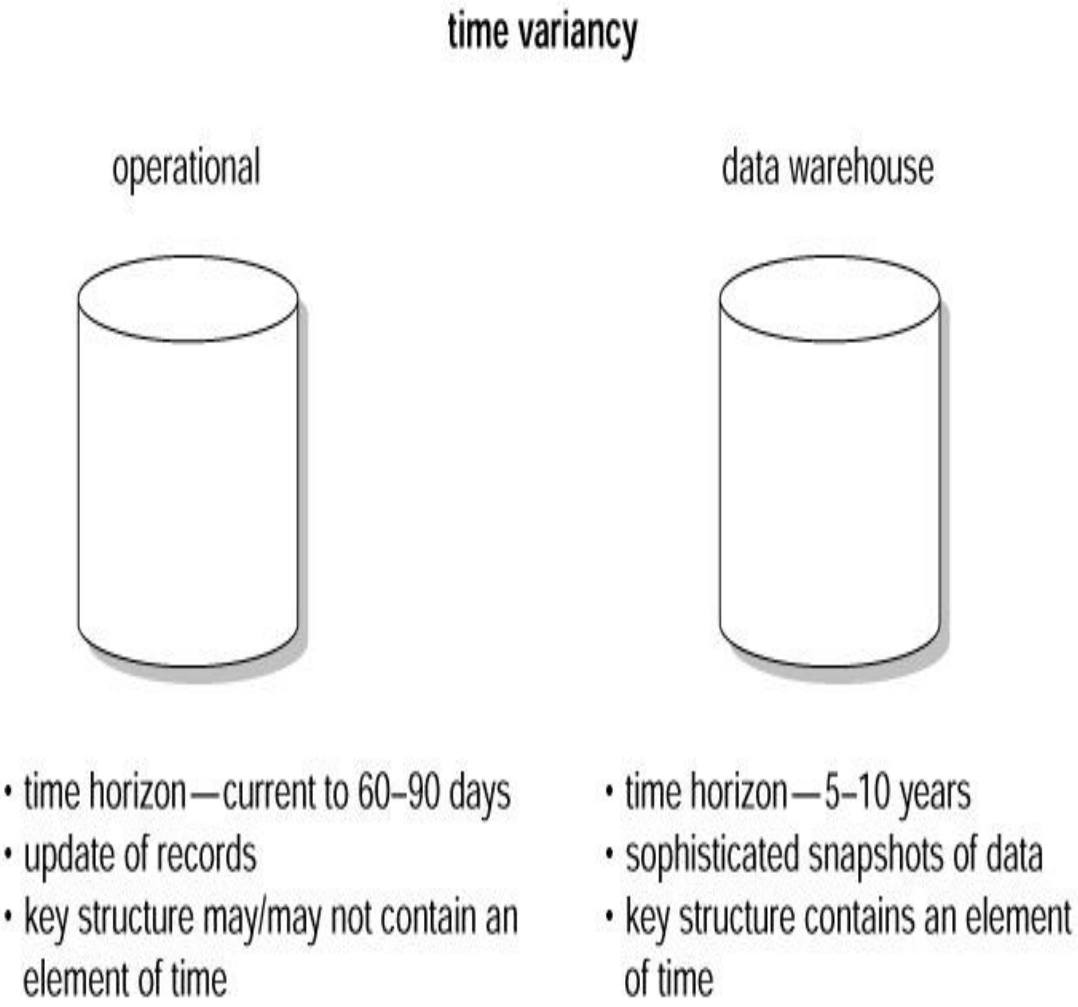
Time Variant

- Most business analysis has a time component

- Trend Analysis (historical data is required)



Time-variant Data Collections



- time horizon—current to 60–90 days
- update of records
- key structure may/may not contain an element of time

- time horizon—5–10 years
- sophisticated snapshots of data
- key structure contains an element of time

Time variancy implies that every unit of data in the data warehouse is accurate as of some one moment in time. In some cases, a record is time stamped. In other cases, a record has a date of transaction. But in every case, there is some form of time marking to show the moment in time during which the record is accurate. A 60-to-90-day time horizon is normal for operational systems; a 5-to-10-year time horizon is normal for the data warehouse. As a result of this difference in time horizons, the data warehouse contains *much* more history than any other environment.

The goals of a Data Warehouse

- The data warehouse must make an organization's information easily accessible.
- The data warehouse must present the organization's information consistently.
- The data warehouse must be adaptive and resilient to change.
- The data warehouse must be a secure bastion that protects our information assets.
- The data warehouse must serve as the foundation for improved decision making.
- The business community must accept the data warehouse if it is to be deemed successful.

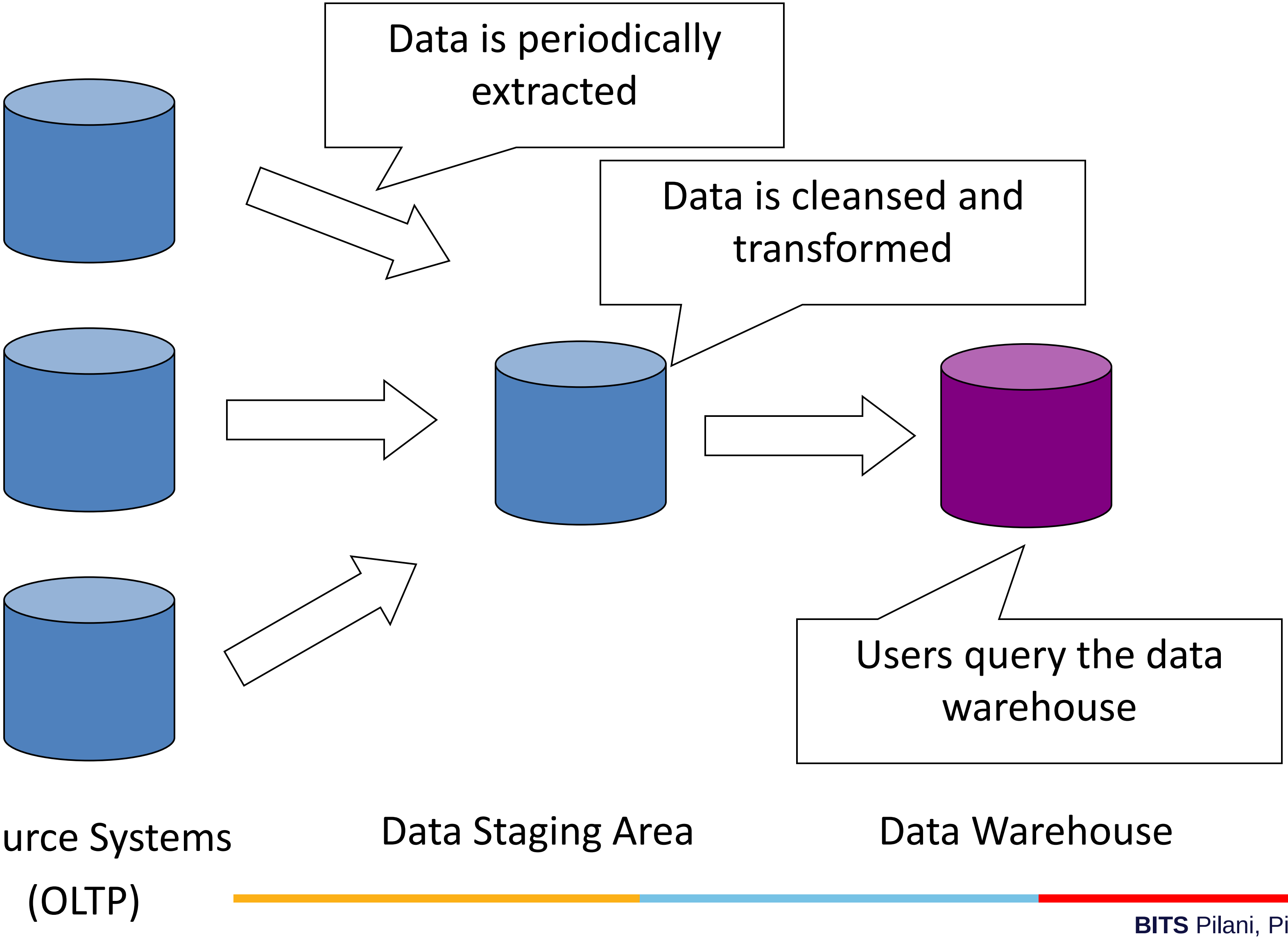
Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations (slice-dice, drilling, pivoting, etc)
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

Problems with Data Warehousing

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long-duration projects
- Complexity of integration

Loading the Data Warehouse



Data is periodically extracted

Data is cleansed and transformed

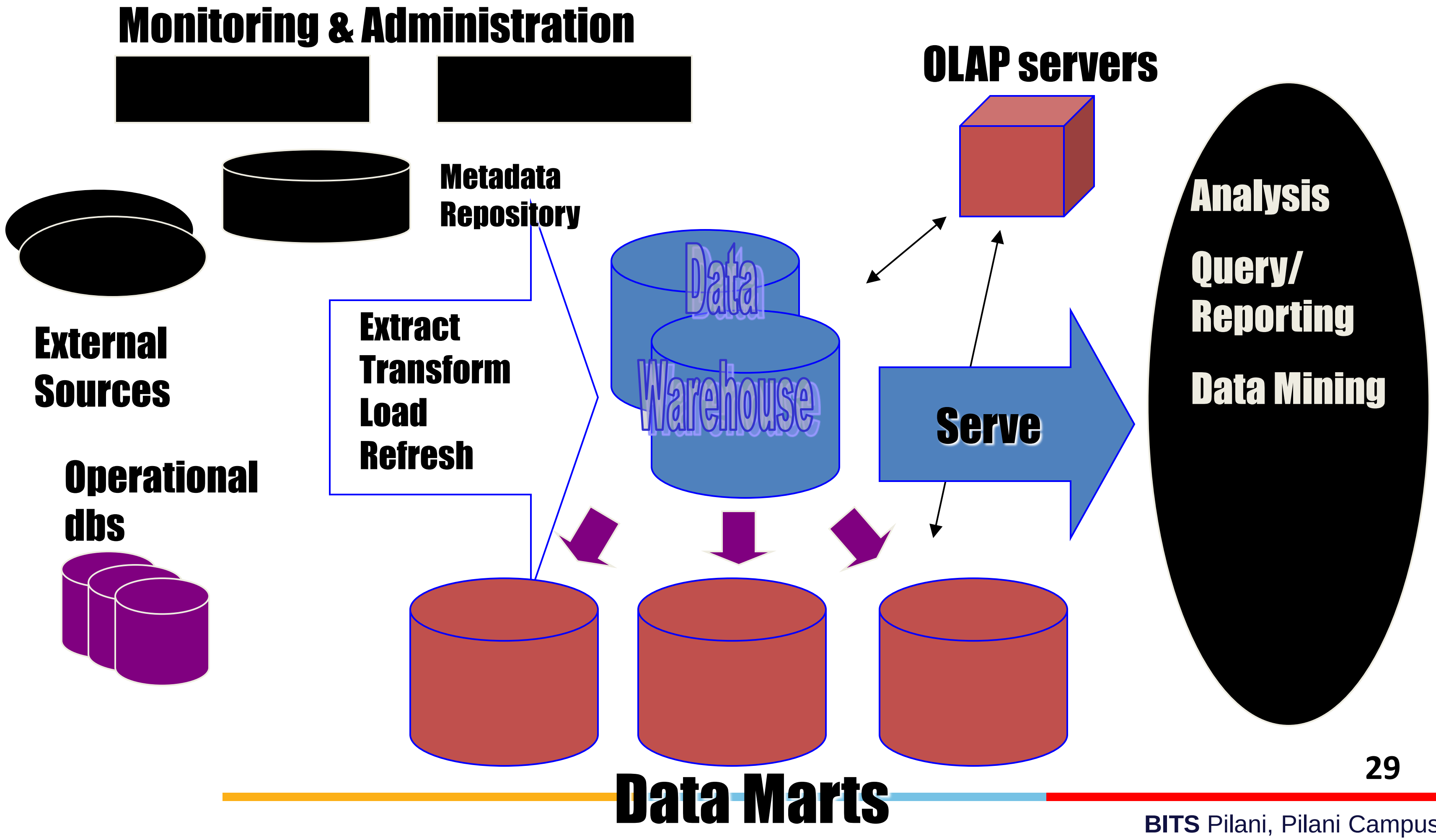
Users query the data warehouse

Source Systems (OLTP)

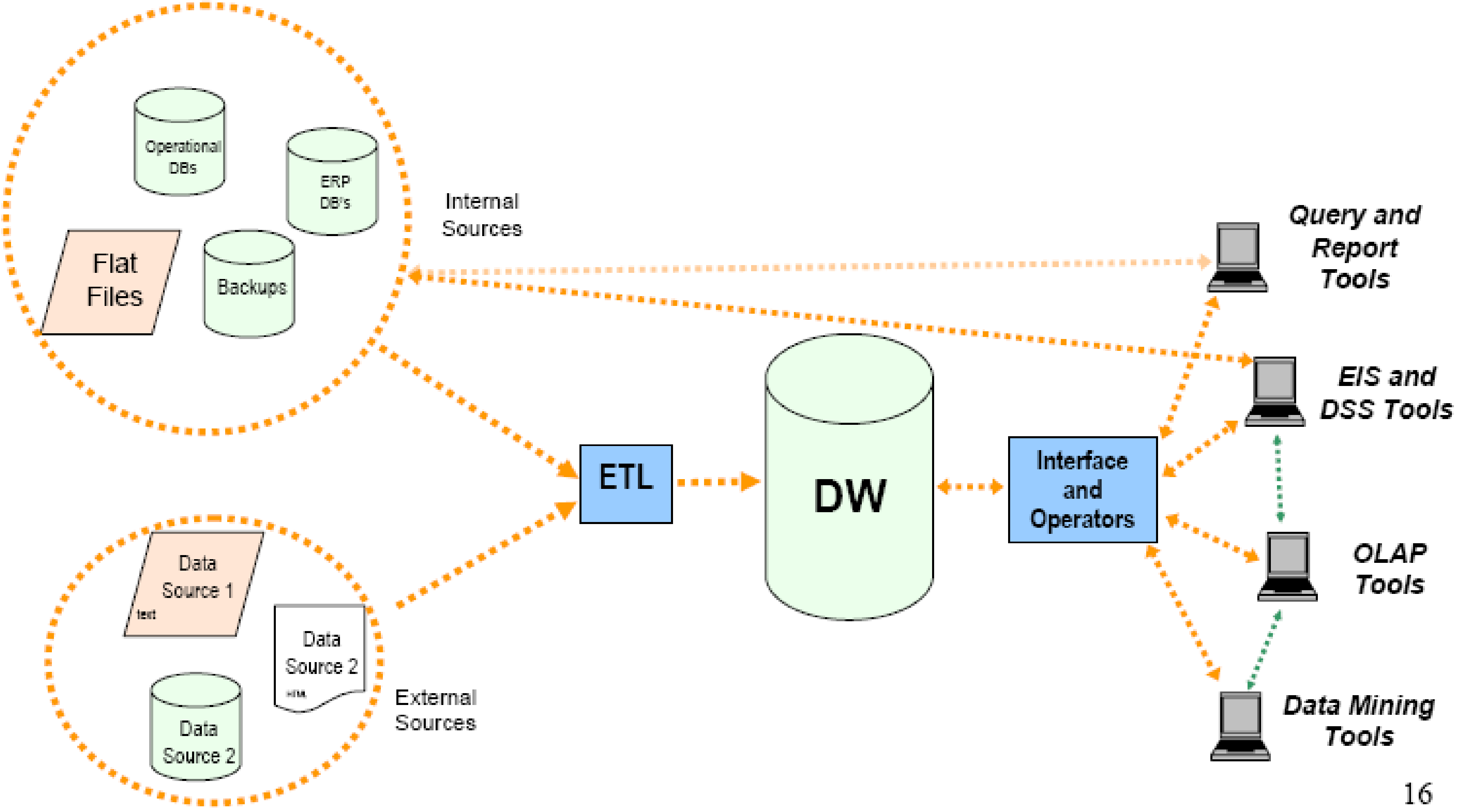
Data Staging Area

Data Warehouse

Data Warehousing Architecture



Data Warehousing Architecture


Operational DBs
ERP DB's
Flat Files
Backups
Internal Sources
Data Source 1
text
Data Source 2
HTML
Data Source 2
External Sources
ETL
DW
Interface and Operators
Query and Report Tools
EIS and DSS Tools
OLAP Tools
Data Mining Tools
16