Data warehousing and OLAP

Module-1

Agenda

- Data warehouse: Introduction to DW,
- DW architecture,
- ETL process, Top- down and Bottom-up approaches, characteristics, and benefits of Data Mart.
- Dimensional Modeling:
- Star, Snowflake, and Fact Constellation Schemas OLAP in the data warehouse: major features and functions, OLAP models- ROLAP and MOLAP, and the difference between OLAP and OLTP.

A Sales Manager wants to know….

Which are our lowest/highest margin customers ?

Who are my customers and what products are they buying?

What is the most effective distribution channel?

What product promotions have the biggest impact on revenue?

Which customers are most likely to go to the competition ?

What impact will new products/services have on revenue and margins?

Data, Data everywhere yet ...

- I can’t find the data I need
  - data is scattered over the network
  - many versions, subtle differences
- I can’t get the data I need
  - need an expert to get the data
- I can’t understand the data I found
  - available data poorly documented
- I can’t use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

What is a Data Warehousing?

Information
Data

A process of transforming data into information and making it available to users in a timely enough manner to make a difference

[Forrester Research, April 1996]

What is a Data warehouse

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

[Barry Devlin]

Data Warehouse

- A data warehouse is a

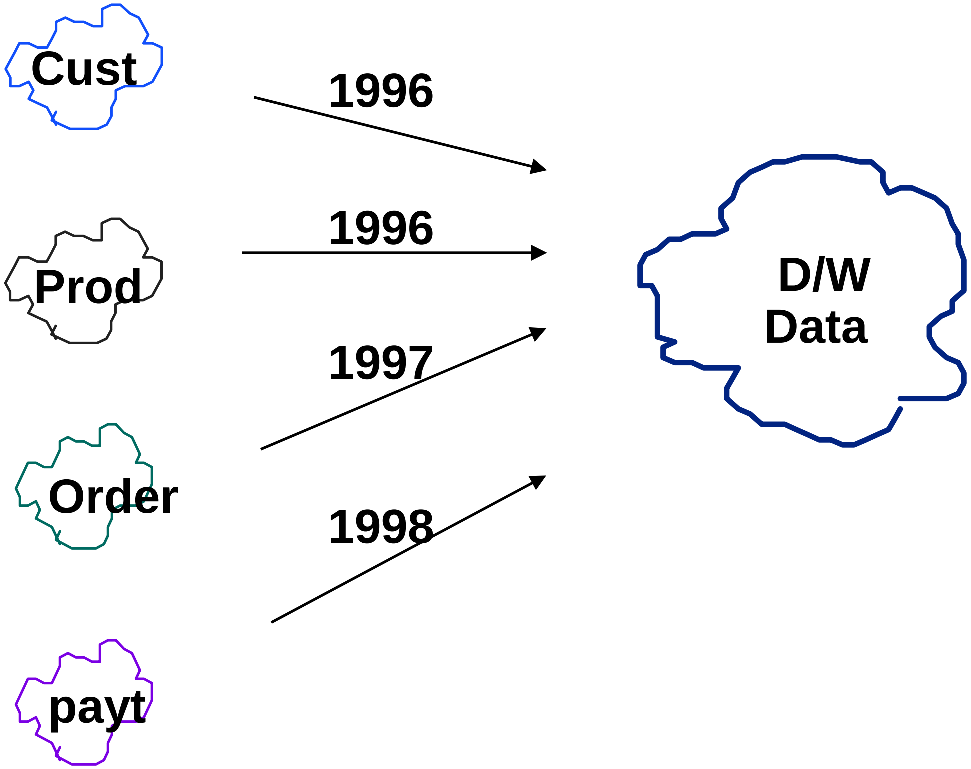
1. subject-oriented
2. integrated
3. time-varying
4. non-volatile

collection of data that is used primarily in organizational decision making.

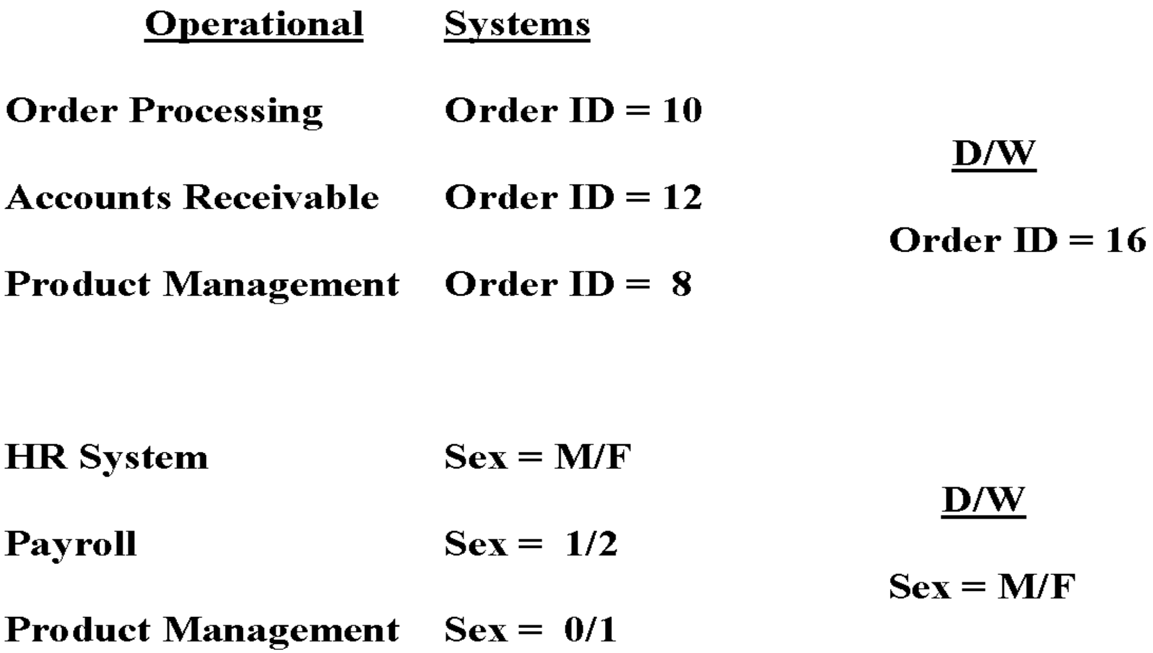
-- Bill Inmon, Building the Data Warehouse

Subject Oriented: Data is organized by business topic not by customer_id

Data is Integrated and Loaded by Subject



- **Integrate:** Integrated means that data are stored as a single unit not as a collection of files that may have different structures .
- Eg:



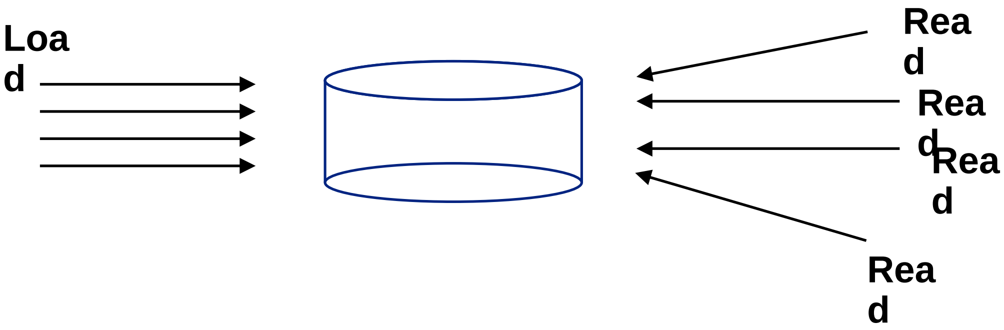
Operational	Systems	D/W
Order Processing	Order ID = 10	Order ID = 16
Accounts Receivable	Order ID = 12	
Product Management	Order ID = 8	
		D/W
HR System	Sex = M/F	Sex = M/F
Payroll	Sex = 1/2	
Product Management	Sex = 0/1	

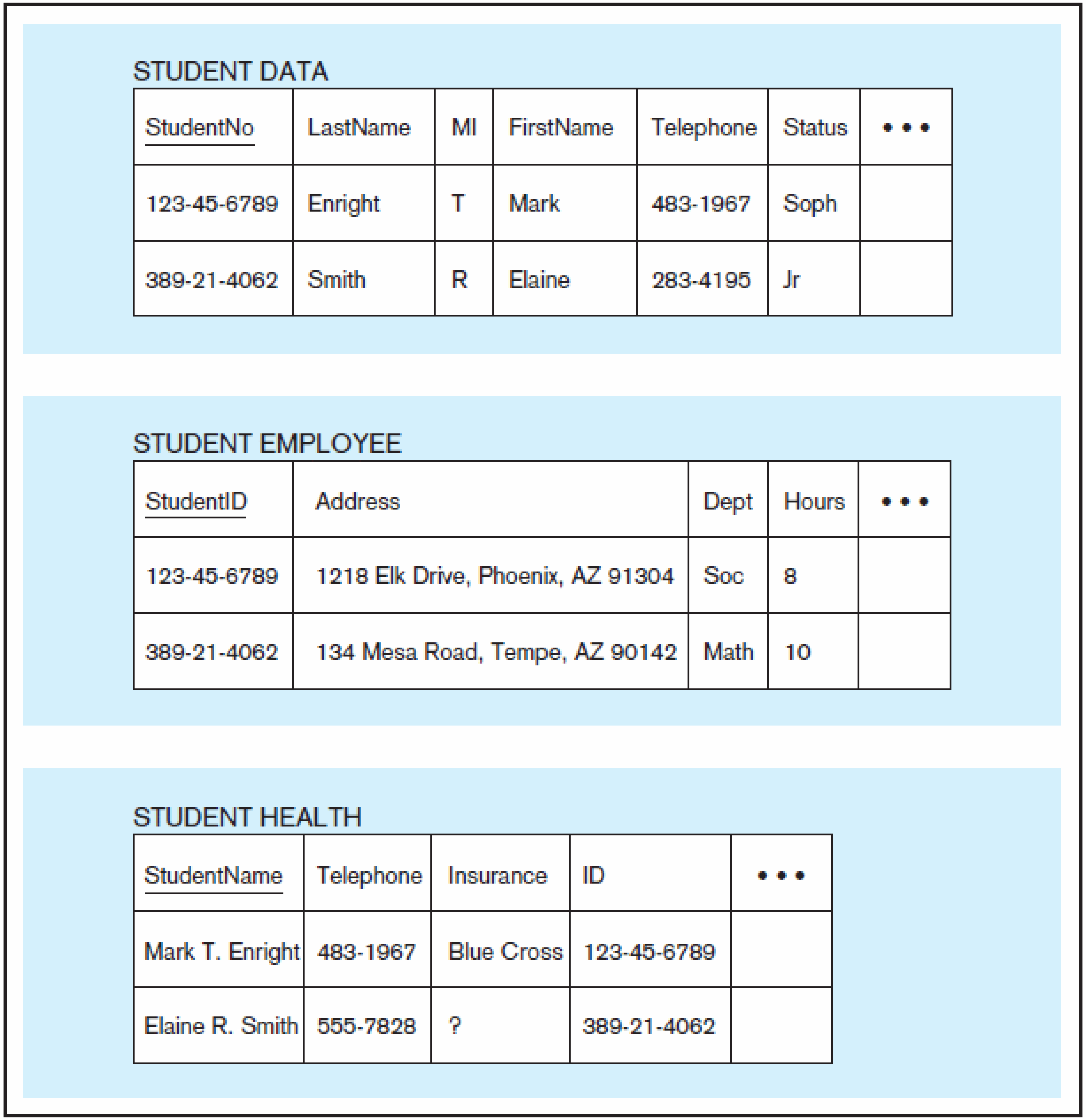
- **Time Variant:** it means that a time dimension is explicitly included in the data so that trends and changes over time can be studied .
- Data elements won’ t change
- i.e a single ,complete and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in business context
- Eg: Designated Time Frame (3 - 10 Years)
- Key Includes Date

Non-Volatile: It means that the data don’t keep changing, new data may be added on a scheduled basic but old data aren’t discarded

Data Warehouse

- **No Data Update**





STUDENT DATA

StudentNo	LastName	MI	FirstName	Telephone	Status	• • •
123-45-6789	Enright	T	Mark	483-1967	Soph	
389-21-4062	Smith	R	Elaine	283-4195	Jr	

STUDENT EMPLOYEE

StudentID	Address	Dept	Hours	• • •
123-45-6789	1218 Elk Drive, Phoenix, AZ 91304	Soc	8	
389-21-4062	134 Mesa Road, Tempe, AZ 90142	Math	10	

STUDENT HEALTH

StudentName	Telephone	Insurance	ID	• • •
Mark T. Enright	483-1967	Blue Cross	123-45-6789	
Elaine R. Smith	555-7828	?	389-21-4062	

Figure :Examples of heterogeneous data

Organizational Trends Motivating Data Warehouses

- No single system of records
- Multiple systems not synchronized
- Organizational need to analyze activities in a balanced way
- Customer relationship management
- Supplier relationship management

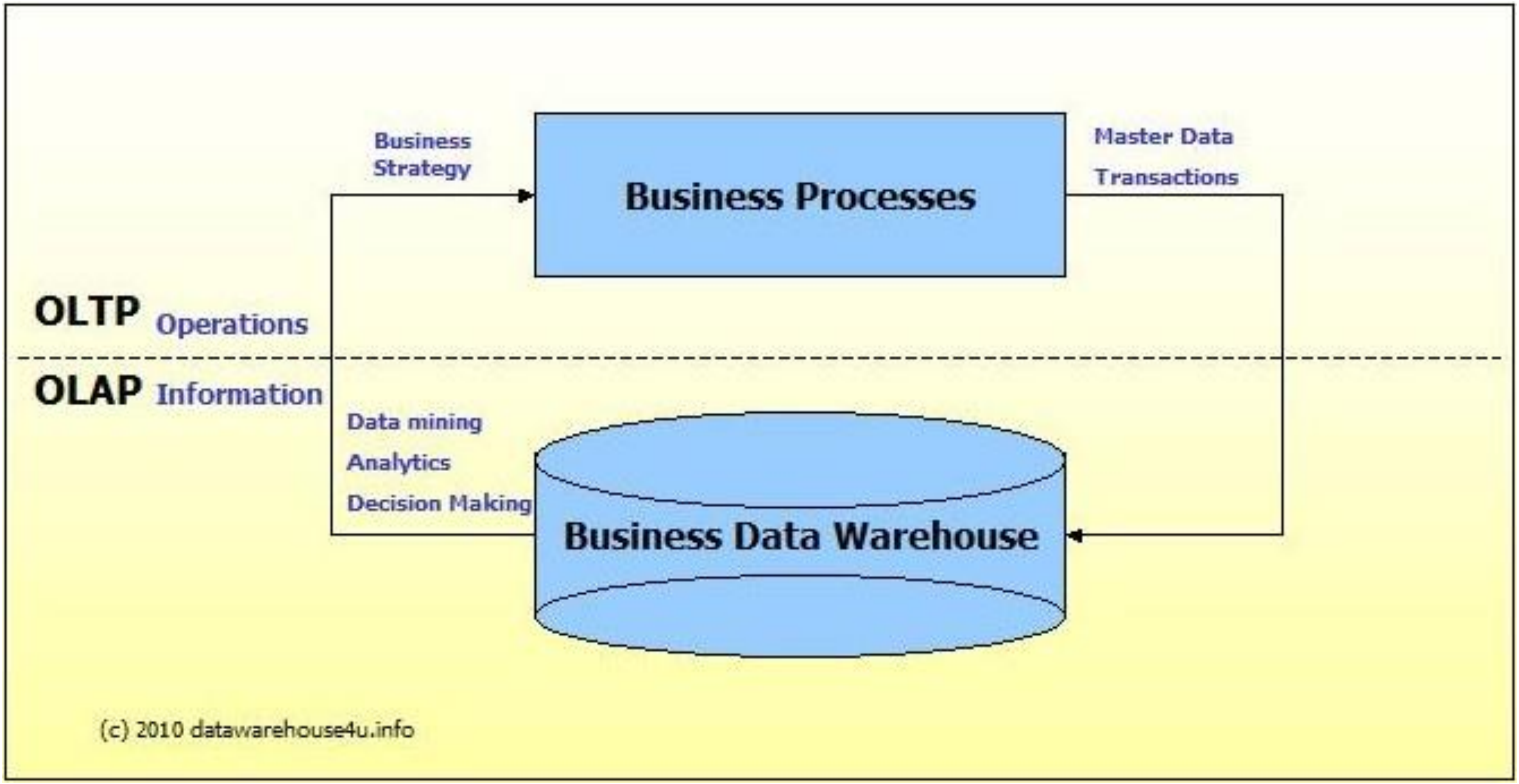
Separating Operational and Informational Systems

- **Operational system** – a system that is used to run a business in real time, based on current data; also called a system of record

- **Operational system** – a system that is used to run a business in real time, based on current data; also called a system of record

ODS Vs DW

ODS	DATAWARE HOUSE
Stores current operational data	Stores historic and current data
Perform numerous quick and simple queries on small amounts of data	Perform complex queries on large amounts of data
Contains small amount of transactional data	contains large amounts of static data
Day To Decisions, Current operational results, Tactical reporting,	Long-Term Decisions, Strategic reporting, Trend detection
Near to Normal form	Star Schema
Frequency of Load: Twice Daily , Daily, Weekly	Frequency of Load: Weekly, Monthly, Quarterly



(c) 2010 datawarehouse4u.info

OLTP vs. Data Warehouse

- OLTP systems are tuned for known transactions and workloads while workload is not known a priori in a data warehouse.

- Special data organization, access methods and implementation methods are needed to support data warehouse queries

    - e.g., *average amount spent on phone calls between 9AM-5PM in Pune during the month of December*

OLTP vs. Data Warehouse

- OLTP
  - Application Oriented
  - Used to run business
  - Detailed data
  - Current up to date
  - Isolated Data
  - Repetitive access
  - Clerical User

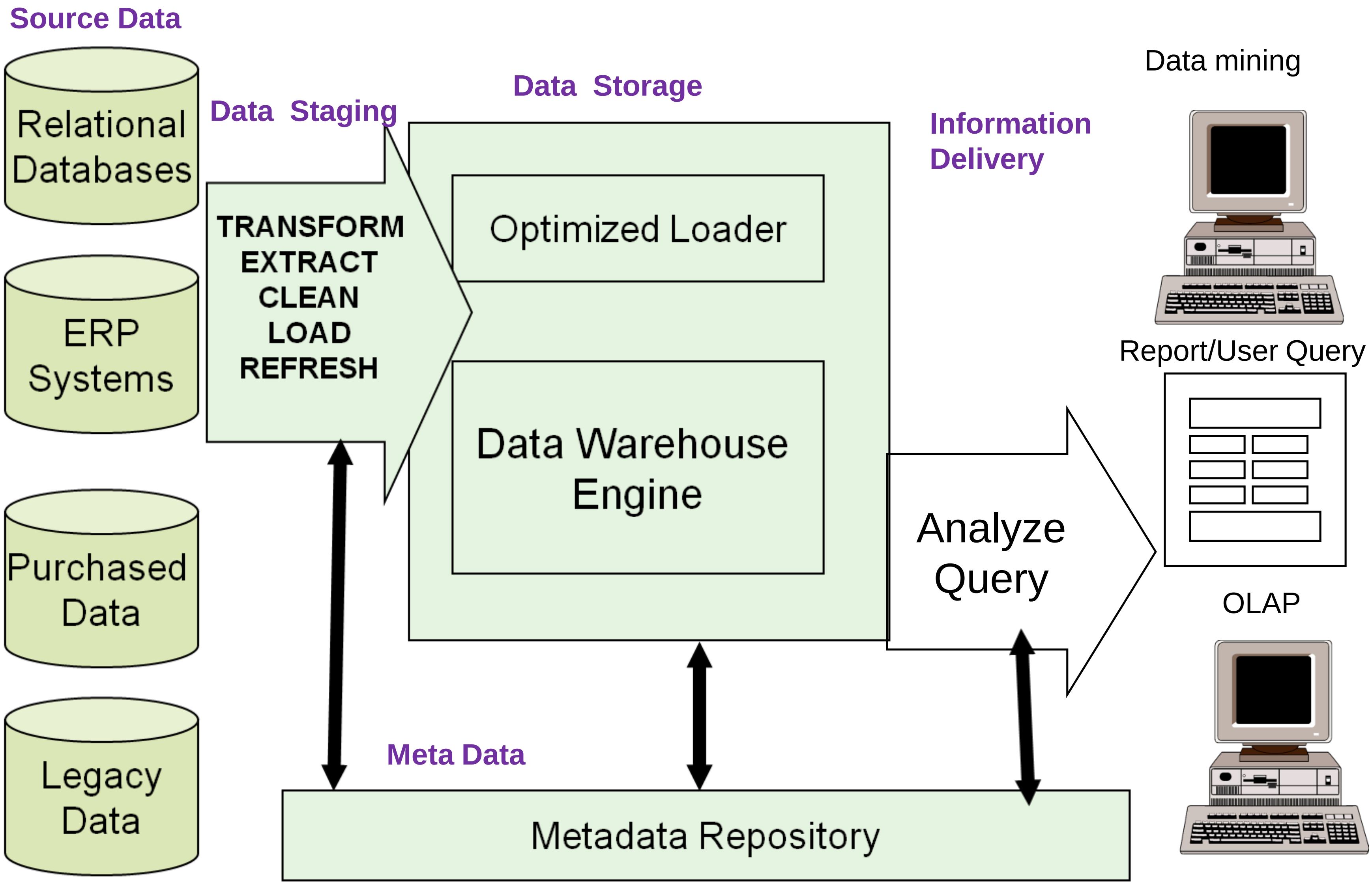
- Warehouse (DSS)
  - Subject Oriented
  - Used to analyze business
  - Summarized and refined
  - Snapshot data
  - Integrated Data
  - Ad-hoc access
  - Knowledge User (Manager)

Data Warehouse Architectures

1. Independent Data Mart
2. Dependent Data Mart and Operational Data Store
3. Logical Data Mart and Real-Time Data Warehouse
4. Three-Layer architecture

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

Data Warehouse Architecture



Overview of the Components

1. Source data component
2. Data staging component
3. Data storage Component
4. Information delivery component
5. Meta data component
6. Management and control component

Overview of the Components

1. **Source data component:**
   - *Production Data* : data comes from operational systems of the Enterprise
   - *Internal data:* in every organization keeps some private data
   - *Archived data*: it comes from back ups

Overview of the Components: Getting data into the DWH

Data staging component

- 5 steps of data staging

1. **Extraction**
2. **Transformation**
3. Cleansing
4. **Loading**
5. Summarization

Overview of the Components: Getting data into the DWH

Data staging component

- Extracting
  - Capture of relevant data from operational source in “as is” status

- Sources for data generally in legacy mainframes in, IMS,, DB2; more data today in relational databases on Unix

Overview of the Components: Getting data into the DWH

Data staging component

- **Transformation:**

  In the case Multiple input sources to a data warehouse, inconsistency can sometimes make data unusable. ***Transformation*** is the process of dealing with these inconsistencies

  Eg: Use of different names/formats

  Mumbai, Bombay

  Cust_id, C_id

  dd/mm/yy, mm/dd/yy

Overview of the Components: Getting data into the DWH

Data staging component

- Cleansing
  - It is necessary to go through data entered into DWH and make it error free. this process is called ***Data cleansing.***
  - This include missing data , incorrect data in one source, inconsistent data and conflicting data when two or more sources are involved.
  - Clean data is vital for the success of the warehouse
  - ***Eg:****Seshadri, Sheshadri, Sesadri, Seshadri S., Srinivasan Seshadri, etc. are the same person*

Overview of the Components: Getting data into the DWH

Data staging component

- **Loading**
  - It implies physical movement of data from the computers storing the source databases to that which will store the data warehouse.
  - Most common channel for the data movement process is a high-speed communication link.
  - It is always necessary to close off access to the DWH when the loading is taking place.

Overview of the Components: Getting data into the DWH

Data staging component

- Summarization : In which any desired summaries of the data warehouse data are precalculated for later use
  - Once the DWH database has been loaded it is possible to create summaries
  - Eg: base customer (1985-87)
      - custid, from date, to date, name, phone, dob
    - base customer (1988-90)
      - custid, from date, to date, name, credit rating, employer
    - customer activity (1986-89) -- monthly summary
    - customer activity detail (1987-89)
      - custid, activity date, amount, clerk id, order no

Overview of the Components: Getting data into the DWH

Data staging component

- Summarized data stored : advantages
  1. reduce storage costs
  2. reduce CPU usage
  3. increases performance since smaller number of records to be processed
  4. design around traditional high level reporting needs

Overview of the Components

Data storage Component

The data storage for data warehouse is a separate repository

Propagate updates on source data to the warehouse

- periodically (e.g., every night, every week) or after significant events
- refresh policy set by administrator based on user needs and traffic
- possibly different policies for different sources
- This function is time consuming.

Overview of the Components

Information delivery component

- In order to provide information to the wide community of data warehouse users the information delivery component includes different methods .
- E g: Online ,intra net, internet, email

Overview of the Components

Meta data component

Meta data are data that describe data in the data warehouse

- Administrative metadata
  - source databases and their contents
  - gateway descriptions
  - warehouse schema, view & derived data definitions
  - dimensions, hierarchies
  - pre-defined queries and reports
  - data mart locations and contents
  - data partitions
  - data extraction, cleansing, transformation rules
  - data refresh
  - user profiles, user groups
  - security: user authorization, access control

- **Meta data component**
- Business meta data
  - business terms and definitions
  - ownership of data
- operational metadata
  - data lineage: history of migrated data and sequence of transformations applied
  - currency of data, active, archived,
- End- user meta data.

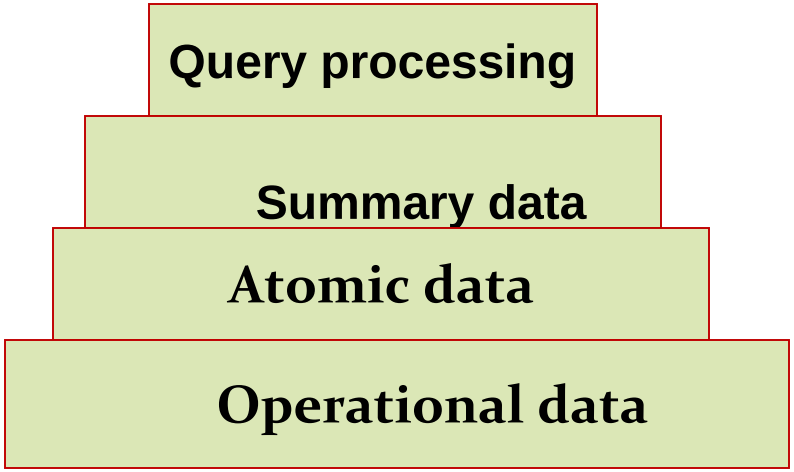
Management and control component

- This component of the data warehouse sits on the top of all other components. it coordinates services and activities with within the DWH

Content of D Warehouse Database

Operational data to warehouse data
4 data levels:

1. Operational data:
2. Atomic data
3. Summary data
4. Query processing



Eg1: My checking account balance right now is “ 1200/-Rs”
Eg2: My checking account balance at the end of January was “25000/-Rs”
Eg3: At the end of January 2011 we have 12500 customers
Eg4: Our customers in Mumbai zone grew by 10% during last three months

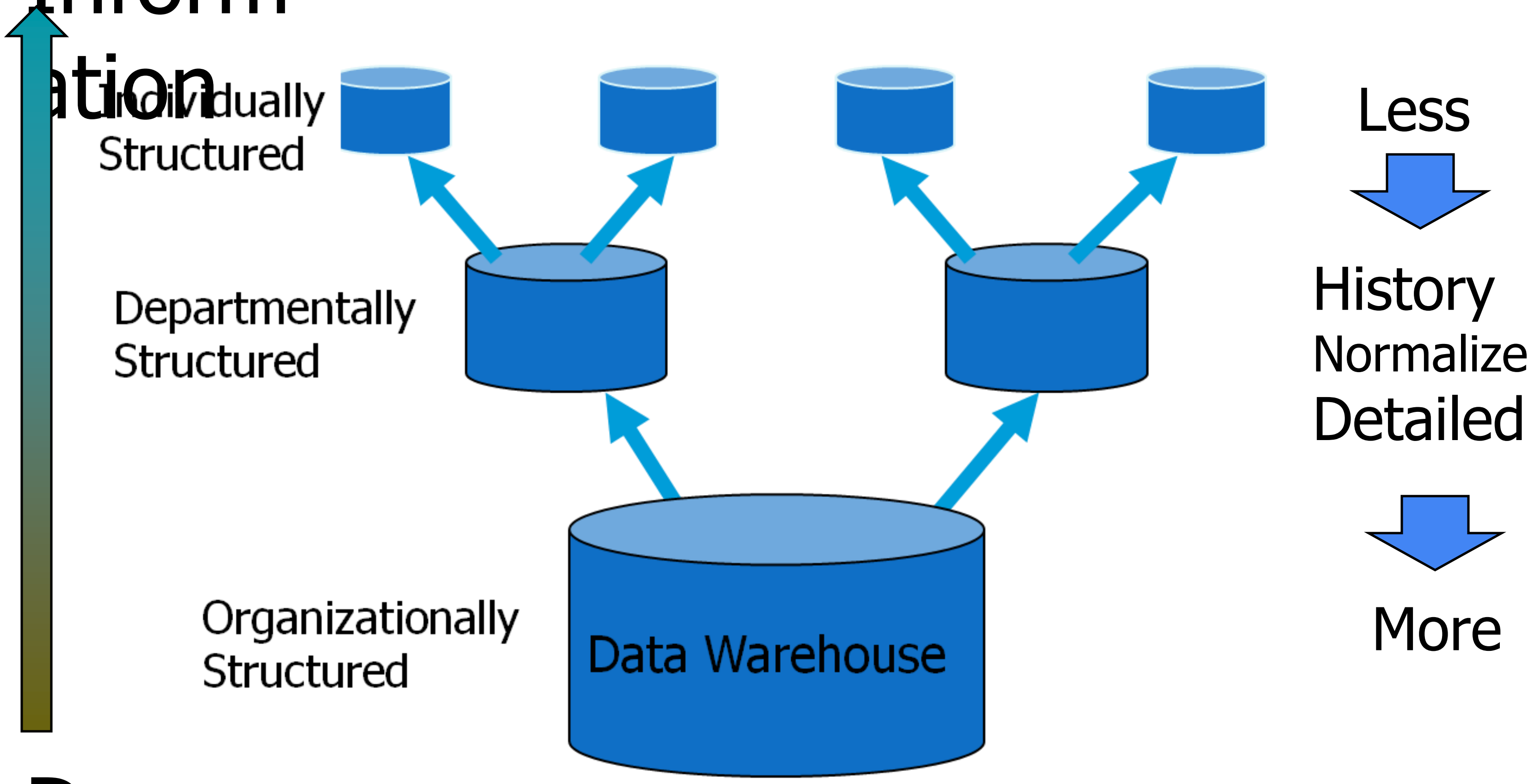
Data Warehouse vs. Data Marts

Data Mart

- Data mart
- ***Data mart***→ a subset of a data warehouse that supports the requirements of particular department or business function.
- Data mart focuses on only the requirements of users associated with one department or business function
- Data marts do not normally contain detailed operational data, unlike data warehouses
- As data marts contain less data compared with data warehouses, data marts are more easily understood and navigated

From the Data Warehouse to Data Marts

Inform ation



Individually Structured

Departmentally Structured

Organizationally Structured

Less

History
Normalized
Detailed

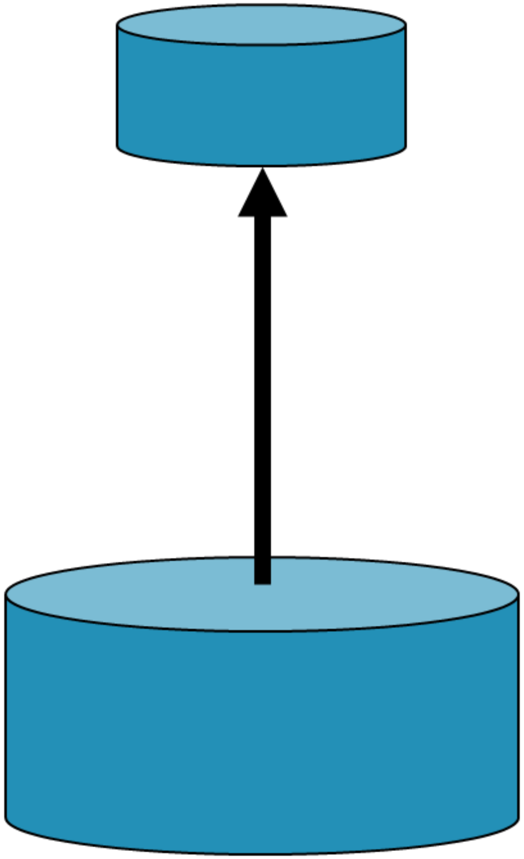
More

Da ta

Reasons for creating a data mart

1. To give users access to the data they need to analyze most often
2. To provide data in a form that matches the collective view of the data by a group of users in a department or business function
3. To improve end-user response time due to the reduction in the volume of data to be accessed
4. To provide appropriately structured data as dictated by the requirements of end-user access tools
5. data cleansing, loading, transformation, and integration are far easier, and hence implementing and setting up a data mart is simpler.
6. The cost of implementing data marts is normally less
7. The potential users of a data mart are more clearly defined and can be more easily targeted to obtain support for a data mart project

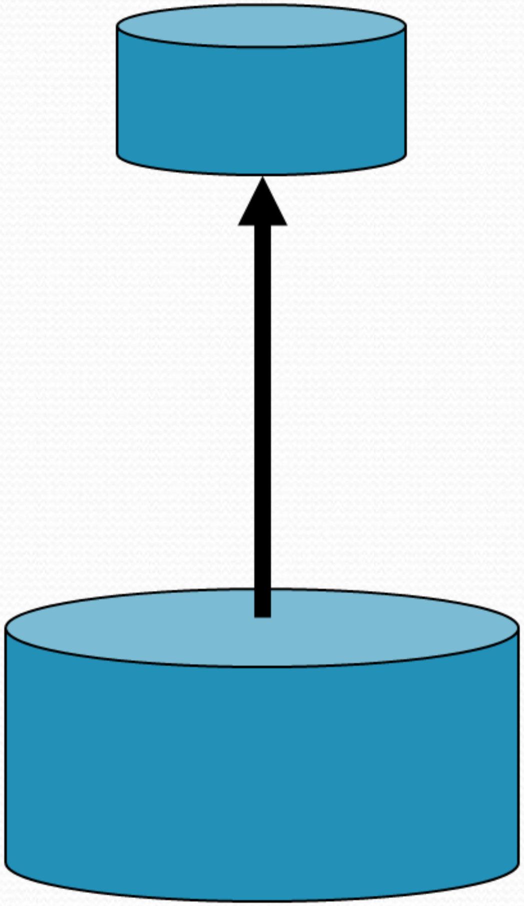
Data Warehouse and Data Marts



OLAP
Data Mart
Lightly summarized
Departmentally structured

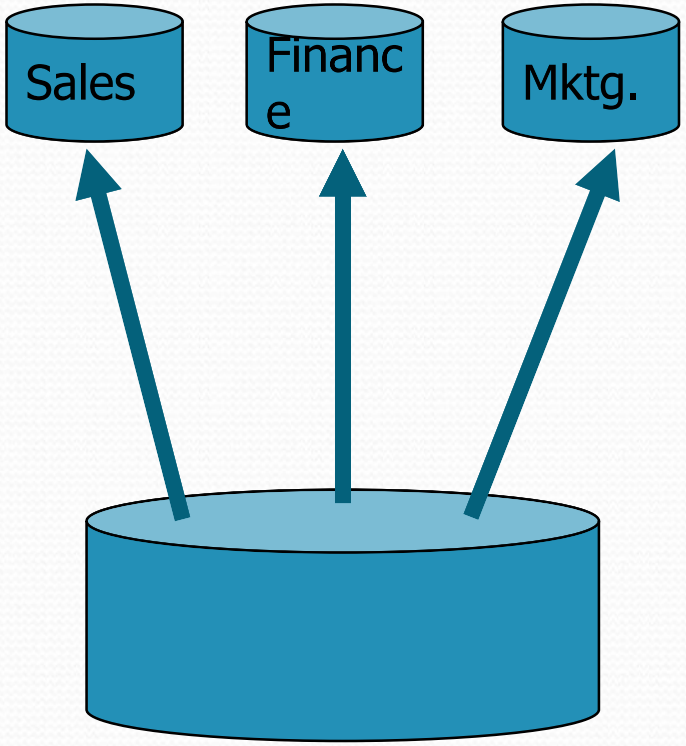
Organizationally structured
Atomic
Detailed Data Warehouse Data

Characteristics of the Departmental Data Mart



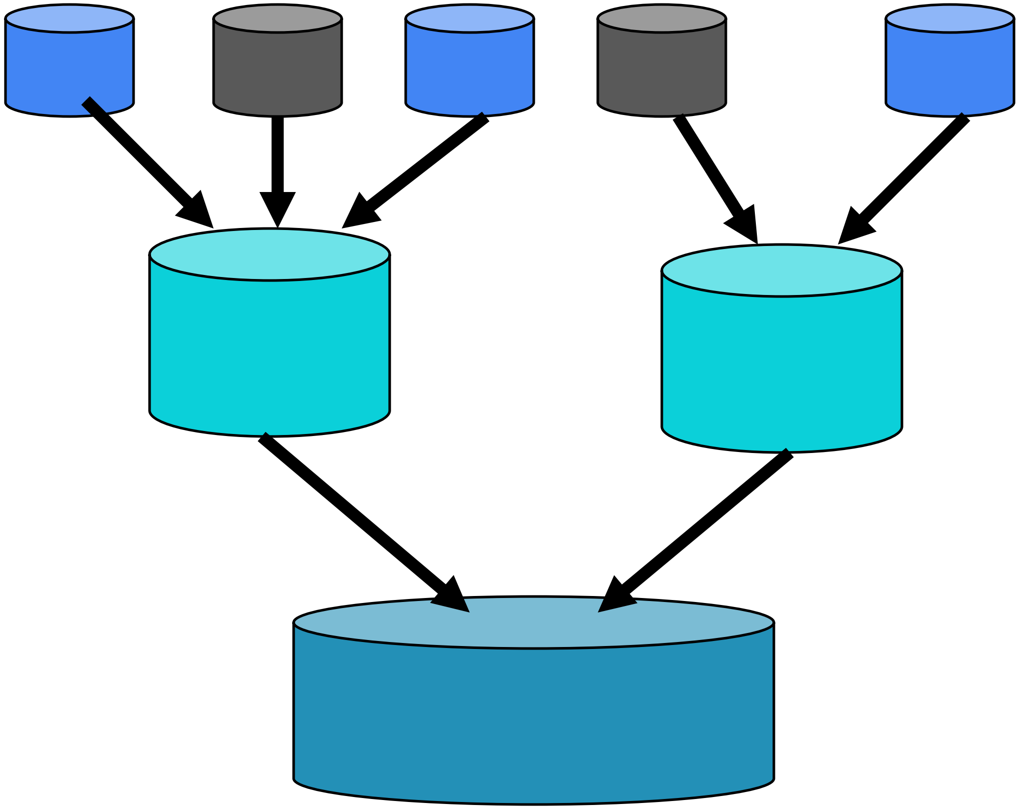
- OLAP
- Small
- Flexible
- Customized by Department
- Source is departmentally structured data warehouse

Techniques for Creating Departmental Data Mart



- OLAP
- Subset
- Summarized
- Superset
- Indexed
- Arrayed

Data Mart Centric

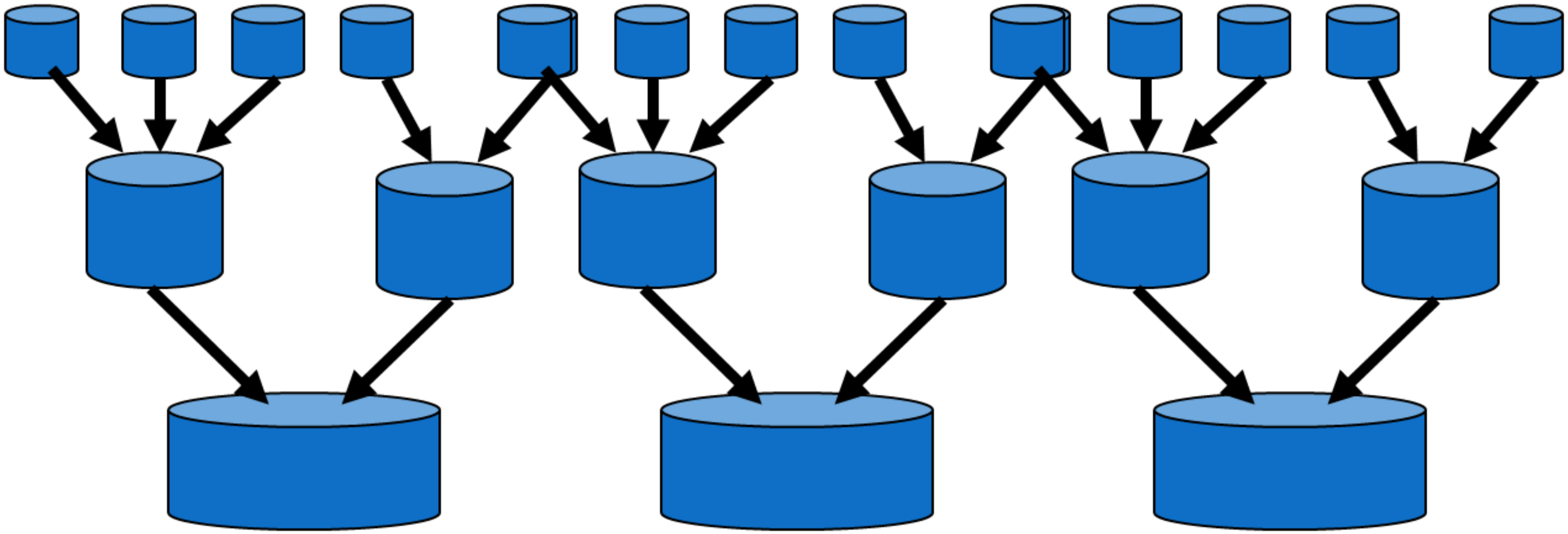


Data Sources

Data Marts

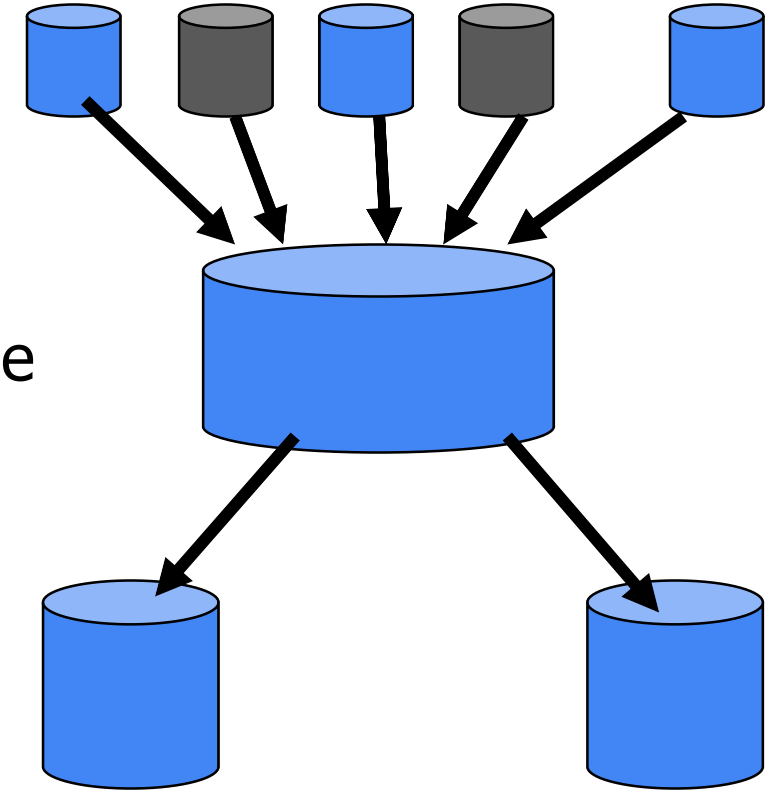
Data Warehouse

Problems with Data Mart Centric Solution



If you end up creating multiple warehouses, integrating them is a problem

True Warehouse



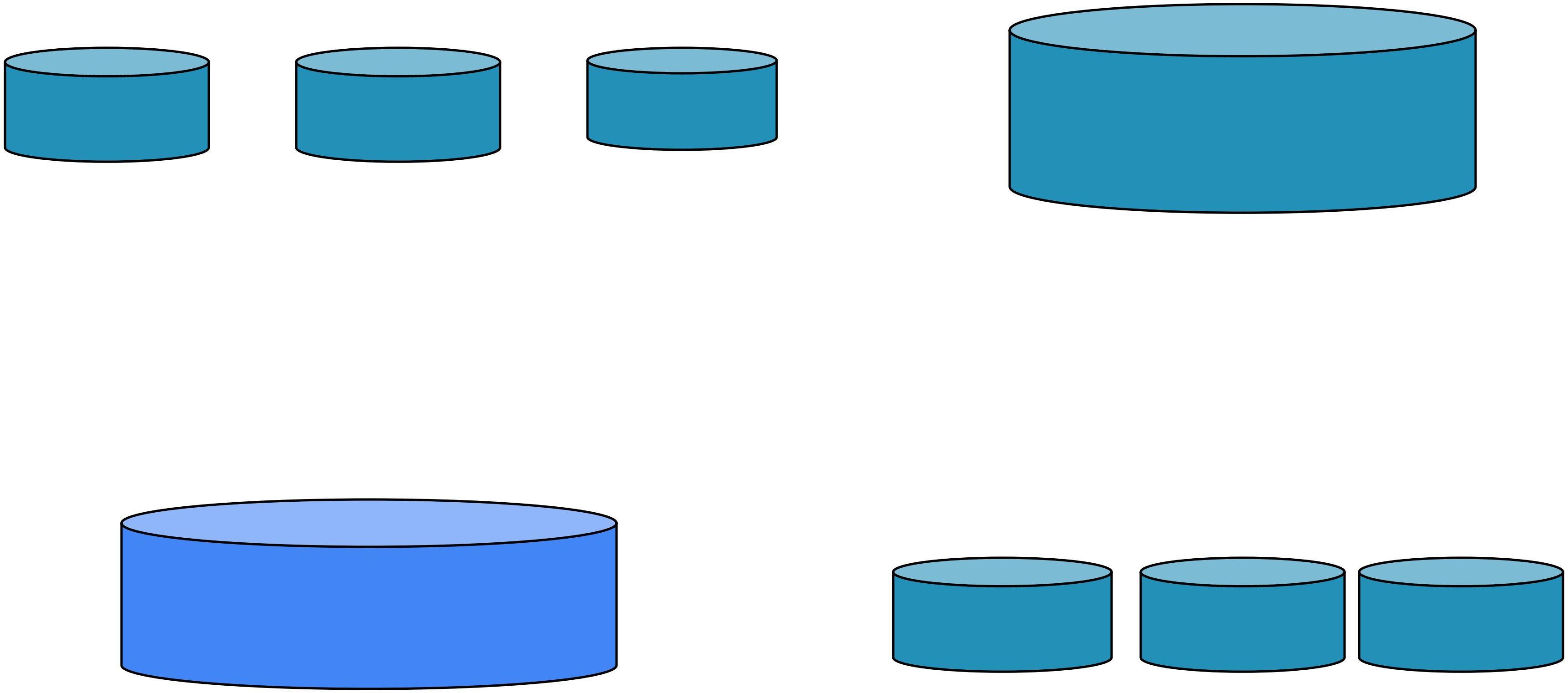
Data Sources

Data Warehouse

Data Marts

Bill Inman Vs Ralph Kimball

- Top -down Vs Bottom -up w.r.t Data Mart



A Comparison DM vs DWH

DATA MART	DATAWARE HOUSE
Limited subject areas (one or two)	Many subject areas
One data mart for one business theme/ subject area	Enterprise repository with multiple data marts
Has limited dimensions and measures depends on the subject area	Has all dimensions and measures required
Time to build is short	Long time activity
Can be built as smaller scale DW	Large scale

•OLAP is a valuable tool for analyzing a wide assortment of business data.
•It can be used to track and monitor a company's day-to-day operations, as well as for forecasting and planning purposes.
•OLAP provides a multitude of ways to parse and present data, making it ideal for business intelligence, decision support and data mining.