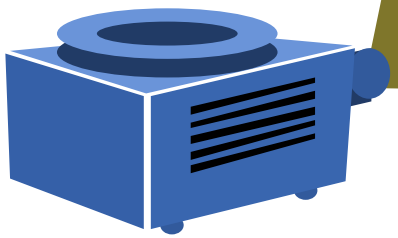
Data Warehouse definition



What is Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses

Data Warehouse: Definition

- Data Warehouse: An enterprise-wide structured repository of subject-oriented, time-variant, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data.
  (Oracle Data Warehouse Method)

Data Warehousing has the Following characteristics:

1. A central database that is loaded from multiple operational databases for the purpose of end-user access and decision support.
2. A data warehouse differs from an operational system in that the data it contains is normally static and updated in a scheduled manner through massive loading procedures.

Data Warehousing characteristics: *Continued*

3. A data warehouse is developed to accommodate random, ad hoc queries and to allow users to ‘drill down’ to minute levels of detail.

Very Large Data Bases

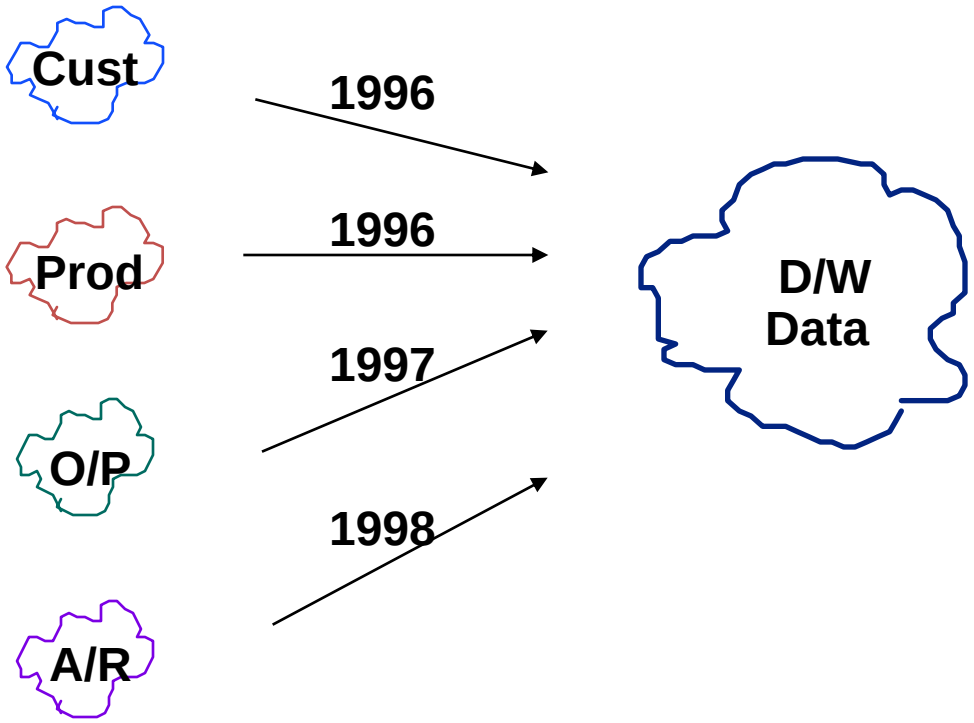
- Terabytes -- 10^12 bytes: Walmart -- 24 Terabytes
- Petabytes -- 10^15 bytes: Geographic Information Systems
- Exabytes -- 10^18 bytes: National Medical Records
- Zettabytes -- 10^21 bytes: Weather images
- Zottabytes -- 10^24 bytes: Intelligence Agency Videos

Features of DW

- Definition: A Data Warehouse is a subject oriented, integrated, nonvolatile and time variant collection of data to support decision making.
- Subject-oriented: data is stored as critical business subjects (not by Application).
- Integrated: all relevant data from various applications.
- Time-variant: historical and current data (time element)
- Nonvolatile data: periodic updates – no deletes.
- Data granularity: level of detail (efficient to keep data summarized at different levels).

Subject Oriented

Data is Integrated and Loaded by Subject



DW is Subject Oriented!

- Data is organized around major subject areas of an enterprise, and is therefore useful for an enterprise-wide understanding of those subjects
- **E.g.** a banking operational system keeps independent records of customer savings, loans, and other transactions. A warehouse pulls this independent data together to provide customer financial information (fees, profits, losses …)
- Examples of subject areas
  - Customer Financial Information
  - Toll calls made in the telecommunications industry company
  - Airline passenger booking information
- Data from operational systems must be transformed so that is consistent and meaningful in the DW

Subject Orientation

Insurance Co. Operational Databases

Data Warehouse

auto

life

health

casualty

claim

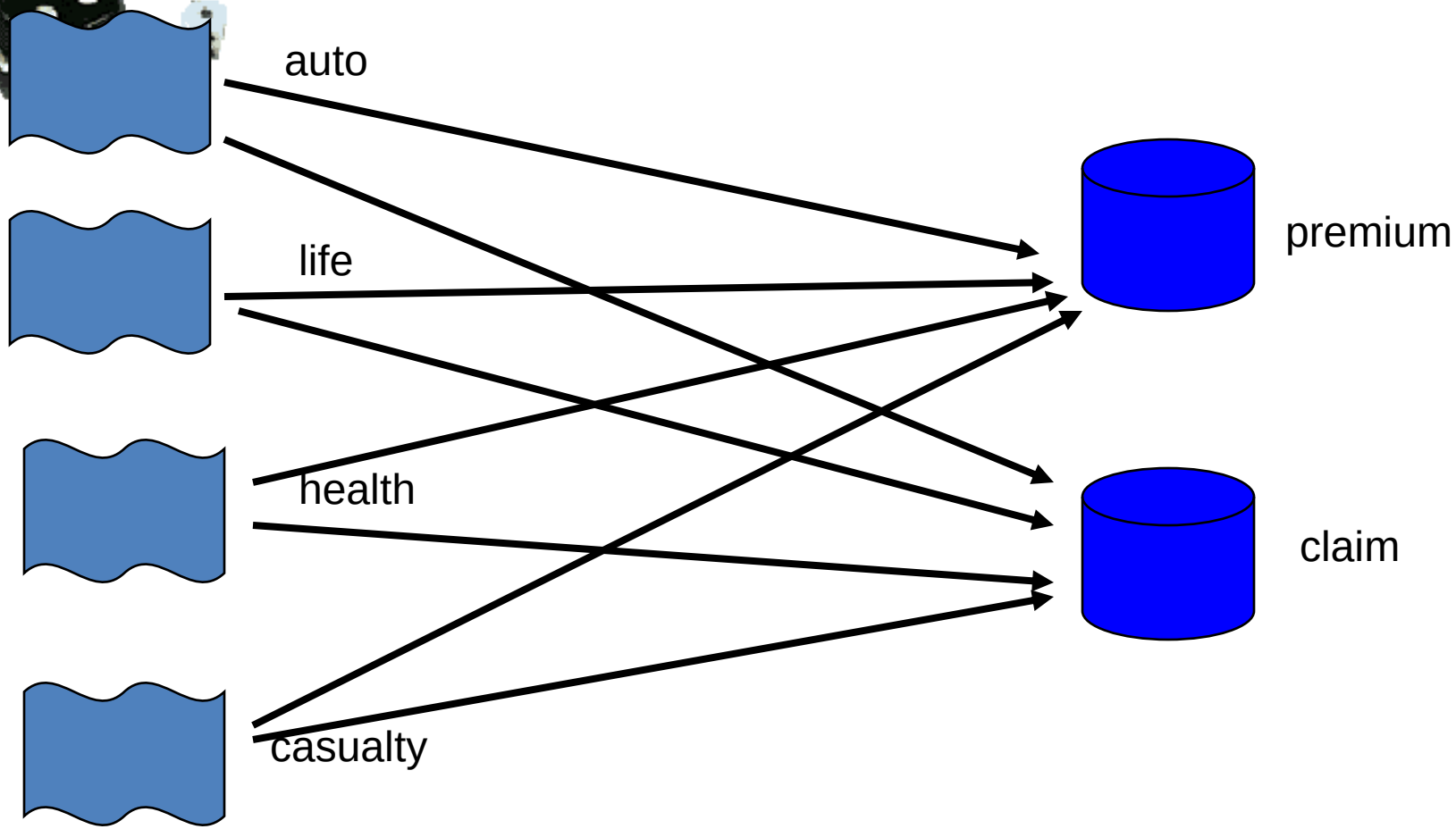
Subject

Applications

Subject Orientation

Insurance Co. Operational Databases

Data Warehouse



Applications

Subjects

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)
  - Making use of snapshots over past and current periods
- 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”.
  - Analysis of the past, relates to the present, and forecasts the future

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.

Integrated

Operational	Systems	
Order Processing	Order ID = 10	D/W
Accounts Receivable	Order ID = 12	Order ID = 16
Product Management	Order ID = 8	

HR System	Sex = M/F	D/W
Payroll	Sex = 1/2	Sex = M/F
Product Management	Sex = 0/1	

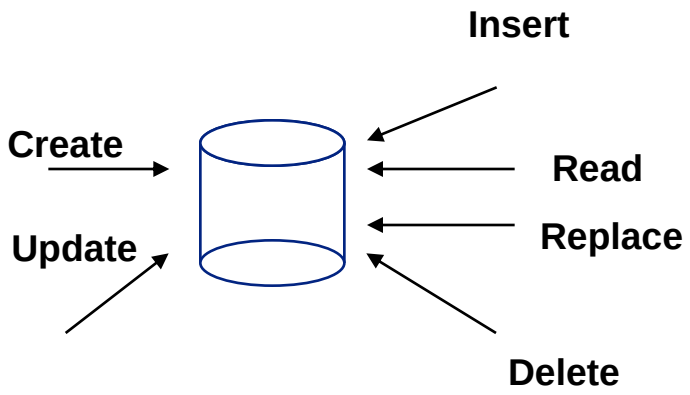
Data Warehouse - Non-Volatile

- 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
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

Non-Volatile

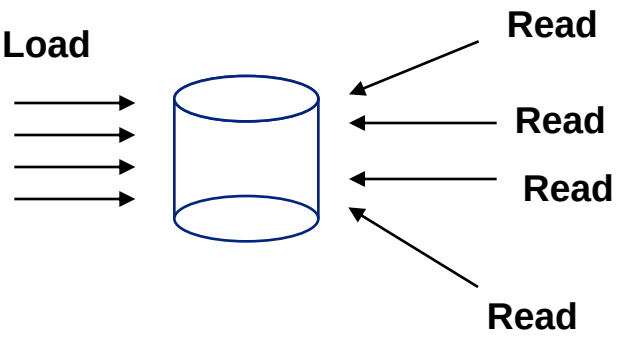
Operational System

- **“CRUD” Actions**



Data Warehouse

- **No Data Update**



Data Granularity in Warehouse

- Summarized data stored
  - reduce storage costs
  - reduce cpu usage
  - increases performance since smaller number of records to be processed
  - design around traditional high level reporting needs
  - tradeoff with volume of data to be stored and detailed usage of data

Data Warehouse - Granularity

- An operational system, data is usually kept at lowest level of detail of data.
- In Data Warehouse, more efficient to keep data summarized at different levels.
  - Sales summarized daily, monthly or quaterly

Granularity in Warehouse

- Can not answer some questions with summarized data
  - Did Anand call Seshadri last month? Not possible to answer if total duration of calls by Anand over a month is only maintained and individual call details are not.
- Detailed data too voluminous

Granularity in Warehouse

- Tradeoff is to have dual level of granularity
  - Store summary data on disks
    - 95% of DSS processing done against this data
  - Store detail on tapes
    - 5% of DSS processing against this data

The Benefits of Data Warehouse

- **Identify hidden business opportunities**

A data warehouse performs a second, and very valuable function by searching data for trends and abnormalities which users may not know to look for.

For example: Assisting companies in spotting sales trends, and detecting erroneous or fraudulent billings.

The Benefits of Data Warehouse

- **Precision Marketing**

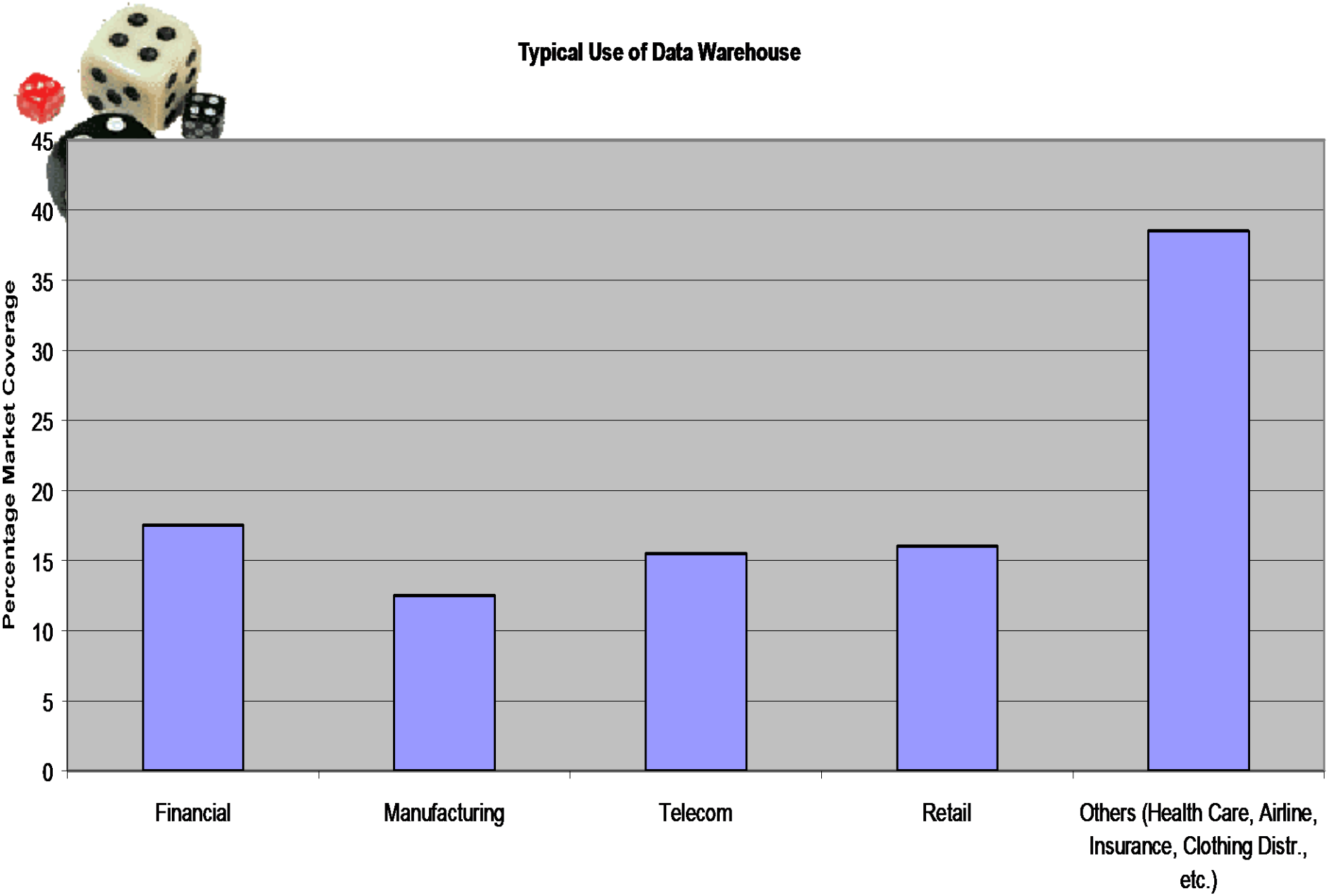
A data warehouse can aid in detecting segments of the marketplace (geographically and demographically) which remain untapped, and help show the best way to reach out to these potential customers (rapid response to market and technology trends).

Data Warehouse Purpose

- Data warehouse is designed and implemented to answer these TWO fundamental questions:
  - Who is buying what?
  - When and where are they doing so?
- More specific
  - Who [which customer] is buying [buying / using / delivering / shipping / ordering / returning] what [products / services] from where [outlet / store / clinic / branch] on what occasion [when], how [credit card / cash / check / exchange / debit] and why [causation]?

Some Uses of a Data Warehouse

- Airlines for aircraft deployment, analysis of route profitability, frequent flyer promotions, and maintenance
- Banking for promotion of products and services, and customer service
- Health care for cost reduction
- Investment and insurance companies for customer analysis, risk assessment, and portfolio management
- Retail stores for buying pattern analysis, promotions, customer profiling, and pricing
- Telecommunications for product and service promotions.

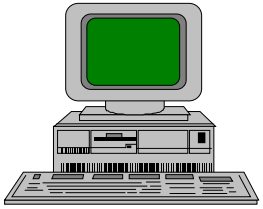


Source: Oracle Corp. 1999

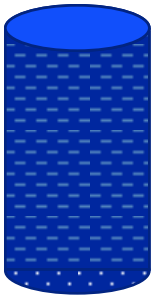
Data Warehouse Concepts

Sources of Data Warehouse Data

Archives (Historic Data)



Current Systems of Record (Recent History)

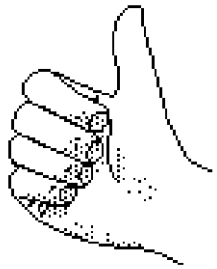


Enterprise Data Warehouse

Operational Transactions (Future Data Source)

Appropriate Uses of Data Warehouse Data

- **Produce Reports For Long Term Trend Analysis**
- **Produce Reports Aggregating Enterprise Data**
- **Produce Reports of Multiple Dimensions (Earned revenue by month by product by branch)**



Inappropriate Uses of Data Warehouse Data

- Replace Operational Systems
- Replace Operational Systems' Reports
- Analyze Current Operational Results

Data Warehouses and Data Marts

- A Data Mart is a logical subset of the complete data warehouse.
- Approaches towards building a Data warehouse:
  - Top-down approach – big picture approach, build the overall, big enterprise wide data warehouse.
  - Bottom-up approach – build departmental data marts one by one.
  - Practical approach:

Data Warehouse vs. Data Marts

- **A Data Mart is a logical subset of the complete data warehouse.**
- **Enterprise warehouse: collects all information about subjects (customers, products, sales, assets, personnel) that span the entire organization.**
  - **Requires extensive business modeling**
  - **May take years to design and build**
- **Data Marts: Departmental subsets that focus on selected subjects: *Marketing data mart: customer, products, sales.***
  - **Faster roll out, but complex integration in the long run.**
- **Virtual warehouse: views over operational dbs**
  - **Materialize some summary views for efficient query processing**
  - **Easier to build**
  - **Requisite excess capacity on operational db servers**

Data Warehouses vs. Data Marts

Enterprise wide

- Union of all data marts
- Data received from staging area
- Queries on presentation resource
- Structure for corporate view of data
- E-R model

Departmental

- Single business process
- Optimal for data access and analysis
- Structure to suit departmental view of data
- Star-Join (facts & dimensions)

Top-Down Approach (Bill Inmon)

- Top-down approach – big picture approach, build the overall, big enterprise wide data warehouse.
- Not union of disparate data marts, inherently architectured
- Centralized rules
- Takes longer to build
- High risk factor
- Needs high level of cross-functional skills
- Need outlay without proof of concept

Bottom-Up Approach (Ralph Kimball)

- Bottom-up approach – build departmental data marts one by one.
- Faster and easier implementation of manageable pieces
- Less risk of failure and allows project team to learn and grow
- Redundant data in every data mart
- Each data mart has its own narrow view of data
- Perpetuates inconsistent and irreconcilable data
- Unmanageable interfaces

Practical Approach

- Best of Top-Down and Bottom-Up approaches

- Practical approach:
  - Plan and define on corporate level.
  - Gather requirements at the overall level.
  - Create architecture for complete warehouse
  - Conform and standardize data content.
  - Implement as a series of supermarts, one at a time.
  - Supermarts are carefully architected data marts.
  - Ensure data types, field lengths for a data element must mean the same thing in every supermart.

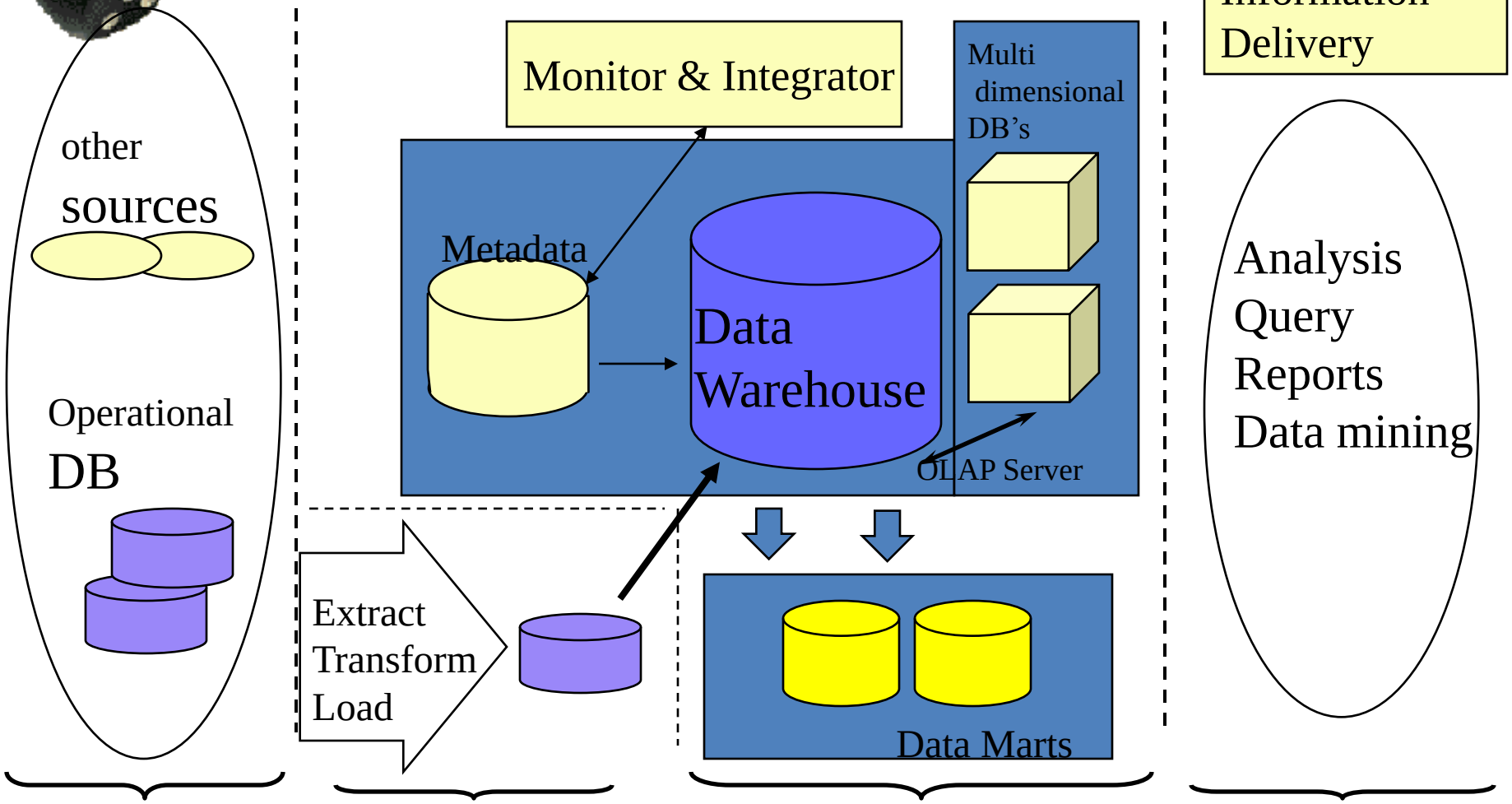
The Design Process

- From software engineering point of view
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around
- Typical data warehouse design process
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

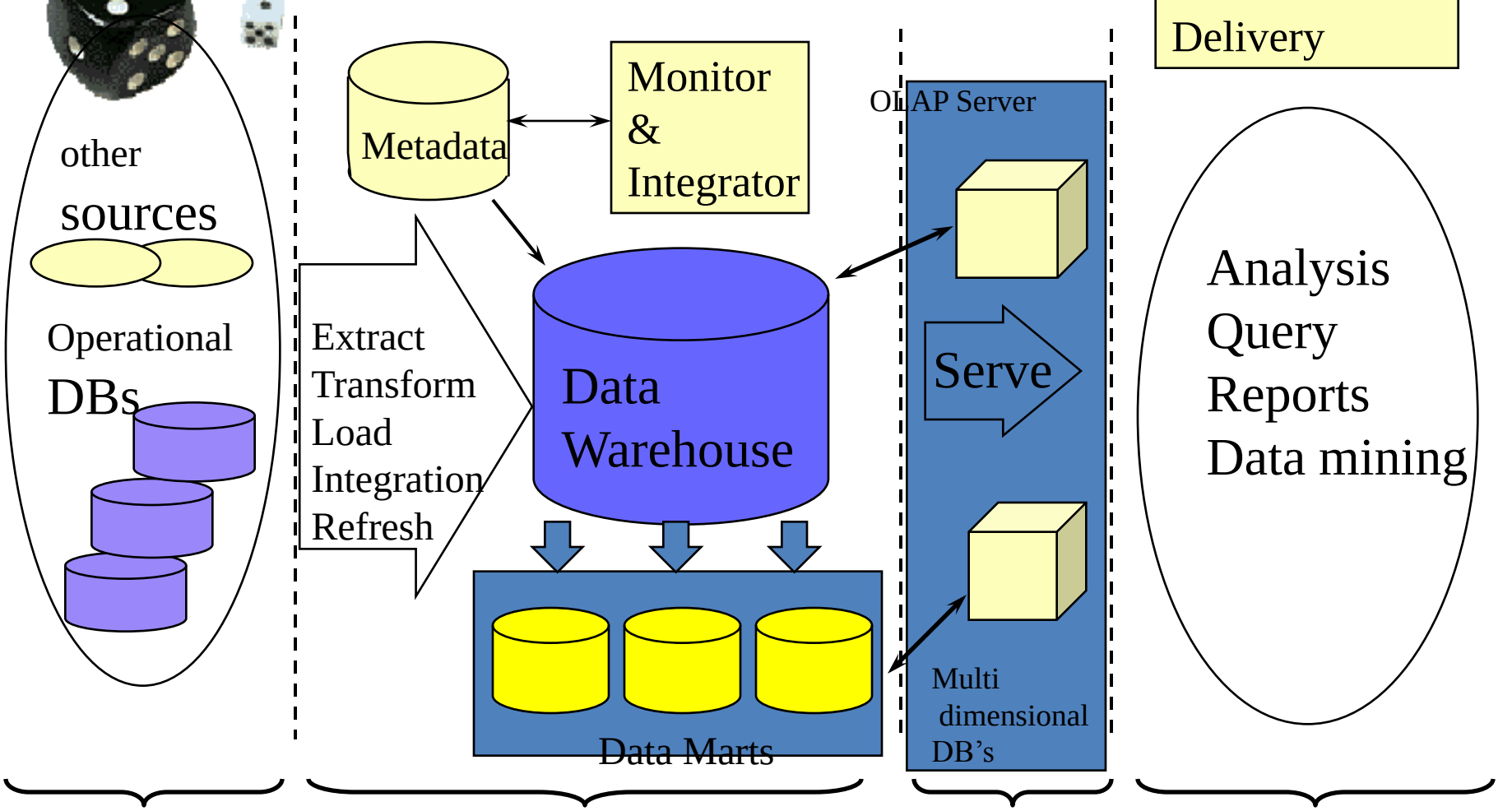
Overview of the Components

- Source Data Component
  - Production Data
  - Internal Data
  - Archived Data
  - External data
- Data Staging Component
  - Data extraction
  - Data transformation
  - Data loading
- Data Storage Component
- Information Delivery Component
- Meta data and Management and control Component

Building Blocks (A)



Building Blocks (B)



Data Have Data -- The Metadata

- The name suggests some high-level technological concept, but it really is fairly simple. **Metadata is "data about data".**
- With the emergence of the data warehouse as a decision support structure, the metadata are considered as much a resource as the business data they describe.
- **Metadata are abstractions** -- they are high level data that provide concise descriptions of lower-level data.

Metadata in Action – The Author's View

The metadata are essential ingredients in the transformation of raw data into knowledge. They are the "keys" that allow us to handle the raw data.

For example, a "line" or row in a sales database may contain:

1023 K596 111.21

This is mostly meaningless until we consult the metadata (in the data directory) that tells us:

store number =1023,

product= K596,

and sales = $111.21

© 2003, Prentice-Hall

Better end user data access and analysis tools can help users figure out how to get information they need out of the warehouse, but only good, easily accessible metadata can help them figure out what is available in the data warehouse and how to ask for it.

Metadata

- Metadata Component
  - Similar to data catalog in a RDMS
  - Information on logical data structures
  - Information on the files and addresses
  - Information on the indexes
- Types of Metadata
  - Operational Metadata (data structures from ODS)
  - Extraction and Transformation (extraction frequencies, methods)
  - End-User Metadata (navigational map of the data warehouse)
- Significance
  - Act as glue to connect all parts of Data Warehouse
  - Information on the contents and structures to developers
  - Make contents recognizable for end-users

Meta Data Description

- Information about the data warehouse system
  - **Content**
  - **Organizational**
  - **Structural**
  - **Management Information**
  - **Scheduling Information**
  - **Contact Information**
  - **Technical Information**

Why Do You Need Meta Data?

- Share resources
  - Users
  - Tools
- Document system
- Without metadata
  - Not Sustainable
  - Not able to fully utilize resource

Metadata Life Cycle

- Collection - Identify metadata and capture into repository; automate
- Maintenance - Put in place processes to synchronize metadata automatically with changing data architecture; automate
- Deployment - Provide metadata to users in the right form and with the right tools; match metadata offered to specific needs of each audience

Metadata Collection

- Right metadata at the right time
- Variety of collection strategies
- Sources
  - potential sources of data for DW
  - external data
  - data structures
- Data Models - enterprise data model start point
  - import from CASE tool
  - correlate enterprise and warehouse models

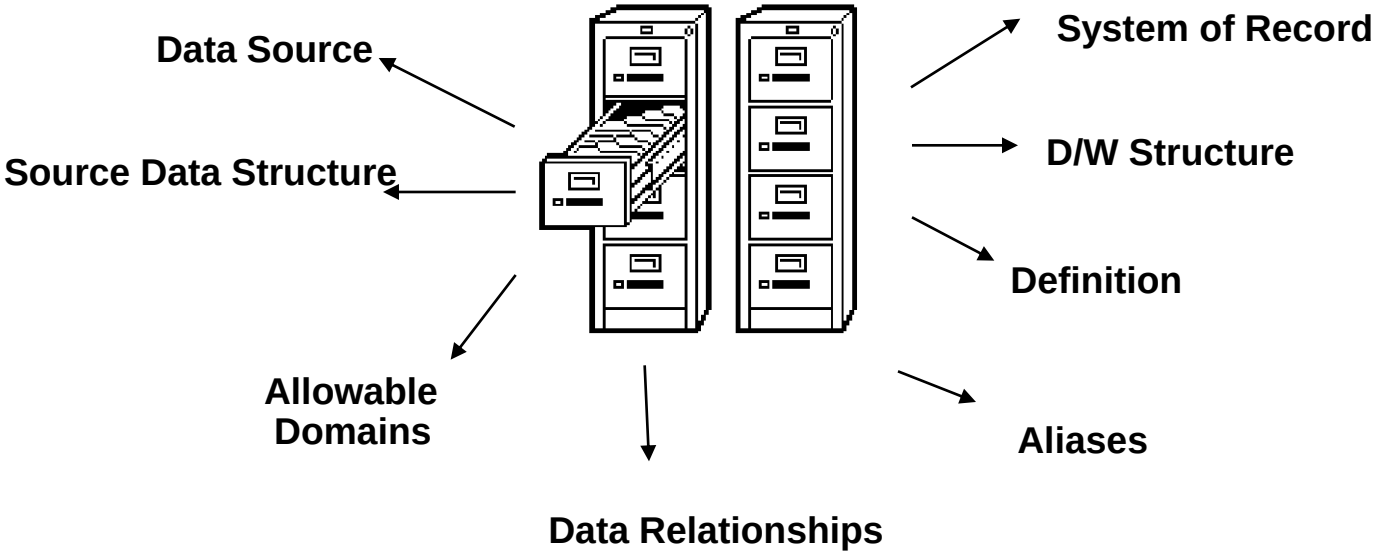
Metadata Extraction

- Regardless of the nature of a query, certain aspects of the metadata are important to all decision-makers. Some of these are:
  - What tables, attributes and keys does the DW contain?
  - Where did each set of data come from?
  - What transformations were applied with cleansing?
  - How have the metadata changed over time?
  - How often do the data get reloaded?
  - Are there so many data elements that you need to be careful what you ask for?

Data Warehouse Concepts

Meta Data - Map of Integration

The Data That Provides the "Card Catalogue" Of References For All Data Within The Data Warehouse



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 (typically multidimensional 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)

OLTP vs Data Warehouse

- OLTP
  - Performance Sensitive
  - Few Records accessed at a time (tens)
  - Read/Update Access
  - No data redundancy
  - Database Size 100MB -100 GB
- Data Warehouse
  - Performance relaxed
  - Large volumes accessed at a time(millions)
  - Mostly Read (Batch Update)
  - Redundancy present
  - Database Size 100 GB - few terabytes

OLTP vs Data Warehouse

- OLTP
  - Transaction throughput is the performance metric
  - Thousands of users
  - Managed in entirety
- Data Warehouse
  - Query throughput is the performance metric
  - Hundreds of users
  - Managed by subsets

To summarize …

- OLTP Systems are used to *"run"* a business

- The Data Warehouse helps to *"optimize"* the business

Data Warehouses vs. Data Marts

Enterprise wide

- Union of all data marts
- Data received from staging area
- Queries on presentation resource
- Structure for corporate view of data
- E-R model

Departmental

- Single business process
- Optimal for data access and analysis
- Structure to suit departmental view of data
- Star-Join (facts & dimensions)