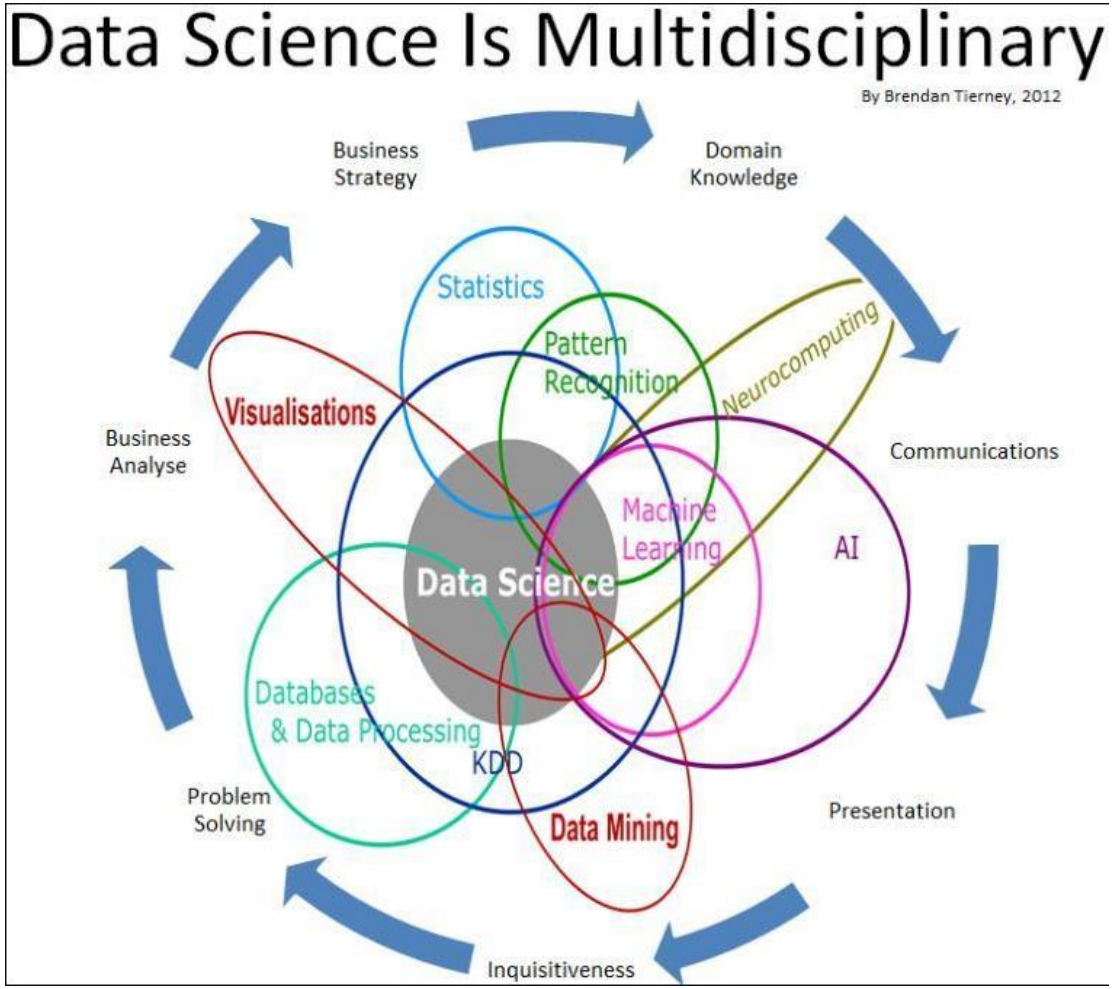
- Explain the basic concepts and importance of data warehousing.
- Differentiate transactional database and data warehousing.
- Explain the importance of integration across the organization through Data.
- Discuss Data Warehouse Design.

Where does it fit?



BI Tools Landscape

- Self-Service BI

Involves free-form reporting and analysis

Enables you to integrate data from disparate sources and drill-down and understand the root cause for data anomalies.

You can perform your own reporting and analysis without relying on IT or others.

Corporate BI

- Formatted reports that are typically based upon approved corporate data, and then shared more broadly with managers, teams, or departments.
- IT oversees the distribution and monitoring of the reporting environment and building of the structured data layer upon which the reports are built.

Advanced Analytics

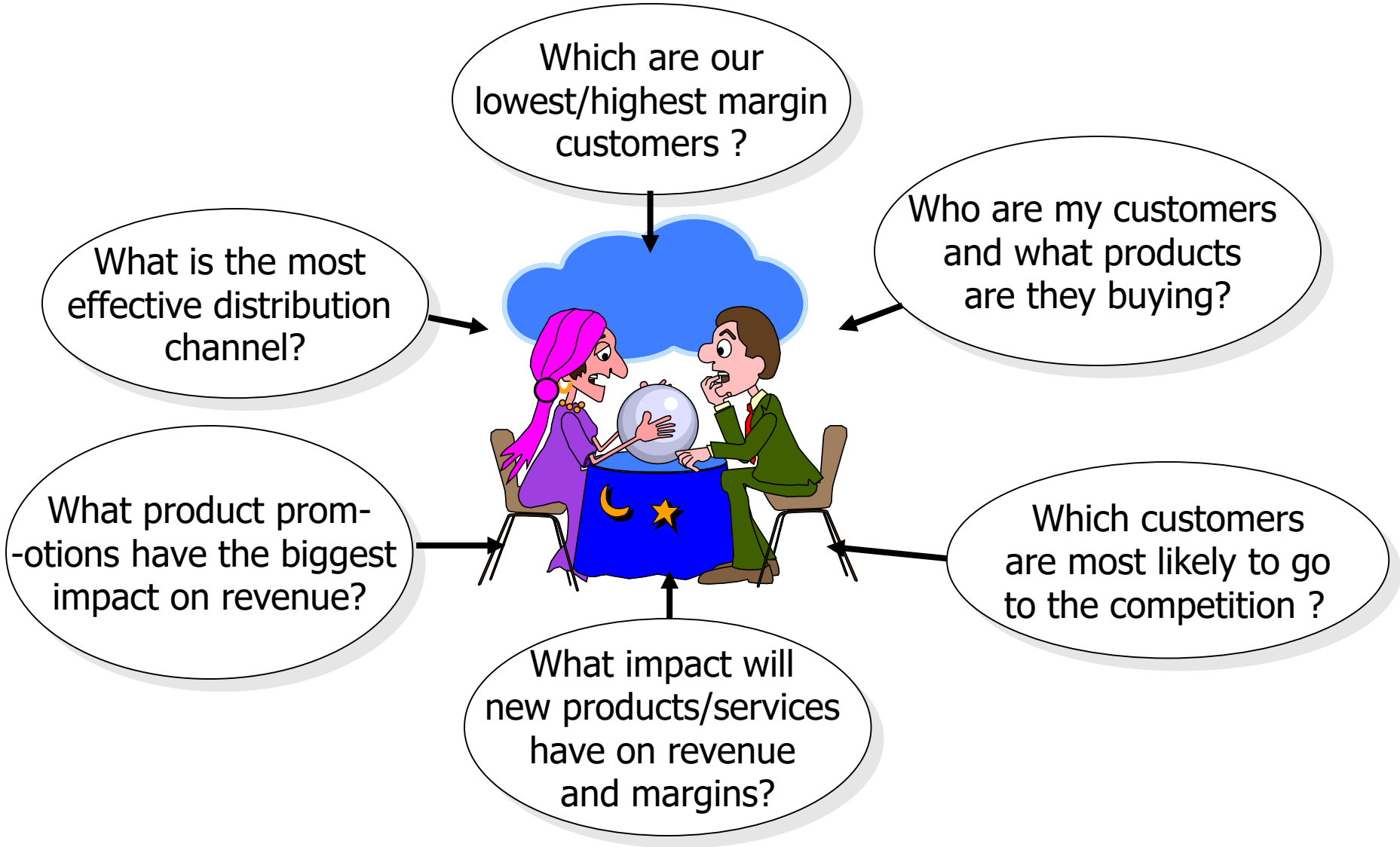
- Tools and techniques used to forecast future outcomes and behaviors
- Data mining uses mathematical analysis to derive patterns and trends that exist in data
- Patterns cannot be discovered by traditional data exploration
- relationships are too complex and there is too much data.

- Database and Data Warehousing
- History of data warehousing
- Evolution in organization use of data warehouses
- Data Warehouse Architecture
- Benefits of data warehousing
- Strategic uses of data warehousing
- Disadvantages of data warehouses
- Data mart
- OLAP
- Data warehousing integration
- Business intelligence

The Difference…

- DWH Constitute Entire Information Base For All Time..
- Database Constitute Real Time Information…
- DWH Supports DM And Business Intelligence.
- Database Is Used To Running The Business
- DWH Is How To Run The Business

A producer wants to know….



Hard/Infeasible Queries for OLTP

- Why not use the existing databases (OLTP) for business analysis?
- Business analysis queries

  In the past five years, which product is the most profitable?

  Which public holiday we have the largest sales?

  Which week we have the largest sales?

  Does the sales of dairy products increase over time?

  Difficult to express these queries in SQL

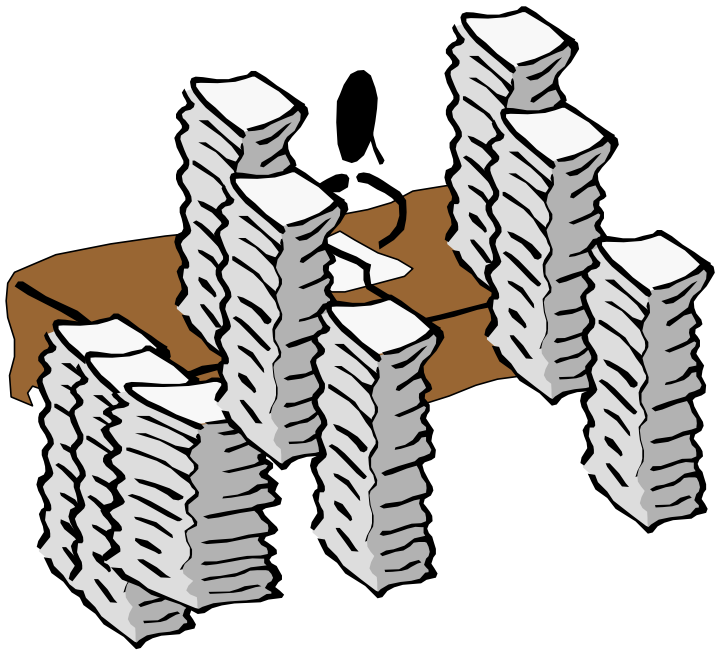
  3rd query: may extract the “week” value using a function

  But the user has to learn many transformation functions …

  4th query: use a “special” table to store IDs of all dairy products, in advance

  There can be many different dairy products; there can be many other product types as well …

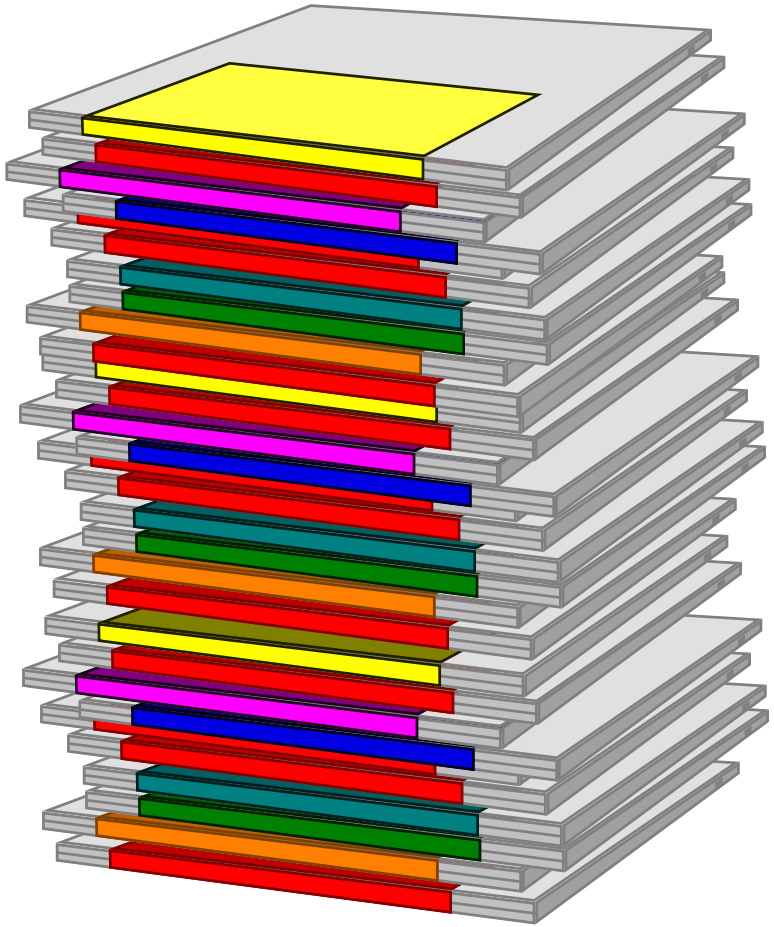
Unstructured , Inconsistent Data



- **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 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.



What is Data Warehousing?

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

Data Warehousing -- a process

- It is a relational or multidimensional database management system designed to support management decision making.
- A data warehousing is derived from transaction data specifically structured for querying and reporting.
- Technique for assembling and managing data from various sources for the purpose of answering business questions. Thus making decisions that were not previous possible

Data warehousing is …

- **Subject Oriented:** Data that gives information about a particular subject instead of about a company's ongoing operations.
- **Integrated:** Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.
- **Time-variant:** All data in the data warehouse is identified with a particular time period.
- **Non-volatile:** Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.
- Data warehousing is combining data from multiple and usually varied sources into one comprehensive and easily manipulated database.
- Common accessing systems of data warehousing include queries, analysis and reporting.
- Because data warehousing creates one database in the end, the number of sources can be anything you want it to be, provided that the system can handle the volume, of course.
- The final result, however, is homogeneous data, which can be more easily manipulated.

Aggregate Example

- Aggregate Example
- Imagine 1 bio. sales rows, 1000 products, 100 locations

CREATE VIEW TotalSales (pid, locid, total) AS SELECT s.pid, s.locid, SUM(s.sales) FROM Sales s GROUP BY s.pid, s.locid

The materialized view has 100'000 rows

- Query rewritten to use view

SELECT p.category, SUM(s.sales) FROM Products p, Sales s WHERE p.pid=s.pid GROUP BY p.category

Rewritten to

SELECT p.category, SUM(t.total) FROM Products p, TotalSales t WHERE p.pid=t.pid GROUP BY p.category

Query becomes 10'000 times faster!

Pre-Aggregation Choices

- Full pre-aggregation: (all combinations of levels)
- Fast query response
- Takes a lot of space/update time (200-500 times raw data)
- No pre-aggregation
- Slow query response (for terabytes…)
- Practical pre-aggregation: chosen combinations

A good compromise between response time and space use

- Most (R)OLAP tools now support practical preaggregation

IBM DB2 UDB

Oracle 9iR2

MS Analysis Services

Hyperion Essbase (DB2 OLAP Services)

Choosing Aggregate

Using practical pre-aggregation, it must be decided what aggregates to store

This is a non-trivial (NP-complete) optimization problem

Many influencing factors

Space use

Update speed

Response time demands

Actual queries

Prioritization of queries

Index and/or aggregates Only choose an aggregate if it is considerably smaller than available, usable aggregates (factor 3-5-10)

Often supported (semi-)automatically by tool/DBMS Oracle, DB2, MS SQL Server

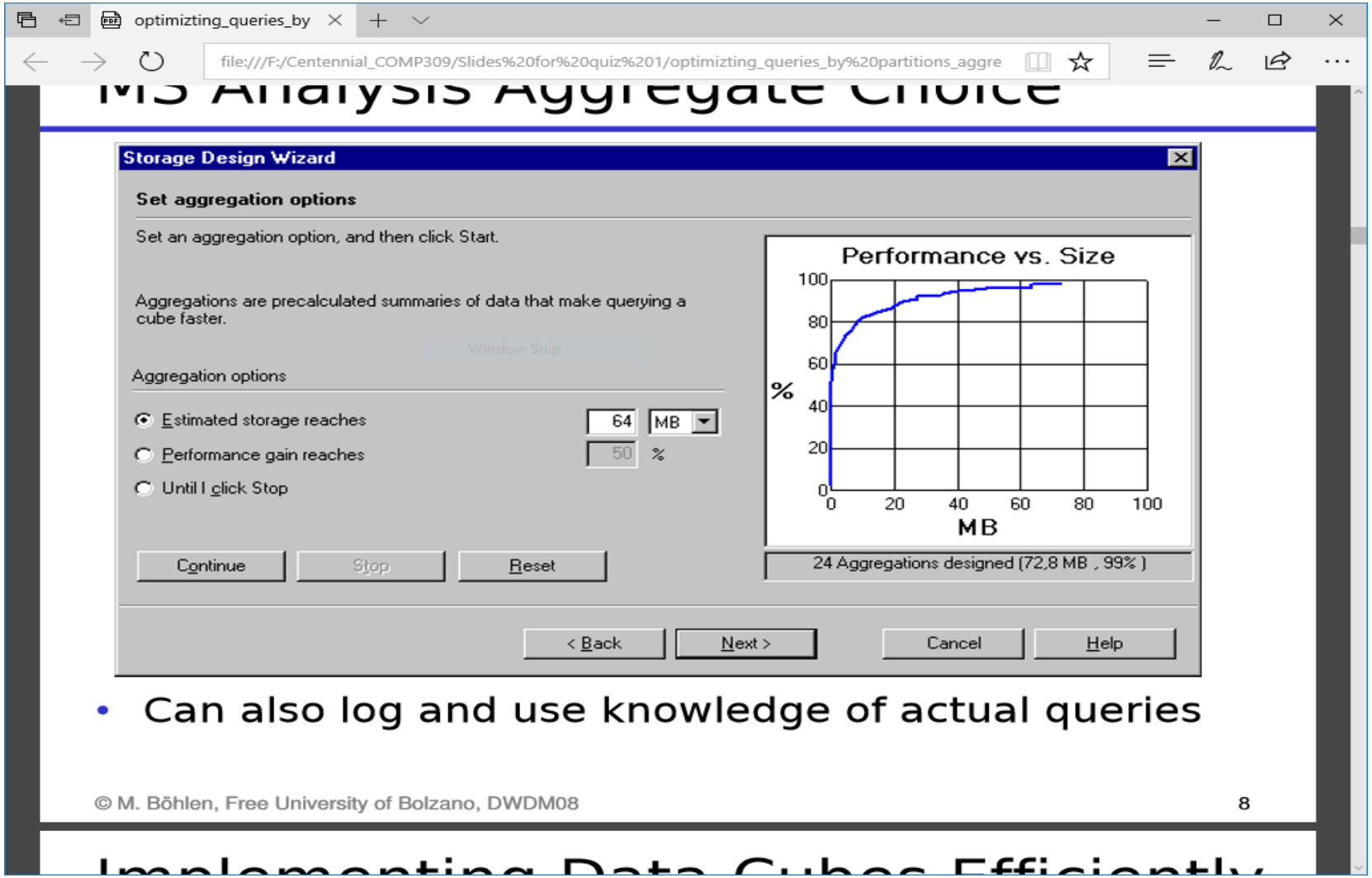
Database explosion

10	**20**
30	**40**

10	**20**	**30**
30	**40**	**70**
40	**60**	**100**

5 extra values needed to hold all possible aggregations on a table that had only 4 values in it originally!

MS Analysis Aggregate Choice



History of data warehousing

- The concept of data warehousing dates back to the late 1980s when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse".
- 1960s - General Mills and Dartmouth College, in a joint research project, develop the terms *dimensions* and *facts*.
- 1970s - ACNielsen and IRI provide dimensional data marts for retail sales.
- 1983 – Tera data introduces a database management system specifically designed for decision support.
- 1988 - Barry Devlin and Paul Murphy publish the article *An architecture for a business and information systems* in *IBM Systems Journal* where they introduce the term "business data warehouse".

OLTP vs Data Warehouse

- **OLTP**
  - Application Oriented
  - Used to run business
  - Detailed data
  - Current up to date
  - Isolated Data
  - Clerical User
  - Few Records accessed at a time (tens)
  - Read/Update Access
  - Volatile Creation/Deletion
  - No data redundancy
  - Database Size 100MB -100 GB
  - Transaction throughput is the performance metric
  - Thousands of users
  - Managed in entirety

- **Warehouse (DSS – Decision Support System)**
  - Subject Oriented
  - Used to analyze business
  - Summarized and refined
  - Snapshot data
  - Integrated Data
  - Knowledge User (Manager)
  - Large volumes accessed at a time (millions)
  - Mostly Read (Batch Update)
  - Redundancy present
  - Database Size 100 GB - few terabytes
  - Query throughput is the performance metric
  - Hundreds of users
  - Managed by subsets

Data Analysis Problems

- The same data found in many different systems
- Example: customer data across different stores and departments
- The same concept is defined differently
- Heterogeneous sources
- Relational DBMS, On-Line Transaction Processing (OLTP)
- Unstructured data in files (e.g., MS Word)
- Legacy systems

Data Analysis Problems (Cont’d)

- Data is suited for operational systems Accounting, billing, etc.
- Do not support analysis across business functions
- Data quality is bad
- Missing data, imprecise data, different use of systems
- Data are “volatile”
- Data deleted in operational systems (6 months)
- Data change over time – no historical information

To summarize …

- OLTP Systems are used to *“run”* a business

- The Data Warehouse helps to *“optimize”* the business

Case Study of an Enterprise

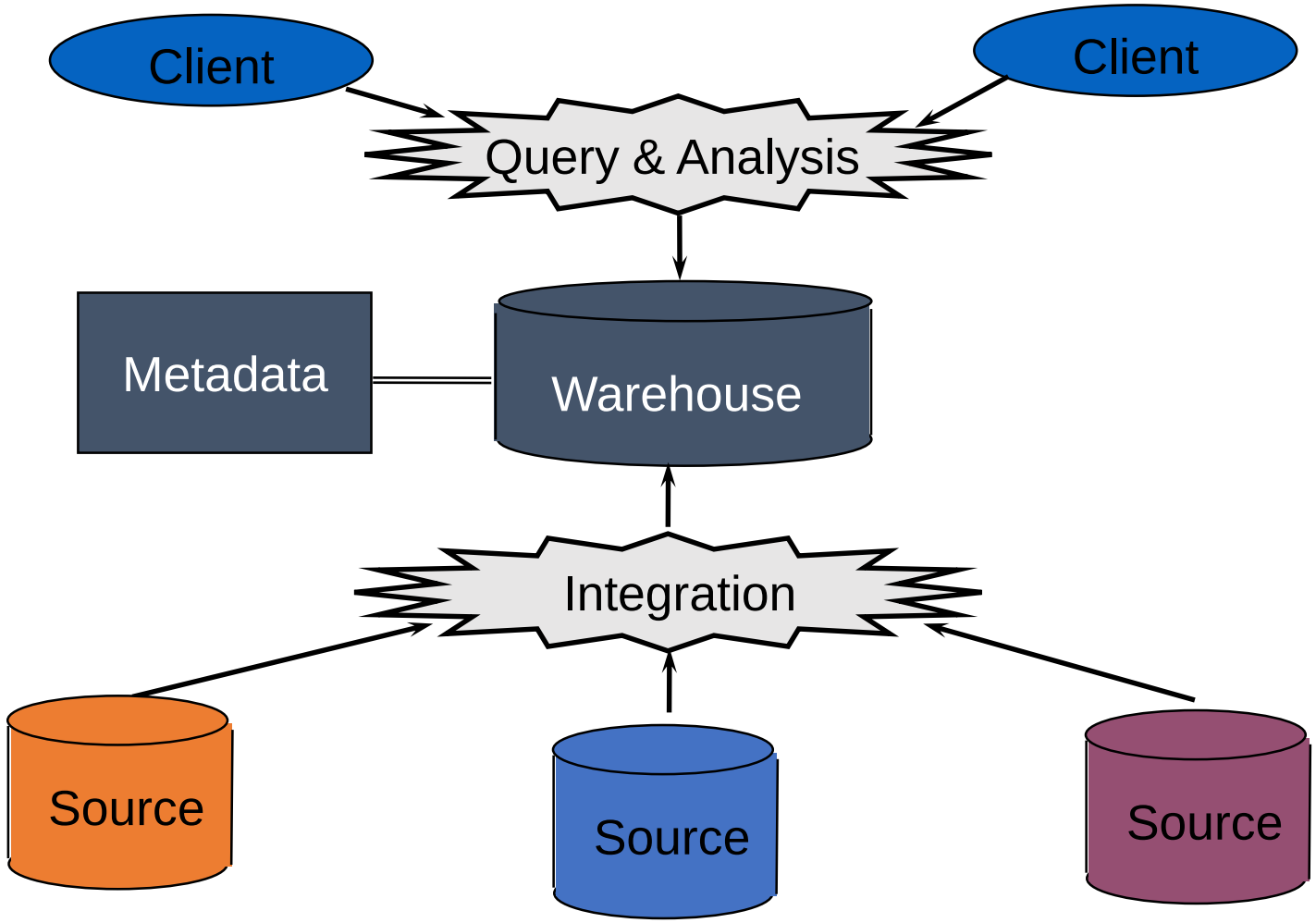
- Example of a chain (e.g., fashion stores or car dealers)
- Each store maintains its own customer records and sales records.
- Hard to answer questions like: “find the total sales of Product X from stores in Aalborg”
- The same customer may be viewed as different customers for different stores; hard to detect duplicate customer information
- Imprecise or missing data in the addresses of some customers
- Purchase records maintained in the operational system for limited time (e.g., 6 months); then they are deleted or archived
- The same “product” may have different prices, or different discounts in different stores

Evolution in organizational use of data warehouses

Organizations generally start off with relatively simple use of data warehousing. Over time, more sophisticated use of data warehousing evolves. The following general stages of use of the data warehouse can be distinguished:

- **Operation Data Store**
  - Data warehouses in this initial stage are developed by simply copying the data off an operational system to another server where the processing load of reporting against the copied data does not impact the operational system's performance.
- **Off line Data Warehouse**
  - Data warehouses at this stage are updated from data in the operational systems on a regular basis and the data warehouse data is stored in a data structure designed to facilitate reporting.
- **Real Time Data Warehouse**
  - Data warehouses at this stage are updated every time an operational system performs a transaction (e.g. an order or a delivery or a booking.)
- **Integrated Data Warehouse**
  - Data warehouses at this stage are updated every time an operational system performs a transaction. The data warehouses then generate transactions that are passed back into the operational systems.

Data Warehouse Architecture



- The data has been selected from various sources and then integrate and store the data in a single and particular format.
- Data warehouses contain current detailed data, historical detailed data, lightly and highly summarized data, and metadata.
- Current and historical data are voluminous because they are stored at the highest level of detail.
- Lightly and highly summarized data are necessary to save processing time when users request them and are readily accessible.
- ***Metadata*** are "data about data". It is important for designing, constructing, retrieving, and controlling the warehouse data.

Technical metadata include where the data come from, how the data were changed, how the data are organized, how the data are stored, who owns the data, who is responsible for the data and how to contact them, who can access the data , and the date of last update.

Business metadata include what data are available, where the data are, what the data mean, how to access the data, predefined reports and queries, and how current the data are.

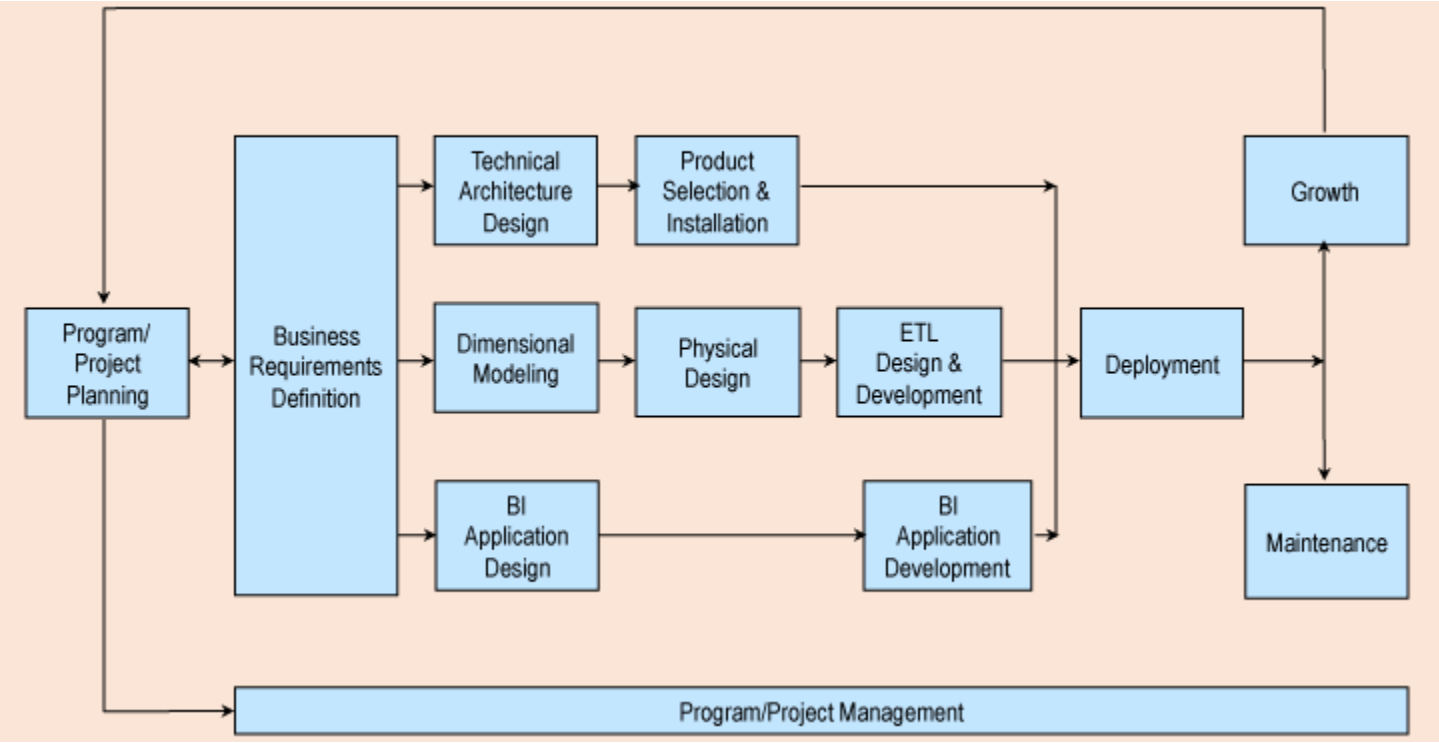
Business advantages

- It provides business users with a “customer-centric” view of the company’s heterogeneous data by helping to integrate data from sales, service, manufacturing and distribution, and other customer-related business systems.
- It provides added value to the company’s customers by allowing them to access better information when data warehousing is coupled with internet technology.
- It consolidates data about individual customers and provides a repository of all customer contacts for segmentation modeling, customer retention planning, and cross sales analysis.
- It removes barriers among functional areas by offering a way to reconcile views from multiple areas, thus providing a look at activities that cross functional lines.
- It reports on trends across multidivisional, multinational operating units, including trends or relationships in areas such as merchandising, production planning etc.

Strategic uses of data warehousing

Industry	Functional areas of use	Strategic use
Airline	Operations; marketing	Crew assignment, aircraft development, mix of fares, analysis of route profitability, frequent flyer program promotions
Banking	Product development; Operations; marketing	Customer service, trend analysis, product and service promotions, reduction of IS expenses
Credit card	Product development; marketing	Customer service, new information service, fraud detection
Health care	Operations	Reduction of operational expenses
Investment and Insurance	Product development; Operations; marketing	Risk management, market movements analysis, customer tendencies analysis, portfolio management
Retail chain	Distribution; marketing	Trend analysis, buying pattern analysis, pricing policy, inventory control, sales promotions, optimal distribution channel
Telecommunications	Product development; Operations; marketing	New product and service promotions, reduction of IS budget, profitability analysis
Personal care	Distribution; marketing	Distribution decisions, product promotions, sales decisions, pricing policy
Public sector	Operations	Intelligence gathering

Life cycle of Data warehouse Design



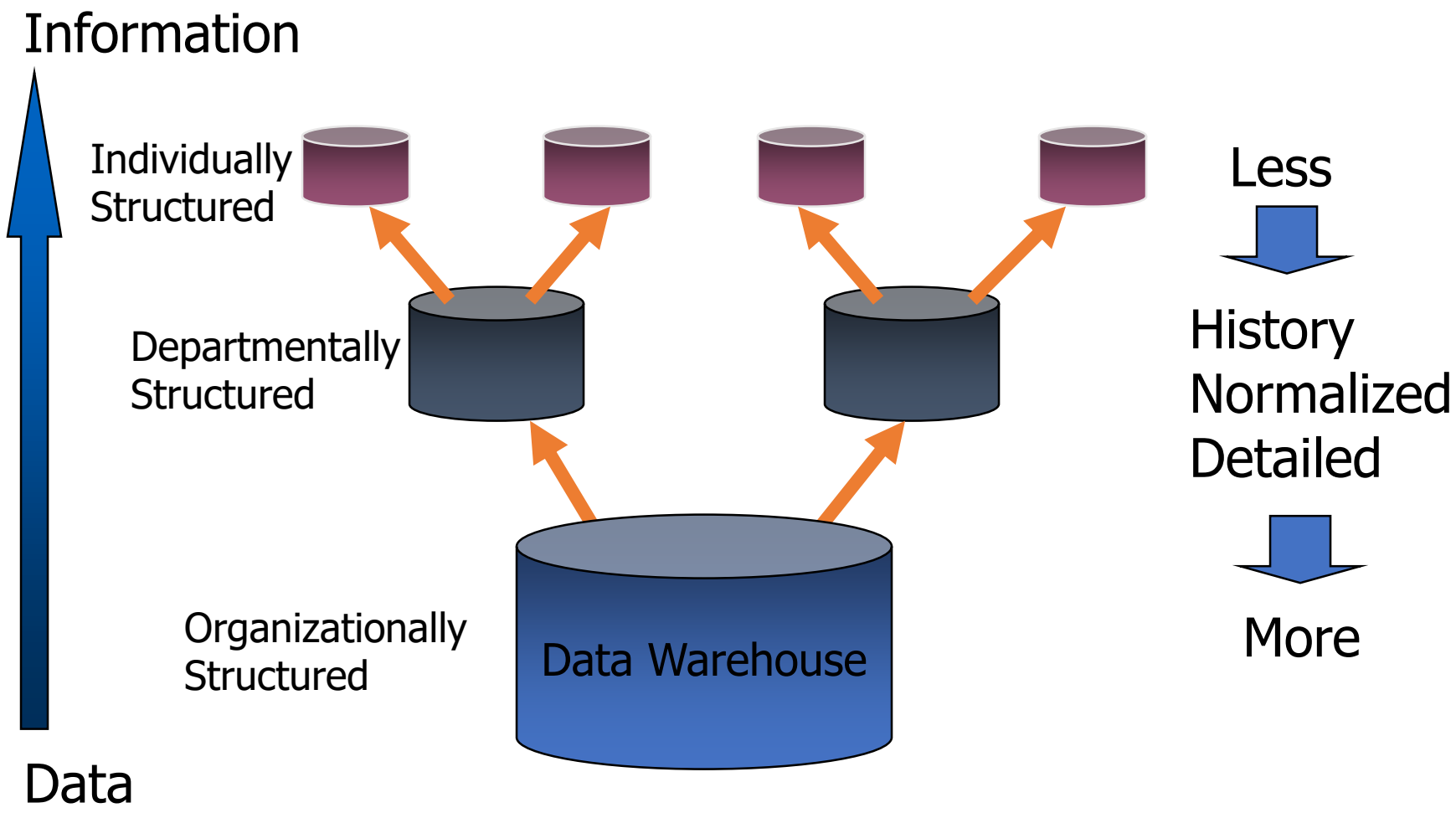
Data Marts

- A data mart is a scaled down version of a data warehouse that focuses on a particular subject area.
- A **data mart** is a subset of an organizational data store, usually oriented to a specific purpose or major data subject, that may be distributed to support business needs.
- Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization.
- Usually designed to support the unique business requirements of a specified department or business process
- Implemented as the first step in proving the usefulness of the technologies to solve business problems

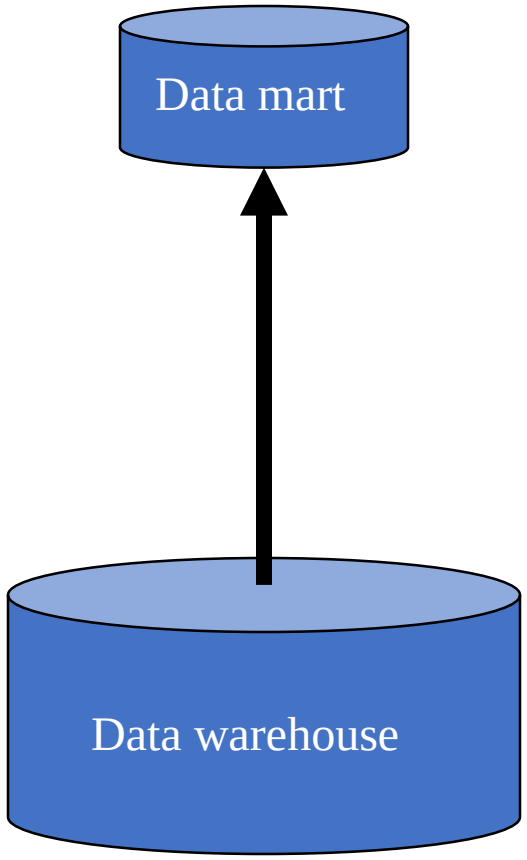
Reasons for creating a data mart

- Easy access to frequently needed data
- Creates collective view by a group of users
- Improves end-user response time
- Ease of creation in less time
- Lower cost than implementing a full Data warehouse
- Potential users are more clearly defined than in a full Data warehouse

From the Data Warehouse to Data Marts



Characteristics of the Departmental Data Mart



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