DATA WAREHOUSING

PRESENTATION PREPARED BY:

Ankur Chandel

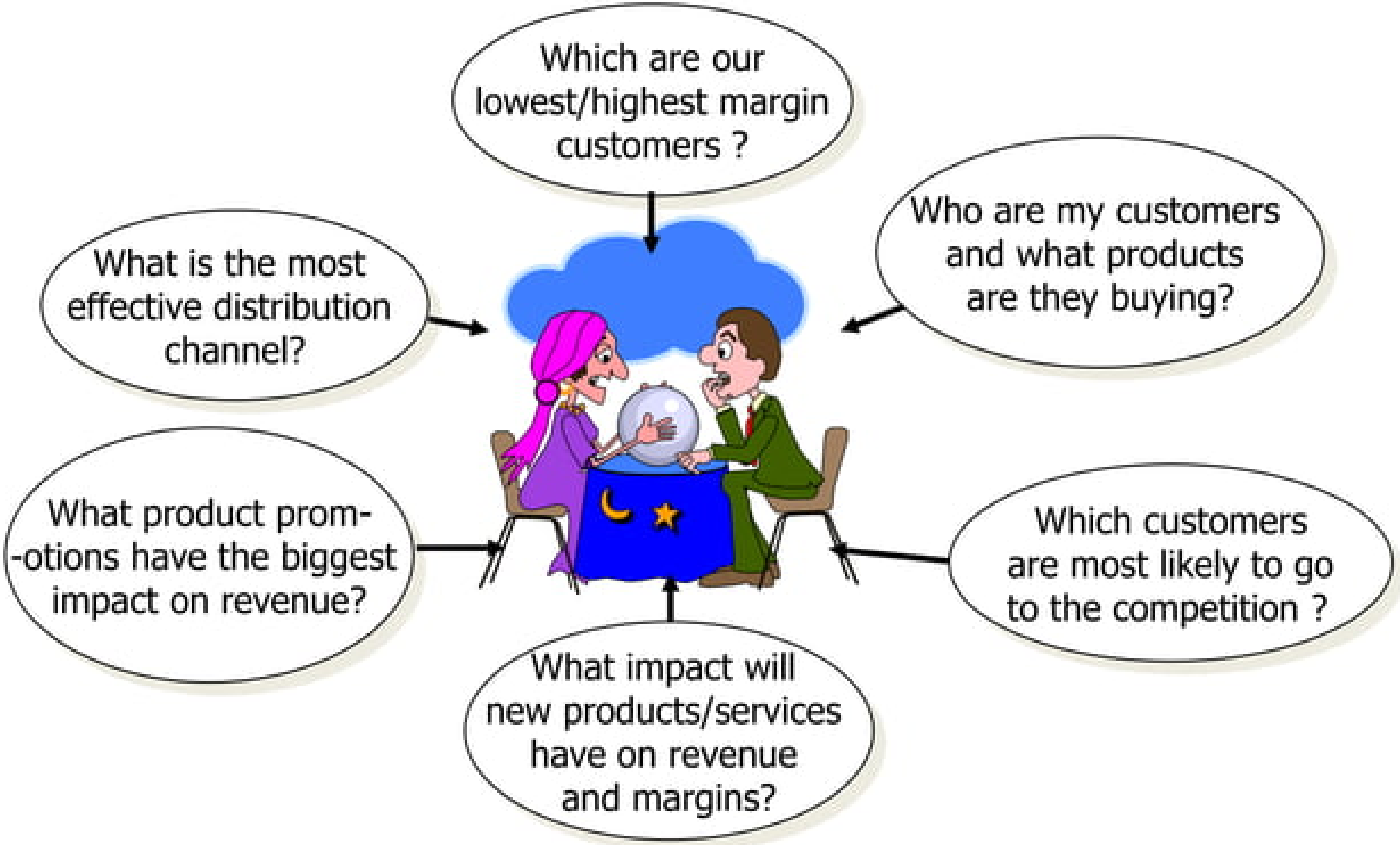
CONTENTS

- 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
- Data mining
- Data mining for decision support
- Text mining
- OLAP
- Data warehousing integration
- Business intelligence

Database and Data Ware Housing….

- 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….



Which are our lowest/highest margin customers ?

What is the most effective distribution channel?

Who are my customers and what products are they buying?

What product promotions have the biggest impact on revenue?

Which customers are most likely to go to the competition ?

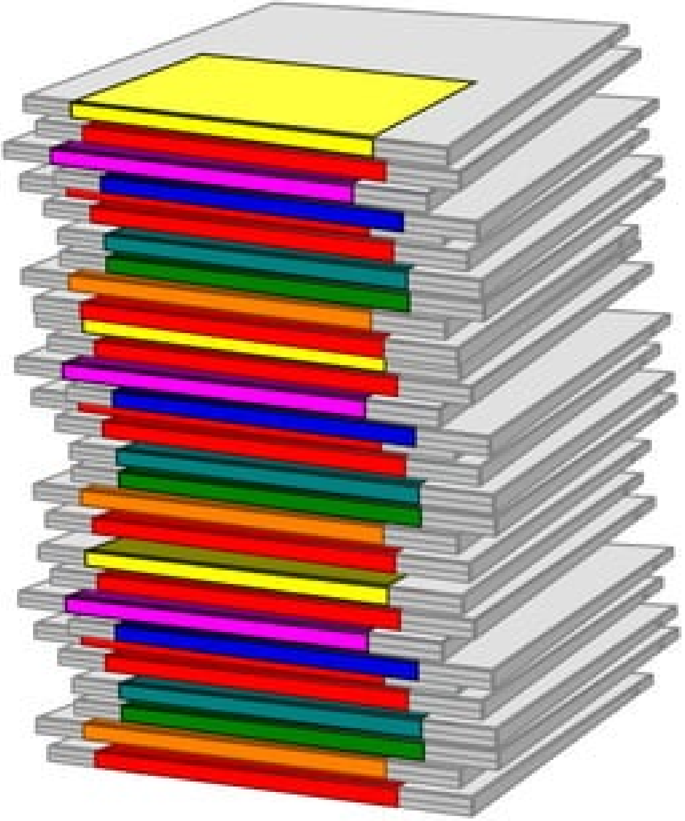
What impact will new products/services have on revenue and margins?

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 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

Information
Data

Data Warehousing -- a process

- It is a relational or multidimensional database management system designed to support management decision making.
- A data warehousing is a copy of 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.

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

OLTP- ONLINE TRANSACTION PROCESSING

- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)
- OLTP systems are tuned for known transactions and workloads while workload is not known a priori in a data warehouse
  - 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
  - Clerical User
  - Few Records accessed at a time (tens)
  - Read/Update Access
  - No data redundancy
  - Database Size 100MB -100 GB
  - Transaction throughput is the performance metric
  - Thousands of users
  - Managed in entirety

- **Warehouse (DSS)**
  - 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

To summarize ...

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

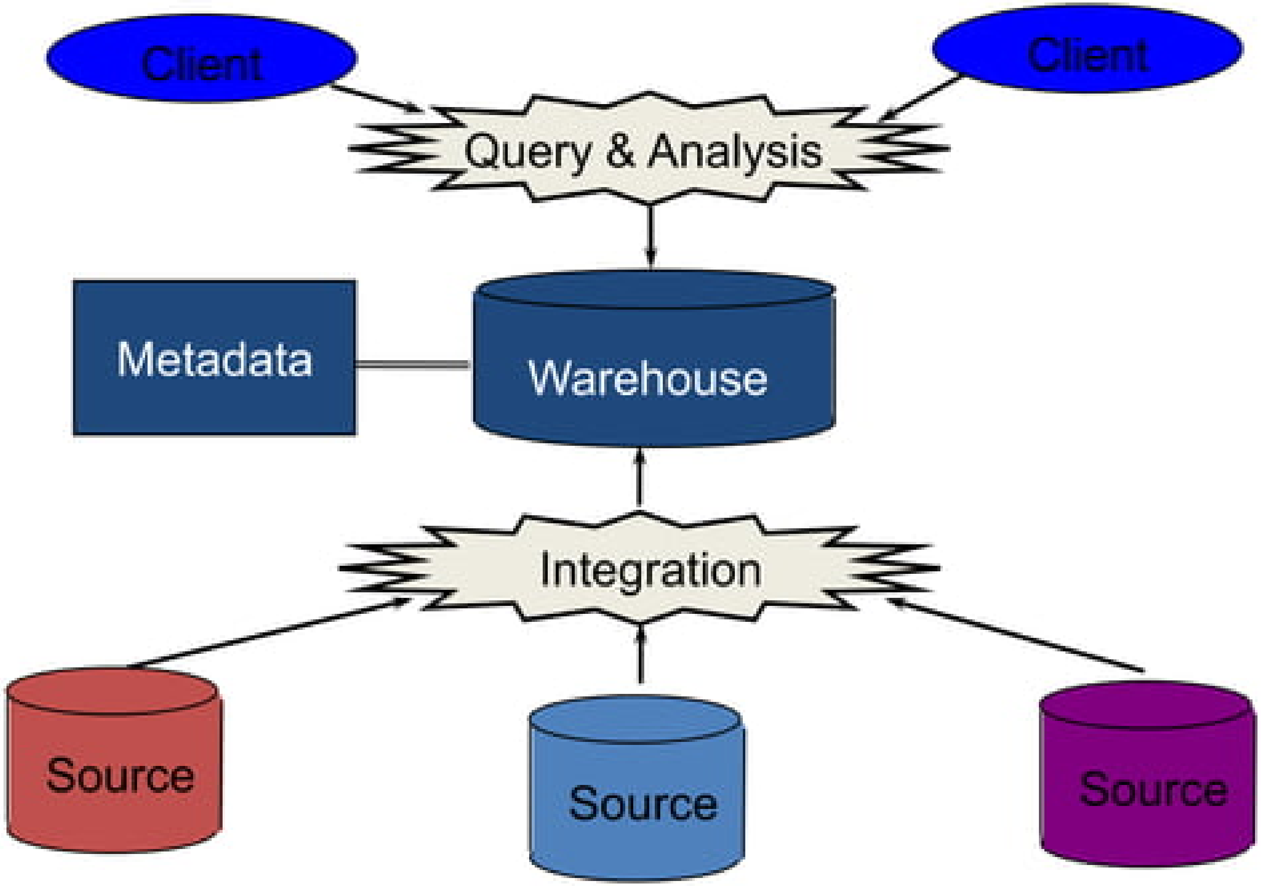
- The Data Warehouse helps to *"optimize"* the business

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:

- **Off line Operational Database**
  - –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.

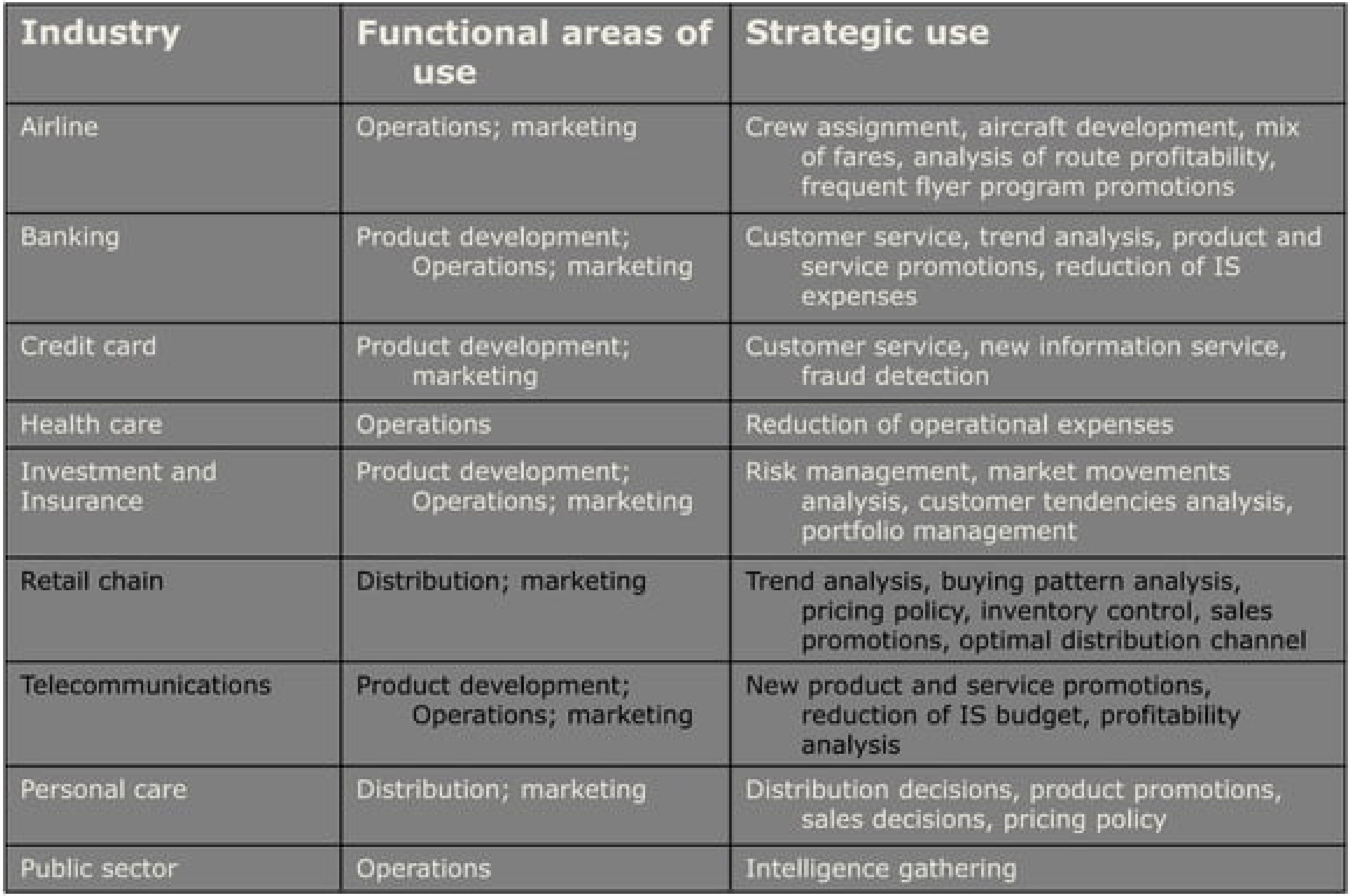
Technical metadata include where the data come from, how the data were changed, how the data are organized, how the data are stored, who owns the data, who is responsible for the data and how to contact them, who can access the data , and the date of last update.

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

Disadvantages of data warehouses

- Data warehouses are not the optimal environment for unstructured data.
- Because data must be extracted, transformed and loaded into the warehouse, there is an element of latency in data warehouse data.
- Over their life, data warehouses can have high costs. Maintenance costs are high.
- Data warehouses can get outdated relatively quickly. There is a cost of delivering suboptimal information to the organization.
- There is often a fine line between data warehouses and operational systems. Duplicate, expensive functionality may be developed. Or, functionality may be developed in the data warehouse that, in retrospect, should have been developed in the operational systems and vice versa.

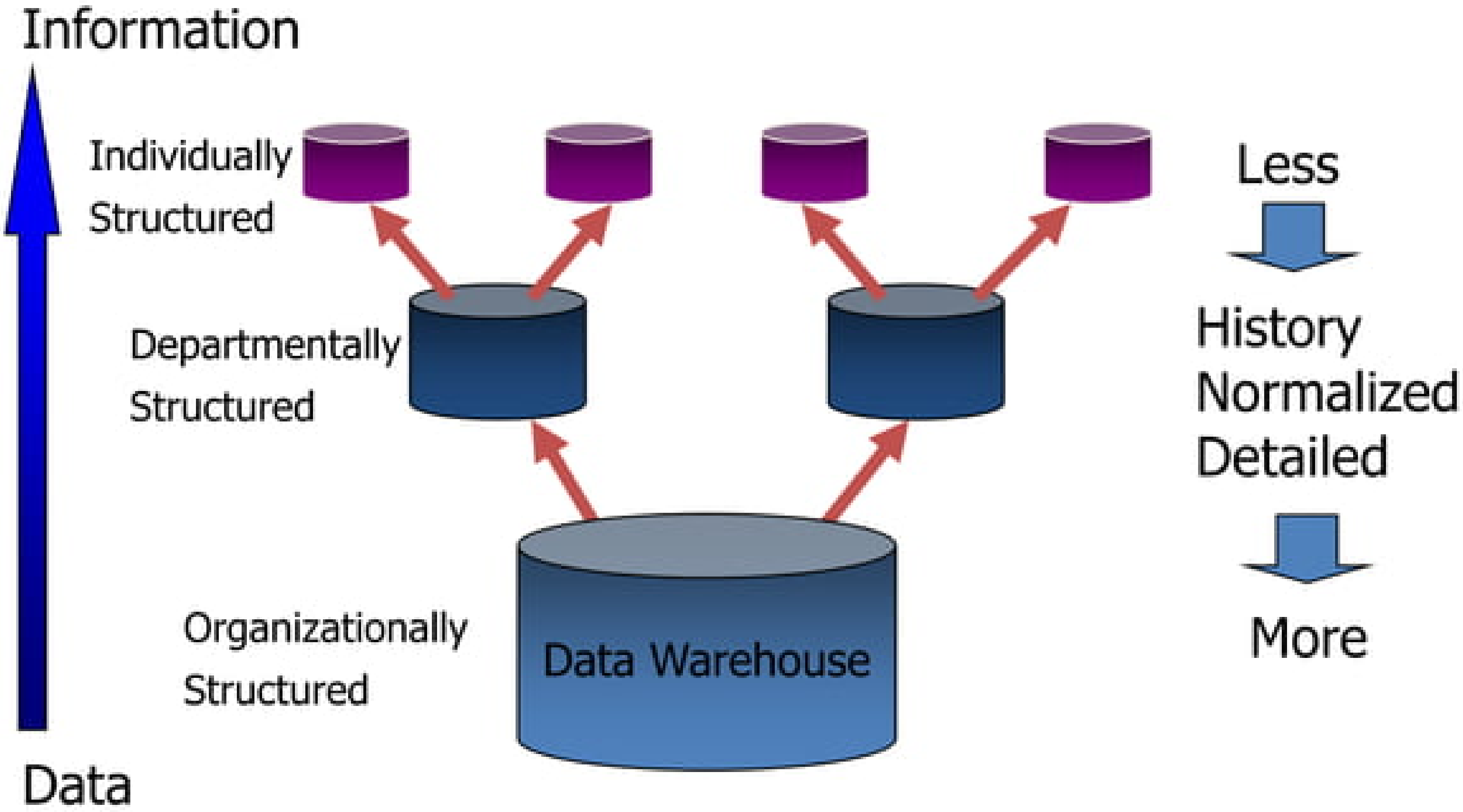
Data Marts

- 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

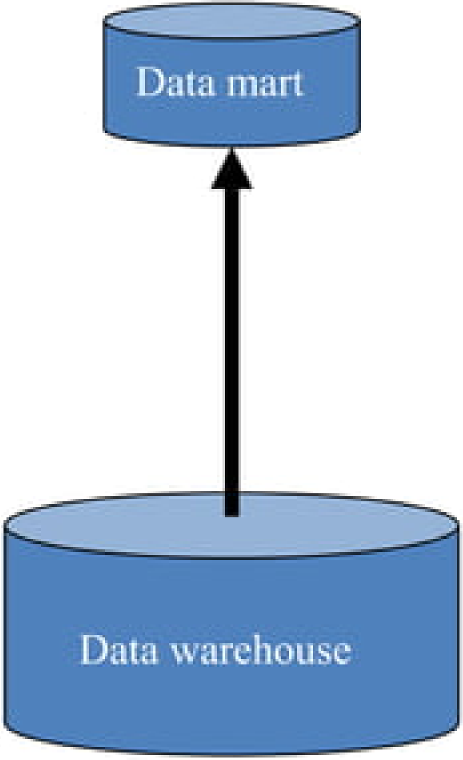
Reasons for creating a data mart

- 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

Data Mining

- **Data Mining** is the process of extracting information from the company's various databases and re-organizing it for purposes other than what the databases were originally intended for.
- It provides a means of extracting previously unknown, predictive information from the base of accessible data in data warehouses.
- Data mining process is different for different organizations depending upon the nature of the data and organization.
- Data mining tools use sophisticated, automated algorithms to discover hidden patterns, correlations, and relationships among organizational data.
- Data mining tools are used to predict future trends and behaviors, allowing businesses to make proactive, knowledge driven decisions.
- For ex: for targeted marketing, data mining can use data on past promotional mailings to identify the targets most likely to maximize the return on the company's investment in future mailings.

Functions

- Classification: It infers the defining characteristics of a certain group
- Clustering: identifies group of items that share a particular characteristic
- Association: identifies relationships between events that occur at one time
- Sequencing: similar to association, except that the relationship exists over a period of time
- Forecasting: estimates future values based on patterns within large sets of data

Characteristics

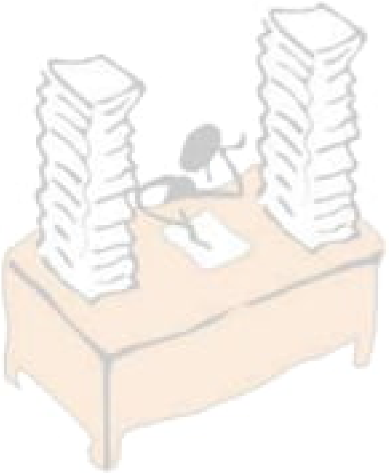
- Data mining tools are needed to extract the buried information “ore”.
- The “miner” is often an end user, empowered by “data drills” and other power query tools to ask ad hoc questions and get answers quickly, with little or no programming skill.
- The data mining environment usually has a client/server architecture.
- Because of the large amounts of data, it is sometimes necessary to use parallel processing for data mining.
- Data mining tools are easily combined with spreadsheets and other end user software development tools, enabling the mined data to be analyzed and processed quickly and easily.
- Data mining yields five types of information: associations, sequences, classifications, clusters and forecasting.
- “Striking it rich” often involves finding unexpected, valuable results.

Common data mining applications

APPLICATION	DESCRIPTION
Market segmentation	Identifies the common characteristics of customers who buys the same products from the company
Customer churn	Predicts which customers are likely to leave your company and go to a competitor
Fraud detection	Identifies which transactions are most likely to be fraudulent
Direct marketing	Identifies which prospects should be included in a mailing list to obtain the highest response rate
Market based analysis	Understands what products or services are commonly purchased together
Trend analysis	Reveals the difference between a typical customer this month versus last month
Science	Simulates nuclear explosions; visualizes quantum physics

Entertainment	Models customer flows in theme parks; analyzes safety of amusement parks rides
Insurance and health care	Predicts which customers will buy new policies; identifies behavior patterns that increase insurance risk; spots fraudulent claims
Manufacturing	Optimizes product design, balancing manufacturability and safety; improves shop-floor scheduling and machine utilization
Medicine	Ranks successful therapies for different illnesses; predicts drug efficacy; discovers new drugs and treatments
Oil and gas	Analyzes seismic data for signs of underground deposits ; prioritizes drilling locations; simulates underground flows to improve recovery
Retailing	Discerns buying-behavior patterns; predicts how customers will respond to marketing campaigns

Data Mining works with Data Warehouse



- *Data Warehousing provides the Enterprise with a memory*

- *Data Mining provides the Enterprise with intelligence*

Data mining for decision support

Two capabilities are provided new business opportunities

- Automated prediction of trends and behavior: for ex, targeted marketing.
- Automated discovery of previously unknown patterns: for ex, detecting fraudulent credit card transactions and identifying anomalous data representing data entry-keying errors.

Data mining tools

IT tools and techniques are used by data miners

- Neural computing: It is a machine learning approach by which historical data can be examined for patterns.

- Intelligent agents: It is the promising approach to retrieve information from the internet or from intranet-based databases.

- Association analysis: An approach that uses a specialized set of algorithms that sort through large data sets and expresses statistical rules among items.

Text mining

- Text mining is the application of data mining to non structured or less structured text files.
- Operates with less structured information
- Frequently focused on document format rather than document content

Text mining helps in….

- Find the “hidden” content of documents, including additional useful relationships
- Relate documents across previously unnoticed divisions (e.g.: discover that customers in two different product divisions have the same characteristics)
- Group documents by common themes (e.g.: identify all the customers of an insurance firms who have similar complaints and cancel their policies)

To summarize ...

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

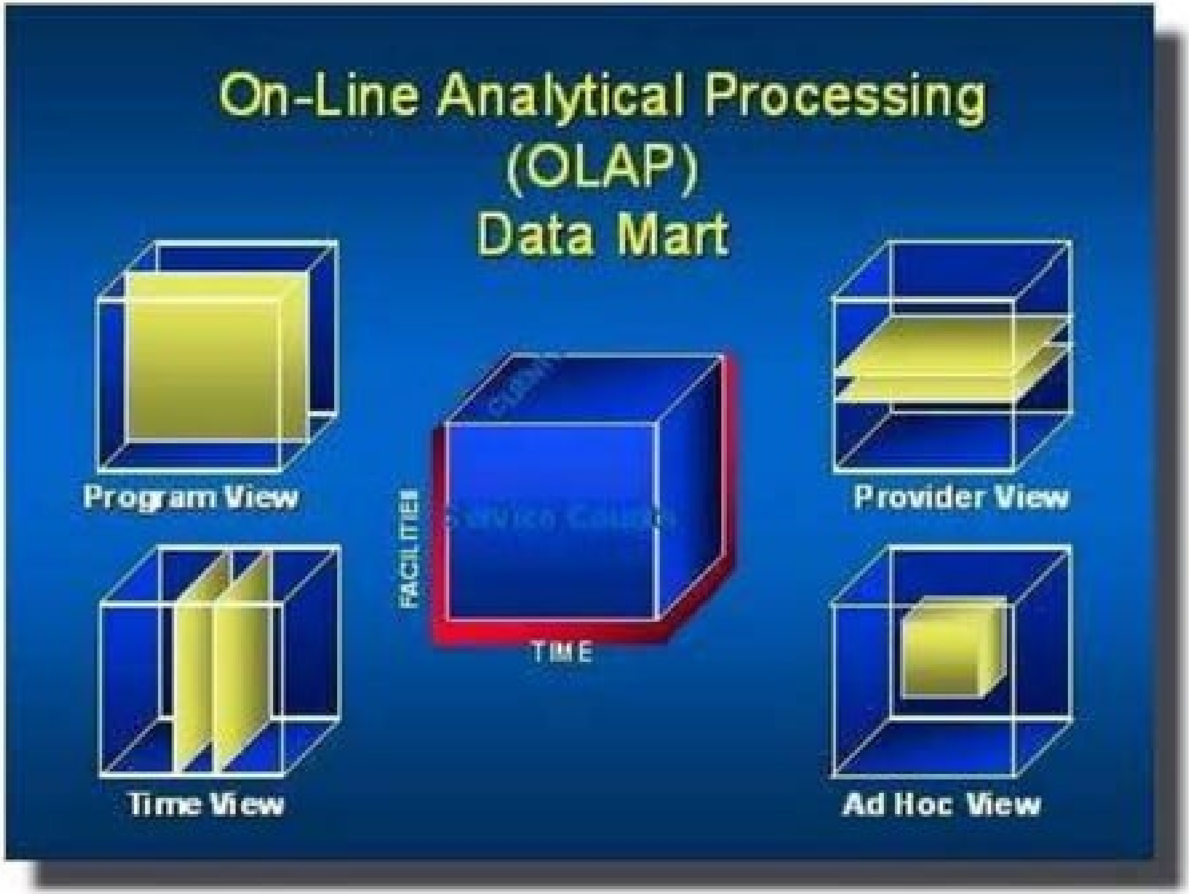
- The Data Warehouse helps to *"optimize"* the business

OLAP

- Online Analytical Processing - coined by EF Codd in 1994 paper contracted by Arbor Software
- Generally synonymous with earlier terms such as Decisions Support, Business Intelligence, Executive Information System
- OLAP = Multidimensional Database

OLAP

- Online analytical processing refers to such end user activities as DSS modelling using spreadsheets and graphics that are done online.
- OLAP involves many different data items in complex relationships.
- Objective of OLAP is to analyze complex relationships and look for patterns, trends and exceptions.



On-Line Analytical Processing (OLAP) Data Mart

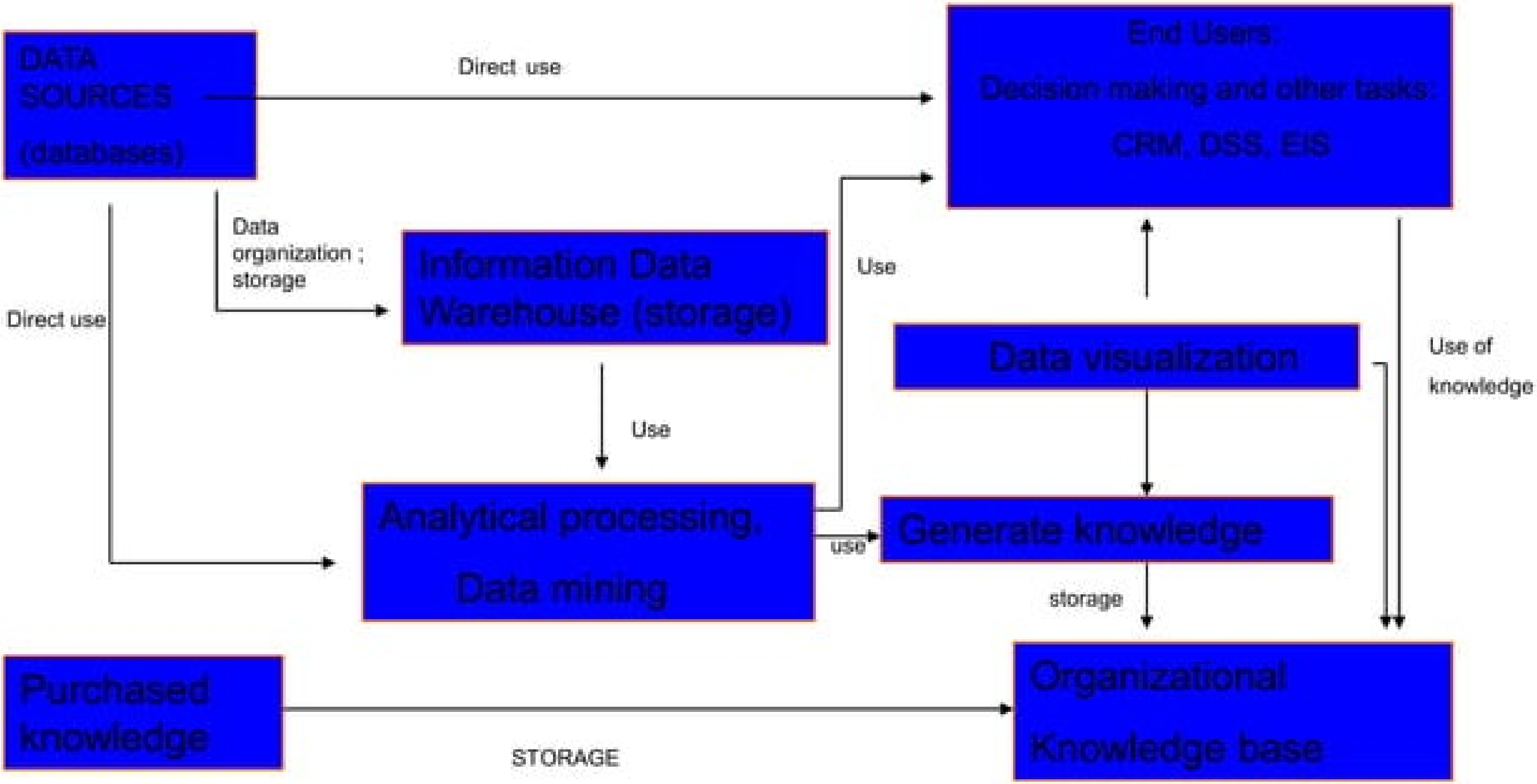
OLAP Is FASMI

- Fast
- Analysis
- Shared
- Multidimensional
- Information

Strengths of OLAP

- It is a powerful visualization paradigm
- It provides fast, interactive response times
- It is good for analyzing time series
- It can be useful to find some clusters and outliers
- Many vendors offer OLAP tools such as brio.com, cognus.com, microstrategy.com etc and it is possible to access an OLAP database from web.

Data warehousing integration



DATA SOURCES (databases)

Direct use

End Users: Decision making and other tasks: CRM, DSS, EIS

Data organization ; storage

Information Data Warehouse (storage)

Direct use

Use

Use

Data visualization

Use of knowledge

Analytical processing, Data mining

use

Generate knowledge

storage

Purchased knowledge

STORAGE

Organizational Knowledge base

- Businesses run on information and the knowledge of how to put that information to use.

- Knowledge is not readily available, it is continuously constructed from data and/or information, in a process that may not be simple or easy.

- The transformation of data into knowledge may be accomplished in several ways

Data collection from various sources stored in simple databases

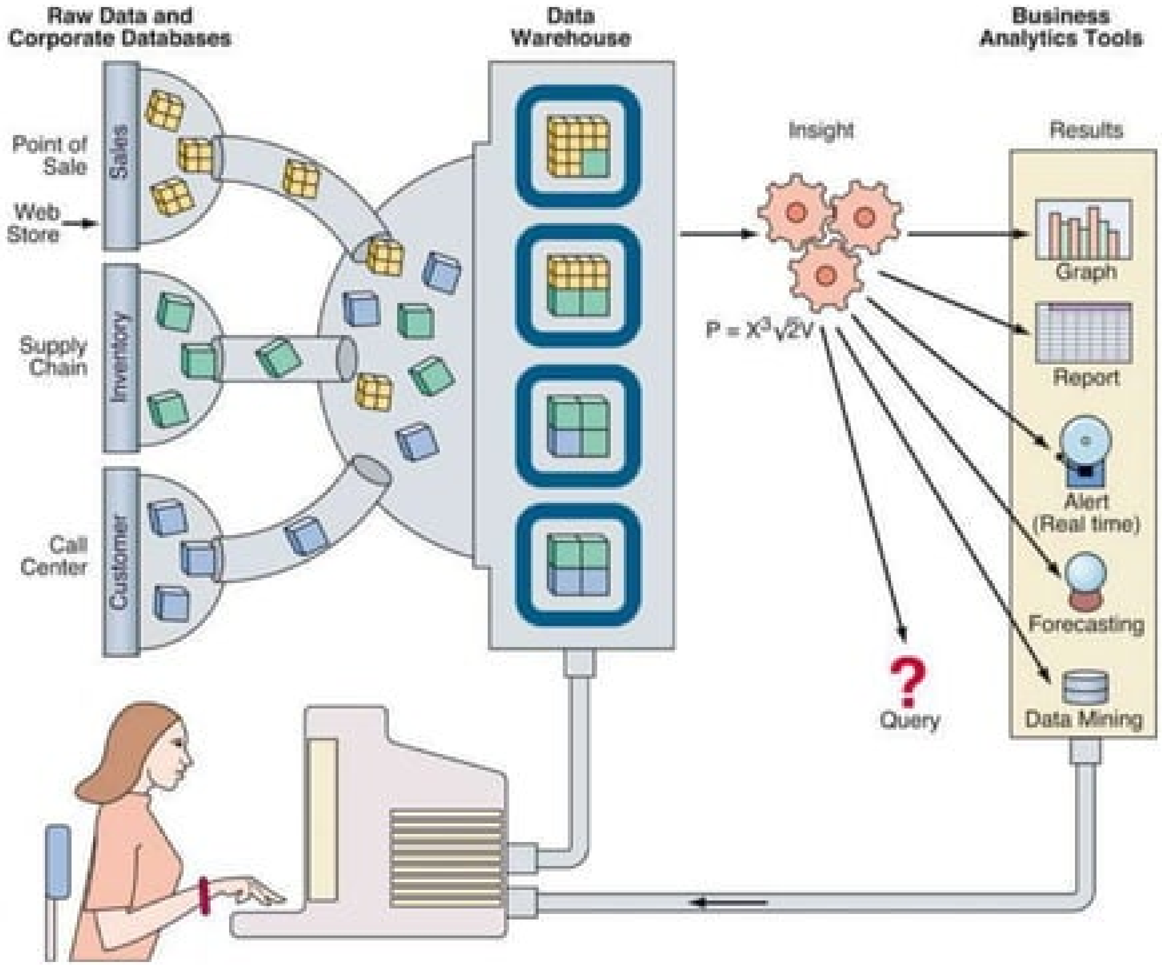
- Data can be processed, organized, and stored in a data warehouse and then analyzed (e.g.) by using analytical processing) by end users for decision support.
- Some of the data are converted to information prior to storage in the data warehouse, and some of the data and/or information can be analyzed to generate knowledge. For example, by using data mining, a process that looks for unknown relationships and patterns in the data, knowledge regarding the impact of advertising on a specific group of customers can be generated.
- This generated knowledge is stored in an organizational knowledge base, a repository of accumulated corporate knowledge and of purchased knowledge.
- The knowledge in the knowledge base can be used to support less experienced and users, or to support complex decision making.

Both the data and the information, at various times during the process, and the knowledge derived at the end of the process, may need to be presented to users.

Data Warehouse for Decision Support

- Putting Information technology to help the knowledge worker make faster and better decisions
- Used to manage and control business
- Data is historical or point-in-time
- Optimized for inquiry rather than update
- Use of the system is loosely defined and can be ad-hoc
- Used by managers and end-users to understand the business and make judgments

Business intelligence and data warehousing



Business Intelligence

- One ultimate use of the data gathered and processed in the data life cycle is for business intelligence.
- Business intelligence generally involves the creation or use of a data warehouse and/or data mart for storage of data, and the use of front-end analytical tools such as Oracle's Sales Analyzer and Financial Analyzer or Micro Strategy's Web.
- Such tools can be employed by end users to access data, ask queries, request ad hoc (special) reports, examine scenarios, create CRM activities, devise pricing strategies, and much more.

How business intelligence works?

- The process starts with raw data which are usually kept in corporate data bases. For example, a national retail chain that sells everything from grills and patio furniture to plastic utensils had data about inventory, customer information, data about past promotions, and sales numbers in various databases.
- Though all this information may be scattered across multiple systems-and may seem unrelated-business intelligence software can being it together. This is done by using a data warehouse.
- In the data warehouse (or mart) tables can be linked, and data cubes are formed. For instance, inventory information is linked to sales numbers and customer databases, allowing for deep analysis of information.

- Using the business intelligence software the user can ask queries, request ad-hoc reports, or conduct any other analysis.
- For example, deep analysis can be carried out by performing multilayer queries. Because all the databases are linked, one can search for what products a store has too much of, determine which of these products commonly sell with popular items, bases on previous sales. After planning a promotion to move the excess stock along with the popular products (by bundling them together, for example), one can dig deeper to see where this promotion would be most popular (and most profitable). The results of the request can be reports, predictions, alerts, and/or graphical presentations. These can be disseminated to decision makers to help them in their decision-making tasks.

More advanced applications of business intelligence include outputs such as

- financial modeling
- budgeting
- resource allocation
- and competitive intelligence.

THANK
YOU