***Prelude/Intro***

**What is data?**

Data progression since the advent of computing

Data --> Information --> Knowledge --> Intelligence

**What is a warehouse?**

It is a facility meant large scale storage

How is data used?

Two perspectives

Software Engineering (Information System development)

End User - Top decision maker of an organization

**What is a Data Warehouse?**

The term data warehouse was introduced by William Inmon known as the father of data warehouse.

The original concept was essentially a historical database containing tables derived from an active operational database.

A large store of data accumulated from a wide range of sources within a company and used to guide management decisions.

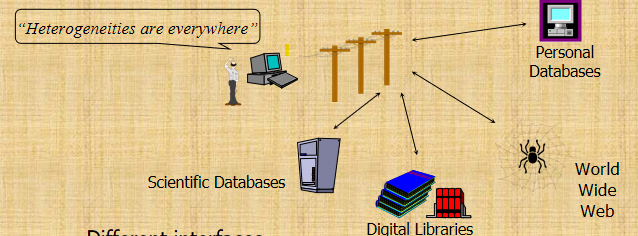
A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a way that can be understood and used in business context.

[Barry Devlin]

Data Warehousing (DW) is a process for collecting and managing data from varied sources to provide meaningful business insights.

 A copy of transaction data specifically structured for query and analysis. [Ralph Kimball]

***Problem: Heterogeneous Information Sources***



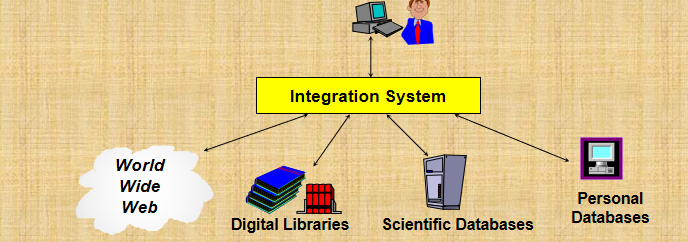
### ****Types of Information Sources:****

* **Personal Databases**: Individual databases that may store information for personal or department-specific use.
* **Scientific Databases**: Databases used for scientific research and data, often specific to certain fields.
* **Digital Libraries**: Collections of digital resources that are typically used for academic, research, or institutional purposes.
* **World Wide Web**: The vast array of information available online, which is diverse in nature and often unstructured.

### ****Challenges Associated with Heterogeneous Information:****

* **Different Interfaces**: Each source may have its own unique interface, making it difficult to access and integrate data seamlessly.
* **Different Data Representations**: Information may be represented in different formats or structures across various sources, leading to inconsistencies.
* **Duplicate and Inconsistent Information**: Redundant data may be present across different sources, and inconsistencies can arise due to varying data entry standards or update cycles.

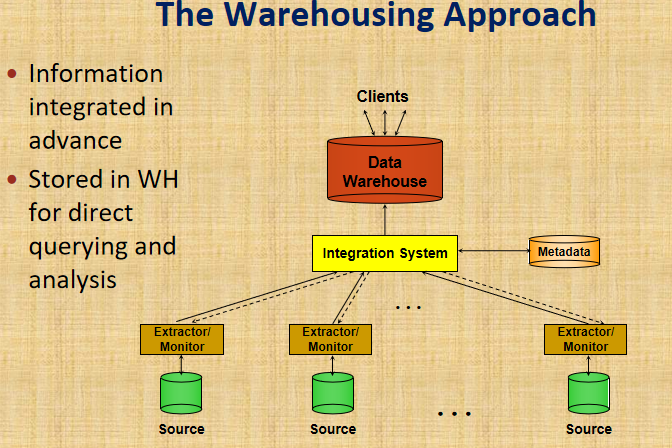
***Goal: Unified Access to Data***



**Integration System**: The central element in the diagram, which serves as a bridge between various types of information sources. This system is designed to:

* **Collect and Combine Information**: It gathers data from diverse sources, such as the World Wide Web, digital libraries, scientific databases, and personal databases.
* **Provide Integrated View, Uniform User Interface**: The integration system offers a cohesive view of all collected data, presenting it through a consistent and user-friendly interface. This helps in overcoming the issue of different interfaces across various sources.
* **Support Sharing**: The system facilitates the sharing of information across different parts of the organization or among users, enhancing collaboration and decision-making.

***The Warehousing Approach***



**Data Warehouse**:

* + A central repository where information is integrated and stored in advance.
  + It allows for direct querying and analysis by clients (end-users).

**Integration System**:

* + Acts as a bridge between various data sources and the data warehouse.
  + It processes and prepares the data for storage in the warehouse, ensuring that it is organized and consistent.

**Extractors/Monitors**:

* + These components are responsible for gathering data from different sources.
  + They extract data, monitor for updates or changes, and feed this information into the integration system.

**Meta data**:

* + Metadata provides information about the data itself, such as structure, format, and meaning, which is essential for efficient querying and data management.

### ****Key Points:****

* **Information Integrated in Advance**: Data is pre-processed and integrated before being stored in the data warehouse, ensuring it is ready for use.
* **Stored in WH for Direct Querying and Analysis**: The data warehouse stores this integrated information, making it easily accessible for querying and analysis by clients.

### ****Process Flow:****

1. **Data Sources**: Data originates from various sources, each with its own extractor/monitor system.
2. **Extractors/Monitors**: These systems extract the necessary data from each source and send it to the integration system.
3. **Integration System**: This system processes and consolidates the extracted data, ensuring it meets the standards for storage.
4. **Data Warehouse**: The processed data is stored in the data warehouse, where it is organized and indexed for easy access.
5. **Clients**: End-users or applications query the data warehouse to retrieve information for analysis, decision-making, or other purposes.

***The Compelling Need For Data Warehousing***

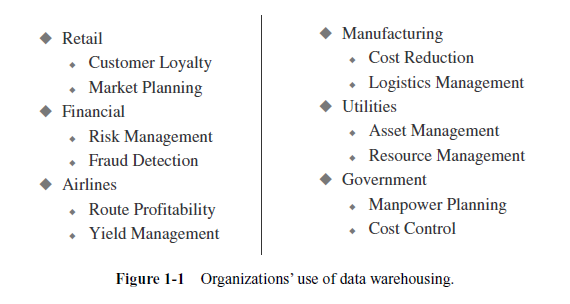
**Introduction:** IT professionals often work on systems like order processing and inventory management that support daily operations. These systems handle data for business activities.

**Role of Operational Systems:** Operational systems are vital for processing data and generating reports to manage daily business functions. However, they don't provide the strategic information executives need for decision-making.

**The Need for Strategic Information:** Executives require specific information for strategic decisions. Operational systems don't meet this need, highlighting the need for a system that provides strategic data.

**Data Warehousing as a Solution:** Data warehousing has become a key tool for business intelligence (BI), offering the strategic information executives require. Since the 1990s, organizations have adopted data warehouses for competitive advantage.

**Conclusion:** Data warehousing is crucial for organizations seeking to leverage business intelligence and make informed decisions, driving long-term success.



***The Escalating Need For Strategic Information?***

**Strategic information:** refers to the critical data and insights that executives and managers within an organization need to make informed decisions that guide the overall direction of the enterprise. This type of information is essential for formulating business strategies, setting objectives, establishing goals, and monitoring results to ensure that the organization remains competitive in its market.

Key Aspects of Strategic Information:

**1. Purpose:**

**- Decision-Making:** Strategic information is primarily used to support decision-making at the highest levels of the organization. It helps in identifying opportunities and threats, assessing competitive positions, and making choices that affect the long-term success of the enterprise.

**- Strategy Formulation**: It aids in the development of business strategies by providing a comprehensive understanding of the market, customer behaviors, industry trends, and internal capabilities.

**2. Examples:**

**- Customer Retention:** Information that helps retain the current customer base.

**- Market Expansion:** Data to support strategies aiming to increase the customer base by a specific percentage over a certain period.

**- Market Share:** Insights on how to gain a larger market share within a set timeframe.

**- Product Quality:** Information related to improving product quality in key product groups.

**- Customer Service:** Data used to enhance customer service levels, particularly in shipments.

**- Product Development:** Information to guide the introduction of new products within a planned timeline.

**- Sales Growth:** Strategic insights to increase sales in specific regions or divisions.

**3. Characteristics of Strategic info:**

**- Integrated:** Strategic information must offer a unified, enterprise-wide perspective, ensuring that all parts of the organization are aligned with the overall strategy.

**- Data Integrity:** The information must be accurate and adhere to established business rules to be reliable.

**- Accessible:** It should be easily accessible, with clear and intuitive access paths, allowing for quick retrieval and analysis.

**- Credible:** Each business factor or metric should have a singular, consistent value, ensuring that decisions are based on trustworthy data.

**- Timely:** The information must be available within the required timeframe to ensure that decisions can be made promptly and effectively.

***The Information Crisis?***

**Key Factors:**

**Vast Data Repositories:** Large-scale companies utilize numerous computer applications and databases that store extensive amounts of customer and financial data, often spanning ten to fifteen years or more.

**Fragmented Data Storage:** This data is scattered across various platforms, including legacy systems, client/server applications, and possibly different locations, making it difficult to manage cohesively.

**Startling Facts:**

* + **Abundance of Data:** Organizations have immense amounts of data.
  + **Ineffective Data Utilization:** Existing IT resources and systems struggle to convert this data into useful strategic information.

**The Information Crisis:** Despite the abundance of information, companies face an information crisis. This isn't due to a lack of data but because the available data isn't readily usable for strategic decision-making. The data required to analyze trends and guide company direction isn't in a suitable format.

**Challenges with Operational Data:** Operational data, which is event-driven, only provides snapshots of specific transactions, making it difficult to track trends over time. For example, operational systems don’t easily allow analysis of a product’s performance over a month, quarter, or year.

**The Solution:** Data warehousing is the solution to this "data glut, knowledge scarcity" problem. It serves as a decision-support system, converting operational data into a format that supports strategic decision-making. Historical data stored in the warehouse is essential for providing the business intelligence needed for effective trend analysis and informed decision-making.

**Why a Data Warehouse?**

**Purpose of a Data Warehouse:**

Businesses require in-depth insights to support strategic decision-making processes. To achieve this, they need answers to critical questions such as:

-What happened? Understanding past events is crucial for analyzing outcomes and performance.

- Why has it happened? Identifying the reasons behind past events helps in improving future strategies.

- What will happen? Predicting future trends and outcomes is essential for proactive decision-making.

- What is happening? Monitoring current activities allows businesses to make timely adjustments.

- What do you want to happen? Setting and aligning goals with strategic objectives ensures long-term success.

**Role of the Data Warehouse:**

A data warehouse serves as a decision support system, specifically designed to provide the strategic information needed to answer these critical questions. It enables enterprises to analyze historical and current data, thereby facilitating informed and effective strategic decision-making.

****Technology Trends****

The entire spectrum of computing has undergone tremendous changes, evolving from:

* **Data** to **Information**
* **Information** to **Knowledge**
* **Knowledge** to **Intelligence**

This progression highlights the explosive growth in technology occurring day by day.

Three critical areas influencing this growth are:

**Computing Technology**: Advances in hardware, software, and networking continue to drive improvements in performance and capabilities.

**Human/Machine Interface**: Innovations in how we interact with machines, such as through touchscreens, voice recognition, and augmented reality, enhance usability and accessibility.

**Processing Options**: The development of various processing methods, including cloud computing, edge computing, and quantum computing, offers diverse solutions for handling complex tasks and large data sets.

***Why have past attempts to provide the marketing department with useful decision support reports ended in failure?***

**Answer:** The past Decision Support Systems failed to meet the needs of the marketing department due to several key reasons:

**Lack of Regular Reporting:** There were no regular, automated reports available that provided the marketing department with the necessary information on a consistent basis.

**Data Fragmentation:** The required data was scattered across multiple applications and platforms, making it difficult and time-consuming to gather and consolidate information.

**Ad-Hoc Reporting Limitations:** The ad-hoc reports produced were initially well-received but eventually fell short as the marketing department requested more specific formats and additional details. The iterative process of creating reports to meet evolving requirements led to inconsistencies and dissatisfaction.

**Inconsistencies in Data:** The data obtained from different applications showed inconsistencies, making it challenging to derive accurate insights and making decision-making difficult.

*****Lack of Integration:*****The absence of an integrated system for data collection and reporting hindered the ability to produce cohesive and reliable analyses.

These issues collectively led to the failures of the past Decision Support Systems, as they were unable to provide the timely, accurate, and comprehensive information needed for effective strategic decision-making.

**History of Decision-Support Systems**

**Ad Hoc Reports**:

* 1. **Description**: IT would create special programs on a case-by-case basis for each report request.
  2. **Characteristics**: Custom-built for specific needs, often involving manual intervention.

**Special Extract Programs**:

* 1. **Description**: IT developed programs to anticipate and generate reports based on expected requests.
  2. **Characteristics**: These programs were designed to handle common reporting needs but still required customization.

**Small Applications**:

* 1. **Description**: Users could specify parameters for generating special reports through these applications.
  2. **Characteristics**: Allowed some level of user interaction and customization but were limited in scope and flexibility.

**Information Centers**:

* 1. **Description**: Centralized locations where users could request ad hoc reports.
  2. **Characteristics**: Served as hubs for obtaining reports but did not provide immediate access to information.

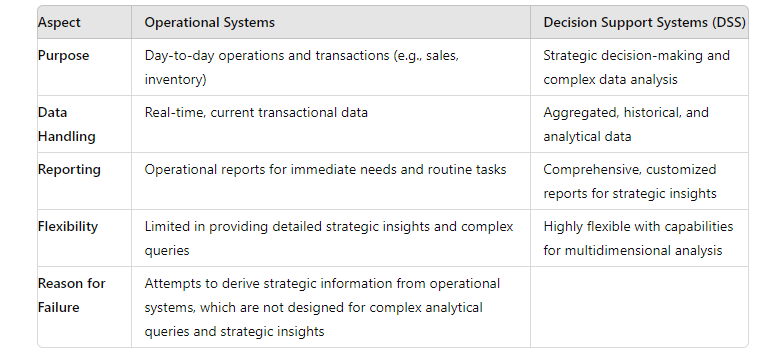
**Decision-Support Systems (DSS)**:

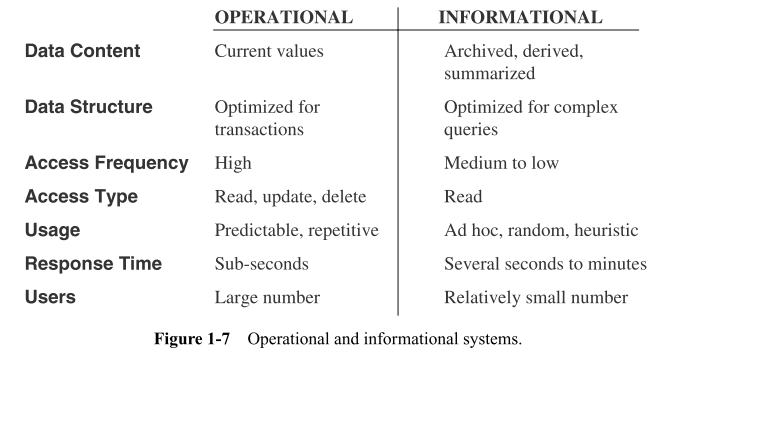
* 1. **Description**: Menu-driven systems that provided online information and the capability to generate special reports.
  2. **Characteristics**: Offered more interactive and user-friendly access to data and reporting features.

**Executive Information Systems (EIS)**:

* 1. **Description**: Aimed to provide strategic information directly to the executive desktop.
  2. **Characteristics**: Focused on delivering high-level insights and summaries for strategic decision-making.

***Operational Versus Decision support systems***





***How are they Different?***

**Operational Systems:** Making The Wheels Of The Business Turn

Take an order

Process a claim

Make a shipment

Generate an invoice

Receive cash

Reserve an airline seat.

DWH: Watching the Wheels Of The Business Turn

Show me the top-selling products.

Show me the problem regions.

Tell me why (drill down)

Show me the highest margins

Alert me when a district sells below target

Different scope different purposes!!

Different purposes

Different scope

Data content is different

Data usage patterns are different

Data access types are different.

***DWH: An Environment, not a Product***

Data warehouse is a computing environment where users can find strategic information.

Flexible and Interactive.

Its 100% user driven.

Responsive to ask-answer-ask-again pattern.

Provide the ability to discover answers to complex and unpredictable questions.

****Functions of Data Warehousing Tools and Utilities****

Data warehousing tools and utilities are essential for managing and optimizing data in a warehouse environment. Their key functions include:

**Data Extraction:** This function involves gathering data from multiple heterogeneous sources. It ensures that data from various origins is collected for integration into the data warehouse.

**Data Cleaning:** This step focuses on finding and correcting errors in the data. Data cleaning is crucial for improving the quality of the data and ensuring accurate analysis.

**Data Transformation:** This process involves converting data from its legacy format into a format compatible with the data warehouse. It ensures that the data is standardized and ready for analysis.

**Data Loading:** This function encompasses sorting, summarizing, consolidating, checking integrity, and building indices and partitions. It prepares the data for efficient querying and reporting within the warehouse.

**Refreshing:** This involves updating the data warehouse with the latest data from the source systems. Regular refreshing ensures that the warehouse reflects the most current information.

**Note:** Data cleaning and data transformation are particularly important as they significantly enhance the quality of data and the results of data mining activities.

***TeraData***

Teradata offers a full suite of service which focuses on Data Warehousing.

Teradata is massively parallel open processing system for developing large-scale data warehousing applications.

Teradata is an open system. It can run on Unix/Linux/Windows server platform.

This tool provides support to multiple data warehouse operations at the same time to different clients.

The system is built on open architecture. So whenever any faster devices are made available, it can be incorporated into the already build architecture.

Teradata supports 50+ petabytes of data.

***CHAPTER NO 02***

***Major Defining Features***

The most popular definition came from Bill Inmon, who provided the following:

“A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process”

*****Data Warehouse Characteristics/defining features and Functions*****

**Characteristics of Data in a Data Warehouse:**

* **Separate:** Data in a data warehouse is distinct from data in operational systems, focusing on a consolidated view of subject areas.
* **Available:** Data is accessible for analysis and reporting purposes.
* **Accessible:** Data is readily available for querying and analysis.
* **Subject-Oriented:** Data warehouses are designed around specific subjects like customers, products, or sales rather than individual applications.
* **Integrated:** Data is integrated from various OLTP (Online Transaction Processing) systems and other sources to provide a unified view.
* **Time-Variant:** Data includes historical snapshots to analyze trends, patterns, and changes over time.
* **Non-Volatile:** Data in a data warehouse is read-only and not subject to frequent changes. It is updated periodically through data refreshes.

*****Operational Systems vs. Data Warehousing*****

* **Operational Systems:** Store data related to individual applications such as order processing or consumer loans. These systems are optimized for transaction processing and current data.
* **Data Warehouses:** Model data around analysis requirements of decision-makers, providing a streamlined view of specific subjects (e.g., sales, shipments, inventory) rather than operational details.

***Data Integration Challenges:***

Data in a data warehouse comes from various sources, including OLTP systems and non-traditional data sources. Challenges include:

* **Different Operational Platforms:** Data may come from various types of systems.
* **Operating Systems and File Layouts:** Variations in how data is formatted and stored.
* **Character Code Representations:** Different methods for encoding characters.
* **Field Naming Conventions:** Inconsistencies in how fields are named across systems.

*****Example of Data Variations:*****

* **Gender Field:**
  + **Application A:** Stores gender as logical values (M or F).
  + **Application B:** Uses numerical values for gender.
  + **Application C:** Represents gender as character values.
* **Date and Balance Fields:** Similar discrepancies may occur in date formats and balance representations.

*****Time-Variant Nature:*****

* **Operational Systems:** Typically store current data, such as account balances and order statuses.
* **Data Warehouses:** Store data as snapshots of past and current periods, allowing for the analysis of historical trends, current information, and future forecasts.

***6. Non-Volatility:***

* **Data Storage:** Data in a data warehouse is not erased when new data is entered. It is read-only and updated periodically.
* **Data Refresh:** Data is moved from operational systems to the data warehouse at specific intervals, not with every business transaction.