Data Mining and Business Intelligence

ITA5007

PROF. E.P.EPHZIBAH

Topic

DATA WAREHOUSING AND BUSINESS INTELLIGENCE

What is a Data Warehouse?

- 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(the father of data warehousing, came up with the concept to develop a data warehouse that identifies the main subject areas)

Data warehousing:

- The process of constructing and using data warehouses.

Data Warehouse—Subject-Oriented

Organized around major subjects, such as customer, product, sales, etc.

Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing

Provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

A data warehouse is subject-oriented since it provides topic-wise information rather than the overall processes of a business. Such subjects may be sales, promotion, inventory, etc.

For example, if you want to analyze your company's sales data, you need to build a data warehouse that concentrates on sales.

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.

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)

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”

Data Warehouse—Nonvolatile

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*

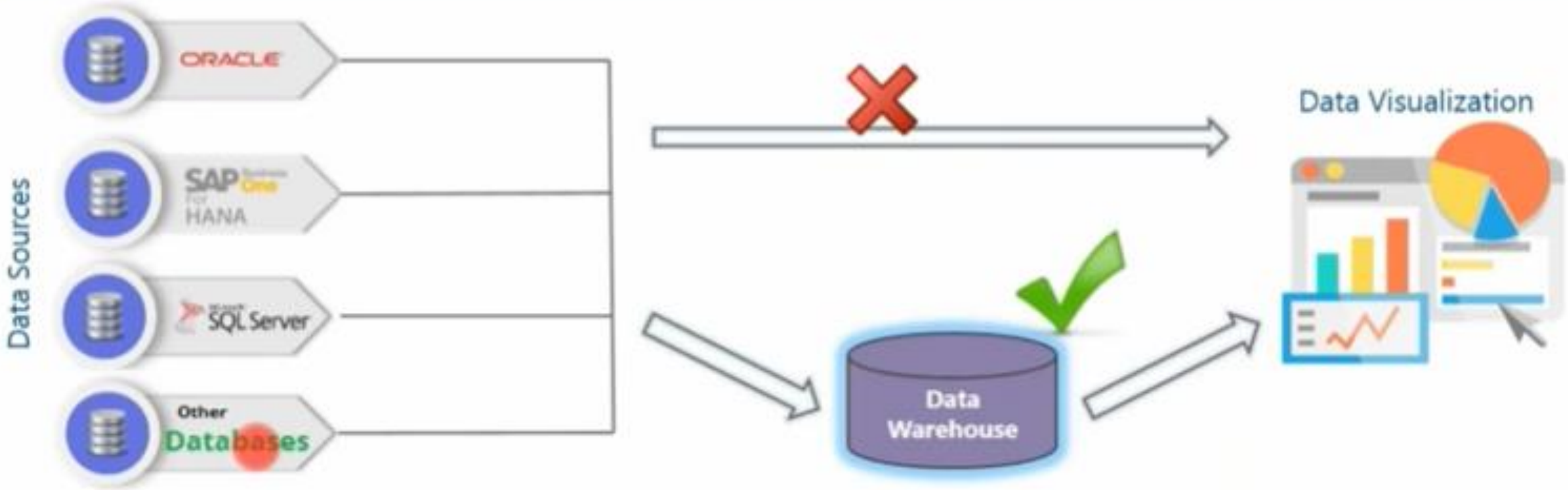
OLTP vs OLAP

OnLine Transaction Processing vs OnLine Analytical Processing

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

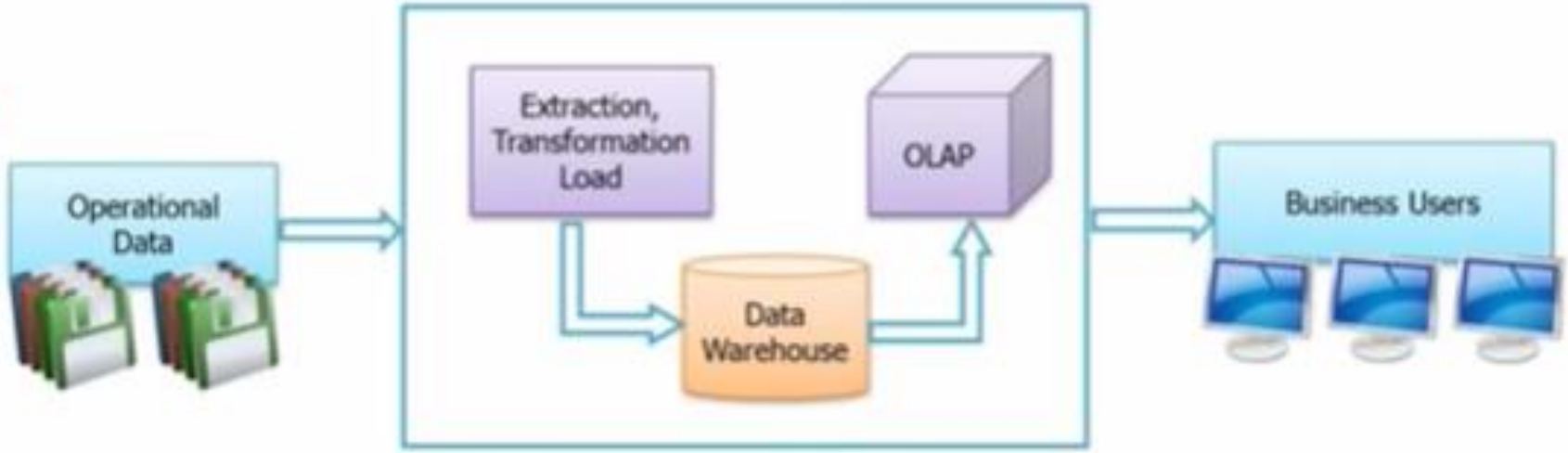
Why Data Warehouse?

- Data collected from various sources & stored in various databases cannot be directly visualized.
- The data first needs to be **integrated** and then **processed** before visualization takes place.

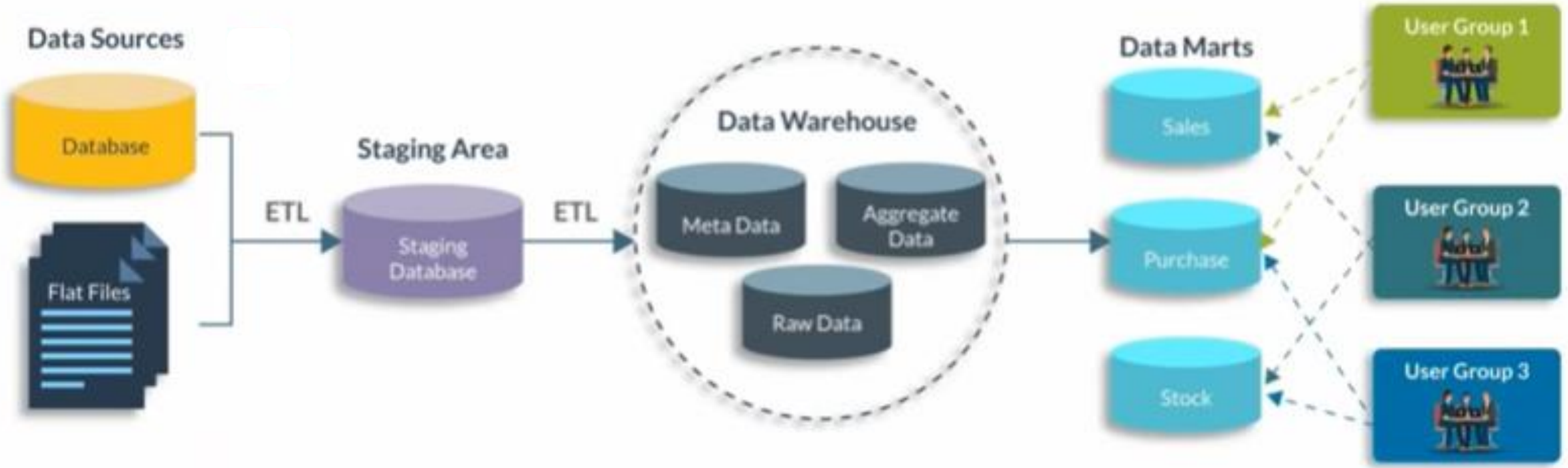


What Is A Data Warehouse?

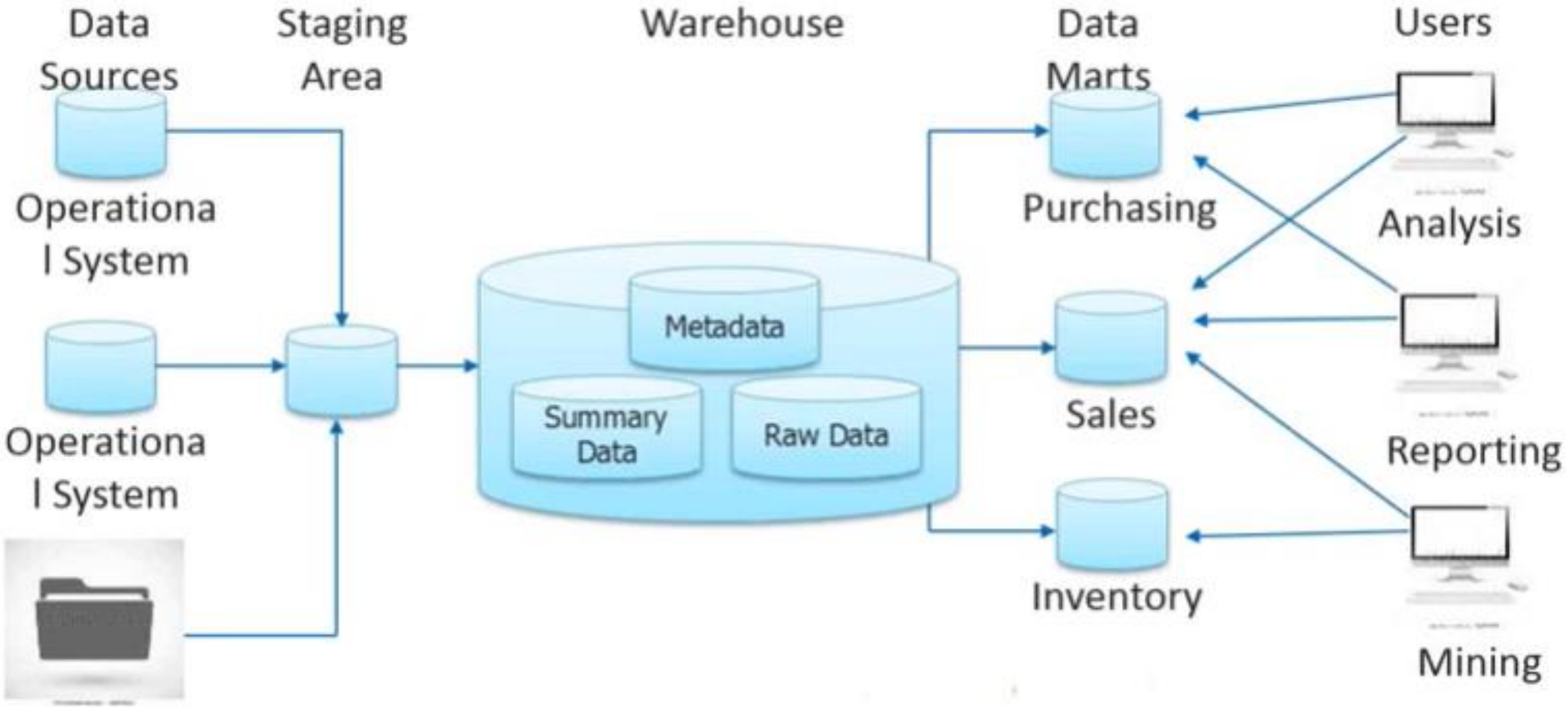
- A central location where consolidated data from multiple locations (databases) are stored.
- DWH is maintained separately from an organization's operational database.
- End users access it whenever any information is needed.
- **Note**:- Data Warehouse is not loaded every time new data is added to database.



Data Warehouse Architecture



Data Ware House Architecture



What is Business Intelligence?

Business intelligence (BI) is the set of techniques and tools for the transformation of raw / production and operational data into meaningful and useful information for business analysis purposes for various level.

Business intelligence (BI) talks about how traditional data which transform into the BI which have multiple initiatives to measure, manage, and improve on the performance of individuals, processes, teams, and business units for the specific business area.

During the operation of business, the following questions must be asked. The functions of monitoring, analyzing, and planning delve into these questions as follows :

- What has happened?
- What is happening?
- Why?
- What will happen?
- What do we want to have happen?

Why Business Intelligence?

Business Intelligence is the activity which contributes to the growth of any company.

What Is Business Intelligence?

BI is the act of transforming raw/ operational data into useful information for business analysis.

How Does It Work?

1. BI based on Data Warehouse technology **extracts** information from a company's operational systems.
2. The data is **transformed** (*cleaned and integrated*), and **loaded** into Data Warehouses.
3. Since this data is credible, it is used for business insights.

What is Business Intelligence (BI)?

Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes.

The main purpose of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make effective and timely decisions.

Business Intelligence

What is Business Intelligence

- Business Intelligence (BI) is about getting the right information, to the right decision makers, at the right time.
- BI is an enterprise-wide platform that supports reporting, analysis and decision making.
- BI leads to:
  - fact-based decision making
  - "single version of the truth"

Business Intelligence

What is Business Intelligence

- Making useful, actionable insight from stored data.
- Allows effective business decisions to be made.
- The act of using historical data to gain new information.
- Techniques include:
  - multidimensional analyses
  - mathematical projection
  - modeling
  - ad-hoc queries
  - 'canned' reporting
  - Dashboards

The 5 Stages of Business Intelligence

- 1.The Data: defining which data will be loaded into the system and analyzed.
  - Where all information is stored
  - Technology dependent
    - MSSQL, MYSQL, Oracle, Red Brick, DB2
    - Often an OLAP type data source
  - Many rows of often summarized data
  - Utilize database queries to retrieve data from the source.
    - SQL – MSSQL and MYSQL
    - PL/SQL – Oracle

- 2.The ETL (Extract, Transform, and Load) Engine: moving the source data to the Data Warehouse.
  - This can be a complex step involving modifications and calculations on the data itself.
  - If this step doesn't work properly, the BI solution simply cannot be effective.
- 3.Data Warehousing:
  - connects electronic data from different operational systems so that the data can be queried and analyzed over time for business decision making.
  - A data warehouse is an analytically oriented, integrated, time-variant, and nonvolatile collection of data that supports decision making processes
  - Large databases that aggregate data collected from multiple sources

- 4.Analytic Engine:
  - analyzes multidimensional data sets found in a data warehouse to identify trends, outliers, and patterns.
  - Data Mining
    - is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.
    - Data mining can be used to uncover patterns in data but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data.
    - Data mining cannot discover patterns that may be present in the larger body of data if those patterns are not present in the sample being "mined".

- 5.Presentation Layer:
  - the dashboards, reports and alerts that present findings from the analysis.
  - Typically Technology Agnostic
  - The presentation layer is for the user.
    - It does not care
    - How?
    - When ?
    - Where?
    - Why?
    - the user accesses the Information just that it is available.

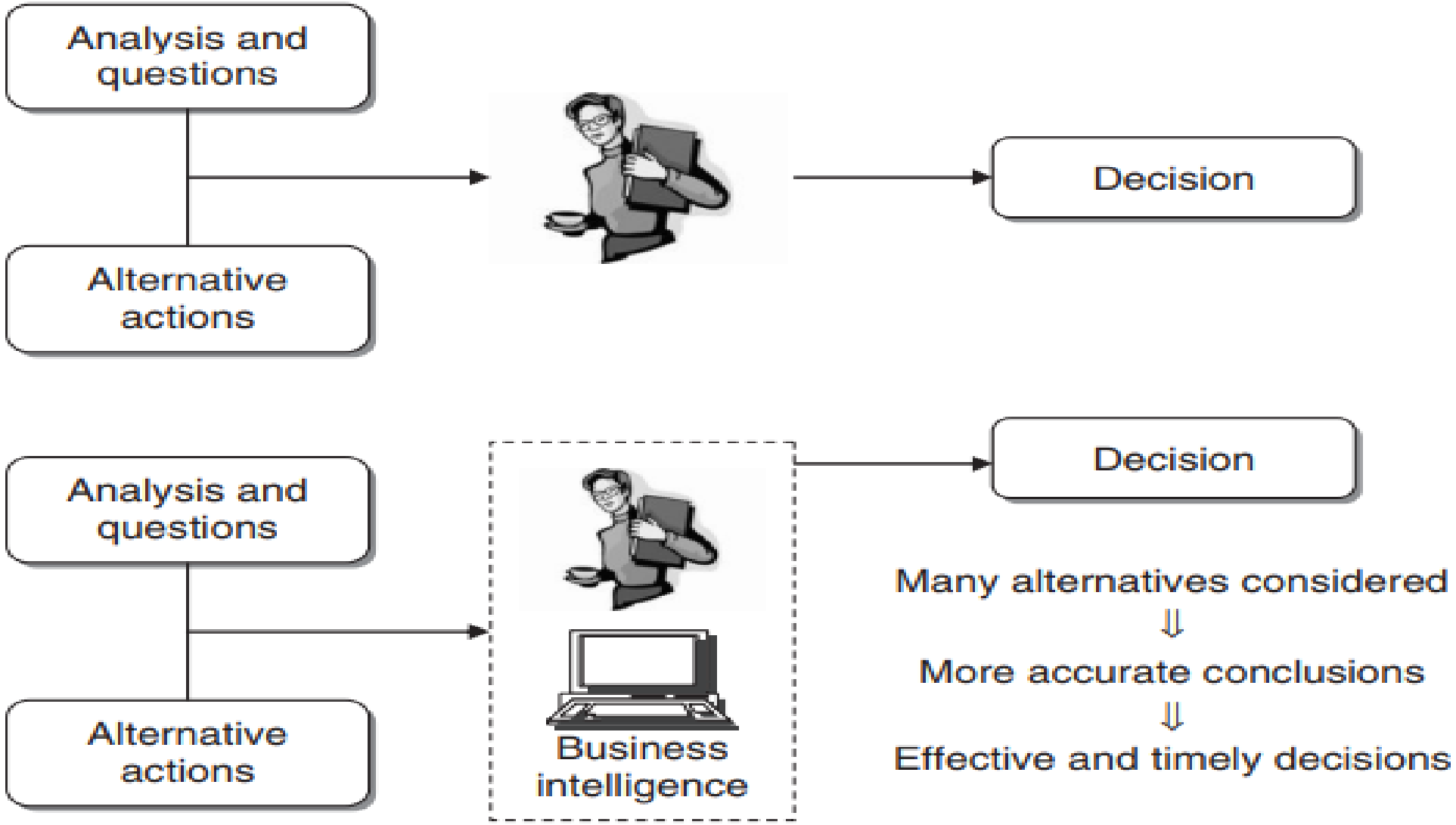


Figure 1.1 Benefits of a business intelligence system