

# Data Mining and Business Intelligence

BMBI

Subject Code	Subject Name	Credits
<b>MCA401</b>	<b>Data Mining and Business Intelligence</b>	<b>04</b>

Subject Code	Subject Name	Teaching Scheme			Credits Assigned			
		Theory	Pract.	Tut	Theory	Pract.	Tut	Total
<b>MCA 401</b>	<b>Data Mining and Business Intelligence</b>	<b>04</b>	--	--	<b>04</b>	--	--	<b>04</b>

Subject Code	Subject Name	Examination Scheme							
<b>MCA 401</b>	<b>Data Mining and Business Intelligence</b>	Theory Marks				TW	Pract.	Oral	Total
		Internal Assessment			End Semester Exam				
		Test1 (T1)	Test2 (T2)	Average of T1 & T2					
		<b>20</b>	<b>20</b>	<b>20</b>	<b>80</b>	--	--	--	<b>100</b>

<b>Sr. No.</b>	<b>Module</b>	<b>Detailed Contents</b>	<b>Hrs</b>
<b>1</b>	<b>Business Intelligence-</b>	Introduction and overview of BI-Effective and timely decisions, Data Information and knowledge, BI Architecture, Ethics and BI. BI Applications- Balanced score card, Fraud detection, Telecommunication Industry, Banking and finance, Market segmentation.	<b>06</b>
<b>2</b>	<b>Prediction methods and models for BI</b>	Data preparation, Prediction methods-Mathematical method, Distance methods, Logic method, heuristic method-local optimization technique, stochastic hill climber, evaluation of models	<b>06</b>
<b>3</b>	<b>BI using Data Warehousing</b>	Introduction to DW, DW architecture, ETL Process, Top-down and bottom-up approaches, characteristics and benefits of data mart, Difference between OLAP and OLTP. Dimensional analysis- Define cubes. Drill- down and roll- up – slice and dice or rotation, OLAP models- ROLAP and MOLAP. Define Schemas- Star, snowflake and fact constellations.	<b>08</b>
<b>4</b>	<b>Data Mining and Preprocessing</b>	Data mining- definition and functionalities, KDD Process, Data Cleaning: - Missing values, Noisy data, data integration and transformations. Data Reduction: - Data cube aggregation, dimensionality reduction- data compression, Numerosity reduction- discretization and concept hierarchy.	<b>06</b>
<b>5</b>	<b>Associations and Correlation</b>	Association rule mining:-support and confidence and frequent item sets, market basket analysis, Apriori algorithm, Incremental ARM, Associative classification- Rule Mining.	<b>06</b>
<b>6</b>	<b>Classification and Prediction</b>	Introduction, Classification methods:-Decision Tree- ID3, CART, Bayesian classification- Baye'stheorem( Naïve Bayesian classification),Linear and nonlinear regression.	<b>08</b>
<b>7</b>	<b>Clustering</b>	Introduction, categorization of Major, Clustering Methods:- partitioning methods- K-Means. Hierarchical- Agglomerative and divisive methods, Model- based- Expectation and Maximization.	<b>08</b>
<b>8</b>	<b>Web mining and Text mining</b>	Text data analysis and Information retrieval, text retrieval methods, dimensionality reduction for text. Web Mining: - web content, web structure, web usage.	<b>04</b>

# Reference books

## References:

- Business Intelligence data mining and optimization for decision making- by Carlo Verzellis ,wiley publication. [Complete book pdf](#)
- Adaptive business Intelligence by ZbigniewMichlewicz, martin Schmidt, matthewmichalewicz, constantinChiriac → [Scanned copy](#)
- Data Mining concepts and techniques second edition by Jiawei Han and MichelineKamber. [Pdf](#)
- Data Mining:” Introductory and Advanced topics” , Pearson Education, by [M.Dunham](#)
- Data warehousing Fundamentals by [PaulrajPonnian](#), John Willey
- Data mining for Business intelligence: concepts, techniques and applications in Microsoft Excel by G. Shumeli, N R Patel, P.C Bruce, Wiley

Subject Code	Subject Name	Credits
<b>MCAL401</b>	<b>Advanced Web Technology and Data Mining and Business Intelligence Lab (AWT and DMBI Lab)</b>	<b>03</b>

Subject Code	Subject Name	Teaching Scheme			Credits Assigned			
		Theory	Pract.	Tut	Theory	Pract.	Tut	Total
<b>MCAL401</b>	<b>Advanced Web Technology and Data Mining and Business Intelligence Lab</b>	--	<b>06</b>	--	--	<b>03</b>	--	<b>03</b>

Subject Code	Subject Name	Examination Scheme							
		Theory Marks				TW	Pract.	Oral	Total
Internal Assessment			End Semester Exam						
Test1 (T1)	Test2(T2)	Average of T1 & T2							
<b>MCA L401</b>	<b>Advanced Web Technology and Data Mining and Business Intelligence Lab</b>	--	--	--	--	<b>25</b>	<b>50</b>	<b>25</b>	<b>100</b>

7	<b>Data Warehousing</b>	<b>Data Warehousing using Oracle</b> <ul style="list-style-type: none"><li>• Setting Up and Starting Warehouse Builder</li><li>• Introducing OWB Architecture and Configuration</li><li>• Defining Source Metadata</li><li>• Ensuring Data Quality Using Data Profiling</li><li>• Defining Staging Metadata and Mapping Tables</li><li>• Deriving Data Rules and Running Correction Mappings</li></ul>	<b>06</b>
		<ul style="list-style-type: none"><li>• Defining a Relational Dimensional Model</li><li>• Handling Slowly Changing Dimensions</li></ul> <b>OLAP with Oracle</b> <ul style="list-style-type: none"><li>• Analytical Queries</li><li>• Grouping Functions</li><li>• Windowing Functions</li><li>• RollUp and Cube</li></ul>	
8	<b>Data Mining</b>	<b>Data Mining Using Weka/R Miner</b> <ul style="list-style-type: none"><li>• Introducing Weka/R Miner</li><li>• The Data Mining Process</li><li>• Using Classification Models</li><li>• Using Regression Models</li><li>• Using Clustering Models</li><li>• Performing Market Basket Analysis</li><li>• Performing Anomaly Detection</li><li>• Deploying Data Mining Results</li></ul>	<b>08</b>
9	<b>BI Tools</b>	<b>Open Source BI Tools</b> <ul style="list-style-type: none"><li>• Preparing Reports</li><li>• Preparing Dashboards</li><li>• Preparing Balanced ScoreCards</li><li>• Analysis of Reports</li></ul>	<b>08</b>

# Chapter 1

---

6Hrs



Educlash  
Just Another Way To Learn

# DMBI : Chapter 1

- Introduction and overview of BI-Effective and timely decisions, Data Information and knowledge, BI Architecture, Ethics and BI.

Business Intelligence : Data Mining and Optimization  
for Decision Making by **Carlo Vercellis** chapter 1

- BI Applications- Balanced score card, Fraud detection, Telecommunication Industry, Banking and finance, Market segmentation



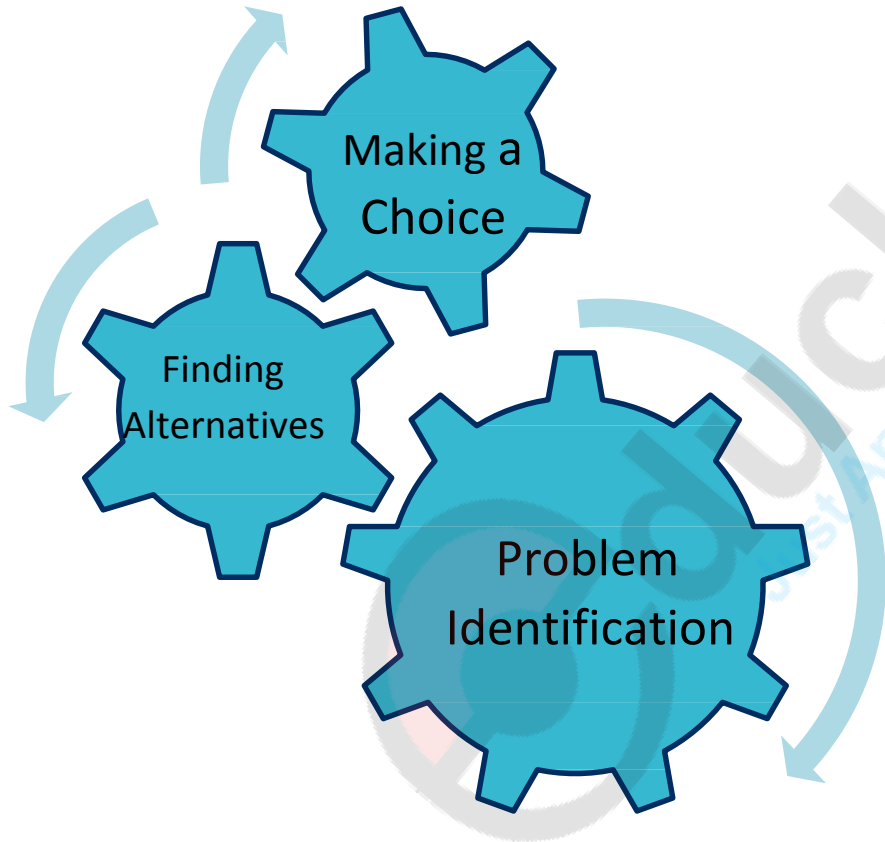
**Business intelligence** may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes.

Business intelligence methodologies are interdisciplinary and broad, spanning several domains of application.

- marketing
- logistics
- accounting and control
- finance
- services
- public administration.

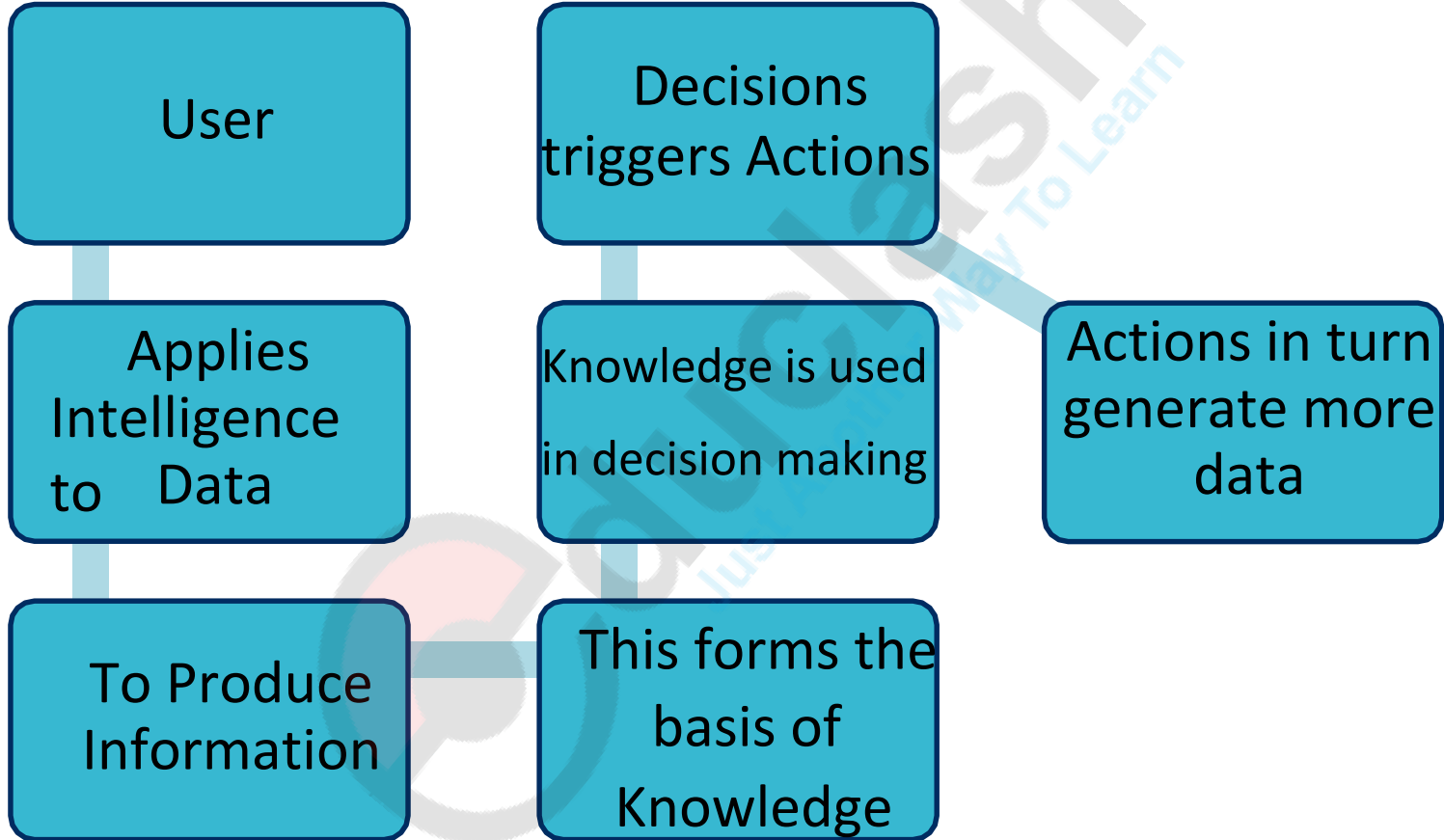


# Decision Making



*“Information and knowledge form the backbone of the decision making process”*

# Decision Action



# Who Makes Decisions?

Decision making at different levels:

– **Operational**

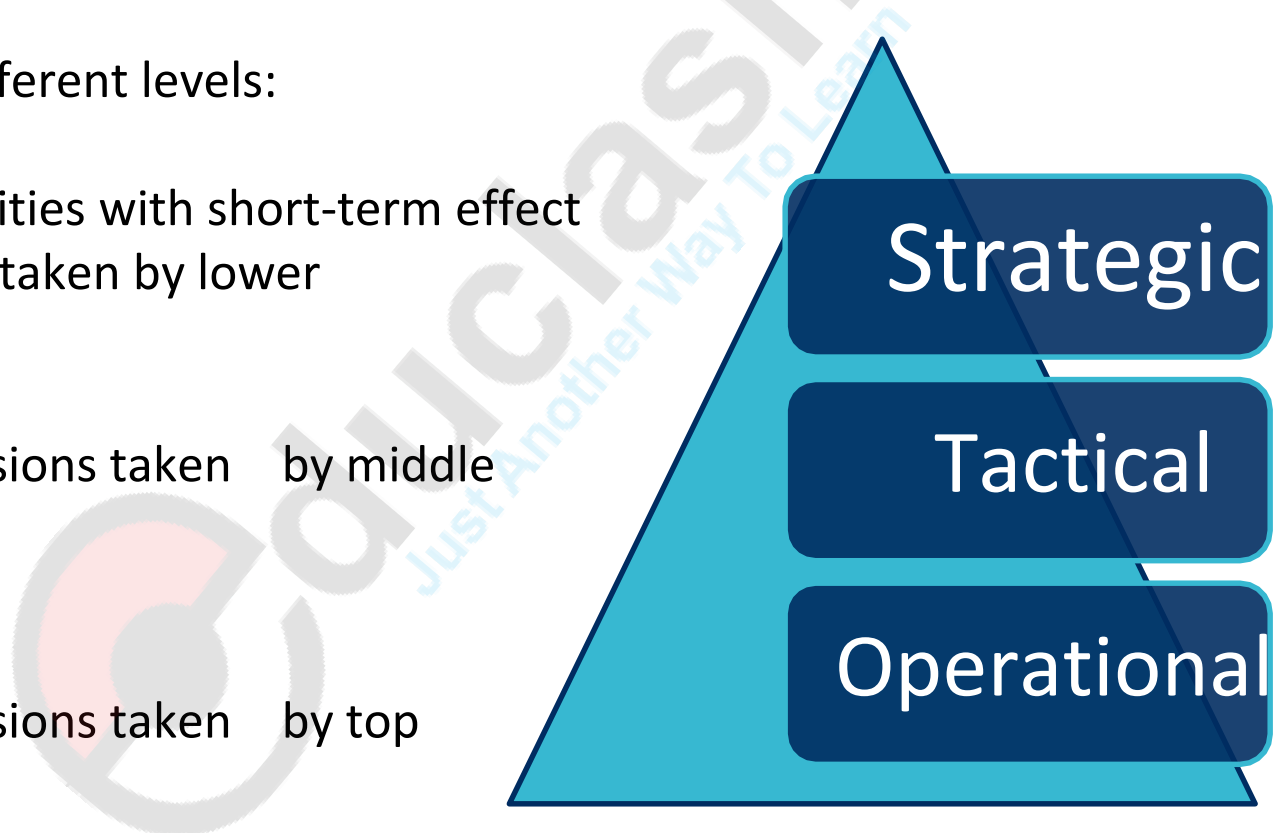
- Related to daily activities with short-term effect
- Structured decisions taken by lower management

– **Tactical**

- Semi-structured decisions taken by middle management

– **Strategic**

- Long-term effect
- Unstructured decisions taken by top management





# What is Business Intelligence

## Technology that Allows:

- Gathering, storing, accessing & analyzing data to help business users make better decisions

## Set of Applications that Allow:

- Decision support systems
- Query and reporting
- online analytical processing (OLAP)
- Statistical analysis, forecasting, and data mining

## Help in analyzing business performance through data-driven insight:

- Understand the past & predict the future

# Business Intelligence

*Information Driven by Online Transaction*

*Processing*

Product Database

- Add a Product Line
- Change a Product Price

Advertising Database

- Change Advertising Timetable
- Increase Radio Advertising Budget

Consumer Demographic Database

- Increase Customer Credit Limit
- Change Salary Level



Data Warehouse

*Business Intelligence Driven by*

*Analytical Processing*

How Many Products Sold Due to TV Ads Last Month

If Inventory Levels Drop by 10%; will Customers Shop Elsewhere?

Which Customer Demographic is Performing Best for Product 'A'



# Data Warehouse

A Data Warehouse Can be Defined as:

A decision support database that is maintained separately from the organization's operational database

A consistent database source that bring together information from multiple sources for decision support queries

Support information processing by providing a solid platform of consolidated, historical data for analysis

*"Data warehousing is the process of constructing and using data warehouses"*

# OLTP – Online Transaction

- Online transaction processing, or OLTP, is a class of information systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing
- Online transaction processing increasingly requires support for transactions that span a network and may include more than one company. For this reason, new online transaction processing software uses client or server processing and brokering software that allows transactions to run on different computer platforms in a network.

## Advantages

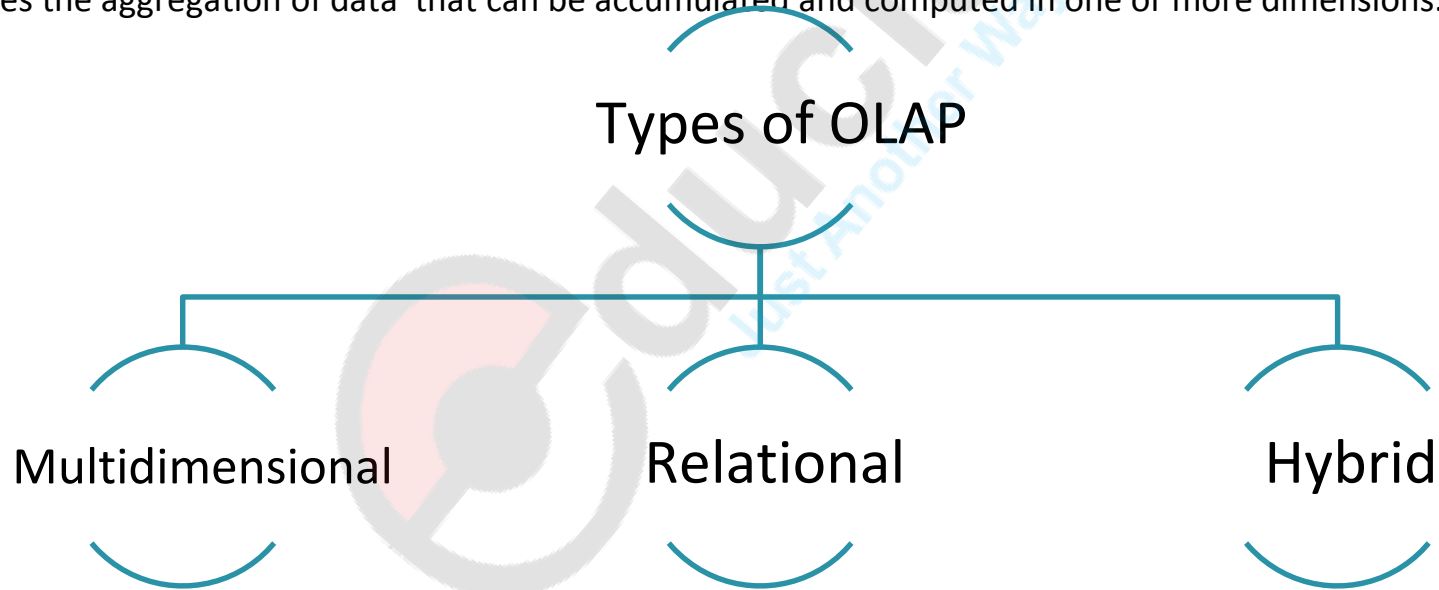
- Reduces paper trails
- Simple & effective
- Highly accurate

## Disadvantages

- Needs technical resources
- Needs maintenance
- Costly

# OLAP – Online Analytical

- Online analytical processing, or OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly
- OLAP tools enable users to analyse multidimensional data interactively from multiple perspectives. OLAP consists of three basic analytical operations: consolidation (roll-up), drill-down, and slicing and dicing. Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions.

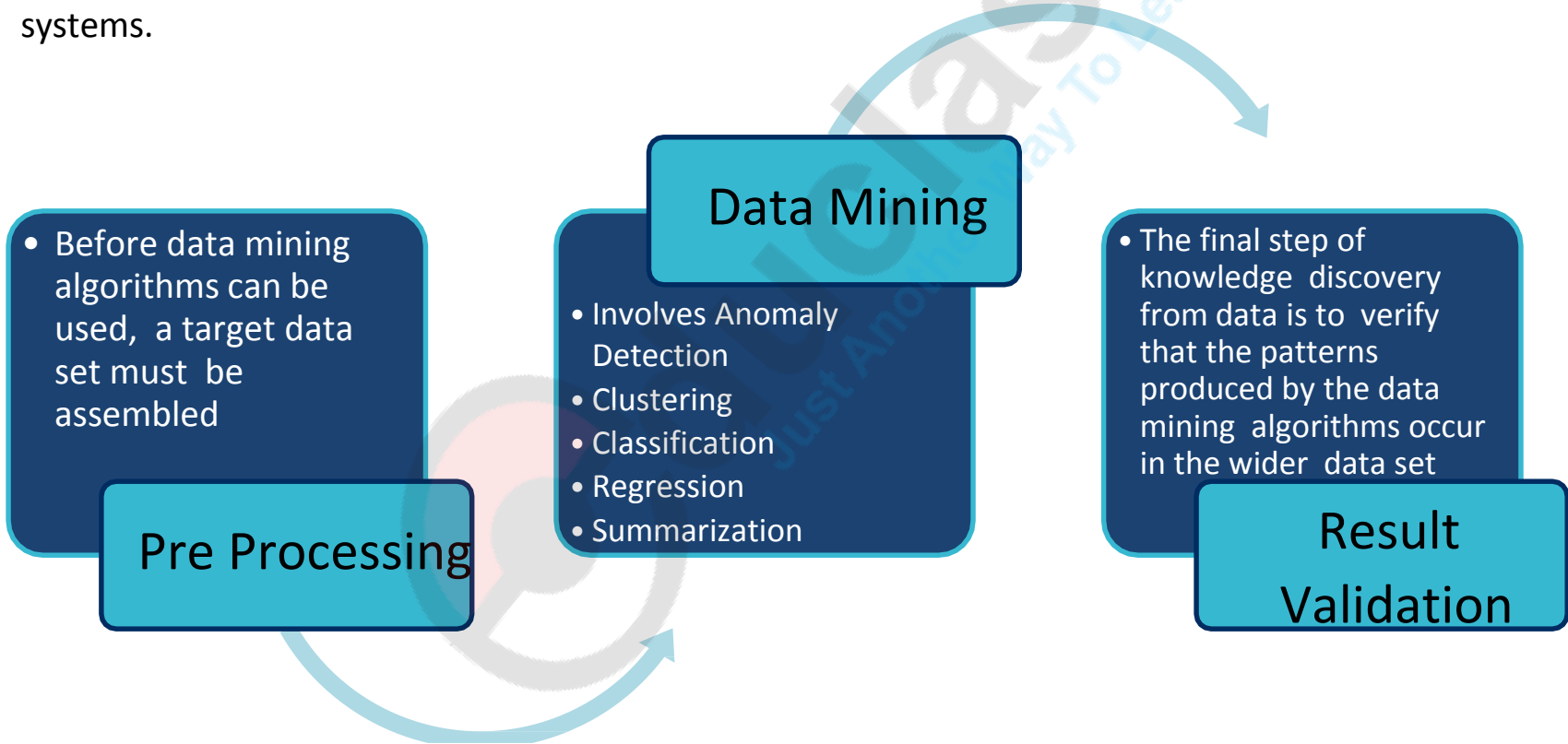


# OLAP Vs. OLTP

	OLTP	OLAP
User	Clerk/IT professional	Knowledge Worker
Function	Day to Day Operations	Decision Support
Database Design	Application Oriented	Subject Oriented
Data	Current, Isolated	Historical, Consolidated
View	Detailed, Flat Relational	Summarized, Multidimensional
Usage	Structured, Repetitive	Ad Hoc
Access	Read/Write	Read

# Data Mining & it's Process

- Data Mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.



# Layers of Business Intelligence



Kash  
Just Another Way To Learn

# Business Intelligence

Presentation Layer

*Reporting*



*Analysis*

Data Warehouse Layer



*Data Warehouse*

Source Layer



*Finance Dept.*



*HR Dept.*



*CRM*



*OS*



*Excel*

BI-Effective and timely decisions  
Data Information and knowledge  
BI Architecture  
Ethics  
BI.

# Chapter 1 Vercellies Book



# Data, information and knowledge



Lash  
Just Another Way To Learn

# Data, information and knowledge

- data originate partly from **internal** transactions of an administrative, logistical and commercial nature and partly from **external** sources
- data cannot be used directly for decision-making purposes
- need to be **processed** by means of appropriate extraction tools and analytical methods so that can be used by decision makers.
- **For example**, for a retailer data refer to primary entities such as customers, points of sale and items, while **sales receipts** represent the commercial transactions.

- **Information :**
  - Information is the **outcome** of extraction and processing activities carried out on data
  - For example, to the sales manager of a retail company, the proportion of sales receipts in the amount of over Rs.1000 per week, represent meaningful pieces of information that can be extracted from raw stored data.
- **Knowledge:**
  - Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions
  - For a retail company, a **sales analysis may detect that a group of customers**, living in an area where a competitor has recently opened a new point of sale, have **reduced their usual amount of business.**
  - **Passive** knowledge extraction: through the analysis criteria suggested by the decision makers
  - **Active** Knowledge extraction: by mathematical models
  - The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as ***knowledge management***

- **knowledge management** methodologies primarily focus on the treatment of information that is usually unstructured, at times implicit, contained mostly in documents, conversations and past experience.
- **Business intelligence** systems are based on structured information, most often of a quantitative nature and usually organized in a database.

However, this distinction is a somewhat fuzzy one:

- for example, the ability to analyze emails and web pages through text mining methods progressively induces business intelligence systems to deal with unstructured information.

Business intelligence analysis can be summarized schematically in the following **main characteristics**.

- First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
- Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.
- Finally, what-if analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters.

# BI Architecture



Uprock  
Just Another Way To Learn  
Uprock  
Uprock

# BI Architecture :three major components.

## Data sources.

**In a first stage, it is necessary to gather and integrate the data** stored in the various primary and secondary sources, which are heterogeneous in origin and type.

The sources consist for the most part of data belonging to operational systems, but may also include unstructured documents, such as emails and data received from external providers.

a major effort is required to unify and integrate the different data sources.

## Data warehouses and data marts.

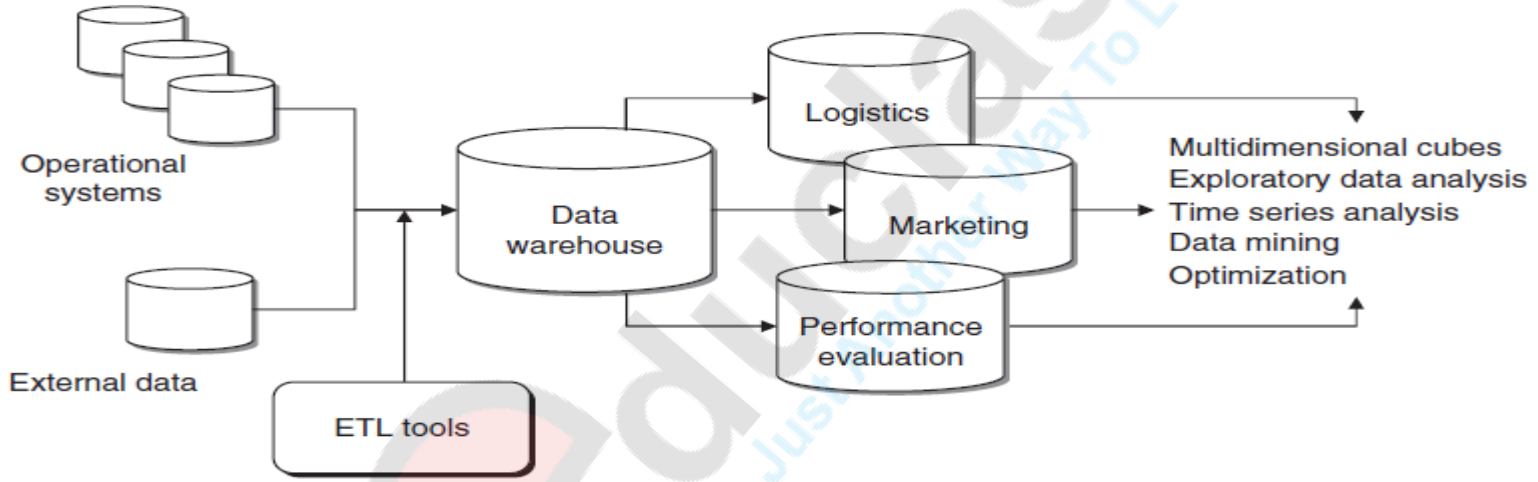
**Using extraction and transformation tools** known as *extract, transform, load (ETL)*, the data originating from the different sources are stored in databases intended to support business intelligence analyses.

These databases are usually referred to as *data warehouses and data marts*.

## Business intelligence methodologies.

**Data are finally extracted and used to feed** mathematical models and analysis methodologies intended to support decision makers.

# BI Architecture

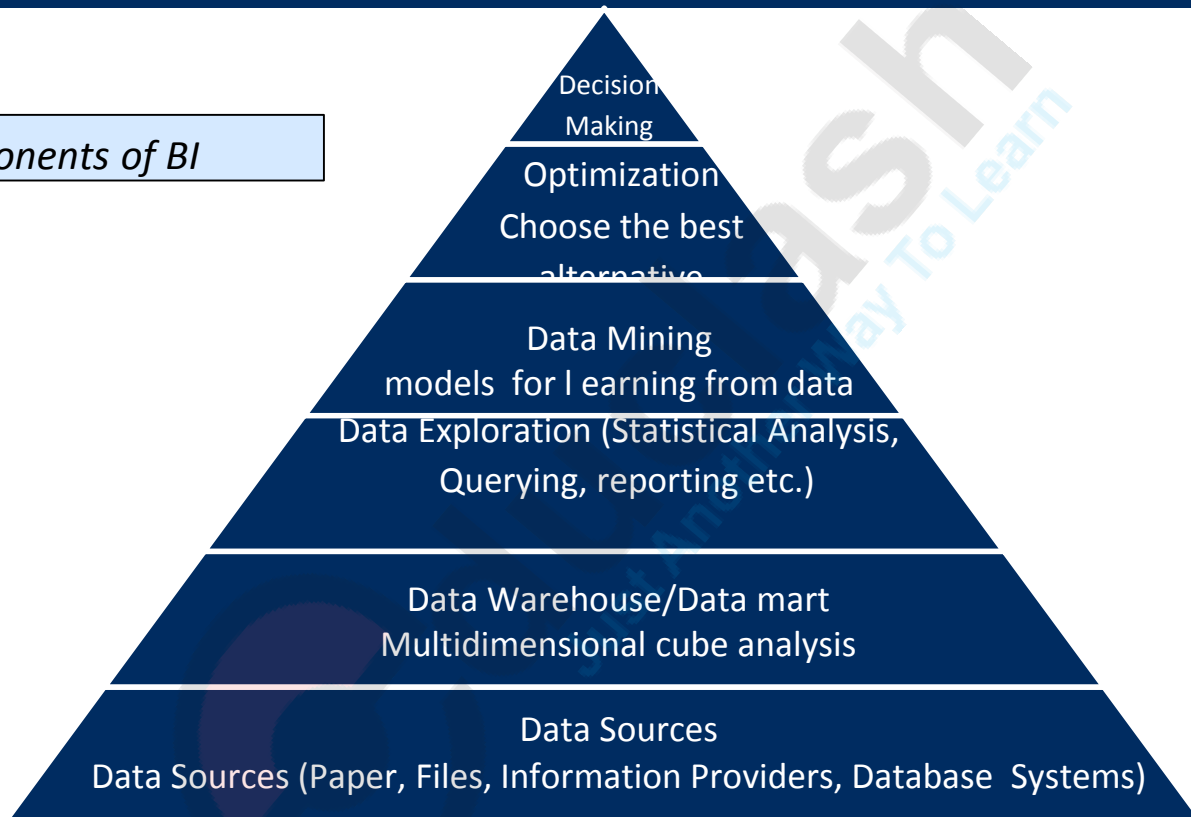






# Building blocks of a Business Intelligence System Process

*Main components of BI*



**Data exploration.** At the third level of the pyramid we find the tools for performing a *passive business intelligence analysis*, which consist of query and reporting systems, as well as statistical methods.

- These are referred to as passive methodologies because decision makers are requested to generate prior hypotheses or define data extraction criteria, and then use the analysis tools to find answers and confirm their original insight.
- For instance, consider the sales manager of a company who notices that revenues in a given geographic area have dropped for a specific group of customers.
- Hence, she might want to bear out her hypothesis by using extraction and visualization tools, and then apply a statistical test to verify that her conclusions are adequately supported by data.

**Data mining. The fourth level includes *active business intelligence methodologies*, whose purpose is the extraction of information and knowledge from data.**

- These include mathematical models for pattern recognition, machine learning and data mining techniques.
- Unlike the tools described at the previous level of the pyramid, the models of an active kind do not require decision makers to formulate any prior hypothesis to be later verified.
- Their purpose is instead to expand the decision makers' knowledge

**Data mining. The fourth level includes *active business intelligence methodologies*, whose purpose is the extraction of information and knowledge from data.**

- These include mathematical models for pattern recognition, machine learning and data mining techniques.
- Unlike the tools described at the previous level of the pyramid, the models of an active kind do not require decision makers to formulate any prior hypothesis to be later verified.
- Their purpose is instead to expand the decision makers' knowledge

# Ethics and BI

- the progress toward the information and knowledge society opens up countless opportunities, but may also generate distortions and risks which should be prevented and avoided by using **adequate control rules** and mechanisms.
- Usage of data by public and private organizations that is improper and does not respect the individuals' right to **privacy** should not be tolerated.
- More generally, we **must guard against the excessive growth of the political and economic power of enterprises** allowing the transformation processes outlined above to exclusively and unilaterally benefit such enterprises themselves, at the expense of consumers, workers and inhabitants.
- it is essential that business intelligence analysts and decision makers abide by the ethical principle of respect for the **personal rights of the individuals**
- There is a diversity of opinion on whether a company should pursue the **short-term maximization of profits** acting exclusively in the interest of shareholders, or should instead adopt an approach that takes into account the **social consequences of its decisions**

# BI Applications



Uprock  
Just Another Way To Learn

## BI Applications

- Balanced score card
- Fraud detection
- Telecommunication Industry
- Banking and finance
- Market segmentation



Educlash  
Just Another Way To Learn

# Why Data Mining?—Potential Applications

- **Market analysis and management**

Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation

- **Risk analysis and management**

Forecasting, customer retention, improved underwriting, quality control, competitive analysis

- **Fraud detection and detection of unusual patterns (outliers)**

## Other Applications

Text mining (news group, email, documents) and Web mining

Stream data mining

Bioinformatics and bio-data analysis



# Ex. 1: Market Analysis and Management

Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies

Target marketing

- Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.

- Determine customer purchasing patterns over time

Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association

Customer profiling—What types of customers buy what products (clustering or classification)

Customer requirement analysis

- Identify the best products for different groups of customers

- Predict what factors will attract new customers

Provision of summary information

- Multidimensional summary reports

- Statistical summary information (data central tendency and variation)

# Ex. 2: Corporate Analysis & Risk Management

## Finance planning and asset evaluation

- cash flow analysis and prediction

- contingent claim analysis to evaluate assets

- cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)

## Resource planning

- summarize and compare the resources and spending

## Competition

- monitor competitors and market directions

- group customers into classes and a class-based pricing procedure

- set pricing strategy in a highly competitive market

# Ex. 3: Fraud Detection & Mining Unusual Patterns

Approaches: Clustering & model construction for frauds, outlier analysis

Applications: Health care, retail, credit card service, telecomm.

Auto insurance: ring of collisions

Money laundering: suspicious monetary transactions

Medical insurance

Professional patients, ring of doctors, and ring of references

Unnecessary or correlated screening tests

Telecommunications: phone-call fraud

Phone call model: destination of the call, duration, time of day or week.

Analyze patterns that deviate from an expected norm

Retail industry

Analysts estimate that 38% of retail shrink is due to dishonest employees