1. **What is OLTP / Online Transaction Processing ?**

- OLTP or Online Transaction Processing is a type of data processing that consists of executing several transactions occurring concurrently—online banking, shopping, order entry, or sending text messages, for example.
- These transactions traditionally are referred to as economic or financial transactions, recorded and secured so that an enterprise can access the information anytime for accounting or reporting purposes.

2. **What is a Data Warehouse?**

"A Data Warehouse is a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management's decisions."

3. **Benefits of Data Warehouse?**
    - Enable workers to make better and wiser decisions
    - Identify hidden business opportunities
    - Blending with the customer
    - Precision Marketing

4. **What are dimensions and attributes?**

- Qualifying characteristics that provide additional perspectives to a given fact
- Dimensions are normally stored in dimension tables

Attributes

- Dimension Tables contain Attributes
- Attributes are used to search, filter, or classify facts
- Dimensions provide descriptive characteristics about the facts through their attributes
- Must define common business attributes that will be used to narrow a search, group information, or describe dimensions. (ex.: Time / Location / Product)
- No mathematical limit to the number of dimensions (3-D makes it easy to model)

5. **Star Schema**

- Every dimension table with its attributes must have an even chance of participating in a query to analyze the attributes in the fact table.
- Each of the dimension tables has a direct relationship with the fact table in the middle.
- Such an arrangement in the dimensional model looks like a star formation, with the fact table at the core of the star and the dimension tables along the spikes of the star.
- The dimensional model is therefore called a STAR schema.

6. Advantages of Star Schema.

- **Easy for Users to Understand:** The STAR schema is intuitively understood by the users. The users themselves will formulate queries through third-party query tools.
- **Optimizes Navigation:** Even when you are looking for a query result that is seemingly complex, the navigation is still simple and straightforward.
- **Most Suitable for Query Processing:** every query is simply executed first by selecting rows from the dimension tables using the filters based on the query parameters and then finding the corresponding fact table rows.
  - This is possible because of the simple and straightforward join paths and because of the very arrangement of the STAR schema.

7. What is Snowflake Schema

- "Snowflaking" is a method of normalizing the dimension tables in a STAR schema.
- When you completely normalize all the dimension tables, the resultant structure resembles a snowflake with the fact table in the middle.

8. What are Advantages and Disadvantages of Snowflake Schema

Advantages

- Small savings in storage space
- Normalized structures are easier to update and maintain

Disadvantages

- Schema less intuitive and end-users are put off by the complexity
- Ability to browse through the contents difficult
- Degraded query performance because of additional joins

9. What is OLAP (On-Line Analytical Processing)?

Definition: On-Line Analytical Processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through **fast, consistent, interactive access** in a wide **variety of possible views** of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

10. What are facts

- Numeric measurements (values) that represent a specific business aspect or activity
- Stored in a fact table at the center of the star scheme
- Contains facts that are linked through their dimensions
- Can be computed or derived at run time
- Updated periodically with data from operational databases

11. OLAP Characteristics.

1. let business users have a multidimensional and logical view of the data in the data warehouse,
2. facilitate interactive query and complex analysis for the users,
3. allow users to drill down for greater details or roll up for aggregations of metrics along a single business dimension or across multiple dimensions,
4. provide ability to perform intricate calculations and comparisons, and
5. present results in a number of meaningful ways, including charts and graphs.

12. OLAP vs OLTP

	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

13. What is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?

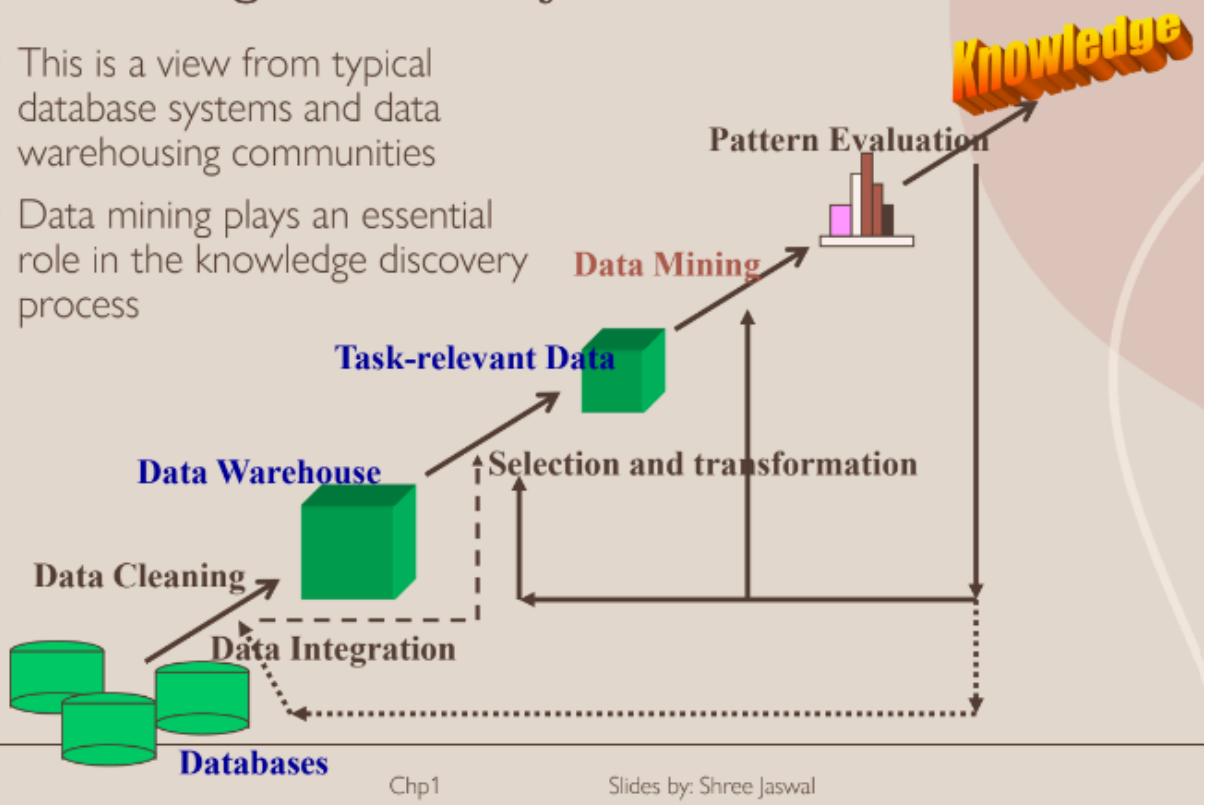
14. KDD Process

KDD Process: Several Key Steps

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



15. Data Mining Functions.

Data Mining Function: (1) Generalization contd.

- **Data discrimination** is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes
  - Eg. Compare 2 groups of customers- those who shop for computer products regularly(more than twice a month) and those who rarely shop for such products(less than 3 times a year)
- Data cube technology for computing used:
  - Drill down on any dimension
  - Discriminant rules: Discrimination descriptions expressed in the form of rules
- Output forms : same as that of data characterization along with discrimination descriptions

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your mart? Eg. Milk & bread
- Association, correlation vs. causality
  - A typical association rule
    - Computer → software [1%, 50%] (support, confidence)
  - Confidence means that if one buys a computer there is a 50% chance that she will buy software too. A 1% support means that 1% of all transactions under analysis show that computer & software are purchased together
- Association rules are discarded as uninteresting if they do not satisfy both a **minimum support threshold** and a **minimum confidence threshold**

Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages,

Data Mining Function: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Data objects are clustered or grouped based on the principle of *maximizing intraclass similarity and minimizing interclass similarity*
- Many methods and applications

Data Mining Function: (5) Outlier Analysis

- Outlier analysis (anomaly mining)
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? – One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

16. Applications of Data Mining.

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

CHAPTER 2

17. What are types of Attributes?

There are different types of attributes

- Nominal
  - Examples: ID numbers, eye color, zip codes
- Binary
  - Symmetric and Asymmetric types
- Ordinal
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- Numeric
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts, height, weight, latitude, longitude, monetary quantities etc.

Discrete Attribute

- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: binary attributes are a special case of discrete attributes

Continuous Attribute

- Has real numbers as attribute values
- Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

NOISE

Noise refers to modification of original values

- Examples: distortion of a person's voice when talking on a poor phone

18.

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

19.

20.

MAJOR TASKS IN DATA PREPROCESSING

Data cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

- Integration of multiple databases, data cubes, files, or notes

Data transformation

- Normalization (scaling to a specific range)
- Aggregation

Data reduction

- Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization: with particular importance, especially for numerical data
- Data aggregation, dimensionality reduction, data compression, generalization

DATA TRANSFORMATION: NORMALIZATION

Particularly useful for classification (NNs, distance measurements, nn classification, etc)

min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that $Max(|v'|)<1$

ORDINAL VARIABLES

An ordinal variable can be discrete or continuous

22.

Order is important, e.g., rank

Q: What is noise?

A: Random error in a measured variable.

Incorrect attribute values may be due to
- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

Other data problems which requires data cleaning
- duplicate records
- incomplete data
- inconsistent data

23.Types of Sampling

TYPES OF SAMPLING

Simple Random Sampling
- There is an equal probability of selecting any particular item
- Sampling without replacement (SRSWOR)
- As each item is selected, it is removed from the population

Sampling with replacement (SRSWR)
- Objects are not removed from the population as they are selected for the sample.
- In sampling with replacement, the same object can be picked up more than once

Cluster sampling: Eg. page

Stratified sampling
- Split the data into several partitions; then draw random samples from each partition

MODULE 3

1. What is Decision Tree?

- A decision tree is a flow chart like tree structure, where
  - Each *internal node* denotes a *test* on an attribute
  - Each *branch* denotes an *outcome* of the test
  - Each *leaf node* represent a *class*
- In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree.

2. What are errors in classification model what is model overfitting

Model Overfitting

- Errors committed by a classification model are of 2 types:
  - Training errors
  - Generalization error
- Training errors is the number of misclassification errors committed on the training dataset
- Generalization errors are the expected errors of the model on previously unseen records
- Both these errors must be low in a good model
- A model that fits the training data too well can have a poorer generalization error than a model with a higher training error. Such a situation is known as **model overfitting**

3. ID3 Algorithm

◦ Constructs a decision tree by using top-down recursive approach.
◦ Main aim is to choose that splitting attribute which is having the highest information gain.

MODULE 4

1. Applications of clustering

- market segmentation.
- social network analysis.
- search result grouping.
- medical imaging.
- image segmentation.
- anomaly detection.
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earthquake epicenters should be clustered along continent faults

2. Types of clustering

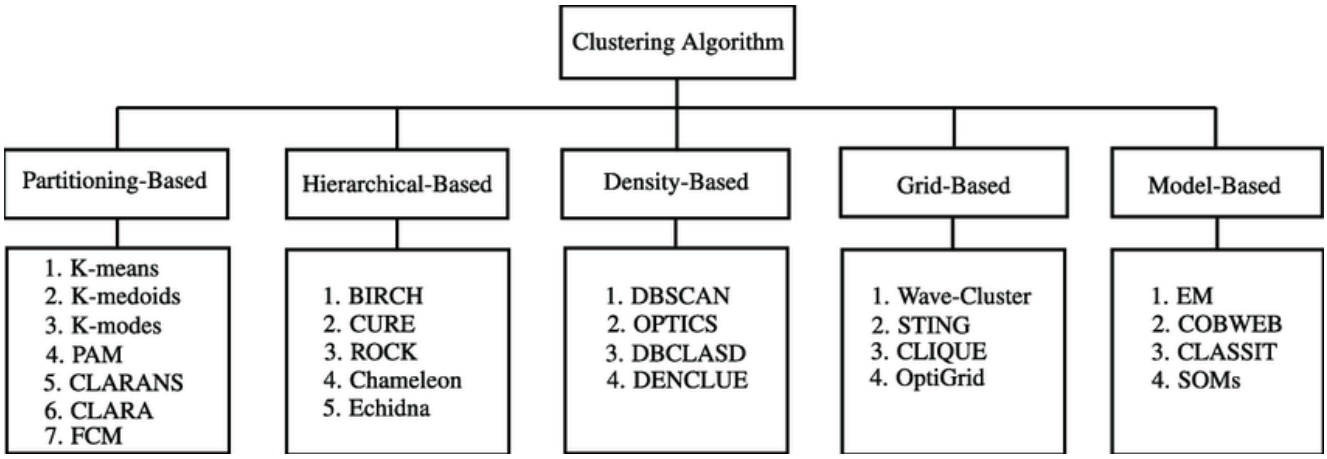
- Partitioning approach: Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors.

  Typical methods: k-means, k-medoids, CLARANS

- Hierarchical approach : Create a hierarchical decomposition of the set of data (or objects) using some criterion either agglomerative or divisive

  Typical methods: Diana, Agnes, BIRCH, ROCK, CAMELEON
- Density-based approach: Based on connectivity and density functions

  Typical methods: DBSCAN, OPTICS, DenClue



3. Agglomerative vs Divisive

Agglomerative(bottomup)	Divisive(topdown)
Start with 1 point (singleton)	Start with a big cluster
Recursively add two or more appropriate clusters	Recursively divide into smaller clusters
Stop when k number of clusters is achieved	Stop when k number of clusters is achieved

Table6.3: Comaprison of agglomerative and divisive clustering

Feature	Agglomerative Hierarchical Clustering	Divisive Hierarchical Clustering
Approach	Bottom-up approach	Top-down approach
Initialization	Each data point starts as its own cluster	All data points start in one cluster
Merging/Splitting Strategy	Merges pairs of clusters iteratively based on similarity	Recursively splits clusters based on dissimilarity
Initial State	Each data point is considered a single-element cluster	All data points belong to one cluster
Steps	Continues merging clusters until all data points belong to one cluster or stopping criterion is met	Continues dividing clusters until each data point is in its own cluster or stopping criterion is met
Complexity (More/Less)	More	Less
Example	Imagine clustering points on a map. At first, each point is its own cluster. Clusters are then merged based on proximity, forming larger clusters until all points are in one cluster.	Imagine a single cluster representing all the points on the map. This cluster is then recursively divided based on dissimilarity until each point is in its own cluster.

4. Advantages and Disadvantages of Hierarchical Clustering

Advantages:

Is simple and outputs a hierarchy, a structure that is more informative

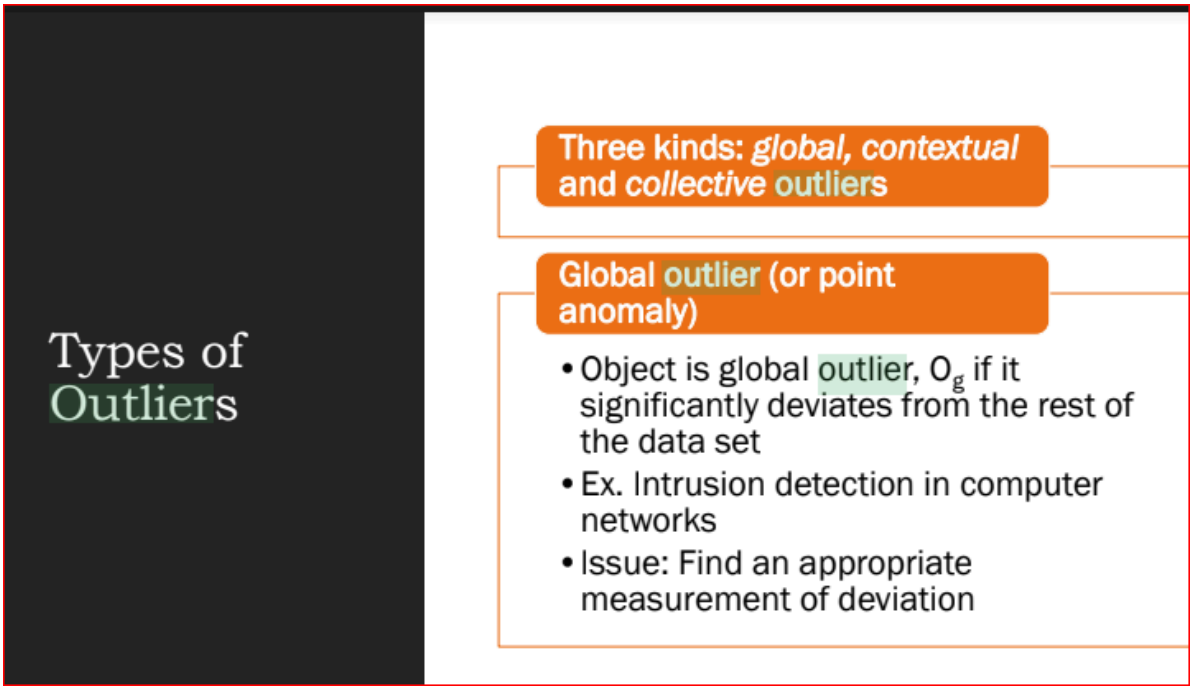
It does not require us to pre-specify the number of clusters

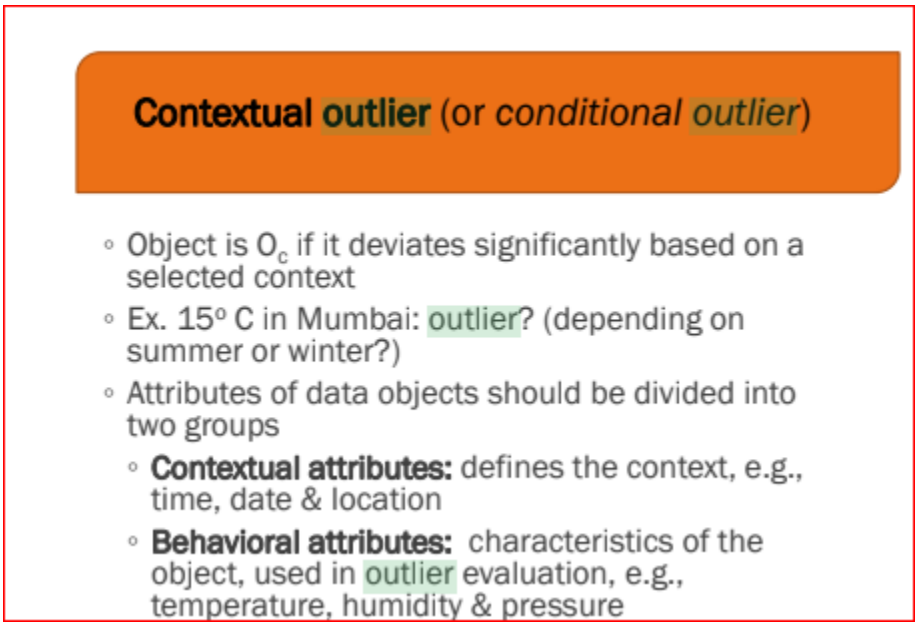
Disadvantages:

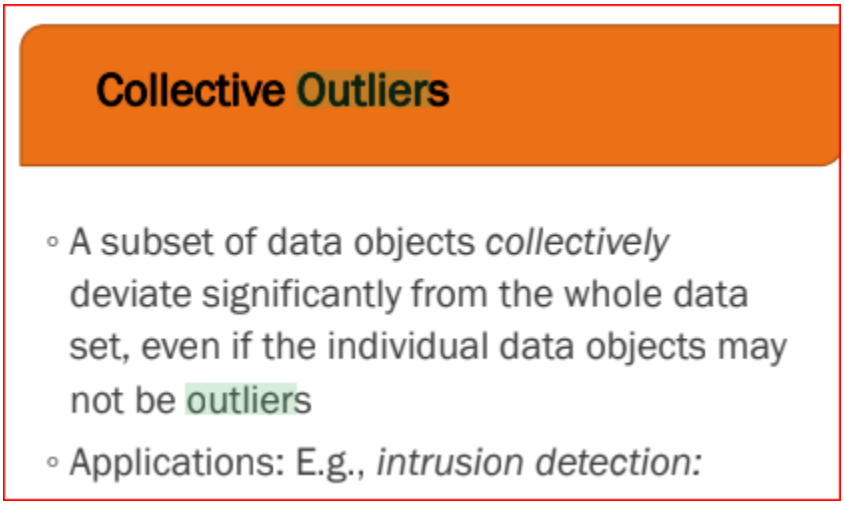
Selection of merge or split points is critical as once a group of objects is merged or split, it will operate on the newly generated clusters and will not undo what was done previously.

Thus merge or split decisions if not well chosen may lead to low-quality clusters

3. What is an outlier? Types of outliers. Describe methods used for outlier analysis







Two ways to categorize outlier detection methods:

◦ Based on whether user-labeled examples of outliers can be obtained:

◦ **Supervised, semi-supervised and unsupervised methods**

◦ Based on assumptions about normal data and outliers:

◦ **Statistical, proximity-based, and clustering-based methods**

MODULE 5

1. What Is Frequent Pattern Analysis?

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining
- Motivation: Finding inherent regularities in data
  - What products were often purchased together?— Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
- Applications
  - Market Basket analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

2. What are Association Rules mining?

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

3. Methods to Improve Apriori's Efficiency

1. Hash-based technique can be used to reduce the size of the candidate k-itemsets, Ck, for k>1.

2. Transaction reduction – a transaction that does not contain any frequent k itemsets cannot contain any frequent k+1 itemsets. Therefore, such a transaction can be marked or removed from further consideration because subsequent scans of the database for j-itemsets, where j>k, will not require it.
3. Partitioning (partitioning the data to find candidate itemsets): A partitioning technique can be used that requires just two database scans to mine the frequent itemsets. It consists of two phases.
4. Sampling (mining on a subset of a given data): The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent itemsets in S instead of D. In this way, we trade off some degree of accuracy against efficiency.

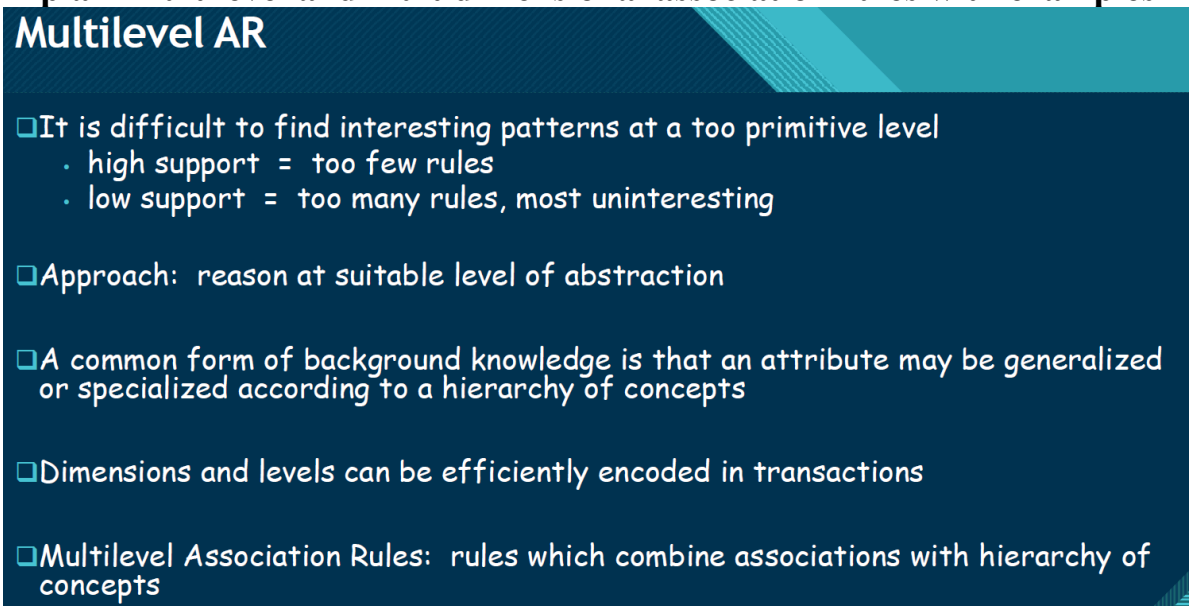
5. Dynamic itemset counting (adding candidate itemsets at different points during a scan): A dynamic itemset counting technique was proposed in which the database is partitioned into blocks marked by start points.

4. FP tree advantage

Advantages of this algorithm:

- Better than Apriori in the generation of candidate (k+1)-itemset from frequent k itemsets
- There is no need to scan the database to find the support (k+1) itemsets (for k>=1). This is because the TID_set of each k-itemset carries the complete information required for counting such support.
- The disadvantage of this algorithm consist in the TID_set being too long, taking substantial memory space as well as computation time for intersecting the long sets.

5. Explain multilevel and multidimensional association rules with examples



Multi-level Association: Redundancy Filtering

- Some rules may be redundant due to "ancestor" relationships between items.
- Example
  - milk ⇒ wheat bread [support = 8%, confidence = 70%]
  - 2% milk ⇒ wheat bread [support = 2%, confidence = 72%]
- We say the first rule is an ancestor of the second rule. So, second rule is redundant
- A rule is redundant if its support is close to the "expected" value, based on the rule's ancestor.

Mining Multi-Dimensional Association

- Single-dimensional rules:

  buys(X, "milk") ⇒ buys(X, "bread")
- Multi-dimensional rules: ≥ 2 dimensions or predicates
  - Inter-dimension assoc. rules (*no repeated predicates*)

    age(X,"19-25") ∧ occupation(X,"student") ⇒ buys(X, "coke")
  - hybrid-dimension assoc. rules (*repeated predicates*)

    age(X,"19-25") ∧ buys(X, "popcorn") ⇒ buys(X, "coke")
- Categorical(nominal) Attributes (e.g. Occupation, brand): finite number of possible values, no ordering among values—data cube approach
- Quantitative Attributes (e.g. age, income): numeric, implicit ordering among values—discretization, clustering, and gradient approaches

Association rules generated from mining data at multiple abstraction levels are called **multiple-level** or **multilevel association rules**. Multilevel association rules can be

...Multidimensional Space 285

mined efficiently using concept hierarchies under a support-confidence framework. In general, a top-down strategy is employed, where counts are accumulated for the calculation of frequent itemsets at each concept level, starting at concept level 1 and working downward in the hierarchy toward the more specific concept levels, until no more frequent itemsets can be found. For each level, any algorithm for discovering frequent itemsets may be used, such as Apriori or its variations.

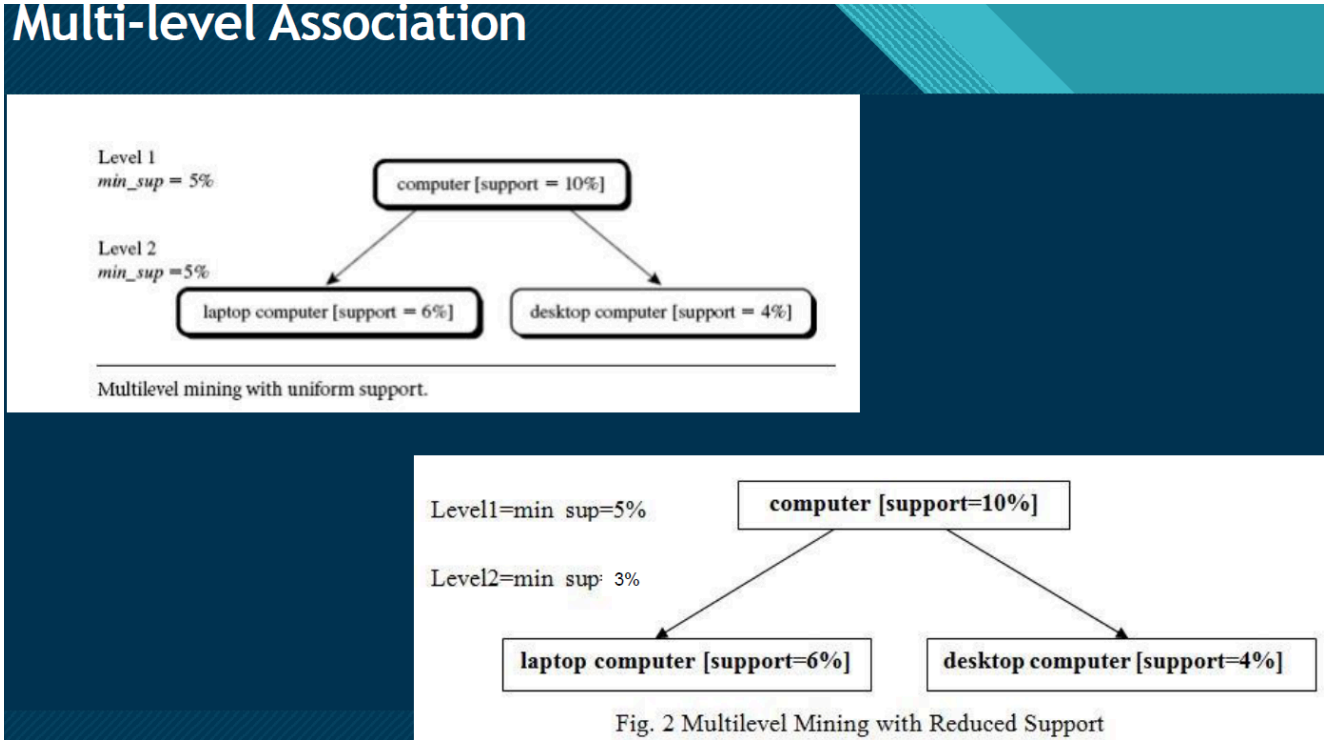
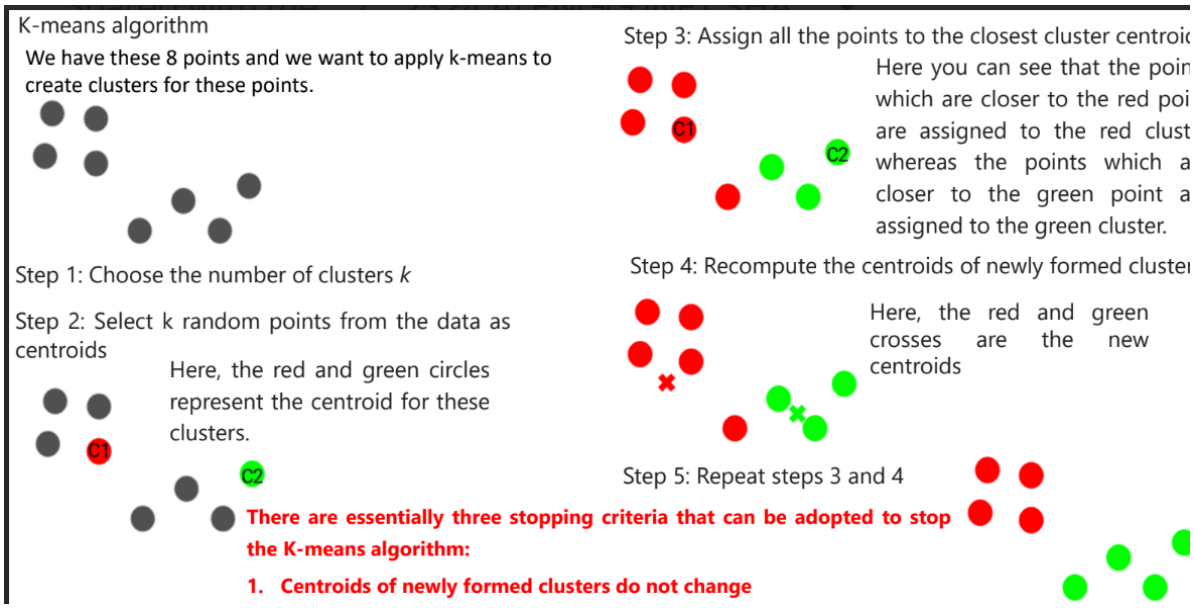


Fig. 2 Multilevel Mining with Reduced Support

6. Explain k means, K mediods algorithm



7.Explain market basket analysis with example

Market basket analysis is a data mining technique used to discover associations between items purchased together. It identifies patterns in consumer behavior by analyzing transactions to uncover relationships between products frequently bought together. This analysis is commonly applied in retail, e-commerce, and marketing to optimize product placement, promotions, and inventory management strategies.

Example:
Consider a grocery store dataset containing transaction records. Each transaction consists of a list of items purchased by a customer. Market basket analysis aims to find associations between these items.

Suppose the analysis reveals the following associations:

1. {Bread} => {Butter}
2. {Milk, Eggs} => {Bread}
3. {Beer} => {Chips}

These associations indicate that:

1. Customers who buy bread are likely to purchase butter as well.
2. Customers who buy milk and eggs together are likely to buy bread too.
3. Customers who buy beer are likely to buy chips.

8.Difference between classification and clustering

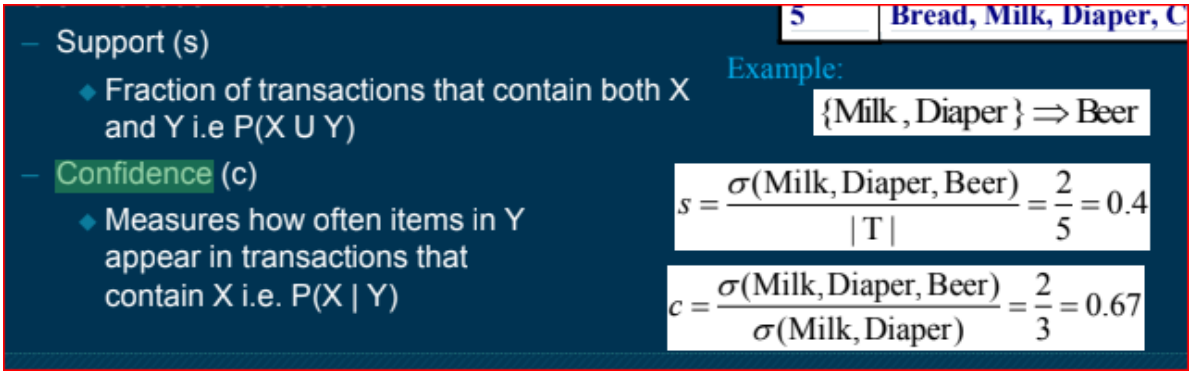
Classification	Clustering
Classification is a supervised learning approach where a specific label is provided to the machine to classify new observations. Here the machine needs proper testing and training for the label verification.	Clustering is an unsupervised learning approach where grouping is done on similarities basis.
Supervised learning approach.	Unsupervised learning approach.
It uses a training dataset.	It does not use a training dataset.
It uses algorithms to categorize the new data as per the observations of the training set.	It uses statistical concepts in which the data set is divided into subsets with the same features.
In classification, there are labels for training data.	In clustering, there are no labels for training data.
Its objective is to find which class a new object belongs to form the set of predefined classes.	Its objective is to group a set of objects to find whether there is any relationship between them.
It is more complex as compared to clustering.	It is less complex as compared to clustering.

Example Algorithms	Logistic regression, Naive Bayes classifier, Support vector machines, etc.	k-means clustering algorithm, Fuzzy c-means clustering algorithm, Gaussian (EM) clustering algorithm, etc.

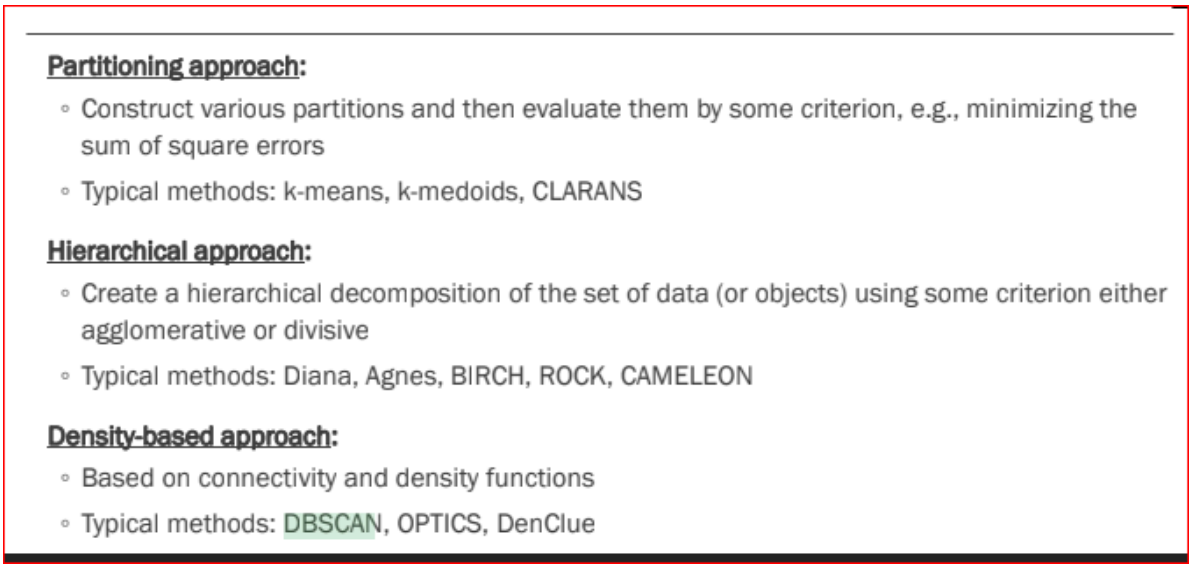
9. Define support and confidence

Support: - Support measures the frequency or occurrence of a particular itemset in a dataset. It indicates how frequently the items in an itemset appear together in the dataset.

Confidence:- Confidence measures the reliability or strength of the association between two itemsets in a rule. It indicates the likelihood that if itemset A appears in a transaction, then itemset B will also appear in the same transaction.



10.Types of clustering



11.Applications of Frequent pattern analysis

Market Basket analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

12. What is Market Basket analysis

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as

purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

13. What is cross marketing

At a fundamental level, cross marketing is a promotional strategy where two or more businesses collaborate to promote their products or services. Cross marketing works because it brings businesses together to help them reach new and broader audiences. It also allows a business to diversify and offer something new to their customers.

Implication means co-occurrence, not causality!

Itemset: A collection of one or more items

Example: {Milk, Bread, Diaper}

•**k-itemset:** An itemset that contains k items

•**Support count (σ):** Frequency of occurrence of an itemset

E.g. σ({Milk, Bread,Diaper}) = 2

•**Support:** Fraction of transactions that contain an itemset

E.g. s({Milk, Bread, Diaper}) = 2/5

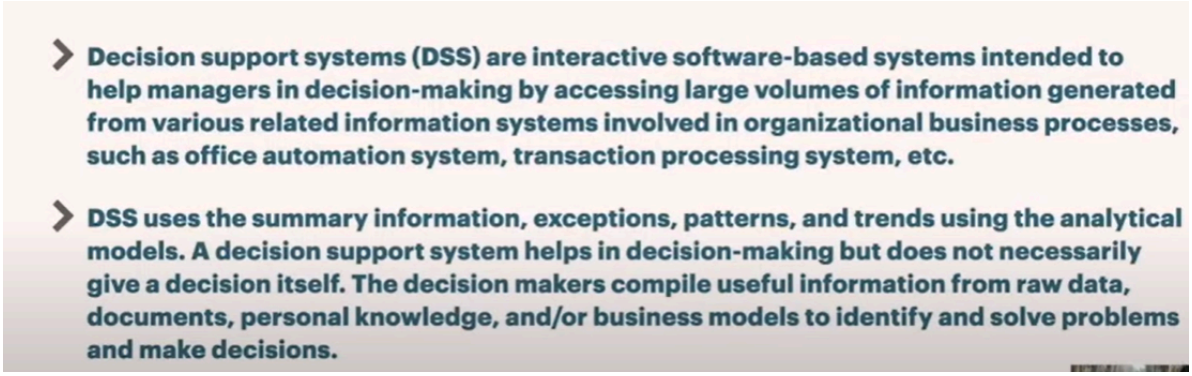
•**Frequent Itemset :** An itemset whose support is greater than or equal to a minsupport threshold

MODULE 6

1.Define Business Intelligence.

- *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.

2.Define Decision support system



3.Applications of BI

1. BI for the finance industry

When implementing BI software into financial institutions, businesses can automate loan applications which ensures compliance. The vast customer data is effectively stored, cleansed and shared without any unnecessary duplication. The information is no longer siloed, but much more easily accessible to the relevant role players. They can use this quality data for financial analytics and business decision-making.

2. BI for hospitality

The restaurant business deals in a lot of receipts, particularly those with several locations. By implementing BI software, restaurateurs can integrate all this data so that they're able to find out which items are more popular, the times or days when the restaurant is full, and who is performing well. This data can then be used to co-ordinate the menu, arrange staff schedules, as well as creating promotions and discounts to attract customers at quieter times.

3. BI for Online Gambling

Online gambling has been experiencing remarkable growth recently. Casino operators use BI tools to gain insights into which games are doing well and what they can do to reach more players. Online casinos operate worldwide and can use data and analytics to understand where your games are popular. Also, understanding players' interests helps online gambling companies design the best games.

4. BI for manufacturing

For manufacturers across multiple locations operating in diverse product offerings, BI is an incredibly useful tool. It can be used to analyse real-time data to forecast market size and demand so that the manufacturing process can be adjusted accordingly. The BI software ensures that all data from across the various operations is centralised which makes for much more operational efficiency.

5. BI for Retail

Traditionally, retail businesses relied on spreadsheets and static reports to make decisions. But with BI gaining prominence, businesses can generate dynamic, real time reports and make informed decisions regarding customization, dynamic pricing, floor plan, inventory management, competitor analysis, customer trends, etc., Retail firms can use Business Intelligence apps to collect and access data about demographics, sales, purchase history, etc. By processing this information, they can make reliable forecasts about customer demand and manage inventory accordingly. BI can help retail businesses track their marketing campaign's performance and make necessary changes. They can leverage the features of BI applications to launch and manage stores in multiple locations.

6. Sales Intelligence

A key application of BI focuses on where your business meets the customer. Business intelligence

collects data on specific KPIs like customer demographics, conversion rates, sales metrics, etc. Then it organizes this data into structured visualizations like graphs, pie charts and scattergrams. Users can identify trends from this data that provide insights into customer behavior and business operations. Knowing the customer means you can better serve them!

4. Draw BI architecture

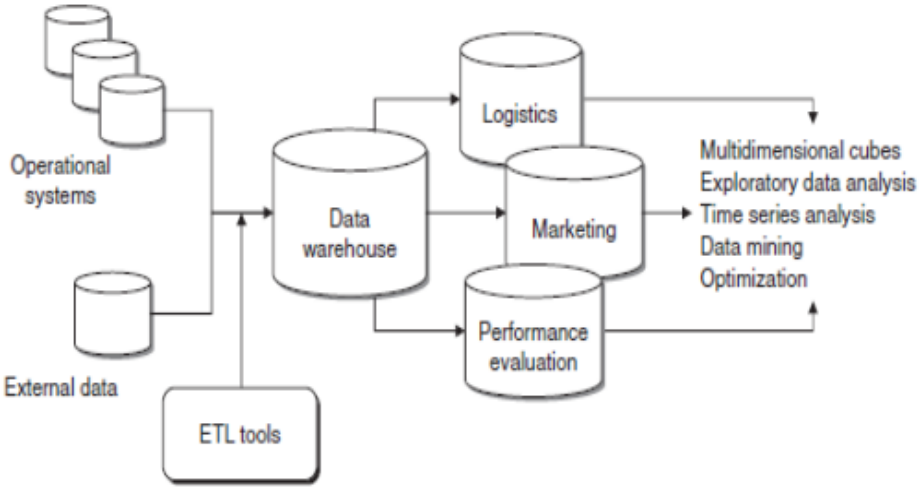


Figure 1.2 A typical business intelligence architecture

5. Data

Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities.

6. Information

Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain.

7. Knowledge

Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions.

9. Components of BI

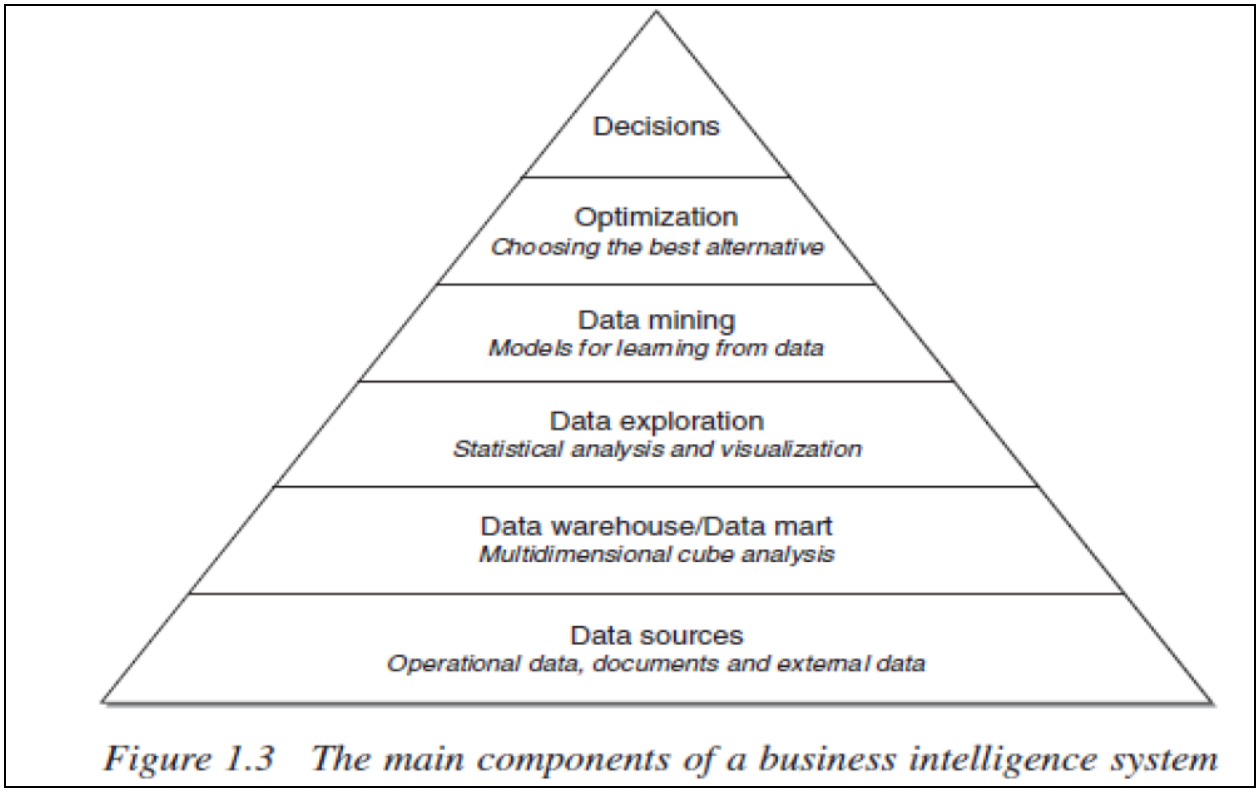


Figure 1.3 The main components of a business intelligence system

10. Cycle of BI

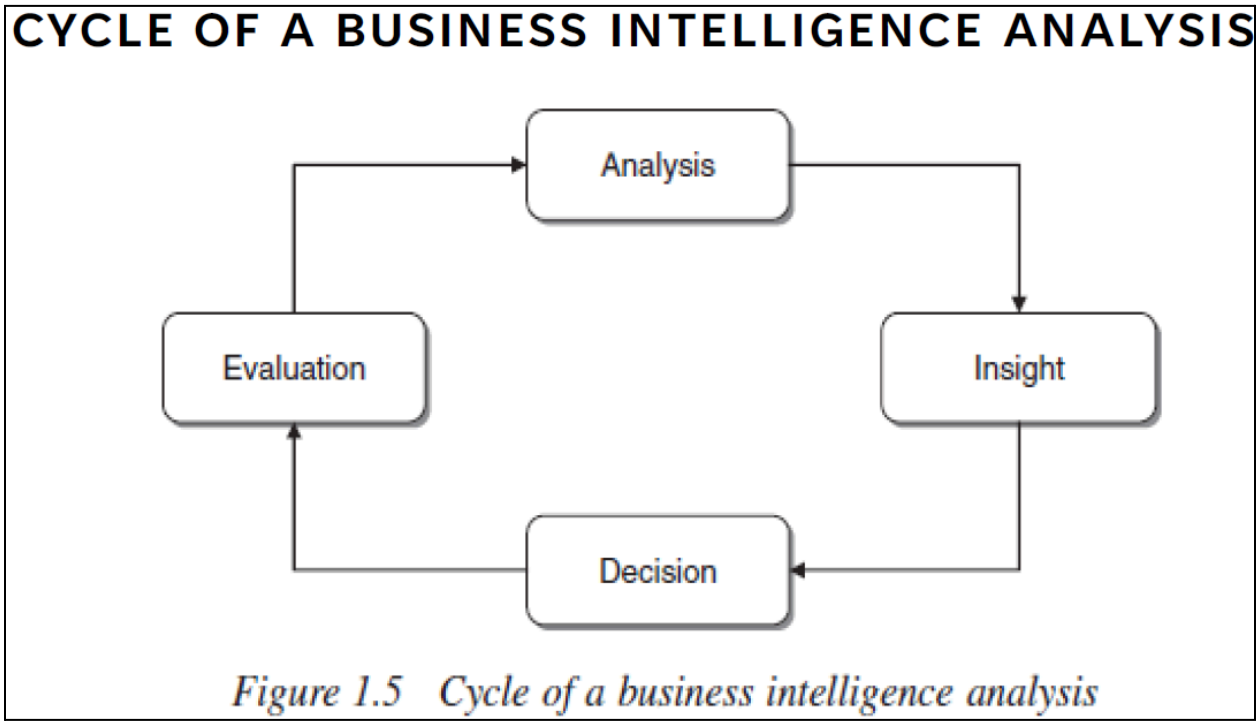


Figure 1.5 Cycle of a business intelligence analysis

11. Types of decisions(based on nature)

1. Structured decisions. A decision is structured if it is based on a well-defined and recurring decision-making procedure.
2. Unstructured decisions. A decision is said to be unstructured if the three phases of intelligence, design and choice are also unstructured.
3. Semi-structured decisions. A decision is semi-structured when some phases are structured and others are not. Most decisions faced by knowledge workers in managing public or private enterprises or organizations are semi-structured.

12. Types of decision (based on scope)

1. Strategic decisions: Decisions are strategic when they affect the entire organization or at least a substantial part of it for a long period of time.
2. Tactical decisions: Tactical decisions affect only parts of an enterprise and are usually restricted to a single department. The time span is limited to a medium-term horizon, typically up to a year.
3. Operational decisions. Operational decisions refer to specific activities carried out within an organization and have a modest impact on the future.

13. Features of DSS

1. Effectiveness
2. Mathematical models
3. Integration in the decision-making process
4. Organizational role
5. Flexibility

14.