DATABASE MANAGEMENT SYSTEM

UNIT-4 (Complete Notes)

By- Bazgha

INDEX

Page No.
Date

Database Security

What is Database Security

- Database security refers to the range of tools, controls and measures designed to establish and preserve database confidentiality, integrity and availability.
- It involves various types or categories of controls, such as technical, procedural / administrative and physical.
- Database security must address and protect the following:
  The data in the database
  The database management system (DBMS)
  Any associated applications
  The physical database server and / or the virtual database server and the underlying hardware
  The computing and / or network infrastructure used to access the database.

Importance of Database Security

- Database security can guard against a compromise of your database, which can lead to financial loss, reputation damage, consumer confidence disintegration, brand erosion, and non-compliance of government and industry regulations.
- Database security safeguards defend against a myriad of security threats and can help protect your enterprise from:

- Deployment failure
- Excessive privileges
- Privilege abuse
- Platform vulnerabilities
- Unmanaged sensitive data
- Backup data exposure
- Weak authentication
- Database injection attacks

What it Controls

Database Security encompasses multiple controls, including system hardening, access, DBMS configuration and security monitoring. These different security controls help to manage the circumventing of security protocols.

System hardening and monitoring
DBMS configuration
Authentication
Authorization
Access control
Database auditing
Backups
Encryption
Application Security

Authentication

Authentication is the process of confirmation that whether the user log in only according to the rights provided to him to perform the activities of database. A particular user can login only up to his privilege but he can't access the other sensitive data. The privilege of accessing sensitive data is restricted by using Authentication.

By using these authentication tools for biometrics such as retina and figure prints can prevent the data base from unauthorized/malicious users.

Authorization

Authorization is a privilege provided by the Database Administrator. Users of the database can only view the contents they are authorized to view. The rest of the database is out of bounds to them.

The different permissions for authorization available are:

- Primary Permission – This is granted to users publicly and directly.
- Secondary Permission – This is granted to groups and automatically awarded to a user if he is a member of the group.
- Public Permission – This is publicly granted to all the users.
- Context sensitive permission – This is related to sensitive content and only granted to a select users.

Categories of Authorization

The categories of authorization that can be given to users are:

- System Administrator – This is the highest administrative authorization for a user. Users with this

authorization can also execute some database administrator commands such as restore or upgrade a database.

- System Control – This is the highest control authorization for a user. This allows maintenance operations on the database but not direct access to data.
- System Maintenance – This is the lower level of system control authority. It is also allows users to maintain the database but within a database manager instance.
- System Monitor – Using the authority, the user can monitor the database and take snapshots of it.

Difference between Authorization & Authentication

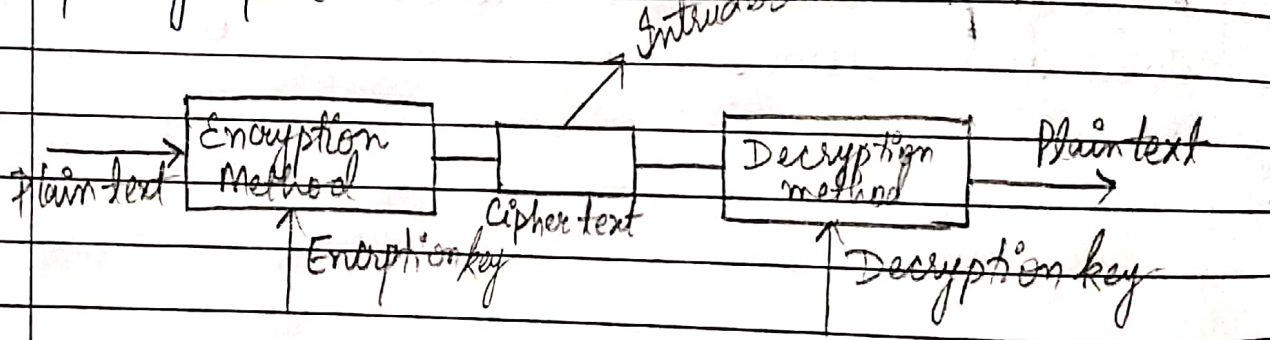
Authentication	Authorization
1. In authentication process, the identity of users are checked for providing the access to the system.	1. While in this process, users or persons are validated.
2. In authentication process, users or persons are verified.	2. While this process, ~~users or persons are validated~~ is done after the authentication process.
3. It is done before the authentication process.	3. While it needs user's privilege or security levels.
4. It needs usually user's login details. Authentication determines whether the person is user or not.	4. While it determines what permission do user have?

Page No.
Date

Encryption

A DBMS can use encryption to protect information in certain situations where the normal security mechanisms of the DBMS are not adequate.

For example, an intruder may steal tapes containing some data or tap a communication line. By storing and transmitting data in an encrypted form, the DBMS ensures that such stolen data is not intelligible to the intruder. Thus, encryption is a technique to provide privacy of data.



In encryption, the message to be encrypted is known as plaintext. The output of the encryption process is known as ciphertext. The process of converting the plaintext to ciphertext is called Encryption. The process of converting ciphertext to plaintext is called Decryption.

Page No.
Date

Access Control

Database access control is a method of allowing access to company's sensitive data only to those people (database users) who are allowed to access to unauthorized persons. It includes two main components i.e. components ie., Authentication and authorization.

Authentication is a method of verifying the identity of a person who is accessing your database. Note that authentication is not enough to protect data. An additional layer of security is required, Authorization, which determines whether a user should be allowed to access the data or make the transaction he or she is attempting without Authentication and Authorization there is no data security.

Any company whose employees connect to the internet, thus every company today, needs some level of access control implemented.

Authentication Factor
→ Password or Pin
→ Biometric measurement (fingerprint & retina scan)
→ Card or key

Types of Access Control

1) Physical access control
Physical access control restricts entry to campuses, buildings, rooms and physical IT assets.

2) Logical access control
Logical access control limits connections to computer networks, system files and data.

The primary concerns of an access control system are the following:

1) Prevent access: In the absence of any privilege, ensure that the subject cannot access the object.
   subject → a process executing on behalf of user.
   object → a piece of data or a resource.

2) Determine access: decide whether the subject has access, according to some policy, to take an action with an object.

3) Grant access: give a subject access to an object.

4) Revoke access: remove a subject's access to an object.

5) Audit access: Determines which subject can access an object or which object a subject can access.

Need for access control

Access control regulates which users, applications and devices can view, edit, add and delete resources in an organization's environment.

Controlling access is one of the key practices to protect sensitive data from theft, misuse, abuse, and any other threats.

There are two levels of access control:

1) Physical: Limits access to offices, rooms and physical IT assets.
2) Logical: Limits connections to computer networks, digital infrastructure, system files and data.

There are several logical access control models: Mandatory, discretionary, role-based, attribute-based, etc. The process of choosing and deploying an access control model ~~like~~ looks different for each organization. This choice depends on:

i) The nature of the protected data
ii) IT requirements and Industry standards.
iii) The number of employees.
iv) The cyber security budget.

There are types of access control:

DAC → Discretionary access control
MAC → Mandatory access control
RBAC → Role Based access control
ABAC → Attribute Based access control

Discretionary Access Control (DAC)

Discretionary access control is a type of security access control that grants or restricts object (data) access via an access policy determines by an object's owner group.

In simple words the owner of the object specifies which subject can access the object. This model is called discretionary because the control of access is based on the discretion of the owner.

DAC mechanism controls are defined by user identification with supplied credentials during authentication, such as username and password. In DAC, the owner determine object access privileges.

Most operating systems such as windows, linux and most of unix are based on DAC models.

For example: In unix rwxr-xr-x file.txt, meaning that the owner of file.txt may read, write or execute it, and that other users may read or execute the file but not write it.

DAC attributes include:

- User may transfer object ownership to another user.
- After several attempts, authorization failures restrict user access.
- Unauthorized users are blind to object characteristics, such as file size, file name.
- User may determine the access type of the users.

DAC is easy to implement but has certain disadvantages:

- Inherent vulnerabilities (Trojan horse)

Trojan horse is a program downloaded and installed on a computer that appears harmless, but is in fact malicious. When the user clicks on the email attachment or downloads the free program, the malware that is hidden inside, transferred to the user's computing device.

- Grant and Revoke permission maintainence

When there are many users in a database it becomes difficult to grant to revoke privilege to users.

- Updating the security policy is costly.
- Less secure and harder to control information leakage.

Advantages

- Flexible
- Easy to implement
- Use in environments where the sharing of information is more important than protection.
- It is enabling fine-grained control over system objects.

Date

MAC Model

It is a model of access control where the operating system provides users with access based on data confidentially and user clearance levels. In this model access is granted on a need to know basis. users have to prove a need for information before gaining access.

MAC is considered to be the most secure of all access control models as Access rules are manually defined by system administrators and are strictly enforced by the operating system or security kernel.

With MAC, the process of gaining access looks like this:

- The administrator configures access policies and defines security attributes: confidentiality levels, clearances for accessing different projects and types of resources.
- The administrator assigns each subject (user or resource that accesses data) and object (file, database, part, etc) a set of attributes.
- When a subject attempts to access an object, the operating system examines the subject security attributes and decides whether access can be granted.

For example, lets consider data that has the "top secret" confidentiality level and "engineering project" security label. It's available to a set of users that have "top secret" clearance and authorization to access engineering documents. Such users can also access information that requires a lower level of clearance. But employees with lower levels of clearance will not be able to access to information that requires a higher level of clearance.

MAC brings lots of benefits to Cyber Security system. But it has several disadvantages to consider.

Advantages of MAC

- High level of data protection: An administrator defines access to objects and users can't edit access.
- Granular: An administrator sets users access rights and object access parameters manually.
- Immune to trojan Horse attacks: User can't declassify data or share access to classified data.

Disadvantages of MAC

- Maintainability: Manual configuration of security levels and clearances requires constant attention from administrators.
- Scalability: MAC does not scale automatically.

- Not user friendly: Users have to request access to each new piece of data; they can't configure access parameters for their own data.

So, MAC model requires more efforts to work with, than any of other access control models.

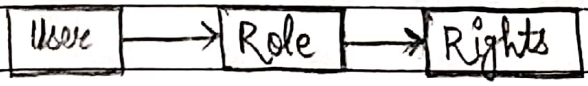
Uses of MAC

- MAC is used by government organizations, militaries and law enforcement institutions. It is reasonable to use MAC in organizations that value data security more than operational flexibility and costs.
- Implementing MAC in a private organization is rare because of the complexity and inflexibility of such a system.
- A pure MAC model provides a high and granular level of security which is difficult to set up and maintain so it is now become common to combine MAC with other access control models.
- For example: Combining it with the role based model speeds up the configuration of user profiles. Instead of defining access rights for each user, an administrator can create user roles. Each organization has users with similar roles and access rights.

Role-Based Access Control [RBAC] Model

Role based access control is an approach to handling security and permissions in which roles and permissions are assigned within an organizations.

It is a way to provide security because it only allows employees to access information they need to do their jobs, while preventing them from accessing additional information that is not relevant to them. An employees role determines the permissions he or she is granted and ensures that lower level employees are ~~employees~~ not able to access sensitive information or perform high-level tasks.

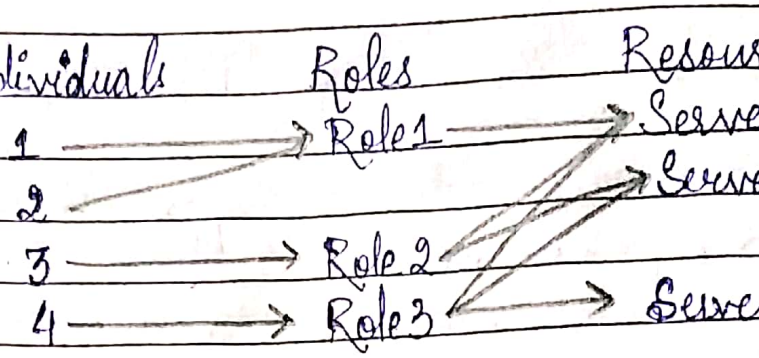


Rules for RBAC to perform

1) A person must be assigned a certain role in order to conduct a certain action called transaction.
2) A user needs a role authorization to be allowed to hold that role.
3) Transaction authorization allows the user to perform certain transactions. User's won't be able to perform transactions other than the one's they are authorized for.

Role-based access control allows you to assign one or more roles per user.

Example:



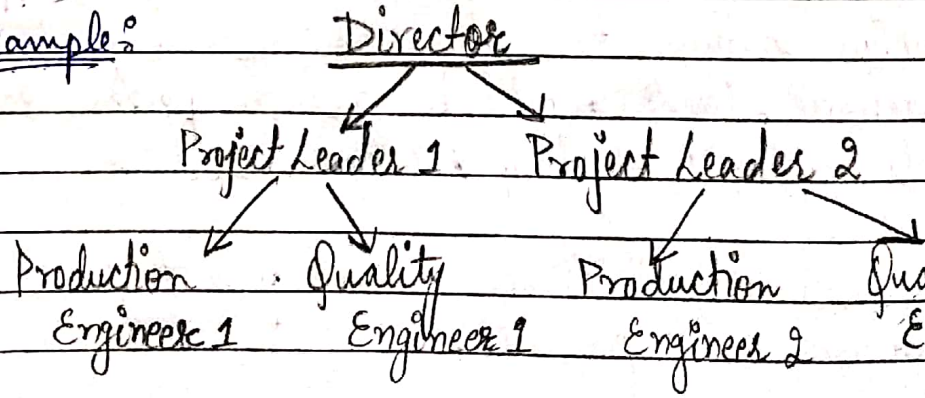
As shown in the above figure, Individual 1 and 2 are given a role 1 and that particular role just have access to server 1 whereas Individual 3 is assigned with role 2 and can access both servers 1 & 2. Individual 4 is given role 3 which can be seen as the senior most role as it can access all 3 servers.

Hierarchical RBAC

A hierarchy is mathematically a partial order of defining a seniority relation between roles, whereby the senior roles acquire the permission of their juniors.

So, a hierarchical RBAC defines inheritance relations between the roles.

Example:



Page No.
Date

Project leader will inherit all the permissions associated with the engineers and in this case Director is considered to be senior to Role B in the role hierarchy. If project leaders are inheriting all the permissions associated with the engineers, then the director will also inherit those permissions by virtue of being senior to the project leaders.

Advantages of RBAC

→ Flexibility: The modifications to the roles can be made as per the requirement.

→ Less errors prone: Assigning permissions individually is a more complex process and is thus more error prone than using role-based access control for assigning.

→ Security: Access permissions are defined exclusively via the role model which prevents you from giving more permissions than needed to individual employees.

→ Transparency: The naming of roles is usually straightforward and thus increases transparency for users.

→ Reducing Costs: By not allowing user access to certain process and applications, companies may conserve or most cost-effectively use resources, such as network bandwidth, memory and storage.

Page No.
Date

Disadvantages of RBAC

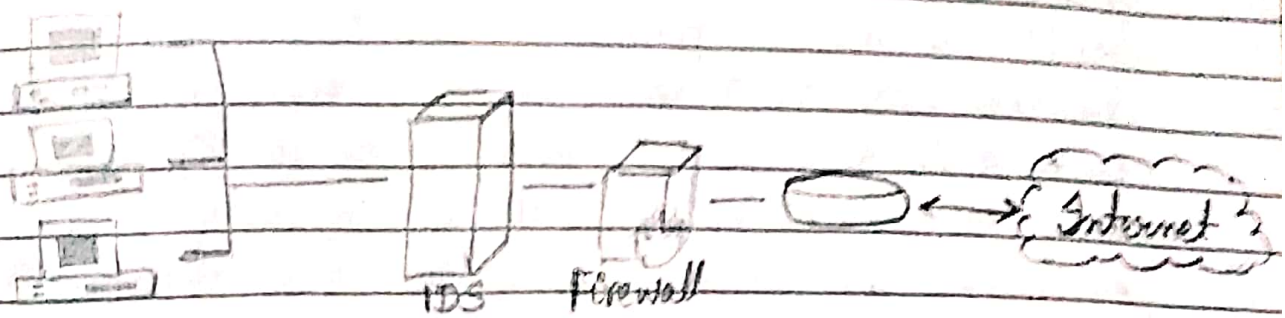
→ Labour-intensive setup : Translating organizational structures into the RBAC model requires a lot of work.

→ Temporary assignments : If a user only needs extended access permissions temporarily. It is easier to forget about them when using RBAC than when assigning permissions individually.

→ Application : In small companies, creating and maintaining roles would be more labour intensive than assigning permissions individually. Therefore, RBAC model is only used when a certain number of roles and employees has been reached. However, even in large companies, RBAC suffers from the drawback that it is easy to end up creating a large number of roles.

Page No.
Date

Intrusion Detection



Intrusion Detection System or IDS is software, hardware or combination of both used to detect intruder activity.

OR

An intrusion detection system (IDS) is a device or software application that monitors a network or system for malicious activity and policy violations. Any malicious traffic or violation is typically reported to an administrator or collected centrally using a security information and event management (SIEM) system.

Classification of Intrusion Detection System ~~(NIDS):~~

IDS are classified into 2 types

1. Network Intrusion Detection System (NIDS):
NIDS are set up at a planned point within the network to examine traffic from all devices on the network. It performs an observation of passing traffic on the entire subnet and matches the traffic that is passed on the subnets to the collection of known attacks. Once an attack is identified or abnormal behaviour is observed, the alert can be sent to the administrator. An example of an NIDS is installing it on the subnet where firewalls are located in order to see if someone is trying crack the firewall.

2. Host Intrusion Detection System (HIDS):
HIDS run on independent hosts or devices on the network. A HIDS monitors the incoming and outgoing packets from the device only and will alert the administrator if suspicious or malicious activity is detected. It takes a snapshot of existing system files and compares it with the previous snapshot. If the analytical system files were edited or deleted, an alert is sent to the administrator to investigate. An example of HIDS usage can be seen on mission critical machines, which are not expected to change their layout.

SNORT :→ It is an open source network intrusion detection system (NIDS) which is available free of cost. The network admin can use it to watch all the incoming packets and find the ones which are dangerous to the system.
→ Available for LINUX and WINDOWS.
→ www.snort.org

Detection Method of IDS:

1. Signature-based Method: Signature-based IDS detects the attacks on the basis of the specific patterns such as number of bytes or number of 1's or number of 0's in the network traffic. It also detects on the basis of the already known malicious instruction sequence that is used by the malware. The detected patterns in the IDS are known as signatures.

It can easily detect the attacks whose pattern (signature) already exists in system but it is quite difficult to detect the new malware attacks as their pattern (signature) is not known.

2. Anomaly-based Method: Anomaly-based IDS was introduced to detect the unknown malware attacks as new malware are developed rapidly. In anomaly-based IDS there is use of machine learning to create a trustful activity model and anything coming is compared with that model and if it is declared suspicious if it is not found in model. Machine learning based method has a better generalized property in comparison to signature-based IDS as these models can be trained according to the applications and hardware configurations.

Intrusion Prevention System (IPS)

Intrusion Prevention System is also known as Intrusion Detection and prevention system. It is a network security application that monitors network or system activities for malicious activity. Major functions of intrusion prevention systems are to identify malicious activity, collect information about this activity, report it and attempt to block or stop it.

Page No.
Date

Classification of Intrusion Prevention System (IPS)

Intrusion Prevention System (IPS) is classified into 4 types:

1. Network-based intrusion prevention system (NIPS): It monitors the entire network for suspicious traffic by analyzing protocol activity.
2. Wireless intrusion prevention system (WIPS): It monitors a wireless network for suspicious traffic by analyzing wireless networking protocols.
3. Network behaviour analysis (NBA): It examines network traffic to identify threats that generate unusual traffic flows, such as distributed denial of service attacks, specific forms of malware and policy violations.
4. Host-based intrusion prevention system (HIPS): It is an inbuilt software package which operates a single host for doubtful activity by scanning events that occur within that host.

Detection Method of Intrusion Prevention System (IPS):

1. Signature-based detection: Signature-based IDS operates packets in the network and compares with pre-built and preordained attack patterns known as signatures.
2. Statistical anomaly-based detection: It monitors network traffic and compares it against an established baseline. The baseline will identify what is normal for that network and what protocols are used. However, it may raise a false alarm if the baselines are not intelligently configured.

3. Stateful protocol analysis detection: This IDS method recognizes divergence of protocols stated by comparing observed events with pre-built profiles of generally accepted definition of not harmful activity.

Capabilities of intrusion detection systems

- Monitoring the operation of routers, firewalls key management servers and files that are needed by other security controls aimed at detecting, preventing or recovering from cyberattacks;
- Providing administrators a way to tune, organize and understand relevant OS audit trails and other logs that are otherwise difficult to track or parse;
- Providing a user-friendly interface so non expert staff members can assist with managing system security;
- Including an extensive attack signature database against which information from the system can be matched;
- Recognizing and reporting when the IDS detects that data files have been altered;
- Generating an alarm and notifying that security has been breached; and
- Reacting to intruders by blocking them or blocking the server.

Page No.
Date

Difference between IDS and IPS

The main difference is an IDS is a monitoring system and an IPS is a control system. Both IDS/IPS read network packets and compare their contents to a database of known threats or baseline activity. However, IDS don't alter network packets while IPS can prevent packets from delivering based on their contents, much like a firewall does with an IP address:

- Intrusion detection systems (IDS): Analyze and monitor traffic for indicators of compromise that may indicate an intrusion or data theft. IDS compares current network activity against known threats, security policy violations and open port scanning. IDS require humans or another system to look at the results and to determine how to respond, making them better as postmortem digital forensics tools. Also, IDS is not inline, so traffic doesn't have to flow through it.
- Intrusion prevention systems (IPS): IPS have detection capabilities too, but will proactively deny network traffic if they believe it represents a known security threats.

Parameter	IDS	IPS
System or Tool	It is a detection and monitoring system or tool.	It is a control system or tool.
Self Decision or Action	It does not take action on their own.	It take the action (accept or reject packet based on rule set).
Work Automatic	It requires human to look the results and action accordingly.	In this virus related database should be update on regular basis.

SQL Injection

SQL injection is a technique used to exploit user data through web page inputs by injecting SQL commands as statements. Basically, these statements can be used to manipulate the application's web servers by malicious users.

- SQL injection is a code injection technique that might destroy your database.
- SQL injection is one of the most common web hacking techniques.
- SQL injection is the placement of malicious code in SQL statements, via web page input.

Examples of SQL injection ~~examples~~

There are a wide variety of SQL injection vulnerabilities, attacks and techniques which arise in different situations. Some common SQL injection examples include:

- Retrieving hidden data, where you can modify an SQL query to return additional results.
- Subverting application logic, where you can change a query to interfere with the application's logic.
- Union attacks, where you can retrieve data from different database tables.
- Examining the database, where you can extract information about the version and structure of the database.
- Blind SQL injection, where the results of a query you control are not returned in the application's responses.

Page No.
Date

SQL Injection Based on 1=1 is Always True

Look at the example, the original purpose of the code was to create an SQL statement to select a user, with a given user id.

If there is nothing to prevent a user from entering "wrong" input, the user can enter some "smart" input like this:

UserId: 105 OR 1

Then, the SQL statement will look like this:

SELECT * FROM Users WHERE UserId = 105 OR 1=1;

The SQL above is valid and will return ALL rows from the "Users" table, since OR 1=1 is always TRUE.

SELECT UserId, Name, Password FROM Users WHERE UserId = 105 or 1=1;

A hacker might get access to all the user names and passwords in a database, by simply inserting 105 OR 1=1 into the input field.

SQL Injection Based on "" = "" is Always True

Here is an example of a user login on a website:

Username: John Doe

Password: myPass

Example:

```
uName = getRequestString("username");
uPass = getRequestString("userpassword");
sql = 'SELECT * FROM Users WHERE Name = '' + uName
    + '' AND Pass = '' + uPass + '' '';
```

Page No.
Date

Result:

SELECT * FROM Users WHERE Name = "John Doe" AND Pass = "my Pass"

A hacker might get access to user names and passwords in a database by simply inserting "OR" "=" into the username or password text box:

User name:

Password:

The code at the server will create a valid SQL statement like this:

Result:

SELECT * FROM Users WHERE Name = "" OR "" = ""
AND Pass = "" or "" = ""

The SQL above is valid and will return all rows from the "Users" table, since OR "" = "" is always TRUE.

Page No.
Date

Types of SQL injection

Blind SQL injection is used when a web application is vulnerable to an SQL injection but the results of the injection are not visible to the attacker. The page with the vulnerability may not be one that displays data but will display differently depending on the results of a logical statement injected into the legitimate SQL statement.

Second-order SQL injection

→ First order SQL injection arises where the application takes user input from an HTTP request and, in the course of processing that request, incorporates the input into an SQL query in an unsafe way.

→ In second-order SQL injection (also known as stored SQL injection), the application takes user input from an HTTP request and stores it for future use. This is usually done by placing the input into a database, but no vulnerability arises at the point where the data is stored. Later, when handling a different HTTP request, the application retrieves the stored data and incorporates it into an SQL query in an unsafe way.

Page No.
Date

How to prevent against SQL Injection attacks

- User input should never be trusted : It must always be sanitized before it is used in dynamic SQL statements.
- Stored procedure : These can encapsulate the SQL statements and treat all input as parameters.
- Bypassed statements : Prepared statements to work by creating the SQL statement first then treating all submitted user data as parameters. This has no effect on the syntax of the SQL statement.
- Regular expressions : These can be used to detect potential harmful code and remove it before executing the SQL statements.
- Database Connection user access rights : Only necessary access rights should be given to accounts used to connect to the database. This can help reduce what the SQL statements can perform on the servers.
- Error messages : These should not reveal sensitive information and when exactly an error occurred. Simple custom error messages such as "Sorry, we are experiencing technical issues. The technical team has been contacted. Please try again later" can be used instead of display the SQL statements that caused the error.

Page No.
Date

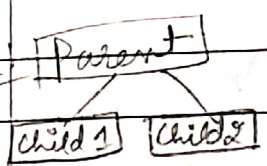
OBJECT ORIENTED DATA MODEL

Object oriented database systems are alternative to relational database and other database systems. In OODB, information is represented in the form of objects.
OODB are exactly same as OOP languages. If we can combine the features of relational model (transaction, concurrency, recovery) to OODB, the resultant model is called as OODB model.

Features

1. Complexity : OODBMS has the ability to represent the complex structure (of object) with multilevel complexity.

2. Inheritance : Creating a new object from an existing object in such a way that new object inherits all characteristics of an existing object.



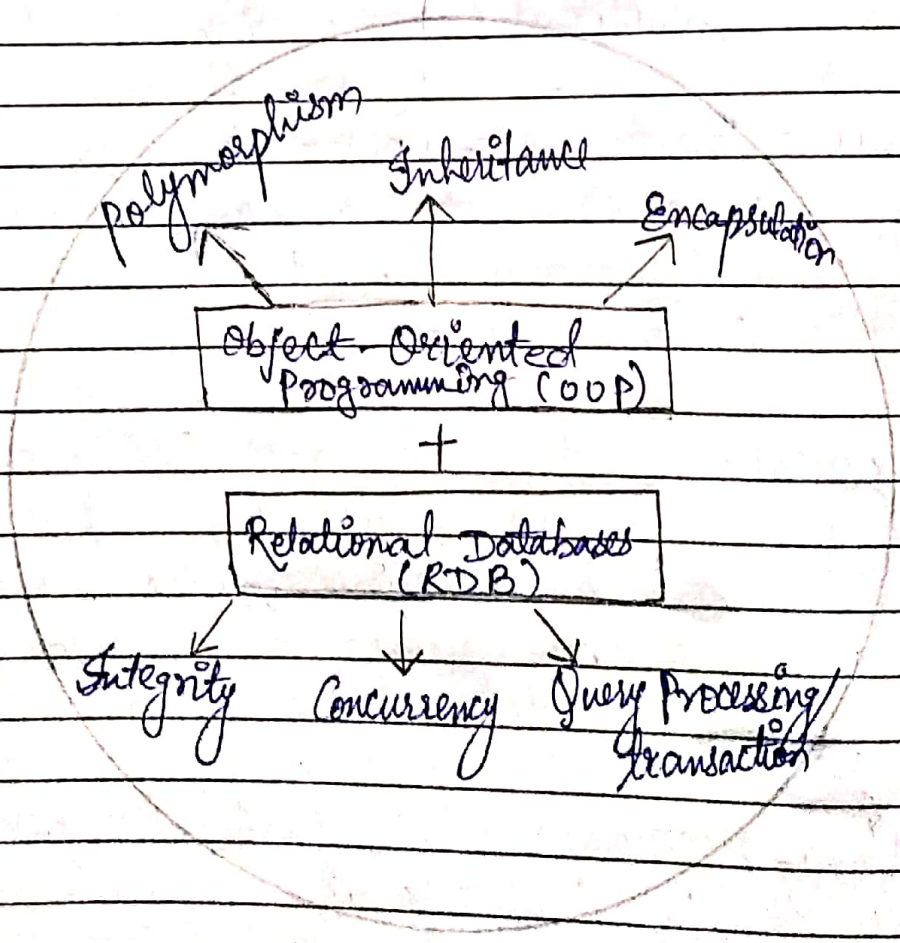
3. Encapsulation : It is an data hiding concept in OOPL which binds the data and functions together which can manipulate data & not visible to outside world.

4. Persistency : OODBMS allows to create persistent object (object remains in memory even after execution. This features can

automatically solve the problem of recovery & concurrency

5. Polymorphism : Polymorphism means having many forms. In simple words, we can define polymorphism as the ability of a message to be displayed in more than one form.

Ex - Area() → at the same time area() of circle, rectangle, square, etc.



Page No.
Date

Object oriented database is product of OOP and RDB (Relational Data base)

Relational DBMS	Object Oriented DBMS
Data Integrity	Object Oriented Interface
Relational Algebra	Object Identifier
Relational Calculus	Inheritance
Persistency	Polymorphism
Integrity	Encapsulation
Security	
Performance	
OR	OR
Relation / Table	Class /
Tuple / Record	Object / Object Instance
Attribute / Column	Variable
Stored Procedure	Method.

Logical Database

Logical database is a special program which retrieves data from various tables, which are interselected and provides a read only view of data to read data from a databases tables we use logical database. A logical database is a hierarchical structure of tables. Logical databases contain open SQL statements that read data from the database. Therefore, you do not need to use SQL in your own language of programs. The logical databases reads the program, stores them in the program if necessary, and then passes them line by line to the application program or the function module.

Functions of logical database :

1. Use the GET statement to process logical databases. Logical database consists of logically related tables group together used for reading and processing data.
2. Preparation of the data records by the logical database and reading of the data records in the actual report are accomplished with the command pair - Put and Get.
3. The three main elements of LDB (logical database) are - structure, selections, database program.

Advantages of logical database :

1. No need of programming for retrieval, meaning for data selection.

2. Easy to use standard user interface, basic check completeness of user input.
3. It offers an easy-to-use selection screen. You can modify the pregenerated selection screen to your needs. It offers check functions to check whether user input is complete, correct.
4. It offers reasonable data selection.
5. It offers and contain central authorization check for database accesses.
6. Meaningfull data selection and good read access performance while retaining the hierarchical data view determined by the application logic.

Disadvantage of logical database.

1. Fast in case of lesser no. of tables, but if the table is in the lowest level of hierarchy, all upper level tables should be read so performance is slower.
2. If you do not specify a logical database in the program attributes, GET events never occur.
3. There is no ENDGET command, so the code block associated with an event ends with the next event statement.

Page No.
Date

Web Databases

A web database is essentially a database that can be accessed from a local network or the Internet instead of one that has its data stored on a desktop or its attached storage. Used for both professional and personal use, they are hosted on websites, and are software as service products, which means that access is provided via a web browser. One of the type of web ~~database~~ browser. One of the type of web database that you may be more familiar with is a relational database. Relational database allows you to store data in groups (known as tables), through its ability to link records together. It uses indexes, and keys, which are added to data, to locate information, fields stored in the database, enabling you to retrieve information quickly.

For eg: Just think about when you online shop and want to have a look on specific product. Typing in keyboard such as "black dress" enables all the black dresses stored on the website to appear right on the browser you are looking on, because the information "black" and "dress" are stored in database entries.

Advantages of Web database

1. Web database applications can be free or require payments usually through monthly subscription. Because of this, you pay for the amount you use. So, whether your

business shrinks or expands, your needs can be accomodated by the amount of server space. You also don't have to fork out for the cost of installing an entire software program.

2. The information is ~~available~~ accessible from almost any device. Having things stored in a cloud means that it is not stuck to one computer. As long as you are granted access, you can technically get a hold of the data from just about any compatible device.
3. It's convenient - Web databases allow users to update information so all you have to do is to create simple web forms.

MY SQL

Something you will see commonly attached to the topic of web databases, and also worth noting due to its use in many high profile websites, such as google, facebook, twitter and wordpress in MySQL.

Disadvantage of Web database

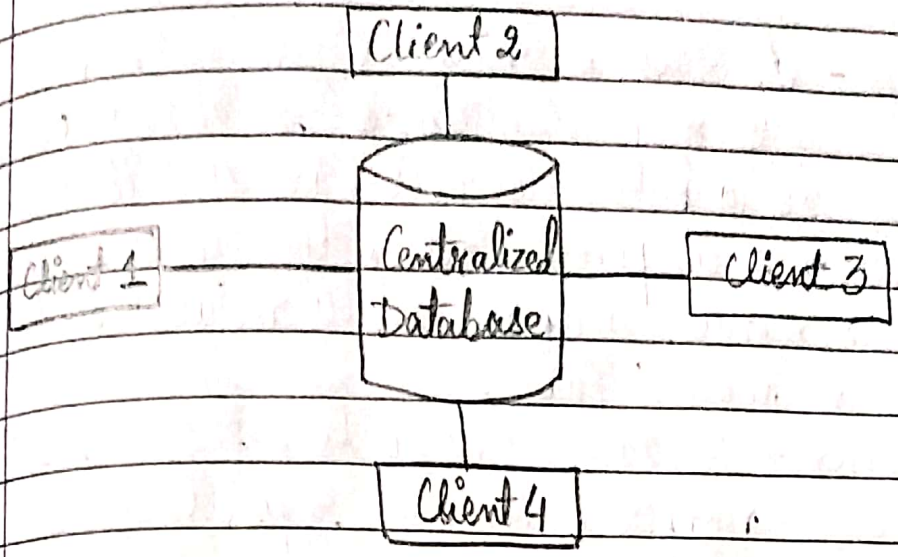
1. No Internet connection = no access
In web database if you don't have a reliable internet connection to support your online database software, you will not be able to access your data and if you are dealing with every sensitive data you can be out of touch with it.

2. Subscription and per user pricing - In web database a user have to pay some cost to use the data and software.

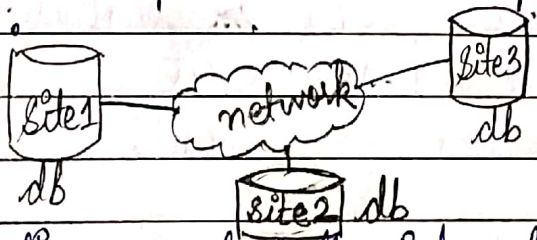
3. Cost of extra storage. Sometimes a user have to pay more money for using even some amount of space and data, according to software demand.

Page No.
Date

Distributed Databases



A distributed database is a collection of multiple interconnected databases, which are spread physically across various locations that communicate via a computer network i.e., Telephone, network applications - peer to peer networks.



Features

- Databases in the collection are logically interrelated with each other. Often they represent a single logical database.
- Data is physically stored across multiple sites. Data in each site can be managed by a DBMS independent of the other sites.
- The processors in the sites are connected via a network. They do not have any multiprocessor configuration.
- A distributed database is not a loosely connected file system.
- Continuous operation.

Page No.
Date

Advantages of Distributed Databases

- More Reliable – In case of database failures, the total system of centralized databases comes to a halt. However, in distributed systems, when a component fails, the functioning of the system continues may be at a reduced performance. Hence, it is more reliable.
- Better Response – If data is distributed in an efficient manner, then user requests can be met from local data itself, thus providing faster response.
- Lower Communication Cost – In distributed database systems, if data is located locally where it is mostly used, then the communication costs for data manipulation can be minimized. This is not feasible in centralized systems.
- Increased availability
- Modular Development.

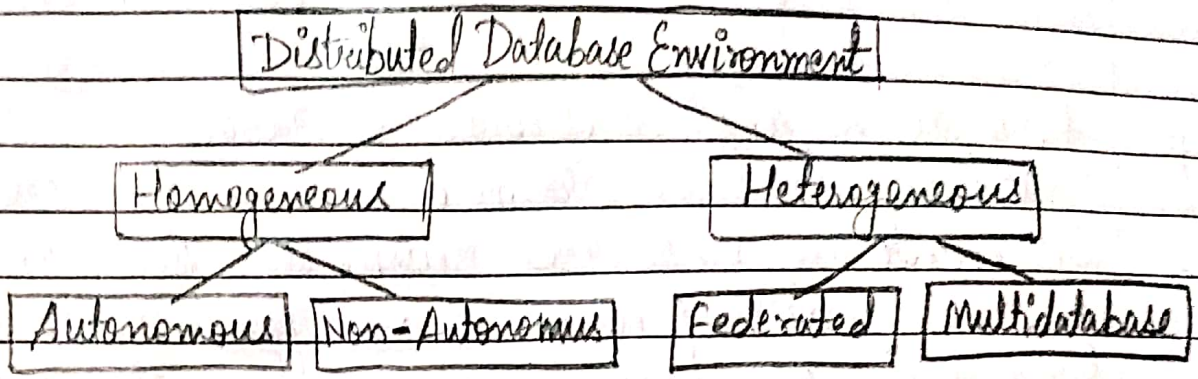
Disadvantages of Distributed Databases

- Need for complex and expensive software – It demands complex and often expensive software to provide data transparency and co-ordination across the several sites.
- Maintenance of data is difficult.
- Data Integrity – The need for updating data in multiple sites pose problems of data integrity.
- Overheads for improper data distribution – Responsiveness of queries is largely dependent upon proper data distribution.

Page No.
Date

Improper data distribution often leads to very slow response to user requests.

Types of Distributed Databases



Homogeneous Distributed Databases

In a homogeneous distributed database, all the sites use identical DBMS and operating systems. Its properties are:

- The sites use very similar software, i.e., same database system software.
- The sites use identical DBMS or DBMS from the same vendor.
- Each site is aware of all other sites and cooperates with other sites to process user requests.

Heterogeneous Distributed Databases

In a heterogeneous distributed database, different sites have different operating systems, DBMS products and data models. Its properties are:

- Different sites use dissimilar schemas and software.
- The system may be composed of a variety of DBMSs like relational, network, hierarchical or object oriented.
- Query processing is complex due to dissimilar schemas.

Q. Why do we need distributed database?

Ans Distributed databases are more reliable than centralized systems. If the WAN goes down, each site can continue processing using its own portion of the database. Only those data manipulation operations that require data not on site will be delayed.

Q. Why we use distributed system?

Ans Distributed computing allows different users or computers to share information. Distributed computing can allow an application on one machine to leverage processing power, memory or storage on another machine. Some applications, such as word processing, might not benefit from distribution at all.

DATA WAREHOUSING

Background

A Database management system (DBMS) stores data in the form of tables, uses ER model. For example a DBMS of college has tables for students, faculty, etc.

A Data Warehouse is separate from DBMS, it stores huge amount of data, which is typically collected from multiple heterogeneous source like files, DBMS, etc. The goal is to produce statistical results that may help in decision makings. For example, a college might want to see quick different results, like how is the placement of CS students has improved over last 10 years, in terms of salaries, counts, etc.

Need of Data Warehouse

An ordinary database can store MBs to GBs of data and that too for a specific purpose. For storing data of TB size, the storage shifted to Data Warehouse. Besides this, a transactional database doesn't offer itself to analytics. To effectively perform analytics, an organisation keeps a central Data Warehouse to closely study its business by organizing, understanding and using its historic data for taking strategic decisions and analyzing trends.

Page No.
Date

Data Warehouse V/S DBMS

Data Warehouse	Database
1. A data warehouse is based on analytical processing.	1. A common database is based on operational or transactional processing. Each operation is an indivisible transaction.
2. A datawarehouse maintains historical data over time. Historical data is the data kept over years and can be used for trend analysis, make future predictions and decision support.	2. Generally, a database stores current and up-to-data which is used for daily operations.
3. A datawarehouse is integrated generally at the organisation level, by combining data from different databases.	3. A database is generally application specific.
4. Constructing a datawarehouse can be expensive.	4. Constructing a database is not so expensive.
5. Example: A datawarehouse integrates the data from one or more databases, so that analysis can be done to get results, such as the best performing school in a city.	5. Example: A database stores related data, such as the student details in a school.

Page No.
Date

Applications of Data Warehousing

Data Warehousing can be applicable anywhere we have huge amount of data and we want to see statistical results that helps..

- Social Media Websites : The social networking websites like facebook, twitter, linkedin, etc are base on analyzing large data sets. These sites gather data related to members, groups, locations, etc and store it in a single central repository. Being large amount of data, Data warehouse is needed for implementing the same.
- Banking : Most of the banks these days use warehouse to see spending patterns of account/car holders. They use this to provide them special offers, deals, etc.
- Government : It uses data warehouse to store and analyze tax payment which is used to detect tax thefts.

There can be many more applications in different sectors like E-Commerce, Telecommunication, Transportation Services, Marketing and distribution, healthcare and retail.

Page No.
Date

Data Warehouse Architecture, Concepts & Components

Data Warehouse Concepts

The basic concept of a Data Warehouse is to facilitate a single version of truth for a company for decision making and forecasting. A data warehouse is an information system that contains historical and communicative data from single or multiple source. Data Warehouse Concepts Simplify the reporting and analysis process of organisations.

Characteristics of Data Warehouse

Data Warehouse Concepts have following characteristics:

Subject-Oriented.
Integrated
Time-variant
Non-volatile

Subject-Oriented

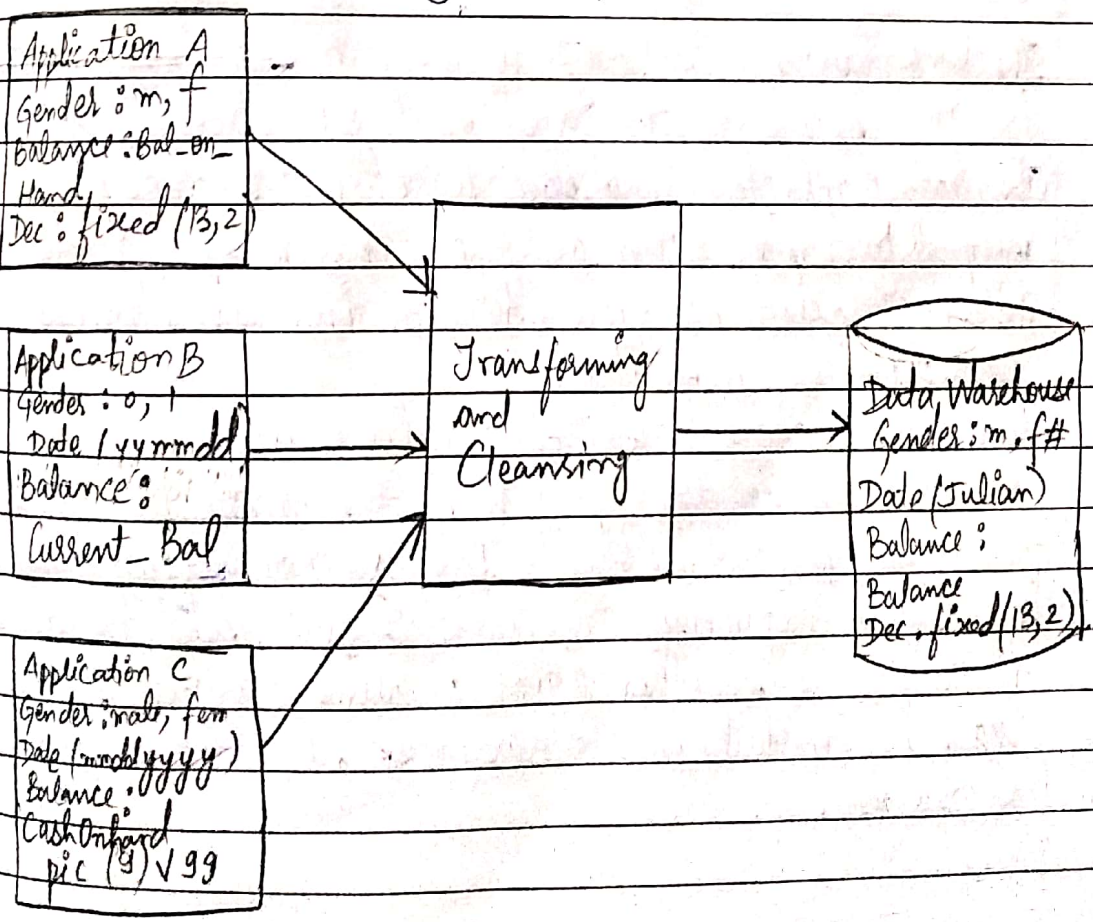
A data warehouse is subject oriented as it offers information regarding a theme instead of companies' ongoing operations. These subjects can be sales, marketing, distributions, etc.

A data warehouse never focuses on the ongoing operations. Instead, it put emphasis on modeling and analysis of data for decision making. It also provides a simple and concise view around the specific subject by excluding data which is not helpful to support the decision process.

Page No.
Date

Integrated

In Data Warehouse, integration means the establishment of a common unit of measure for all similar data from the dissimilar database. The data also needs to be stored in the Data warehouse in common and ~~university~~ universally acceptable manner.
A data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc. Moreover, it must keep consistent naming conventions, format and coding. This integration helps in effective analysis of data. Consistency in naming conventions, attribute measures, encoding structure etc. have to be ensure. Consider the following example:



In the above example, there are three different applications labeled A, B and C. Information stored in these applications are Gender, Date and Balance. However, each application data is stored different way.

- In application A gender field store logical values like M or F.
- In application B gender field is a numerical value.
- In application C, gender field stored in the form of a character value.
- Same is the case with Date and balance.

However, after transformation and cleaning process all this data is stored in common format in the Data Warehouse.

Time-Variant

The time horizon for data warehouse ~~data display time~~ is quite extensive compared with operational systems. The data collected in a data warehouse is recognized with a particular period and offers information from the historical point of view. It contains an element of time, explicitly or implicitly.

One such place where Datawarehouse data display time variance is in the structure of the record key. Every primary key contained with the Data Warehouse should have either implicitly or explicitly an element of time. Like the day, week, month, etc. Another aspect of time variance is that once data is inserted in the warehouse, it can't be updated or changed.

Page No.
Date

Non-Volatile

Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it. Data is read-only and periodically refreshed. This also helps to analyze historical data and understand what & when happened. It does not require transaction process, recovery and concurrency control mechanisms. Activities like delete, update and insert which are performed in an operational application environment are omitted in Data warehouse environment. Only two types of data operations performed in the Data Warehousing are:

1) Data loading
2) Data access

Operational Application	Data Warehouse
→ Complex program must be coded to make sure that data upgrade processes maintain high integrity of the final product.	→ This kind of issues does not happen because data update is not performed.
→ Data is placed in a normalized form to ensure minimal redundancy	→ Data is not stored in normalized form.
→ Technology needed to support issues of transactions, data recovery, rollback, and resolution as its deadlock is quite complex.	→ It offers relative simplicity in technology.

Page No.
Date

Data Warehouse Architecture

Data Warehouse Architecture is complex as it's an information system that contains historical and commutative data from multiple sources. There are 3 approaches for constructing Data Warehouse layers: Single Tier, Two Tier and Three Tier. This 3 tier architecture of Data Warehouse is explained as below.

→ Single-tier architecture: The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice.

→ Two-tier architecture: It is one of the datawarehouse layers which seperates physically available sources and data warehouse. This architecture is not expandable and also not supporting a large number of end-users. It also has connectivity problems because of network limitations.

→ Three-tier architecture: This is the most widely used architecture of data warehouse. It consists of the Top, Middle and Bottom tier.

1) Bottom Tier: The database of the datawarehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed and loaded into this layer using back-end tools.

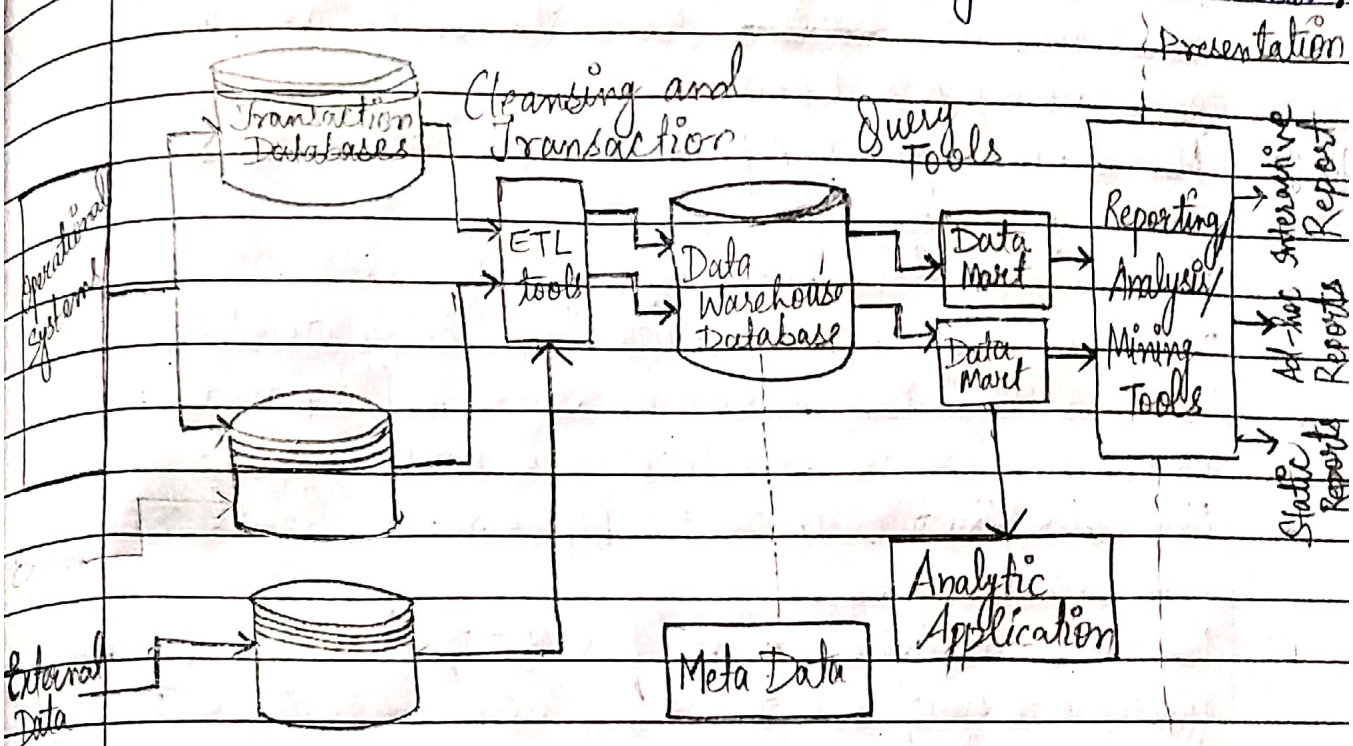
2) Middle Tier: It is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator b/w the end-user and the database.

3) Top-Tier: It is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. It could be Query tools, reporting tools, managed query tools, analysis tools and datamining tools.

Page No.
Date

Data warehouse Components

We will learn about the Datawarehouse Components and Architecture of Data Warehouse with Diagram as shown below:



DATA WAREHOUSE ARCHITECTURE

The data warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key Data Warehousing components to make the entire environment functional, manageable and accessible.

There are mainly 5 Data Warehouse Components:

→ Data Warehouse Database

The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology. Although, this kind of implementation is constrained by the fact that

Page No.
Date

Traditional RDMS system is optimized for transaction database processing and not for data warehousing. For instance, adhoc query, multi-table joins, aggregates are resource intensive and slow down performance.
Hence, alternative approaches to Database are used as listed below:-

- In a datawarehouse, relational databases are deployed in parallel to allow for scalability. Parallel relational databases also allow shared memory or shared nothing model on various multiprocessor configurations or massively parallel processors.
- New index structures are used to bypass relational table scan and improve speed.
- Use of multidimensional database (MDDBs) to overcome any limitations which are placed because of the relational Data Warehouse Models.

Example: Essbase from Oracle.

→ Sourcing, Acquisition, Clean-up and Transformation Tools
The data sourcing, transformation and migration tools are used for performing all the conversions, summarizations and all the changes needed to transform data into a unified format in the datawarehouse. They are also called Extract, Transform and Load (ETL) tools
Their functionality includes:

- Eliminating unwanted data in operational databases from loading into Data warehouse.
- Search and replace common names and definitions for data arriving from different sources.
- Calculating summaries and derived data.

In case of missing data, populate them with defaults.
De-duplicated repeated data arriving from multiple data sources.
These Extract, Transform and Load tools may generate cron jobs, background jobs, Cobol programs, shell scripts, etc that regularly update data in data warehouse. These tools are also helpful to maintain the Metadata.
These ETL tools have to deal with challenges of Database and Data heterogeneity.

→ Metadata

The name MetaData suggests some high-level technological Data Warehousing Concepts. However, it is quite simple. Metadata is data about data which defines the data warehouse. It is used for building, maintaining and managing the data warehouse.

In the Data Warehouse Architecture, meta-data plays an important role as it specifies the source, usage, values and features of data warehouse data. It also defines how data can be changed and processed. It is closely connected to the data warehouse.
For example, a line in sales database may contain:
4030 KJ732 299.90
This is a meaningless data until we consult the Meta that tell us it was
- Model number: 4030
- Sales Agent ID: KJ732
- Total Sales amount of $ 299.90

Page No.
Date

Therefore, Meta Data are essential ingredients in the transformation of data into knowledge.
Metadata helps to answer the following questions
- What tables, attributes and keys does the Data Warehouse contain?
- Where did the data come from?
- How many times do data get reloaded?
- What transformations were applied with cleansing?

Metadata can be classified into following categories:
1. Technical Meta Data: This kind of Metadata contains information about warehouse which is used by Data warehouse designers and administrators.
2. Business Meta Data: This kind of Metadata contains detail that give end-users a way easy to understand information stored in the data warehouse.

→ Query Tools:
One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions. Query tools allows users to interact with the data warehouse system.
These tools fall into four different categories:
1. Query and reporting tools:
It can be further divided into
- Reporting tools
- Managed query tools

Page No.
Date

Reporting Tools can be further divided into production reporting tools and desktop report writer.

1. Report writers : This kind of reporting tool are tools designed for end-users for their analysis.
2. Production reporting : This kind of tools allows organisations to generate regular operational reports. It also supports high volume batch jobs like printing and calculating. Some popular reporting tools are BEio, Business Objects, Oracle, PowerSoft, SAS Institute.

Manage query tools helps end users to resolve snags (complications) in database and SQL and database structure by inserting meta-layer between users and database.

1. Application development tools : Sometimes built-in graphical and analytical tools do not satisfy the analytical needs of an organization. In such cases, custom reports are developed using Application development tools.
2. Data mining tools : It is a process of discovering meaningful new correlation, patterns and trends by mining large amount data. Data mining tools are used to make this process automatic.
3. OLAP tools : These tools are based on concepts of a multidimensional database. It allows users to analyse the data using elaborate and complex multidimensional views.

→ Data warehouse Bus Architecture

It determines the flow of data in your warehouse. The data flow in a data warehouse can be categorized as Inflow, Upflow, Downflow, Outflow and Metaflow. While designing a Data Bus, one needs to consider the shared dimensions, facts across data marts.

Data Marts

It is an access layer which is used to get data out to the users. It is presented as an option for large size data warehouse as it takes less time and money to build. However, there is no standard definition of a data mart is differing from person to person.
In a simple word, it is a subsidiary of a data warehouse. The data mart is used for partition of data which is created for the specific group of users. Data marts could be created in the same database as the Datawarehouse or a physically separate database.

Page No.
Date

Data Warehouse Architecture Best Practices

To design Data Warehouse Architecture, you need to follow below given best practices:

- Use data warehouse models which are optimized for information retrieval which can be the dimensional mode, denormalized or hybrid approach.
- Choose the appropriate designing approach as top down and bottom up approach in Data Warehouse.
- Need to assure that Data is processed quickly and accurately. At the same time, you should take an approach which consolidates data into a single version of the truth.
- Carefully design the data acquisition and cleansing process for data warehouse.
- Design a MetaData architecture which allows sharing of metadata b/w components of data warehouse.
- Consider implementing an ODS model when information retrieval need is near the bottom of the data abstraction pyramid or when there are multiple operational sources required to be accessed.
- One should make sure that the data model is integrated and not just consolidated. In that case, you should consider 3NF data model. It is also ideal for acquiring ETL and Data Cleansing tools.

Page No.
Date

ETL (Extract, Transform & Load) Process in data warehouse

What is ETL?

It is a process that extracts the data from different source systems, then transform the data and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load. It's tempting to think that creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Datawarehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives & is technically challenging. In order to maintain its value as a tool for decision-makers, Data warehouse system needs to change with business changes. ETL is a recurring activity of a datawarehouse system and needs to be agile, automated and well documented.

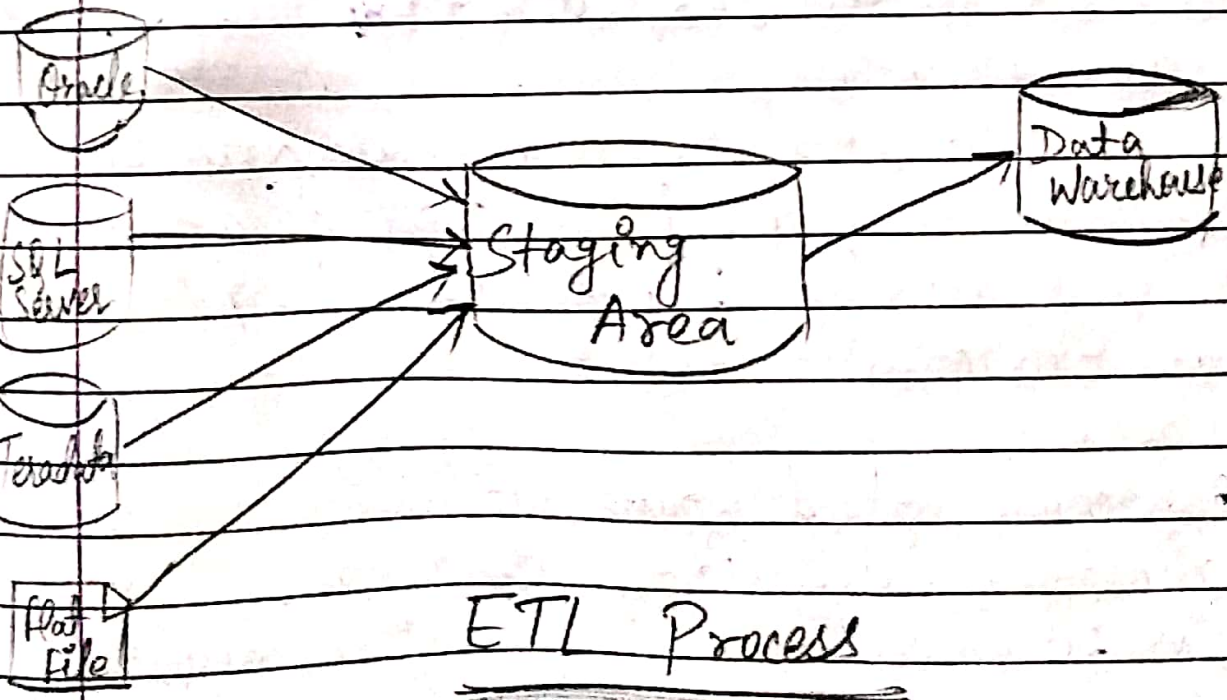
Why do you need ETL?

There are many reasons for adopting ETL in the organization:

- It helps companies to analyze their business ~~questions~~ data for taking critical business decisions.
- Transactional databases cannot answer complex business questions that can be answered by ETL example.
- A data warehouse provides a common data repository
- ETL provides a method of moving the data from various sources into a datawarehouse.
- Well-designed and documented ETL system is almost essential to the success of a Datawarehouse project.

Page No.
Date

- ETL process allows sample data comparison b/w the source and the target system.
- ETL process can perform complex transformations and requires the extra area to store the data.
- ETL helps to migrate data into a Datawarehouse. Convert to the various formats and types to adhere to one consistent system.
- ETL is a predefined process for accessing and manipulating source data into the target database.
- ETL in data warehouse offers deep historical context for the business.
- It helps to improve productivity because it codifies and reuses without a need for technical skills.



ETL Process

Page No.
Date

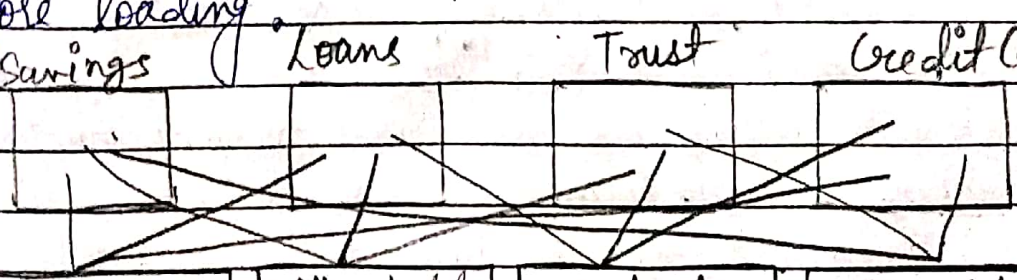
Step 1) Extraction : In this step of ETL architecture, data is extracted from the source system into the staging area. Transformations if any are done in staging are so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Datawarehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data Warehouse.
Data Warehouse needs to integrate systems that have different DBMS, hardware, Operating Systems and communication protocols. Sources could include legacy applications like mainframes, customized applications, Point of contact devices like ATM, Call switches, textfiles, spreadsheets, ERP, data from vendors, partners amongst others.
Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship b/w sources & target data.
Three Data Extraction methods :

1. Full extraction
2. Partial Extraction - without update notification.
3. Partial extraction - with update notification.

Irrespective of the method used, extraction should not affect performance and response time of the source systems. These source systems are live production databases. Any slow down or locking could effect company's bottom line.

Some validations are done during extraction:
Reconcile records with the source data
Make sure that no spam / unwanted data loaded.
Datatype check.
Remove all types of duplicate / fragmented data
Check whether all the keys are in place or not.

Step 2) Transformation: Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.
It is one of the important ETL concepts where you apply a set of functions on extracted data. Data that does not require any transformation is called as direct move or pass through data.
In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name & last name in a table is in different columns. It is possible to concatenate them before loading.



Same data different name	Different data Same Name	Data found here nowhere else	Different keys Same data

Data Integration Issues

Following are data integrity problems

1. Different spelling of the same person like Jon, John, etc.
2. There are multiple ways to denote company name like Google, Google Inc.
3. Use of different names like Cleaveland, Cleveland.
4. There may be a case that different account numbers are generated by various applications for the same customer.
5. In some data required files remains blank.

Validations are done during this stage

- Filtering - Select only certain columns to load.
- Using rules and look up tables for Data standardization
- Character set conversion and encoding handling.
- Conversion of units of measurements like Date Time Conversion, currency conversions, numerical conversions, etc.
- Data threshold validation check. For example, age cannot be more than two digits.
- Data flow validation from the staging area to the intermediate tables.
- Required fields should not be left blank.
- Cleaning (for example, mapping NULL to 0 or Gender Male to "M" and female to "F", etc.)
- Split column into multiples & merging multiple columns into a single column.
- Transposing rows & columns.
- Use lookups to merge data
- Using any complex data validation.

Page No.
Date

Step 3) Loading : Loading data into the target datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period. Hence, load process should be optimized for performance. In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data warehouse admins need to monitor, resume cancel loads ~~process should be optimized for performance~~ as per prevailing server performance.

Types of loading

- Initial load – populating all the Data warehouse tables
- Incremental load – applying ongoing changes as when needed periodically.
- Full Refresh – Erasing the contents of one or more tables and reloading with fresh data.

Load verification

- Ensure that the key field data is neither missing nor null.
- Test modeling views based on the target tables.
- Check that combined values and calculated measures.
- Data checks in dimension table as well as history table.
- Check the BI reports on the loaded fact and dimension table.

Page No.
Date

ELT Tools

There are many Data Warehousing tools available in the market. Here, are some most prominent one:

1. Market Logic: It is a data warehousing solution which makes data integration easier & faster using an array of enterprise features. It can query different types of data like documents, relationships & metadata.
2. Oracle: It is the industry-leading database. It offers a wide range of choice of data warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.
3. Amazon Red Shift: Amazon Redshift is Datawarehouse tool. It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data.

Best practices for ETL process

Following are the best practices for ETL process steps:

→ Never try to cleanse all the data: Every organisation would like to have all the data clean, but most of them are not ready to pay to wait or not ready to wait. To clean it all would simply take too long, so it is better not to try to cleanse all the data.

Page No.
Date

- Never cleanse anything: Always plan to clean something because the biggest reason for building the Data Warehouse is to offer cleaner and more reliable data.
- Determine the cost of cleansing the data: Before cleansing all the dirty data, it is important for you to determine the cleansing cost for every dirty data element.
- To speed up query processing, have auxiliary views and indexes: To reduce storage costs, store summarized data into disk tapes. Also, the trade-off b/w the volume of data to be stored and its detailed usage is required. Trade-off at the level of granularity of data to decrease the storage costs.

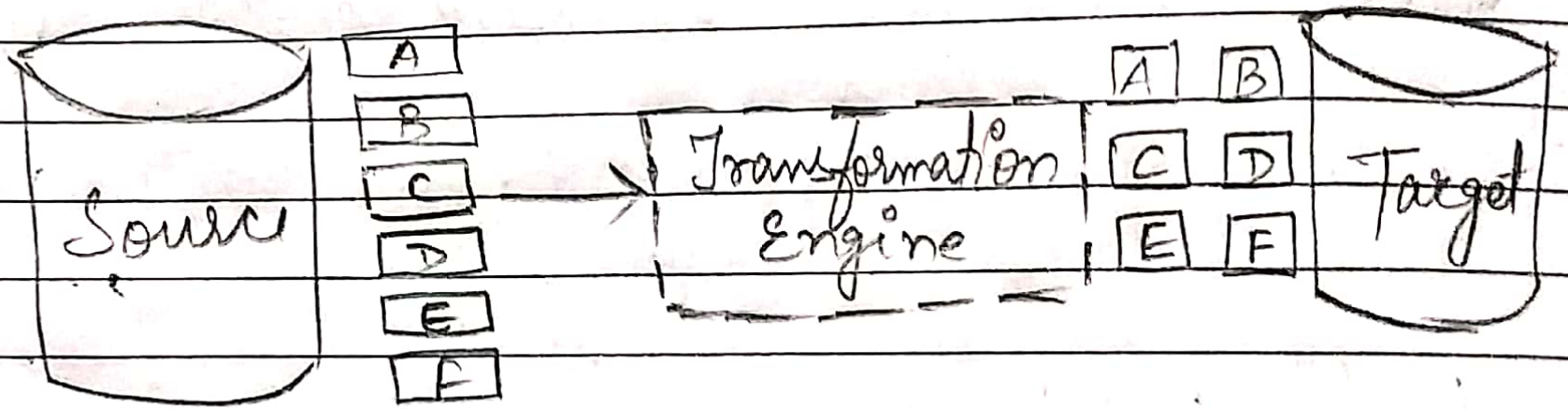
What is ETL?

Extract, Transform & Load. In this process, an ELT tool extracts the data from different RDBMS source systems then transforms the data like applying calculations, concatenations, etc. and then load the data into the Data Warehouse system. In ETL data flows from the source to the target. In ETL process transformation engine takes care of any data changes.

What is ELT?

It is a different method of looking at the tool approach to data movement. Instead of transforming the data before it's written, ELT lets the target system to

do the transformation. The data first copied to the target and then transformed in place. ELT usually used with no-SQL databases like Hadoop clusters, data appliance or cloud installation.



Date :
Page :

Difference between ETL and ELT

- ETL stands for extract, Transform and load while ELT stands for extract, load, Transform.
- ETL loads data first into the staging server and then into the target system whereas ELT loads data directly into the target system.
- ETL model is used for on-premises, relational and structured data while ELT is used for scalable cloud structured and unstructured data sources.
- ETL is mainly used for a small amount of data whereas ELT is used for large amounts of data.
- ETL doesn't provide data lake supports while ELT provides data lake support.
- ETL is easy to implement whereas ELT requires skills to implement and main-tain.

Difference between data warehousing and data mining.

- Data mining is considered as a process of extracting data from large data sets, whereas a data warehouse is the process of pulling all the relevant data together.
- Data mining is the process of analyzing unknown patterns of data, whereas a data warehouse is a technique for collecting and managing data.
- Data mining is usually done by business user with the assistance of engineers. While Data warehousing is a process which needs to occur before any data mining can take place.
- Data mining allows users to ask more complicated queries, which would increase the workload while data warehouse is complicated to implement and maintain.
- Data mining helps to create suggestive patterns of important factors like the buying habits of customers while Data Warehouse is useful for operational business systems like CRM (customer relationship management) systems where warehous is integrated.

Data Mining

Data Mining refers to the extraction of useful information from a bulk of data or data warehouses.

OR

Data Mining is the process of discovering or mining knowledge from a large amount of data.

The result of data mining is the hidden patterns and the knowledge that we gain at the end of the extraction process.

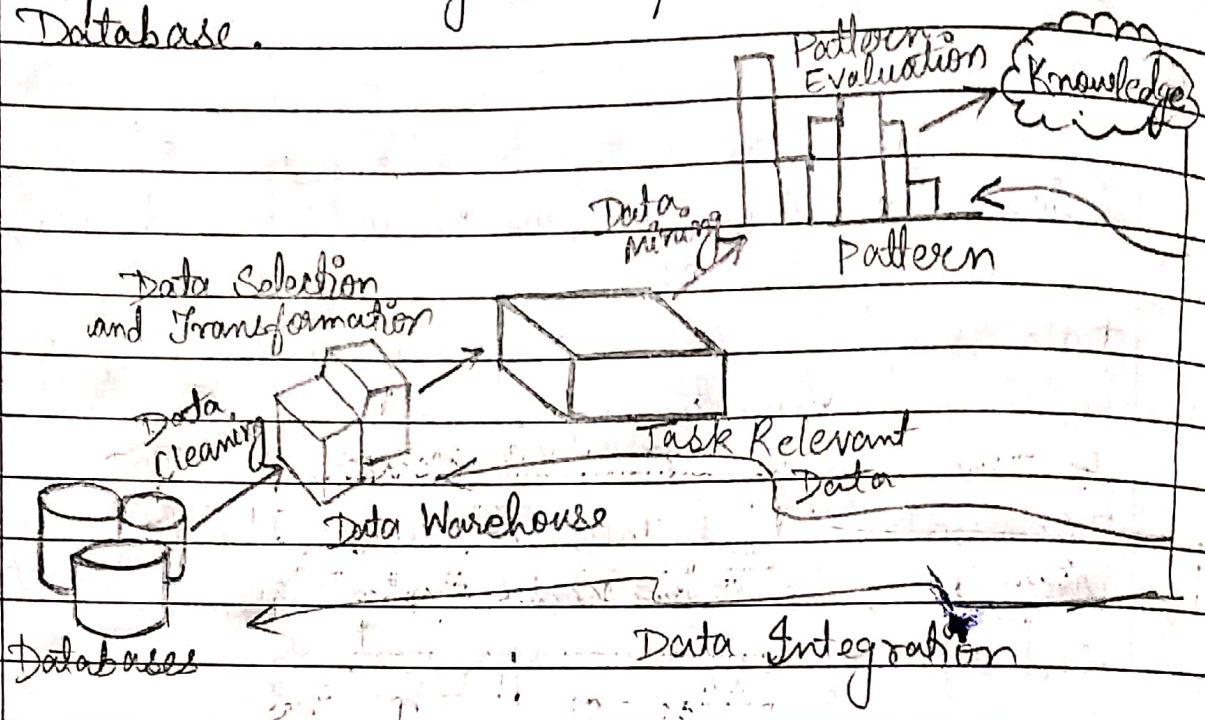
Data mining is also known as Knowledge Discovery from Data (KDD).

Purpose Of Data Mining

With the advancement in technology there is an increase in the size of the database, manual analysis of that data to get meaningful information from different perspectives and dimensions is not possible so we need automatic analysis for that to make our task or process effective and better than before.

Page No.
Date

Steps in Data Mining Process / Phases of KDD in Database.

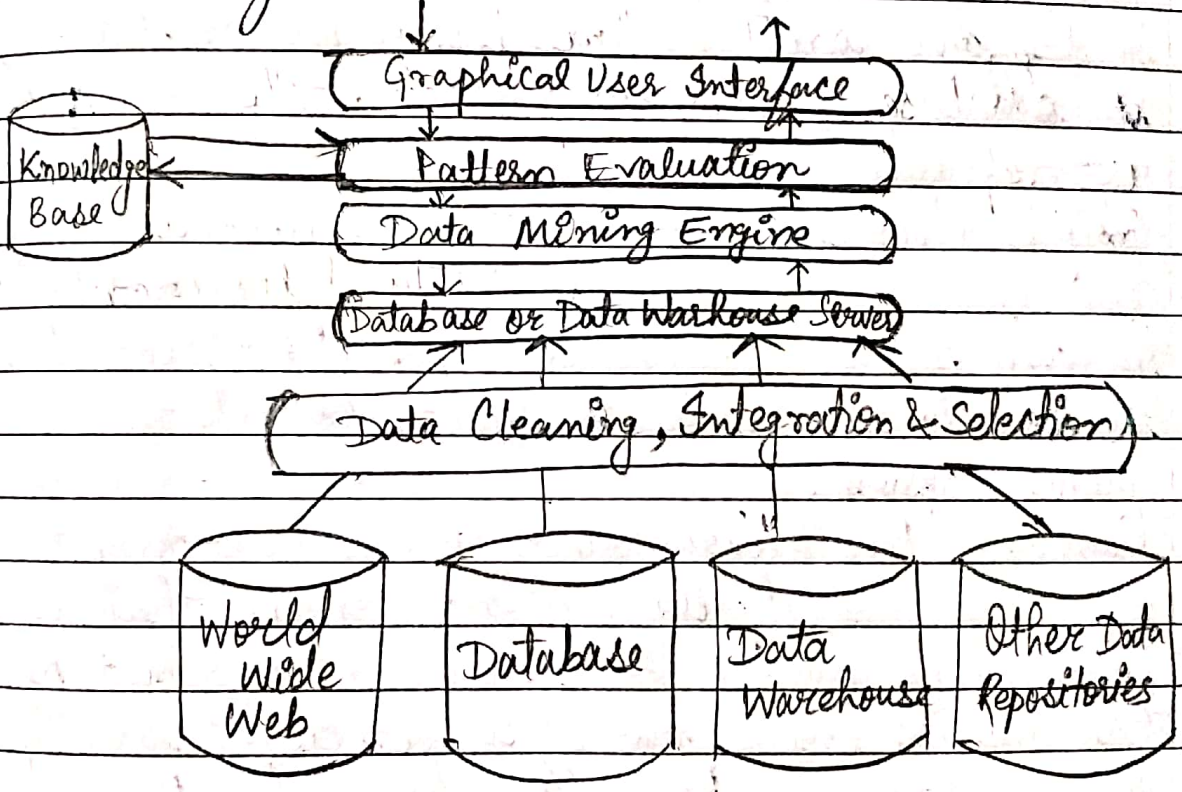


1. Data Cleaning : In this step, the noisy and irrelevant data is removed from the collection.
2. Data Integration : In this step, the heterogeneous data from multiple sources is combined in a common source.
3. Data Selection : In this step, the data relevant to the analysis is decided and retrieved from the data collection.
4. Data Transformation : In this step, the data is transformed into appropriate form required by the mining procedure.
5. Data Mining : In this step, intelligent techniques are applied to extract meaningful information that is useful for us.
6. Pattern Evaluation : In this step, strictly increasing patterns are identified that represent

knowledge based on the given measures.

7. Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools like tables, reports, etc to represent data mining results.

Data Mining Architecture



- Data Sources: Database, World Wide Web (WWW). Data warehouse and other data repositories are data sources. The data in these sources may be in the form of plain text, spreadsheets or in other forms of media like photos or videos. WWW is one of the biggest sources of data.
- Database Server: The database server contains the actual data ready to be processed. It performs the task of handling data retrieval as per the request of the user.

- Data Mining Engine: It is one of the core components of the data mining architecture that performs all kinds of data mining techniques like association, classification, characterization, clustering, prediction, etc.
- Pattern Evaluation Modules: They are responsible for finding interesting pattern in the data and sometimes they also interact with the database servers for producing the result of the user requests.
- Graphic User Interface: Since the user cannot fully understand the complexity of the data mining process so graphical user interface helps the user to communicate effectively with the data mining system.
- Knowledge Base: Knowledge Base is an important part of the data mining engine that is quite beneficial in guiding the search for the result patterns. Data mining engine may also sometimes get inputs from the knowledge base. This knowledge base may contain data from user experiences. The objective of the knowledge base is to make the result more accurate and reliable.

Page No.
Date

Applications of Data Mining

- Biological Analysis: If the doctor has all the patient's information, such as medical records, physical examinations and treatment patterns, allows more accurate diagnosis and effective treatments to be prescribed. It also helps in identifying the medical risks and predicting illnesses of patients.
- Financial Analysis - As banks have transaction details and detailed profiles of their customers, they analyse all this data and try to find out patterns which help them predict that certain customers could be interested in loans, credit cards, insurances, etc. Loan payment prediction and detection of money laundering and other financial crimes is also done with the help of same information.
- Fraud Detection - A supervised method includes collection of sample records which are classified on the basis of fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.
- Intrusion Detection - Intrusion refers to any kind of action that threatens integrity or confidentiality. Data mining helps in intrusion detection by detecting activities different from usual network activities. Measures to avoid intrusion includes user authentication and information protection.
- Research Analysis - Data mining helps researchers in finding any similar data from the database that might bring any change in the research. For this data visualisation is needed.

- Market Basket Analysis - It's widely used in Retail Industry. If you buy a certain item you are more likely to buy group of items related to it. This technique will allow the retailer to understand the purchase behaviours of buyers which will help him in changing the store's layout accordingly to increase his sale.
- Customer Segmentation - Data mining aids in aligning the customers into distinct segments based on their vulnerabilities so that the seller can tailor the needs of the customers accordingly and can give them special offers which will enhance the customer's satisfaction.
- Education - Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in maths subject.

Advantages of Data Mining

- Improvement in the direct mail promotions through targeted use.
- Aids companies to attract and retain customers.
- Helps the companies to improve their relationship with the customers and increase their sales.
- Compresses data into valuable information.
- Detects fraud practices with more accurate analysis.

Page No.
Date

Disadvantages of Data Mining

- It violates user privacy : Data mining process involves several numbers of factors. But while involving those factors, data mining system violates the privacy of its user and that is why it lacks in the matters of safety and security of its users. Eventually, it creates miscommunication between people.
- Additional irrelevant information : The main functions of the data mining systems create a relevant space for beneficial information. But the main problem with these information collections is that there is a possibility that the collection of information processes can be a little overwhelming for all. Therefore, it is very much essential to maintain a minimum level of limit for all the data mining techniques.
- Misuse of information : As it has been explained that in the data mining system the possibility of safety and security measure are really minimal. And that is why some can misuse this information to harm others in their own way.
- Accuracy of data : One of the most possible limitations of this data mining system is that it can provide accuracy of data with its own limits.

Page No.
Date

Difference between Data Mining and Data Warehousing

- Data mining is considered as a process of extracting data from large data sets, whereas a data warehouse is the process of pooling all the relevant data together.
- Data mining is the process of analyzing unknown patterns of data, whereas a Data warehouse is a technique for collecting and managing data.
- Data mining is usually done by business users with the assistance of engineers while Data warehousing is a process which needs to occur before any data mining can take place.
- Data mining allows users to ask more complicated queries which would increase the workload while Data Warehouse is complicated to implement and maintain.
- Data mining helps to create suggestive patterns of important factors like the buying habits of customers while Data Warehouse is useful for operational business systems like CRM (customer relationship management) systems when the warehouse is integrated.

Past Year Questions From UNIT-4

Q. Write about the four types of data warehouse schemas.

Q. What is data stripping?

Q. What do you understand by distributed data processing?

Q. Explain briefly the database security for integrity of database.

Q. What is data warehouse and its usage?

Q. Discuss security issues involved in database design.

Q. Explain web database.

Q. List and discuss different issues related to database security.

REFERENCE

Database Security:

https://drive.google.com/file/d/1KI2HxpKG0GrB-FzE5mHK2UBxYqqVxpH8/view?usp=sharing

DAC Model:
https://drive.google.com/file/d/1Vc8NhYnb7cdECfuySTfn7NY2R56TaAFJ/view?usp=sharing

MAC Model:

https://drive.google.com/file/d/1zzwzYoUsN3vbc489DTt1VgQwEtxrZk50/view?usp=sharing

RBAC Model:

https://drive.google.com/file/d/1M_U_9iJ5iI4kTPyOj9jaj8c7R9UYvOUs/view?usp=sharing

Intrusion Detection:
https://drive.google.com/file/d/1fNA5E_jPnUDAw1OpYmq7Hc14b3E_czsk/view?usp=sharing

SQL Injection:

https://drive.google.com/file/d/1iDFGV7K5AfHFFJhi75hu65LYitFRrkJd/view?usp=sharing

Data Warehouse:

https://drive.google.com/file/d/1Ze-bP6XZXOtKHByLfbAz1qNmoyP7ZBOd/view?usp=sharing

Data Mining:

https://drive.google.com/file/d/1dR-C2XBdbf9WJoW1TdoUSCWhOHKtSkMG/view?usp=sharing

Distributed Database:

https://drive.google.com/file/d/1uVWlsEc6RaG_3horx-qC3S_j8-gCwdoc/view?usp=sharing

Logical and Web Database:

https://drive.google.com/file/d/1lYrlJtM76qJK7ojvWaVdmWqFLRdBG8JY/view?usp=sharing