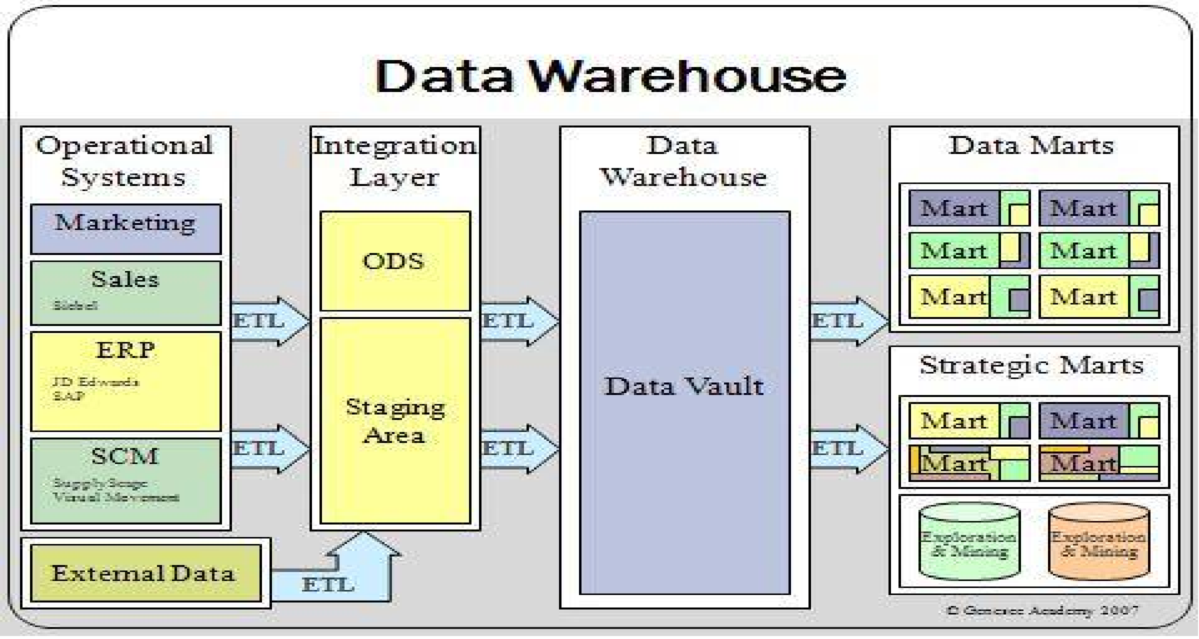
CH-1 : Introduction of Data Warehouse

- A Data Warehouse (DW, DWH), or an Enterprise Data Warehouse (EDW), is a system used for data analysis and reporting.
- Here the data will be integrated from one or more sources and managed centrally. It is one type of house where all data will be stored and managed that will be used for future report generation and use.
- Data warehouses store current and historical data and are used for creating reports for senior management reporting such as annual and quarterly comparisons.
- The data stored in the warehouse is uploaded from the operational systems (such as marketing, sales, and finance etc., shown in the figure below). The data may pass through an operational data store for additional operations before it is used in the Data Warehouse for reporting.
- A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as sales, finance or marketing.
- Data marts are often built and controlled by a single department within an organization.
- Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data.



Operational and Informational systems

- Operational systems are used for data integrity and speed of recording of business transactions through use of database normalization and an entity-relationship model.

- Operational system designers generally follow the Dr. E.F.Codd rules of database normalization in order to ensure data integrity. Dr.E.F. Codd defined five rules of normalization. Fully normalized database designs often result in information from a business transaction being stored in dozens to hundreds of tables.
- Relational databases are efficient at managing the relationships between these tables. The databases have very fast insert/update performance because only a small amount of data in those tables is affected each time a transaction is processed.
- Data warehouses are optimized for analytic access patterns. Analytic access patterns generally involve selecting specific fields and rarely if ever 'select *' as is more common in operational databases.
- Operational systems maintain a snapshot of the business; data warehouses generally maintain an infinite history which is implemented through ETL processes that periodically migrate data from the operational systems over to the data warehouse.
- Information system, an integrated set of components for collecting, storing, and processing data and for deliveringinformation, knowledge, and digital products.
- Business firms and other organizations use on information systems to carry out and manage their operations, interact with their customers and suppliers, and compete in the marketplace.
- Corporations use information systems to reach their customers with targeted messages over the Web, to process financial accounts, and to manage their human resources.
- Governments use information systems used to provide services cost-effectively to citizens. Digital goods, such as electronic books and software, and online services, such as auctions and social networking, are delivered with information systems.

What is Data Warehouse?

- Data warehouse is used to store current and historical data.
- Data warehouse also noun os DW,DWH, or EDW(Enterprise Data Warehouse).
- System used for data analysis and reporting.
- Here the data will be integrated from one or more sources and managed centrally.
- You can say that it is a collection of various databases.
- It is used to store large amount of data or big data.
- Data Warehouse term was first used by “Billinmon” 1990.

- Data Warehouse subject oriented integrated and know collection of data.
- It is used to support in decisionmacking process.
- The data warehouse provides generalize combine & historical data in multi dimension view it also provide “OLAP”(Online Analytical Process).
- The data stored in warehouse is uploaded by different operational system,marketing ,accountetc..here that individual softwere is known as datamart.

Characteristics of Data Warehouse

- Characteristics and functioning of Data Warehousing
  (1) Subject Oriented
  (2) Integrated
  (3) Nonvolatile
  (4) Time Variant

1) Subject Oriented:

- Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented.

2) Integrated:

- Integration is closely related to subject orientation. Data warehouses must put data from different sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

3) 3) Nonvolatile:

- Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

4) Time Variant:

- In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive.

- A data warehouse's focus on change over time is what is meant by the term time variant. Typically, data flows from one or more online transaction processing (OLTP) databases into a data warehouse on a monthly, weekly, or daily basis.
- The data is normally processed in a staging file before being added to the data warehouse. Data warehouses commonly range in size from tens of gigabytes to a few terabytes. Usually, the vast majority of the data is stored in a few very large fact tables.

Explain data warehouse in Today Era or Use of data warehouse in today's market:-

- Data Warehousing is the process of constructing and using the data warehouse.
- In today's trend or era the data warehousing is very much useful.
- Data warehousing involves data cleaning , data integration.

Data Warehouse Advantages:-

- It is very much useful for decision making.
- It is very much useful to support analytic reporting and perform the query.
- In any organization it is helpful to perform production strategies by comparing the new or today's data with old stored data.
- For customer analysis and operations analysis it is helpful here, it analyze the customers choice buying time, relationship with customer, budget cycle etc..are analyze.
- There are many data warehousing applications which are used in different field like:-
    1. Financial Service
    2. Banking Service
    3. Consumer Goods
    4. Retail Sectors
    5. Manufacturing Sectors etc.

Data Warehouse Types:-

1. Information Processing
2. Analytical Processing
3. Data Mining

- **Information Processing:-**
    - Data warehouse allows us to process the information stored in it.
    - The information can be process by query statistical analysis, reporting charts , tables and graphs .
- **Analytical Processing:-**

    - Data warehouse support to analyze the data with basic with operational of OLAP.

- **Data Mining:-**
    - Here, you can discover or get the information or knowledge by finding different model, hidden patterns, classification and prediction.
    - The data mining result can be presented using visualization.

Explain future trends in Data Warehousing.

- Data warehousing is very much useful in the future era.
- For this it use traditional and new improvements because day by day the data will be increased and the market will be grow.
- The future trend is the age of the customer and consumer.
- So, data warehouse is use with business intelligence for growing and successes the business otherwise you will be fail.
- Data warehousing provides you up to date data and analytical information about the customers. So , you can customers using business intelligency(BI).

Future trends of or in DWH?

- The data warehousing is very much useful for business intelligency(BI) , use of Data binding, Store big data, fast analysis and effective business solution.

  There are many softwares which use big data and provide fast result and reach more customers for ex. Agilesoftwere.
- The agile combine with BI and provide the growth of your business faster.
- For the next future which is related to customers and business the DWH is used some future development which are as under.
- The cloud more and more people and business are storing data on the cloud.
- The cloud based computing offers ability to access more data from different sources without the need of changing the data movements and duplication.
- So, the cloud is major factor in future of DWH.

The next generation of data

- In today in feature area there is lange change in data storage or in data mining related to big data.
- The latest concept is internet of thing that is the next generation of data.it also include real time data and string data.

Activity(speed)

- In today area it you want to get success you have to implement such new things like data minding, analyses, BI, IT & diff. new models or software.

How new datawarehousinghelpful :-

- The new datawarehousing solutions ofter business more powerful,achieve real time data by connecting live data with previously stored historical data.
- The new datawarehousing techniques and software provides solutions for business which are under:
    - Instand of storing data in hirachicly files and folders. Noe, datawarehouse allows raw data to be stored in natural formate until it is require.
    - New datawarehousing allows for taster data collection and analysis from organization and department.
    - Use IOT is the major game changer concept in the datawarehousing. It share and store the data across multiple devices and provide streaming data.

Ex. of the future of datawarehousing :-

- Different companies like sap are working for the future of DWH it lound the new concept called SAP BW/
- AWS and other tools also works for future datawarehousing. It combine the historical data and streaming data for better Implementation.
- Hadoop and spark framework provides good programming on data.
  Note:-SAP(system , application & products)
  AWS(amazon web services)

4. Explain: Data warehouse Architecture

- A data warehouse architecture include different aspect of the data warehouse.
- It will include operational data store(ods), different data marts. ETL tools[Extract, Transform & Load] and verity of internal & external dta sources.
- The data warehouse architecture include different layers with are as under:
    1. Data Source Layer
    2. Data Staging Layer
    3. Data Store Layer
    4. Data Presentation Layer

1. **Data Source Layer:-**

- The data source layer of DWH architecture is include original data which is collected from verity of internal and external sources like database.
- Data source include three types of data.
    - Operational data:
      (Product data, Marketing data, HR data, Stock data, etc.)
    - Social data:
      (Website data, Contents, Page, etc.)
    - Third party data:
      (Sarveydata, Other unstructured data like scan imager, Voice _recording, Text, etc.)

2. **Data Staging Layer:-**

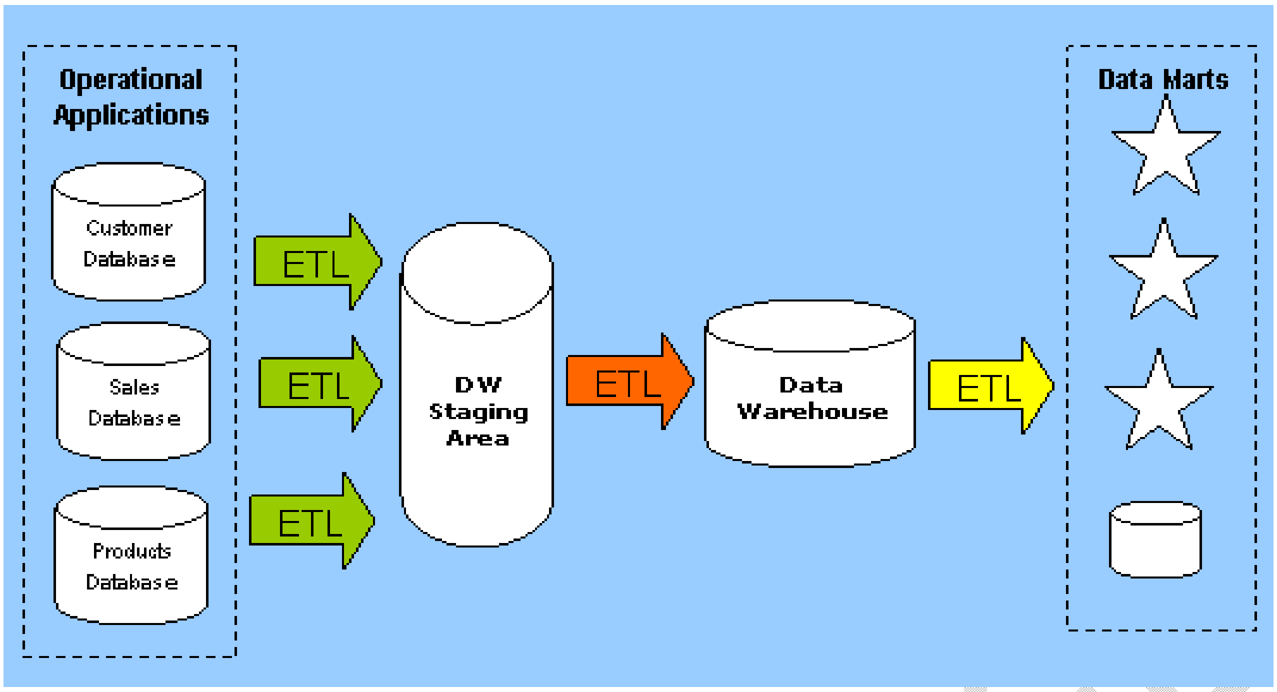
- The data staging layer is between data source and the datawarehouse.
- This layer include ETL(Extract transform & load) tools where the data is extracted from different sources in different formats then after those data will be transform in a proper manner and find result data will be loaded into the data warehouse.
- In the staging layer the data staging concept apply on to the data for quality check then after it will be move into the data warehouse.
- The poor data will generate wrong information and the result is poor business decision making. So, here all types of data will divided into different stages.

3. **Data Storage Layer:-**

- The data storage layer include the data which is changed are passed from the staging area.
- Here, it include different data marts and operational data source for storing the data which relatedto your business.

4. **Data Presentation Layer:-**

- The presentation layer is where the users intract with the cleaned data.
- This layer of the datawarehouse architecture provide users with the ability to query the data for the product and the service.



1. Explain: Data Flow Architecture

- The datawarehouse system has two main architectures:
  1) The datawarehouse architecture (system architecture)
  2) Data flow archotecture
- The system architecture is about physical configuration of the servers, networks, software, and clients.
- The data flow architecture is about how the data stores, arrenged I system, how the data flows from the source to the users etc.
- The data flow architecture include some purpose and some components to perform the movement of data which are as under:
  1. NBS (Normalized Data Store)
  2. ODS (Operational Data Store)
  3. DDS (Dimensional Data Store)
  4. MDB (Multidimensional Database)
- The data flow architecture is define or determine first before system architecture design.
- The data flow architecture is a configuration of data stores within datawarehouses system. By arranging how the data flows from the source systems to the applications used by users.
- Data flow architecture also include how the data flows are controlled, managed and monitor as well as provide the quality of data.

What is Data Store?

- Data store is important component of data flow architecture.it is like a database or files that contain the data, arrange proper formate& perform the process. It include the following:

1. **A Stage:-**

- It is the internal data store used for transformation and preparing the data.
- It will get the data from sources before the data is loaded in the datawarehouse.

2. **NDS(Normalized Data Store):-**

- It is the internal master data store in the form of one or more normalized relational database.
- It is used to integrate the data from different source database.

3. **ODS(Operational Data Store):-**

- It is a hybrid data store in the form of one or more normalized relational database,includetransaction data and latest master data.
- It support the operations in applications.

4. **DDS(Dimentional Data Store):-**

- A dimensional data store is a user side data store in the form of one or more relational database where the data is arranged in dimensional.
- DDS is very much usefull to support analysis queries.
- A dimensional database is a denormalize database that contain dact tables and other dimension tables used for majorment of business events.

5. **MDB(Multidimentional Database):-**

- Multidimensional database is a form of database where the data is stored in cells and the position of each cell is defined by number of variables called dimensions.
- Some applications require the data in multidimensional database format instead of relational database.
- MDB is populated from DDS.
- A data flow architecture show how the data is store from source to destination via possing different phoses like:
  1. Validate the data.
  2. Arrange the data with sorting.
  3. Update the data as per requirements.
  4. Filter the data as per needed.
  5. Generate the reports related to data.

Online Transaction Processing Design

Transaction processing system databases should be designed to promote:

- **Good data placement.**

  The number of users modifying data all over the database that determine the access patterns of the data and place frequently accessed data together. Use file groups and RAID (redundant array of independent disks) systems to assist in this.

- **Short transactions to minimize long-term locks and improve concurrency.**
  Avoid user interaction during transactions. Whenever possible, execute a single stored procedure to process the entire transaction. The order in which you reference tables within your transactions can affect concurrency.
- **Online backup.**
  OLTP systems are often characterized by continuous operations (24 hours a day, 7 days a week) for which downtime is kept to an absolute minimum. Although Microsoft® SQL Server™ 2000 can back up a database while it is being used, schedule the backup process to occur during times of low activity to minimize effects on users.
- **High normalization of the database.**
  Reduce redundant information as much as possible to increase the speed of updates and hence improve concurrency. Reducing data also improves the speed of backups because less data needs to be backed up.
- **Little or no historical or aggregated data.**
  Data that is rarely referenced can be archived into separate databases, or moved out of the heavily updated tables into tables containing only historical data. This keeps tables as small as possible, improving backup times and query performance.
- **Careful use of indexes.**
  Indexes must be updated each time a row is added or modified. To avoid over-indexing heavily updated tables, keep indexes narrow. Use the Index Tuning Wizard to design your indexes.
- Optimum hardware configuration to handle the large numbers of concurrent users and quick response times required by an OLTP system.

Decision Support System

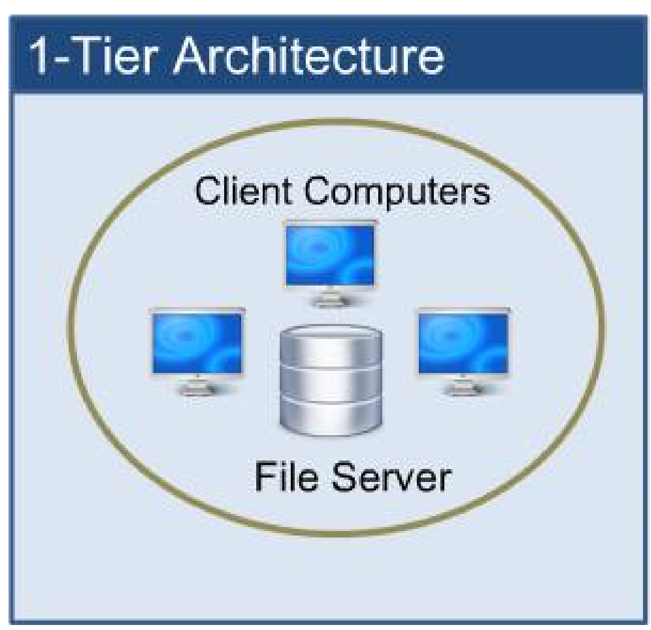
- Decision-support database applications are used for data queries that do not change data. For example, a company can periodically summarize its sales data by date, sales region, or product and store this information in a separate database to be used for analysis by senior management.
- To make business decisions, users need to be able to determine trends in sales quickly by querying the data based on various criteria. However, they do not need to change this data. The tables

in a decision-support database are heavily indexed, and the raw data is often preprocessed and organized to support the various types of queries to be used. Because the users are not changing data, concurrency and atomicity issues are not a concern; the data is changed only by periodic, bulk updates made during off-hour, low-traffic times in the database.

Data warehouse system architecture (Two-Tiered and Three-Tiered)

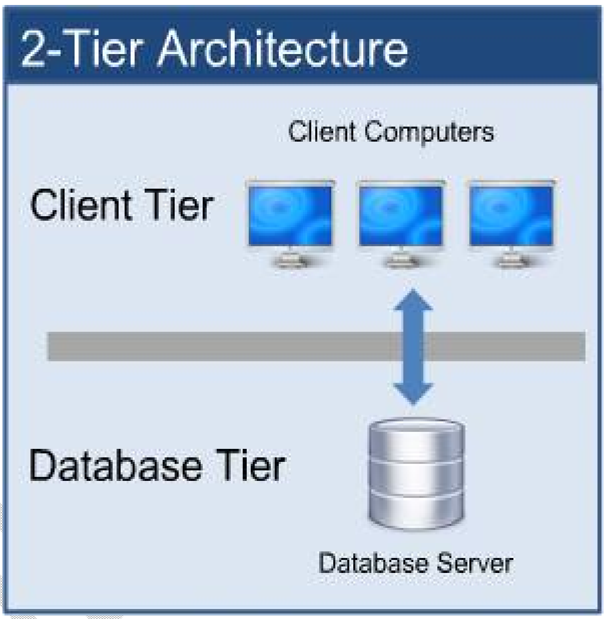
There are three major tiers to the software:

- User Interface (UI). This is what you see when you work with the software. You interact with it. There might be buttons, icons, text boxes, radio buttons, etc. The UI passes on clicks and typed information to the Business Logic tier.
- Business Logic (BL). The business logic is code that is executed to accomplish something. When a user clicks a button it will trigger the BL to run some code. The BL can send information back to the UI, so the user can see the result of clicking a button or typing something in a field. For instance when you enter something in a cell in Excel, the BL will recalculate other cells once you hit Enter and the UI will present the new information to you. The BL also needs to be able to store and retrieve data and that is handled in the Database tier.
- Database (DB). The database is where the data is stored and where the BL can retrieve it again.
- **1-Tier Architecture**
- This architecture has the UI, the BL, and the DB in one single software package. Software applications like MS Access, MS Excel, QuickBooks, and Peachtree all have the same in common: the application handles all three tiers (BL, UI, and DB). The data is stored in a file on the local computer or a shared drive. This is the simplest and cheapest of all the architectures, but also the least secure. Since users have direct access to the files, they could accidentally move, modify, or even worse, delete the file by accident or on purpose. There is also usually an issue when multiple users access the same file at the

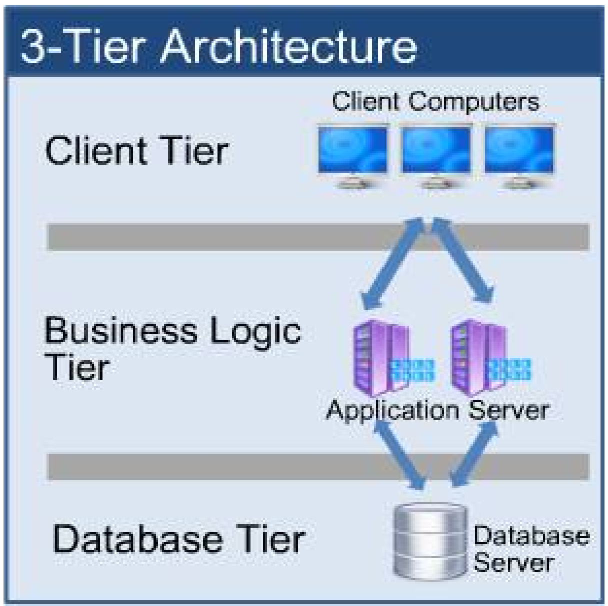


same time: In many cases only one can edit the file while others only have read-only access.

- Another issue is that 1-tier software packages are not very scalable and if the amount to data gets too big, the software may be very slow or stop working.
- So 1-tier architecture is simple and cheap, but usually unsecured and data can easily be lost if you are not careful.
- **2-Tier Architecture**
- This architecture is also called Client-Server architecture because of the two components: The client that runs the application and the server that handles the database back-end. The client handles the UI and the BL and the server handles the DB. When the client starts, it establishes a connection to the server and communicates as needed with the server while running the client. The client computer usually can't see the database directly and can only access the data by starting the client. This means that the data on the server is much more secure. Now users are unable to change or delete data unless they have specific user rights to do so.



- The client-server solution also allows multiple users to access the database at the same time as long as they are accessing data in different parts of the database. One other huge benefit is that the server is processing data (DB) that allows the client to work on the presentation (UI) and business logic (BL) only. This mean that the client and the server are sharing the workload and by scaling the server to be more powerful than the client, you are usually able to load many clients to the server allowing more users to work on the system at the same time and at a much greater speed.
- **3-Tier Architecture**
- In this architecture all three tiers are separated onto different computers. The UI runs on the client (what the user is working with). The BL is running on a separate server, called the business logic tier, middle tier, or service tier. Finally the DB is running on its own database server.



- In the client-server solution the client was handling the UI and the BL that makes the client “thick”. A thick client means that it requires heavy traffic with the server, thus making it difficult to use over slower network connections like Internet and Wireless (4G, LTE, or Wi-Fi).
- By introducing the middle tier, the client is only handling presentation logic (UI). This means that only little communication is needed between the client and the middle tier (BL) making the client “thin” or “thinner”. An example of a thin client is an Internet browser that allows you to see and provide information fast and almost with no delay.
- As more users access the system a three-tier solution is more scalable than the other solution because you can add as many middle tiers (running on each own server) as needed to ensure good performance (N-tier or multiple-tier).
- Security is also the best in the three-tier architecture because the middle tier protects the database tier.
- There is one major drawback to the N-tier architecture and that is that the additional tiers increase the complexity and cost of the installation.

	1-Tier	**2-Tier**	**Multi-Tier**
Benefits	Very simple Inexpensive No server needed	Good security More scalable Faster execution	Exceptional security Fastest execution “Thin” client Very scalable
Issues	Poor security Multi user issues	More costly More complex “Thick” client	Very costly Very complex
Users	Usually 1)	2-100	50-2000 (+)