Unit 1 : Introduction to BI

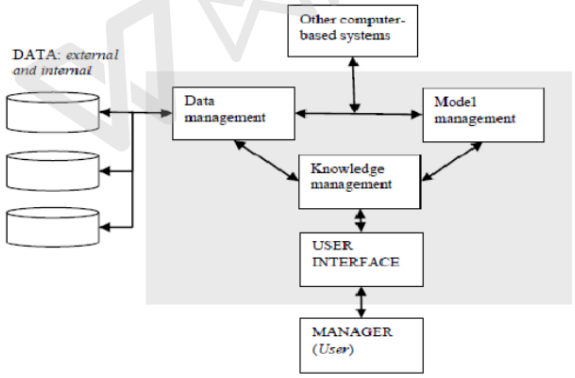
1. What is Decision Support system? (Appeared in May 2016)

- Little (1970) defines DSS as a "model-based set of procedures for processing data and judgments to assist a manager in his decision-making."
- Decision support systems (DSS) are interactive software-based systems intended to help managers in decision-making by accessing large volumes of information generated from various related information systems involved in organizational business processes, such as office automation system, transaction processing system, etc.
- A decision support system (DSS) is a computer program application that analyzes business data and presents it so that users can make business decisions more easily.
- Typical information that a decision support application might gather and present would be:
  - Comparative sales figures between one week and the next
  - Projected revenue figures based on new product sales assumptions
- DSS uses the summary information, exceptions, patterns, and trends using the analytical models. A decision support system helps in decision-making but does not necessarily give a decision itself. The decision makers compile useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions.
- **Types of DSS:**

1. Communication Driven DSS:
   Its purpose are to help conduct a meeting, or for users to collaborate. The most common technology used to deploy the DSS is a web or client server. Examples: chats and instant messaging software's, online collaboration and net-meeting systems.
2. Data Driven DSS:
   Examples: computer-based databases that have a query system to check (including the incorporation of data to add value to existing databases.

3. Document Driven DSS:
The purpose of such a DSS is to search web pages and find documents on a specific set of keywords or search terms. The usual technology used to set up such DSS are via the web or a client/server system.
4. Knowledge Driven DSS:
It is essentially used to provide management advice or to choose products/services.
It works exactly like Data mining. Tools used for building Knowledge-Driven DSS are sometimes called Intelligent Decision Support methods.
5. Model Driven DSS:
Model-driven DSS are complex systems that help analyse decisions or choose between different options. These are used by managers and staff members of a business, or people who interact with the organization, for a number of purposes depending on how the model is set up - scheduling, decision analyses etc.

Components of DSS:



1. Data Management Subsystem:
The data management subsystem includes a database that extracts relevant data for the situation and is managed by software called the database management system (DBMS).
The data management subsystem can be interconnected with the corporate data warehouse, a repository for corporate relevant decision-making data. Usually the data are stored or accessed via a database Web server.

2. Model Management Subsystem:
This is a software package that includes financial, statistical, management science, or other quantitative models that provide the system's analytical capabilities and appropriate software management. Modeling languages for building custom models are also included. This software is often called a model base management system (MBMS). This component can be connected to corporate or external storage of models. Model solution methods and management systems are implemented in Web development systems (like Java) to run on application servers.

3. User Interface Subsystem:
The user communicates with and commands the DSS through this subsystem. The user is considered part of the system. Researchers assert that some of the unique contributions of DSS are derived from the intensive interaction between the computer and the decision-maker. The Web browser provides a familiar, consistent graphical user interface structure for most DSS.

4. Knowledge based management subsystem:
It provides intelligence to augment the decision-maker's own. It can be interconnected with the organization's knowledge repository (part of a knowledge management system), which is sometimes called the organizational knowledge base.

2. What is Data - Information - Knowledge - Decision making - Action cycle.

A. Data:

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- non operational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

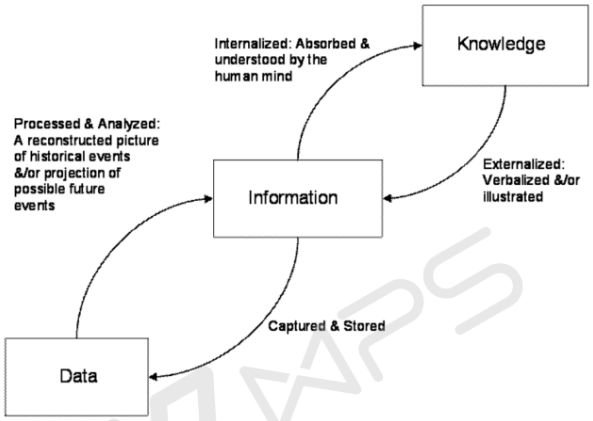
B. Information:

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

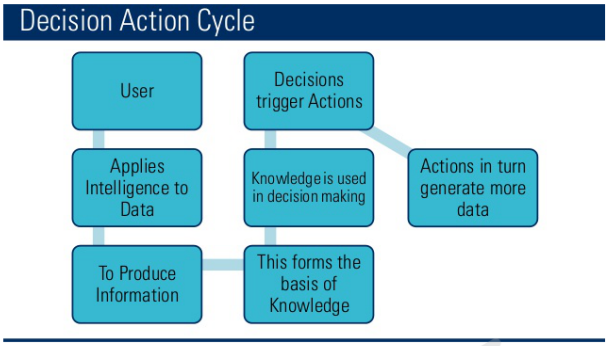
C. Knowledge:

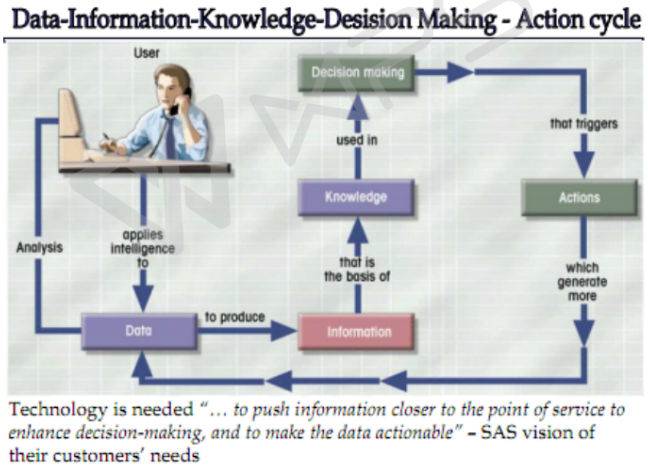
Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Relationship between Data, Information and Knowledge:



(Explain the diagram in your words, briefly in 4-5 lines)





Technology is needed *"… to push information closer to the point of service to enhance decision-making, and to make the data actionable"* – SAS vision of their customers' needs

3. What is Business Intelligence. State importance and need.

The term Business Intelligence (BI) refers to technologies, applications and practices for the collection, integration, analysis, and presentation of business information. The purpose of Business Intelligence is to support better business decision making.

Business intelligence (BI) is a collection of technical and process innovations across the data warehousing and business intelligence space. Proactive BI focuses on decision-making acceleration by leveraging existing BI infrastructure to identify, calculate, and distribute up-tothe-moment, mission-critical information. Through the application of these techniques and technologies, the reach and value of data warehouse and BI systems can be increased by one or more orders of magnitude. Business success today requires intelligent data use.

Importance:

A. Gain Insights into Consumer Behaviour:

One of the main advantages of investing in BI and skilled personnel is the fact that it will boost your ability to analyze the current consumer buying trends. Once you understand what your consumers are buying, you can use this information to develop products that match the current consumption trends and consequently improve your profitability.

B. To Improve Visibility:

If you want to improve your control over various important processes in your organization, you should consider investing in a good BI system. This will improve the visibility of these processes and make it possible to identify any areas that need improvement. Moreover, if you currently have to skim through hundreds of pages in your detailed periodic reports to assess the performance of your organisation's processes, you can save time and improve productivity by having skilled intelligence analysts using relevant BI software. Good BI should be at the heart of every organization as it can provide increased control. Visibility is essential.

C. To Turn Data into Actionable Information:

A BI system is an analytical tool that can give you the insight you need to make successful strategic plans for your organization. This is because such a system would be able to identify key trends and patterns in your organisation ' s data and consequently make it easier for you to make important connections between different areas of your business that may otherwise seem unrelated. As such, a BI system can help you understand the implications of various organizational processes better and enhance your ability to identify suitable opportunities for your organization, thus enabling you to plan for a successful future.

D. To Improve Efficiency:

One of the most important reasons why you need to invest in an effective BI system is because such a system can improve efficiency within your organization and, as a result, increase productivity. You can use business intelligence to share information across different departments in your organization. This will enable you to save time on reporting processes and analytics. This ease in information sharing is likely to reduce duplication of roles/duties within the organization and improve the accuracy and usefulness of the data generated by different departments. Furthermore, information sharing also saves time and improves productivity.

E. To Gain Sales & Market Intelligence:

Whether you are a sales person or a marketer, you probably like to keep track of your customers — probably using a CRM to help you. CRM stands for Customer Relationship Management. It refers to software that handles all aspects of an organization's interactions with its customers. In other words, it collects the data about your customer and tries to make sense of it, presents it to you in various tables and charts. That may include the entire sales cycle, from winning new customers, to servicing and tracking existing customers, to providing post-sales services. CRM systems are now more involved in decision-support processes than ever before. In our next articles, we will look at how BI can help increase

your sales efficiently and gain a further insight into your current market as well as market entry support.

F. To Gain Competitive Intelligence:

BI can also be used to gain an insight into what your competitors are doing. This strengthens your company's ability to make decisions and plan for the future.

4. Data Warehousing definition and characteristics:

A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources. A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels. Data is populated into the DW through the processes of extraction, transformation and loading.

A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or adhoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Characteristics:

A. Subject-oriented.

Data are organized by detailed subject (e.g., by customer, policy type, and claim in an insurance company), containing only information relevant for decision support. Subject orientation enables users to determine not only how their business is performing, but why. A data warehouse differs from an operational database in that most operational databases have a product orientation and are tuned to handle transactions that update the database; subject orientation provides a more comprehensive view of the organization.

B. Integrated.

Data at different source locations may be encoded differently. For example, gender data may be encoded as 0 and 1 in one place and "m" and "f" in another. In the warehouse they are scrubbed (cleaned) into one format so that they are standardized and consistent. Many organizations use the same terms for data of different kinds. For example, "net sales" may mean net of commission to the marketing department but gross sales returns to the accounting department. Integrated data resolve inconsistent meanings and provide

uniform terminology throughout the organization. Also, data and time formats vary around the world.

C. Time-variant (time series).

The data do not provide the current status. They -are-kept for five or ten years or more and are used for trends, forecasting, and comparisons. There is a temporal quality to a data warehouse. Time is the one important dimension that all data warehouses must support. Data for analysis from multiple sources contain multiple time points (e.g., daily, weekly, monthly views).

D. Nonvolatile.

Once entered into the warehouse, data are read-only, they cannot be changed or updated. Obsolete data are discarded, and changes are recorded as new data. This enables the data warehouse to be tuned almost exclusively for data access. For example, large amounts of free space (for data growth) typically are not needed, and database reorganizations can be scheduled in conjunction with the load operations of a data warehouse.

E. Summarized

Operational data are aggregated, when needed, into summaries.

F. Not normalized

Data in a data warehouse are generally not normalized and highly redundant.

G. Sources.

All data are present; both internal and external.

H. Metadata.

Metadata (defined as data about data) are included.

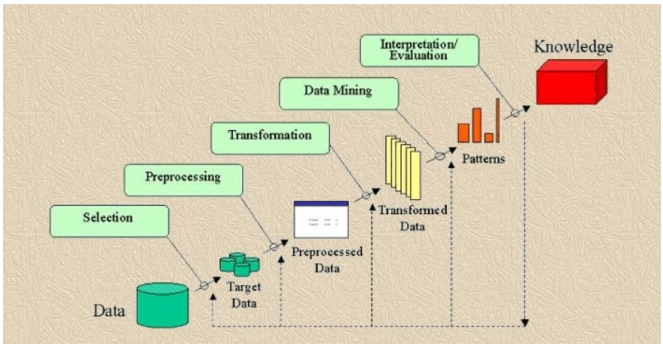
Unit 1: Introduction to BI part II

5. Knowledge discovery in databases(KDD).

The term *Knowledge Discovery in Databases*, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database.



The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

Before starting the process, an unsaid step is Developing an understanding of

- the application domain
- the relevant prior knowledge
- the goals of the end-user

Step 1- Selection

Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

Step 2- Preprocessing

Data cleaning and preprocessing.

- Removal of noise or outliers.
- Collecting necessary information to model or account for noise.
- Strategies for handling missing data fields.
- Accounting for time sequence information and known changes.

Step 3- Transformation

Data reduction and projection.

- Finding useful features to represent the data depending on the goal of the task.
- Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

Step 4- Data Mining

- Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
- Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

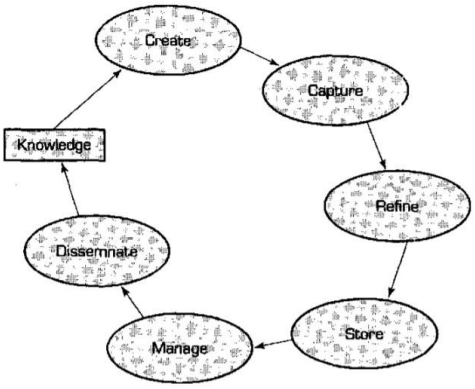
Step 5- Interpretation and Evaluation

Out of the patterns found from above step, various findings are noted down. And patterns are mined for final results that is Knowledge. All the knowledge is then consolidated.

6. Knowledge Management Cycle.

A functioning knowledge management system follows six steps in a cycle (Figure 9.2). The reason for the cycle is that knowledge is dynamically refined over time. The knowledge in a good KM system is never finished because the environment changes over time, and the knowledge must be updated to reflect the changes. The cycle works as follows:

1. **Create knowledge:** Knowledge is created as people determine new ways of doing things or develop know-how. Sometimes external knowledge is brought in. Some of these new ways may become best practices.



2. **Capture knowledge:** New knowledge must be identified as valuable and be represented in a reasonable way.
3. **Refine knowledge:** New knowledge must be placed in context so that it is actionable. This is where human insights (tacit qualities) must be captured along with explicit facts.
4. **Store knowledge:** Useful knowledge must then be stored in a reasonable format in a knowledge repository so that others in the organization can access it.
5. **Manage knowledge:** Like a library, the knowledge must be kept current. It must be reviewed to verify that it is relevant and accurate.
6. **Disseminate knowledge:** Knowledge must be made available in a useful format to anyone in the organization who needs it, anywhere and anytime.

As knowledge is disseminated, individuals develop, create, and identify new knowledge or update old knowledge which they replenish into the system. Knowledge is a resource that is ncft consumed when used, though it can age. (For example, driving a car in 1900 was different from driving one now, but many of the basic principles still apply.) Knowledge must be updated. Thus, the amount of knowledge grows over time.

Normalization with example.

Normalization is a database design technique which organizes tables in a manner that reduces redundancy and dependency of data.

It divides larger tables to smaller tables and links them using relationships.

1st NF:

- All values in the cell should be atomic that is single valued.

- The table must have a primary key.

2nd NF:

- It should be in 1st NF.

- All non key attributes should be dependent on Key attributes.

3rd NF:

- It should be in 2nd NF.

- It has no transitive functional dependencies.

Check the normalization example in excel sheet.

BI Unit 1- Part 3
Topic: Business Intelligence Architecture

Presented by: Drashti Shrimal

BI Architecture

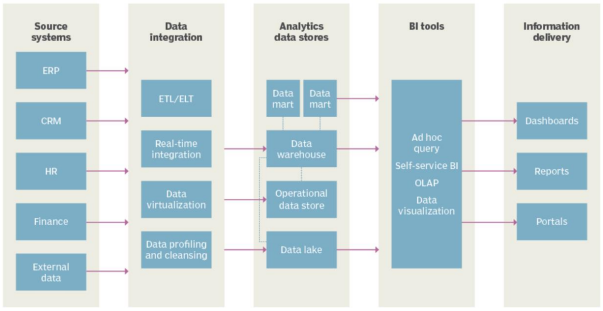
Topics :

BI architecture

Data preprocessing: Cleaning - Missing values, inconsistent values, noisy data.

Data preprocessing: Transformation

Data preprocessing: Reduction



BI Architecture

- Source Systems :

Systems from where the data is collected for analysis.

- Data Integration :

Merging the data collected.

- Analytic stores :

Creating various sub parts of whole data like Data Marts, Data Views.

- BI Tools :

Mining tools which perform analysis on the analytics stores.

- Information Delivery :

Various tools to display reports of analysis.

Data Pre-processing - Cleaning

- <u>Missing values :</u>

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

a. Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

b. Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

Data Pre-processing - Cleaning

- Noisy values :

Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task.

Data Pre-processing - Cleaning

- Inconsistent values :

Examples - Male/Female in one set and M/F in other.

- Redundant values :

Copies of data all over the set.

- Outliers :

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

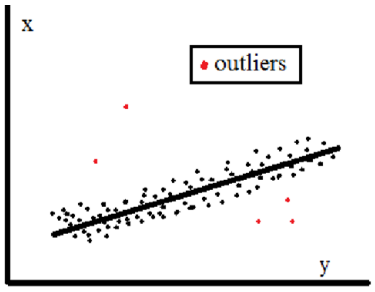
Data Pre-processing - Transformation

- This step is taken in order to transform the data in appropriate forms suitable for mining process.
- This involves following ways:

Normalization

Attribute Selection

Outliers



Data Pre-processing -Reduction

- Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

Data Binning

- Data binning, bucketing is a data pre-processing method used to minimize the effects of small observation errors.
- The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin.
- This has a smoothing effect on the input data and may also reduce the chances of overfitting in case of small datasets

Use the following data for age :
N = 13, 12, 16, 19, 20, 14, 9, 10, 21, 24, 22, 23, 27, 25, 28
Use depth = 3

Solution:

Step 1 : Sort your data
N = 9, 10, 12, 13, 14, 16, 19, 20, 21, 22, 23, 24, 25, 27, 28

Step 2 : Partition your data in different bins of equal size using the formula (N/Depth)
No of bins = N/D i.e 15/3 = 5 Bins

Step 3 : Partition Bins
Bin 1 = 9,10,12
Bin 2 = 13,14,16
Bin 3 = 19,20,21
Bin 4 = 22,23,24
Bin 5 = 25,27,28

Step 4 : Smooth each Bin
Bin 1 = (9 + 10 + 12)/3 = 31/3
Bin 2 = (13 + 14 + 16)/3 = 43/3
Bin 3 = (19 + 20 + 21)/3 = 60/3 = 20
Bin 4 = (22 + 23 + 24)/3 = 69/3 =23
Bin 5 = (25 + 27 + 28)/3 = 80/3

Step 5 : Replace new values in Bins
Bin 1 = 31/3 , 31/3, 31/3 (10.33)
Bin 2 = 43/3, 43/3, 43/3 (14.33)
Bin 3 = 20,20,20
Bin 4 = 23,23,23
Bin 5 = 80/3, 80/3, 80/3 (26.66)

Step 6 : Final Output
Smoothed Data = { 10.33, 10.33, 10.33, 14.33, 14.33, 14.33, 20, 20, 20, 23, 23, 23, 26.66, 26.66, 26.66}

Use the following data for age :
N = 15,11,10,20,21,26,19,18,14,22,12
Use depth according to you.

Solution:

Step 1 : Sort your data
N = 10,11,12,14,15,18,19,20,21,22,26

Step 2 : Partition your data in different bins of equal size using the formula (N/Depth)
Total number of bins = 11
To use Depth = 3, we need to add 1 element in the set of numbers to make it divisible by 12.
Thus No of bins = Elements / depth
i.e No of bins = 12/3 = 4

Step 3 : Partition Bins
Bin 1 = 10,11,12
Bin 2 = 14,15,18
Bin 3 = 19,20,21
Bin 4 = 22,26,x
x can be next element of 26 i.e 27

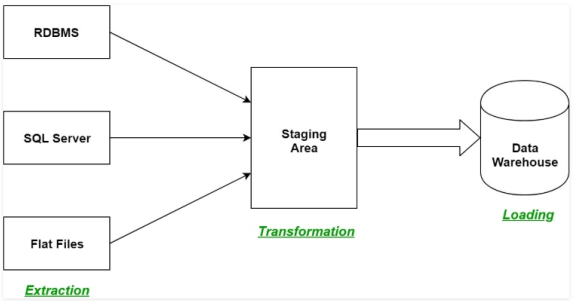
Step 4 : Smooth each Bin = Sum of Elements /Depth
Bin 1 = (10+11+12) /3
Bin 2 = (14+15+18) /3
Bin 3 = (19+20+21) /3
Bin 4 = (22+26+27) /3

Step 5 : Replace new values in Bins
Bin 1 = (10+11+12) /3 = 11
Bin 2 = (14+15+18) /3 = 15.66
Bin 3 = (19+20+21) /3 = 20
Bin 4 = (22+26+27) /3 = 25

Step 6 : Final Output :
Smoothed Data = {11,11,11,15.66,15.66,15.66,20,20,20,25,25,25}

BI - UNIT II

1. Explain ETL process.



1. **Extraction:**

The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML and flat files into the staging area. It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also. Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult. Therefore, this is one of the most important steps of ETL process.

2. **Transformation:**

The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes/tasks:

- Filtering – loading only certain attributes into the data warehouse.
- Cleaning – filling up the NULL values with some default values, mapping U.S.A, United States and America into USA, etc.

- Joining – joining multiple attributes into one.
- Splitting – splitting a single attribute into multiple attributes.
- Sorting – sorting tuples on the basis of some attribute (generally key-attribute).

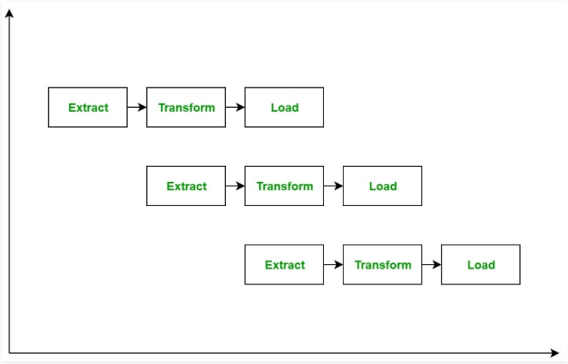
3. **Loading:**

The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals. The rate and period of loading solely depends on the requirements and varies from system to system.

ETL process can also use the pipelining concept i.e. as soon as some data is extracted, it can transformed and during that period some new data can be extracted. And while the transformed data is being loaded into the data warehouse, the already extracted data can be transformed.

ETL Tools: Most commonly used ETL tools are Sybase, Oracle Warehouse builder, CloverETL and MarkLogic.

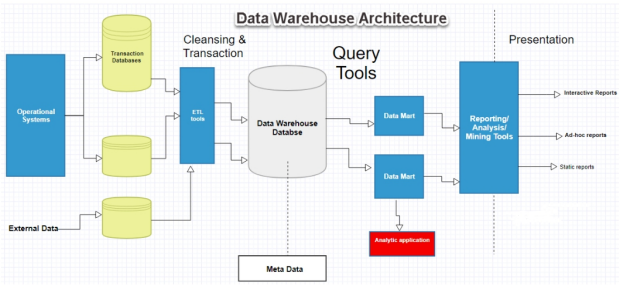
The block diagram of the pipe-lining of ETL process is shown below:



2. Explain the logical architecture of DWH.

Data warehouse is an information system that contains historical and commutative data from single or multiple sources. It simplifies reporting and analysis process of the organization.

It is also a single version of truth for any company for decision making and forecasting.



There are 5 main components :

1. Data Warehouse Database

The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology.

2. Sourcing, Acquisition, Clean-up and Transformation Tools (ETL)

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 data warehouse. They are also called Extract, Transform and Load (ETL) Tools.

3. Metadata

The name Meta Data 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.

4. Query Tools

One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions. Query tools allow users to interact with the data warehouse system.

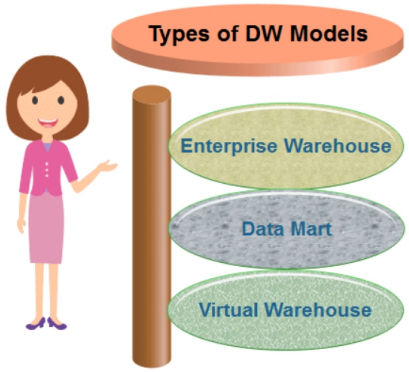
These tools fall into four different categories:

- Query and reporting tools
- Application Development tools
- Data mining tools
- OLAP tools

5. Data warehouse Bus Architecture

Data warehouse Bus determines the flow of data in your warehouse. The data flow in a data warehouse can be categorized as Inflow, Upflow, Downflow, Outflow and Meta flow.

3. DW models.



1. Enterprise Warehouse

An Enterprise warehouse collects all of the records about subjects spanning the entire organization. It supports corporate-wide data integration, usually from one or more operational systems or external data providers, and it's cross-functional in scope. It generally contains detailed information as well as summarized information and can range in estimate from a few gigabyte to hundreds of gigabytes, terabytes, or beyond.

An enterprise data warehouse may be accomplished on traditional mainframes, UNIX super servers, or parallel architecture platforms. It required extensive business modeling and may take years to develop and build.

2. Data Mart

A data mart includes a subset of corporate-wide data that is of value to a specific collection of users. The scope is confined to particular selected subjects. For example, a marketing data mart may restrict its subjects to the customer, items, and sales. The data contained in the data marts tend to be summarized.

Data Marts is divided into two parts:

Independent Data Mart: Independent data mart is sourced from data captured from one or more operational systems or external data providers, or data generally locally within a different department or geographic area.

Dependent Data Mart: Dependent data marts are sourced exactly from enterprise data-warehouses.

3. Virtual Warehouses

Virtual Data Warehouses is a set of perception over the operational database. For effective query processing, only some of the possible summary vision may be materialized. A virtual warehouse is simple to build but required excess capacity on operational database servers.

BI Unit 2- Part 2
Topic: Business Intelligence Architecture

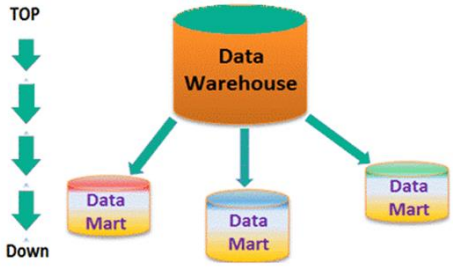
Topics :
- Differentiate between DWH and Data Marts
- ELT
- Differentiate between ELT and ETL
- Advantages of ELT
- Data Lakes

Presented by: Drashti Shrimal

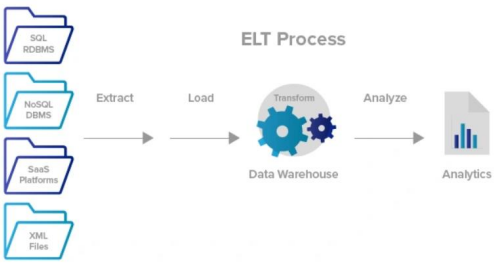
Differentiate between DWH and Data Marts

Data Mart	Data Warehouse
• Often holds only one subject area- for example, Finance, or Sales	• Holds multiple subject areas
• May hold more summarized data (although many hold full detail)	• Holds very detailed information
• Concentrates on integrating information from a given subject area or set of source systems	• Works to integrate all data sources
• Is built focused on a dimensional model using a star schema.	• Does not necessarily use a dimensional model but feeds dimensional models

Differentiate between DWH and Data Marts



ELT



ELT

- ELT stands for "Extract, Load, and Transform." In this process, data gets leveraged via a data warehouse in order to do basic transformations.
- That means there's no need for data staging.
- ELT uses cloud-based data warehousing solutions for all different types of data - including structured, unstructured, semi-structured, and even raw data types.

ELT

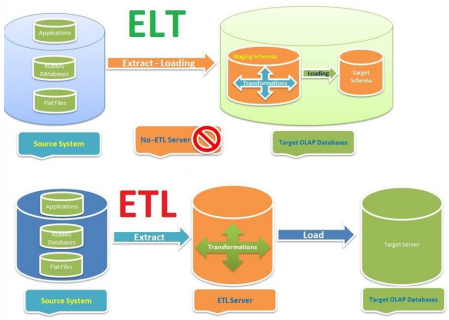
- The primary advantage of ELT over ETL relates to flexibility and ease of storing new, unstructured data.
- Other benefits:

#1: High Speed

#2: Low-Maintenance

3#: Quicker Loading

ELT and ETL



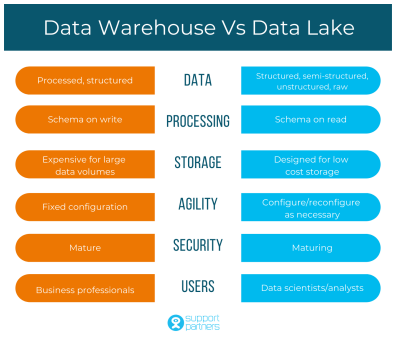
ELT and ETL

- ETL stands for Extract, Transform, and Load, while ELT stands for Extract, Load, and Transform.
- In ETL, data flow from the data source to staging to the data destination.
- ELT lets the data destination do the transformation, eliminating the need for data staging.
- ETL can help with data privacy and compliance, cleansing sensitive data before loading into the data destination, while ELT is simpler and for companies with minor data needs.

Data Lake

- A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale.
- Examples. Many companies use cloud storage services such as Google Cloud Storage.

DWH and Data Lake



DWH and Data Lake