UNIT-2

INTRODUCTION:

Data has been the buzzword for ages now. Either the data being generated from large-scale enterprises or the data generated from an individual, each and every aspect of data needs to be analysed to benefit yourself from it.

Why is Data Analytics important?

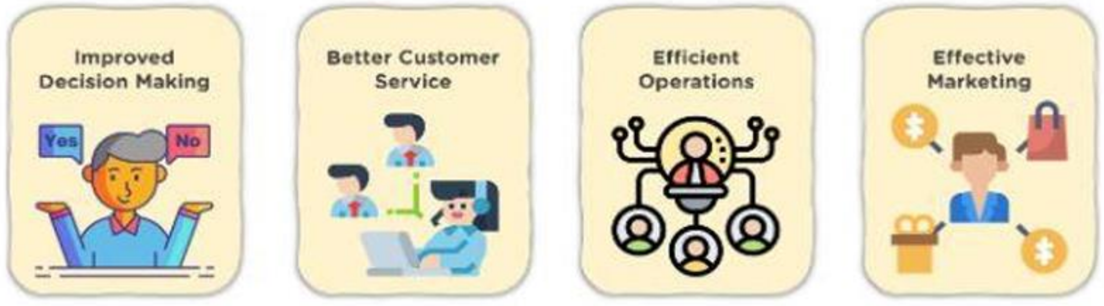
Data Analytics has a key role in improving your business as it is used to gather hidden insights, generate reports, perform market analysis, and improve business requirements.

What is the role of Data Analytics?

- **Gather Hidden Insights** – Hidden insights from data are gathered and then analysed with respect to business requirements.
- **Generate Reports** – Reports are generated from the data and are passed on to the respective teams and individuals to deal with further actions for a high rise in business.
- **Perform Market Analysis** – Market Analysis can be performed to understand the strengths and weaknesses of competitors.
- **Improve Business Requirement** – Analysis of Data allows improving Business to customer requirements and experience.

Ways to Use Data Analytics

Now that you have looked at what data analytics is, let’s understand how we can use data analytics.



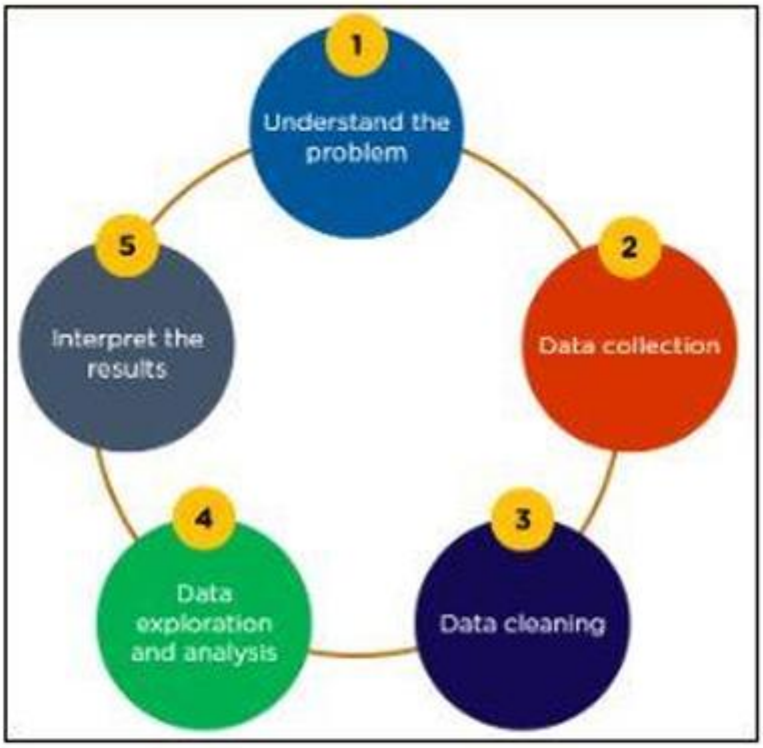
1. **Improved Decision Making**: Data Analytics eliminates guesswork and manual tasks. Be it choosing the right content, planning marketing campaigns, or developing products. Organizations can use the insights they gain from data analytics to make informed decisions. Thus, leading to better outcomes and customer satisfaction.

2. **Better Customer Service**: Data analytics allows you to tailor customer service according to their needs. It also provides personalization and builds stronger relationships with customers. Analysed data can reveal information about customers' interests, concerns, and more. It helps you give better recommendations for products and services.

3. **Efficient Operations:** With the help of data analytics, you can streamline your processes, save money, and boost production. With an improved understanding of what your audience wants, you spend lesser time creating ads and content that aren't in line with your audience's interests.

4. **Effective Marketing:** Data analytics gives you valuable insights into how your campaigns are performing. This helps in fine-tuning them for optimal outcomes. Additionally, you can also find potential customers who are most likely to interact with a campaign and convert into leads.

Steps Involved in Data Analytics:



Next step to understanding what data analytics is to learn how data is analyzed in organizations. There are a few steps that are involved in the data analytics lifecycle. Below are the steps that you can take to solve your problems.

Data Analytics process steps

1. **Understand the problem**: Understanding the business problems, defining the organizational goals, and planning a lucrative solution is the first step in the analytics process. E-commerce companies often encounter issues such as predicting the return of items, giving relevant product recommendations, cancellation of orders, identifying frauds, optimizing vehicle routing, etc.

2**. Data Collection**: Next, you need to collect transactional business data and customer-related information from the past few years to address the problems your business is facing. The data can have information about the total units that were sold for a product, the sales, and profit that were made, and also when was the order placed. Past data plays a crucial role in shaping the future of a business.

3. **Data Cleaning**: Now, all the data you collect will often be disorderly, messy, and contain unwanted missing values. Such data is not suitable or relevant for performing data analysis. Hence, you need to clean the data to remove unwanted, redundant, and missing values to make it ready for analysis.

4. **Data Exploration and Analysis**: After you gather the right data, the next vital step is to execute exploratory data analysis. You can use data visualization and business intelligence tools, data mining techniques, and predictive modelling to analyze, visualize, and predict future outcomes from this data. Applying these methods can tell you the impact and relationship of a certain feature as compared to other variables.

Below are the results you can get from the analysis

- You can identify when a customer purchases the next product.
- You can understand how long it took to deliver the product.
- You get a better insight into the kind of items a customer looks for, product returns, etc.
- You will be able to predict the sales and profit for the next quarter.
- You can minimize order cancellation by dispatching only relevant products.
- You'll be able to figure out the shortest route to deliver the product, etc.

5. **Interpret the results**: The final step is to interpret the results and validate if the outcomes meet your expectations. You can find out hidden patterns and future trends. This will help you gain insights that will support you with appropriate data-driven decision making.

The tools used in Data Analytics

With the increasing demand for Data Analytics in the market, many tools have emerged with various functionalities for this purpose. Either open-source or user-friendly, the top tools in the data analytics market are as follows.

R programming – This tool is the leading analytics tool used for statistics and data modelling. R compiles and runs on various platforms such as UNIX, Windows, and Mac OS. It also provides tools to automatically install all packages as per user-requirement.

Python – Python is an open-source, object-oriented programming language that is easy to read, write, and maintain. It provides various machine learning and visualization libraries such as Scikit-learn, TensorFlow, Matplotlib, Pandas, Kera's, etc. It also can be assembled on any platform like SQL server, a MongoDB database or JSON.

Tableau Public – This is a free software that connects to any data source such as Excel, corporate Data Warehouse, etc. It then creates visualizations, maps, dashboards etc .with real-time updates on the web.

QlikView – This tool offers in-memory data processing with the results delivered to the end-users quickly. It also offers data association and data visualization with data being compressed to almost 10% of its original size.

SAS – A programming language and environment for data manipulation and analytics, this tool is easily accessible and can analyse data from different sources.

Microsoft Excel – This tool is one of the most widely used tools for data analytics. Mostly used for clients' internal data, this tool analyses the tasks that summarize the data with a preview of pivot tables.

RapidMiner – A powerful, integrated platform that can integrate with any data source types such as Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase etc. This tool is mostly used for predictive analytics, such as data mining, text analytics, machine learning.

KNIME – Konstanz Information Miner (KNIME) is an open-source data analytics platform, which allows you to analyse and model data. With the benefit of visual programming, KNIME provides a platform for reporting and integration through its modular data pipeline concept.

Open Refine – Also known as Google Refine, this data cleaning software will help you clean up data for analysis. It is used for cleaning messy data, the transformation of data and parsing data from websites.

Apache Spark – One of the largest large-scale data processing engines, this tool executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk. This tool is also popular for data pipelines and machine learning model development.

Data Analytics Applications

Data analytics is used in almost every sector of business, let's discuss a few of them:

1. **Retail:** Data analytics helps retailers understand their customer needs and buying habits to predict trends, recommend new products, and boost their business. They optimize the supply chain, and retail operations at every step of the customer journey.
2. **Healthcare**: Healthcare industries analyse patient data to provide lifesaving diagnoses and treatment options. Data analytics help in discovering new drug development methods as well.
3. **Manufacturing**: Using data analytics, manufacturing sectors can discover new cost-saving opportunities. They can solve complex supply chain issues, labour constraints, and equipment breakdowns.
4. **Banking sector:** Banking and financial institutions use analytics to find out probable loan defaulters and customer churn out rate. It also helps in detecting fraudulent transactions immediately.
5. **Logistics**: Logistics companies use data analytics to develop new business models and optimize routes. This, in turn, ensures that the delivery reaches on time in a cost-efficient manner.

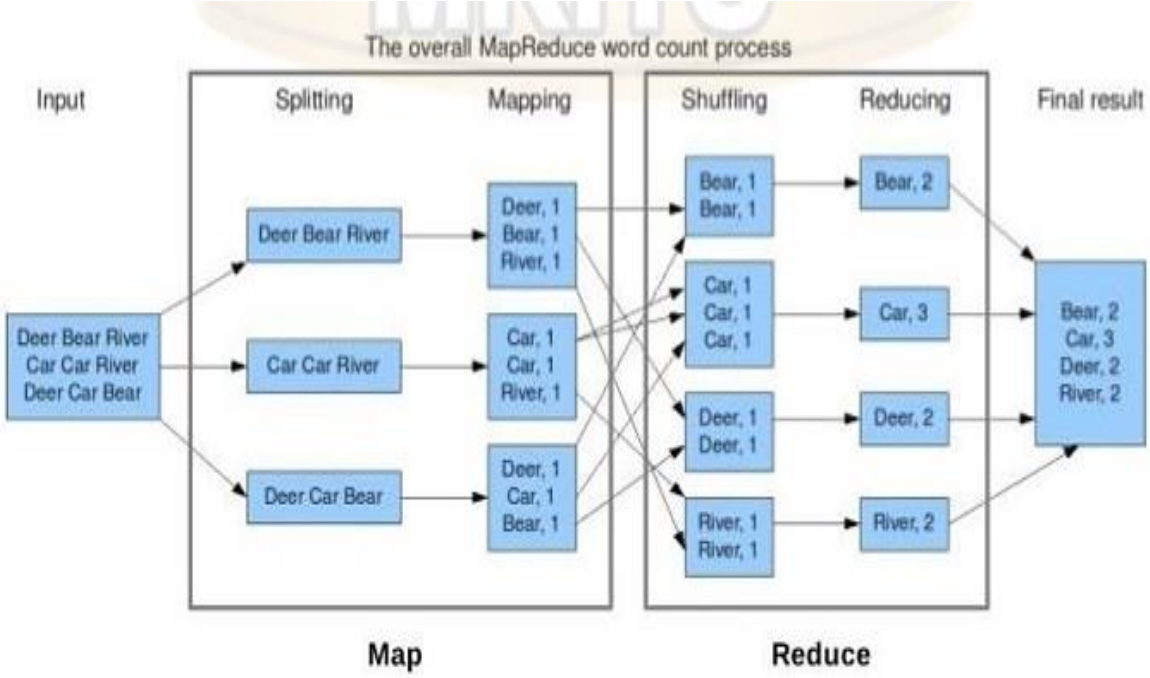
Need for business Modelling

Using big data as fundamental factor of making decision which need new capability, most firms are far away from accessing all data resources. Companies in various sectors have acquired crucial insight from the structured data collected from different enterprise systems and anatomize by commercial database management systems.

1.) Facebook and Twitter to standard the instantaneous influence on campaign and to examine consumer opinion about their products

2.) Some companies, like Amazon, eBay, and Google, considered as early commandants, examining factors that control performance to define what raise sales revenue and user interactivity.

2.3.1 Utilizing Hadoop in Big Data Analytics.



Hadoop is an open-source software platform that enables processing of large data sets in distributed computing environment", it discusses some concepts according to big data, the rules for building, organizing and analysing huge data-sets in the business environment, they offered 3 architecture layers and also, they indicate some graphical tools to explore and represent unstructured-data, the authors specified how the famous companies could improve their business. E.g.: Google, Twitter and Facebook show their attention in processing big data within cloud-environment

The Map() step: Each worker node applies the Map() function to the local data and writes the output to a temporary storage space. The Map() code is run exactly once for each K1 key value, generating output that is organized by key values K2. A master node arranges it so that for redundant copies of input data only one is processed.

The Shuffle()step: The map output is sent to the reduce processors, which assign the K2 key value that each processor should work on, and provide that processor with all of the map generated data associated with that key value, such that all data belonging to one key are located on the same worker node.

The Reduce() step: Worker nodes process each group of output data(per key) in parallel, executing the user provided Reduce() code; each function is run exactly once for each K2 key value produced by the map step. Produce the final output: The MapReduce system collects all of the reduce outputs and sorts them by K2 to produce the final out-come.

Fig.2.4 shows the classical "word count problem" using the MapReduce paradigm. As shown in Fig.2.4, initially a process will split the data into a subset of chunks that will later be processed by the mappers. Once the key/values are generated by mappers, a shuffling process Is used to mix(combine) these key values (combining the same keys in the same worker node). Finally, the reduce functions are used to count the words that generate a common output as a result of the algorithm. As a result of the execution or wrappers/reducers, the output will generate a sorted list of word counts from the original text input.

2.3.2 The Employment of Big Data Analytics on IBM.

IBM and Microsoft are prominent representatives. IBM represented many big data options that enable users to storing, managing, and analysing data through various resources; it has a good rendering on business-intelligence also healthcare areas. Compared with IBM, also Microsoft showed powerful work in the area of cloud computing activities and techniques another example is Facebook and Twitter, who are collecting various data from user's profiles and using it to increase their revenue

2.3.3 The Performance of Data Driven Companies.

Big data analytics and Business intelligence are united fields which became widely significant in the business and academic area, companies are permanently trying to make insight from the extending the three V's (variety, volume and velocity) to support decision making.

Databases & Types of Data and variables

Data Base: A Database is a collection of related data.

Database Management System: DBMS is a software or set of Programs used to define, construct and manipulate the data.

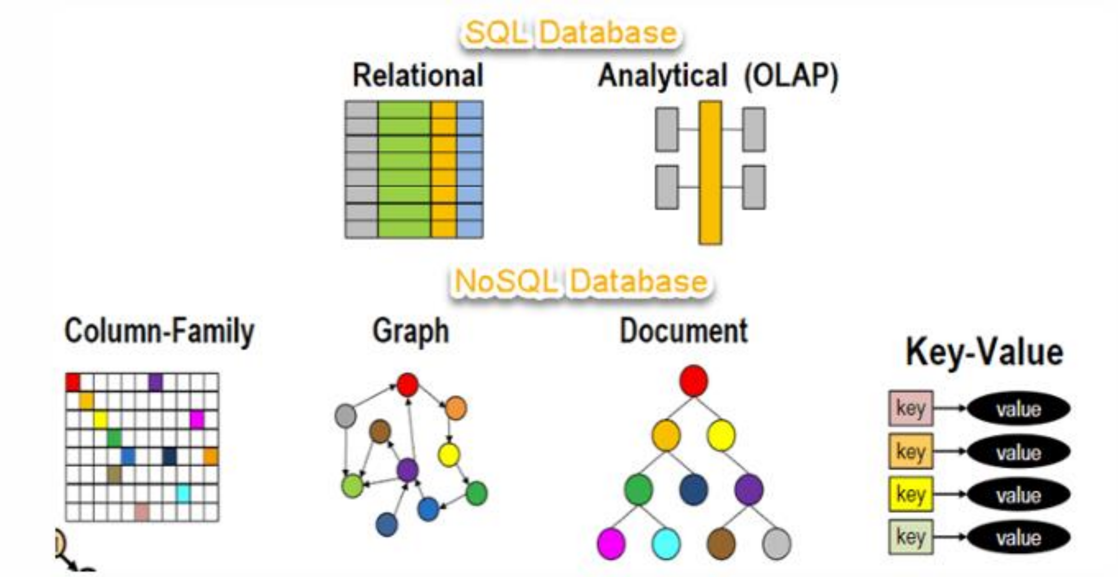
Relational Database Management System: RDBMS is a software system used to maintain relational databases. Many relational database systems have an option of using the SQL.

NoSQL

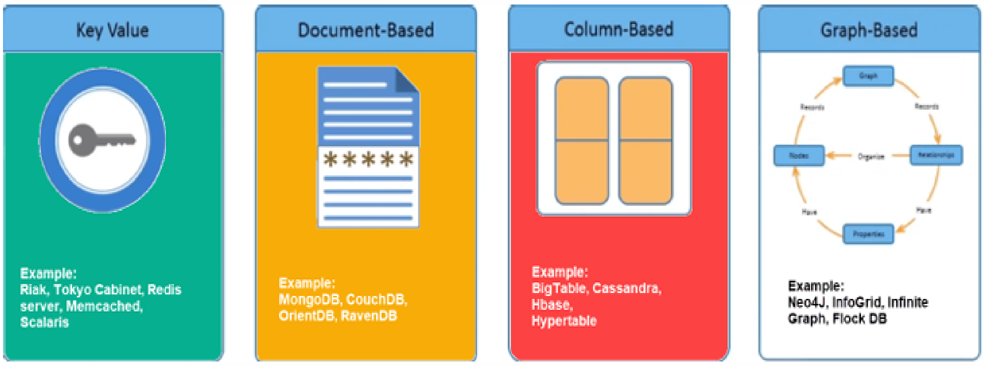
NoSQL Database is a non-relational Data Management System, that does not require a fixed schema. It avoids joins, and is easy to scale. The major purpose of using a NoSQL database is for distributed data stores with humongous data storage needs. NoSQL is used for Big data and real-time web apps. For example, companies like Twitter, Facebook and Google collect terabytes of user data every single day.

NoSQL database stands for “Not Only SQL” or “Not SQL.” Though a better term would be “NoREL”, NoSQL caught on. Carl Strozz introduced the NoSQL concept in 1998.

Traditional RDBMS uses SQL syntax to store and retrieve data for further insights. Instead, a NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data.



Types of NoSQL Databases:


Key Value
Example:
Riak, Tokyo Cabinet, Redis server, Memcached, Scalaris
Document-Based
*****
Example:
MongoDB, CouchDB, OrientDB, RavenDB
Column-Based
Example:
BigTable, Cassandra, Hbase, Hypertable
Graph-Based
Graph
Records
Records
Nodes
Organize
Relationships
Have
Have
Properties
Example:
Neo4J, InfoGrid, Infinite Graph, Flock DB

- Document-oriented: JSON documents MongoDB and CouchDB
- Key-value: Redis and DynamoDB
- Wide-column: Cassandra and HBase
- Graph: Neo4j and Amazon Neptune

Relational Databases (SQL)	Non-relational Databases (NoSQL)
Oracle	MongoDB
MySQL	couchDB
SQL Server	BigTable

Differences between SQL and NoSQL

The table below summarizes the main differences between SQL and NoSQL databases.

	SQL Databases	NoSQL Databases
Data Storage Model	Tables with fixed rows and columns	Document: JSON documents, Key-value: key-value pairs, Wide-column: tables with rows and dynamic columns, Graph: nodes and edges
Development History	Developed in the 1970s with a focus on reducing data duplication	Developed in the late 2000s with a focus on scaling and allowing for rapid application change driven by agile and DevOps practices.
Examples	Oracle, MySQL, Microsoft SQL Server, and PostgreSQL	Document: MongoDB and CouchDB, Key-value: Redis and DynamoDB, Wide-column: Cassandra and HBase, Graph: Neo4j and Amazon Neptune
Primary Purpose	General purpose	Document: general purpose, Key-value: large amounts of data with simple lookup queries, Wide-column: large amounts of data with predictable query patterns, Graph: analyzing and traversing relationships between connected data
Schemas	Rigid	Flexible
Scaling	Vertical (scale-up with a larger server)	Horizontal (scale-out across commodity servers)
Multi-Record ACID Transactions	Supported	Most do not support multi-record ACID transactions. However, some—like MongoDB—do.
Joins	Typically required	Typically not required

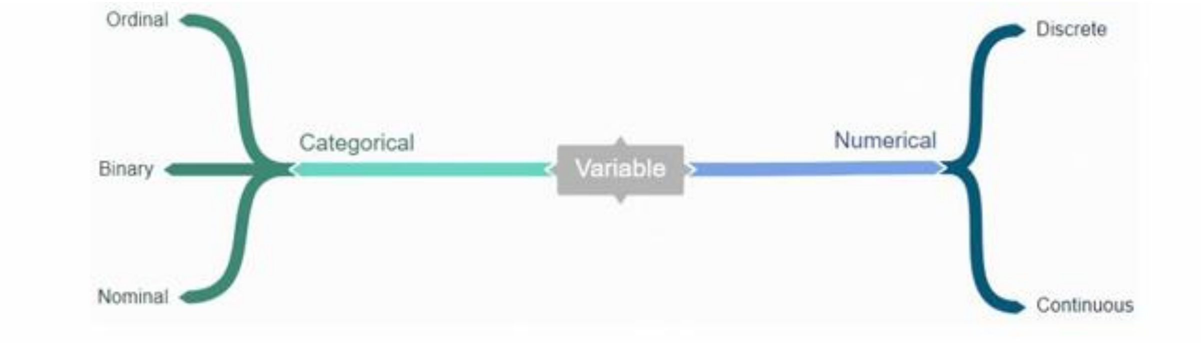
Variables:

Data consist of individuals and variables that give us information about those individuals. An individual can be an object or a person. A variable is an attribute, such as a measurement or a label.

Two types of Data

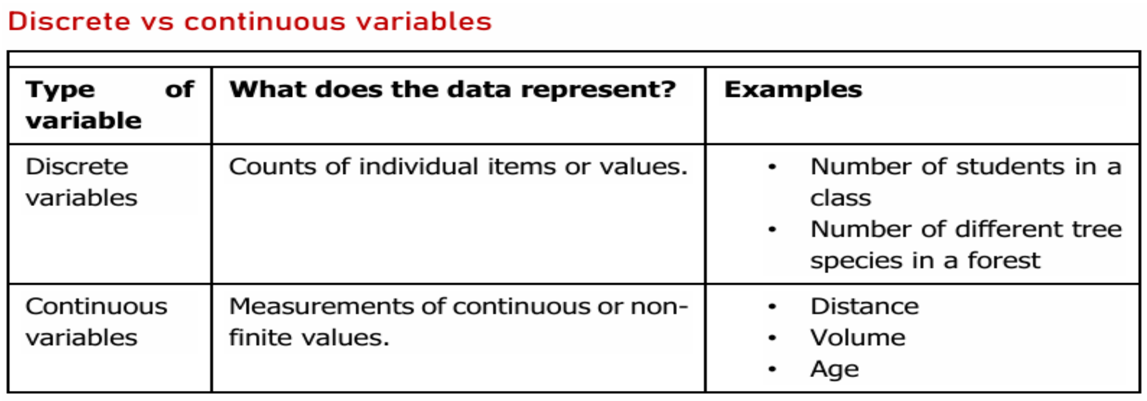
1.Quantitative data (Numerical)

2.Categorical data



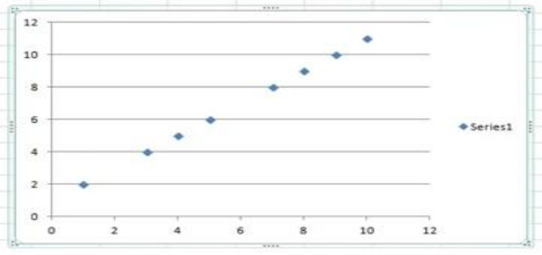
Quantitative Variables: Quantitative data, contains numerical that can be added, subtracted, divided, etc.

There are two types of quantitative variables: discrete and continuous.

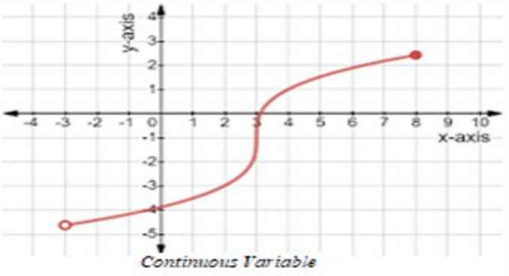


Discrete vs continuous variables

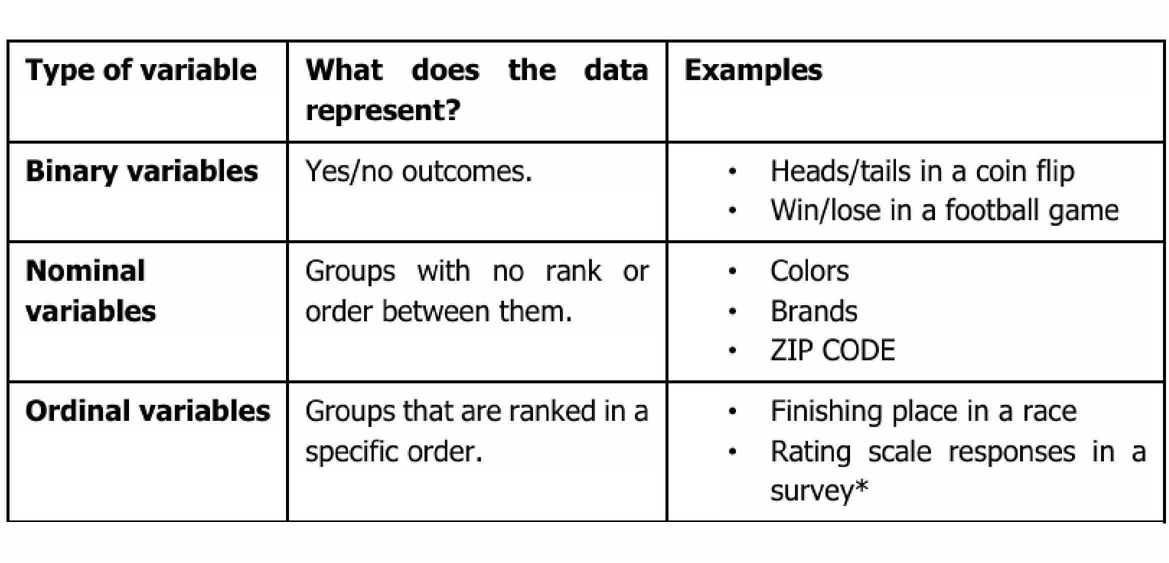
Type of variable	What does the data represent?	Examples
Discrete variables	Counts of individual items or values.	• Number of students in a class • Number of different tree species in a forest
Continuous variables	Measurements of continuous or non-finite values.	• Distance • Volume • Age



Discrete variables on a scatter plot.



Categorical variables: Categorical variables represent groupings of some kind. They are sometimes recorded as numbers, but the numbers represent categories rather than actual amounts of things. There are three types of categorical variables: binary, nominal, and ordinal variables.



Type of variable	What does the data represent?	Examples
Binary variables	Yes/no outcomes.	• Heads/tails in a coin flip • Win/lose in a football game
Nominal variables	Groups with no rank or order between them.	• Colors • Brands • ZIP CODE
Ordinal variables	Groups that are ranked in a specific order.	• Finishing place in a race • Rating scale responses in a survey*

Missing Imputations:

Imputation is the process of replacing missing data with substituted values.

Types of missing data

Missing data can be classified into one of three categories

1. MCAR

Data which is Missing Completely At Random has nothing systematic about which observations are missing values. There is no relationship between missingness and either observed or unobserved covariates.

2. MAR

Missing At Random is weaker than MCAR. The missingness is still random, but due entirely to observed variables. For example, those from a lower socioeconomic status may be less willing to provide salary information (but we know their SES status). The key is that the missingness is not due to the values which are not observed. MCAR implies MAR but not vice-versa.

3. MNAR

If the data are Missing Not At Random, then the missingness depends on the values of the missing data. Censored data falls into this category. For example, individuals who are heavier are less likely to report their weight. Another example, the device measuring some response can only measure values above .5. Anything below that is missing.

There can be two types of gaps in Data:

1. Missing Data Imputation

2. Model based Technique

Imputations: (Treatment of Missing Values)

1. Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

2. Fill in the missing value manually: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.

3. Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like "Unknown" or -∞. If missing values.

Missing Imputations:

Imputation is the process of replacing missing data with substituted values.

Types of missing data

Missing data can be classified into one of three categories

1. MCAR

Data which is Missing Completely At Random has nothing systematic about which observations are missing values. There is no relationship between missingness and either observed or unobserved covariates.

2. MAR

Missing At Random is weaker than MCAR. The missingness is still random, but due entirely observed variables. For example, those from a lower socioeconomic status may be less willing to provide salary information (but we know their SES status). The key is that the missingness is not due to the values which are not observed. MCAR implies MAR but not vice-versa.

3. MNAR

If the data are Missing Not At Random, then the missingness depends on the values of the missing data. Censored data falls into this category. For example, individuals who are heavier are less likely to report their weight. Another example, the device measuring some response can only measure values above .5. Anything below that is missing.

There can be two types of gaps in Data:

1. Missing Data Imputation

2. Model based Technique

Imputations: (Treatment of Missing Values)

1. Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

2. Fill in the missing value manually: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.

3. Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like "Unknown" or -∞. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common-that of "Unknown." Hence, although this method is simple, it is not foolproof.

4. Use the attribute mean to fill in the missing value: Considering the average value of that particular attribute and use this value to replace the missing value in that attribute column.

5. Use the attribute mean for all samples belonging to the same class as the given tuple:

For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

6. Use the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

Types of Data Models

Data modeling can be achieved in various ways. However, the basic concept of each of them remains the same. Let's have a look at the commonly used data modeling methods:

Hierarchical model

As the name indicates, this data model makes use of hierarchy to structure the data in a tree-like format as shown in figure 2.6. However, retrieving and accessing data is difficult in a hierarchical database. This is why it is rarely used now.

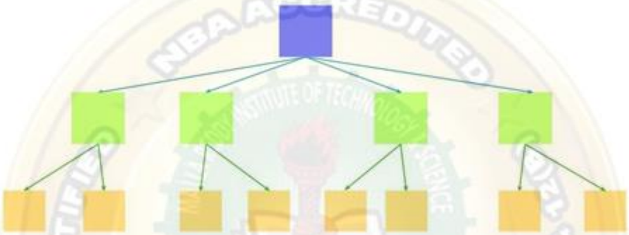
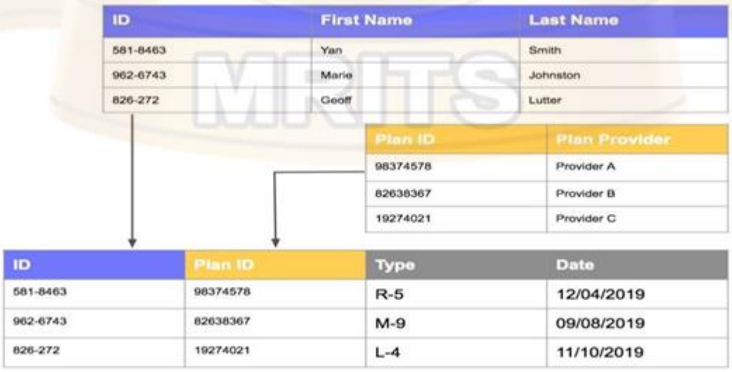


Fig 2.6: Hierarchical Model Structure

Relational model

Proposed as an alternative to hierarchical model by an IBM researcher, here data is represented in the form of tables. It reduces the complexity and provides a clear overview of the data as shown below in figure 2.7.



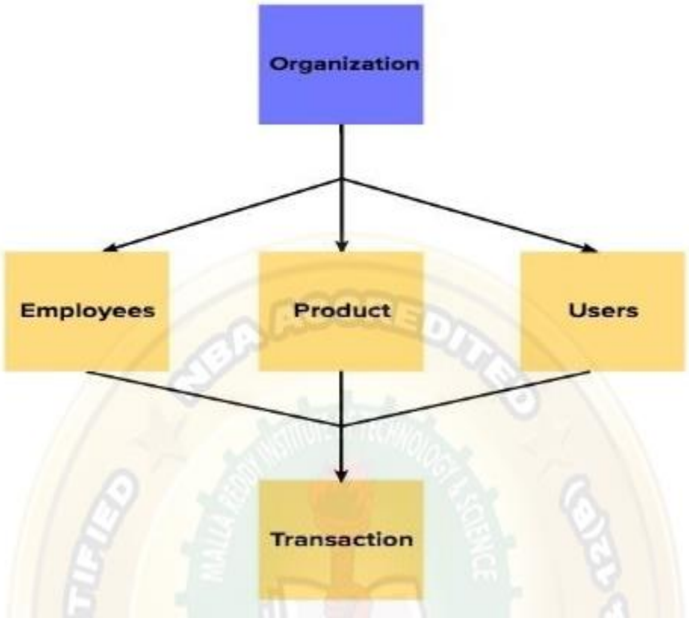
ID	First Name	Last Name
581-8463	Yan	Smith
962-6743	Marie	Johnston
826-272	Geoff	Lutter

Plan ID	Plan Provider
98374578	Provider A
82638367	Provider B
19274021	Provider C

ID	Plan ID	Type	Date
581-8463	98374578	R-5	12/04/2019
962-6743	82638367	M-9	09/08/2019
826-272	19274021	L-4	11/10/2019

Network model

The network model is inspired by the hierarchical model. However, unlike the hierarchical model, this model makes it easier to convey complex relationships as each record can be linked with multiple parent records as shown in figure 2.8. In this model data can be shared easily and the computation becomes easier.



Entity-relationship model

Entity-relationship model, also known as ER model, represents entities and their relationships in a graphical format. An entity could be anything – a concept, a piece of data, or an object.

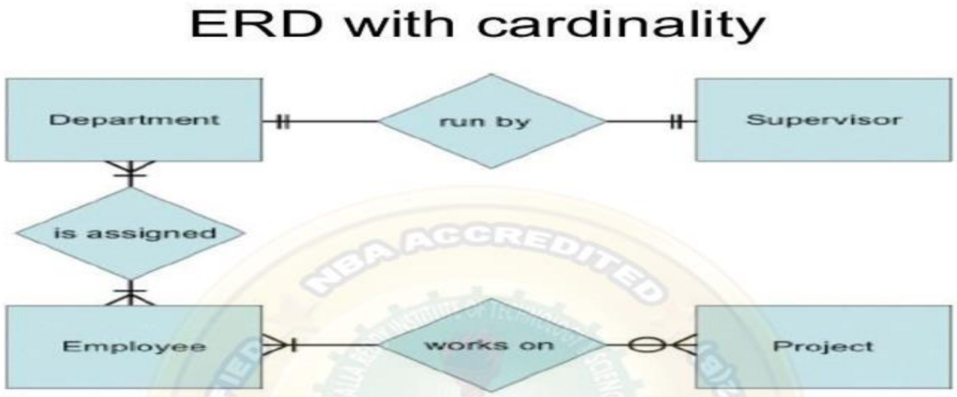


Fig 2.10: Entity Relationship Diagram

The entity relationship diagram explains relation between variables and with their primary key and foreign key as shown in figure 2.10. along with this it also explains the multiple instances of relation between tables.

Now that we have a basic understanding of data modeling, let's see why it is important.