Unit-1

Data and architecture design:

Data architecture in Information Technology is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations.

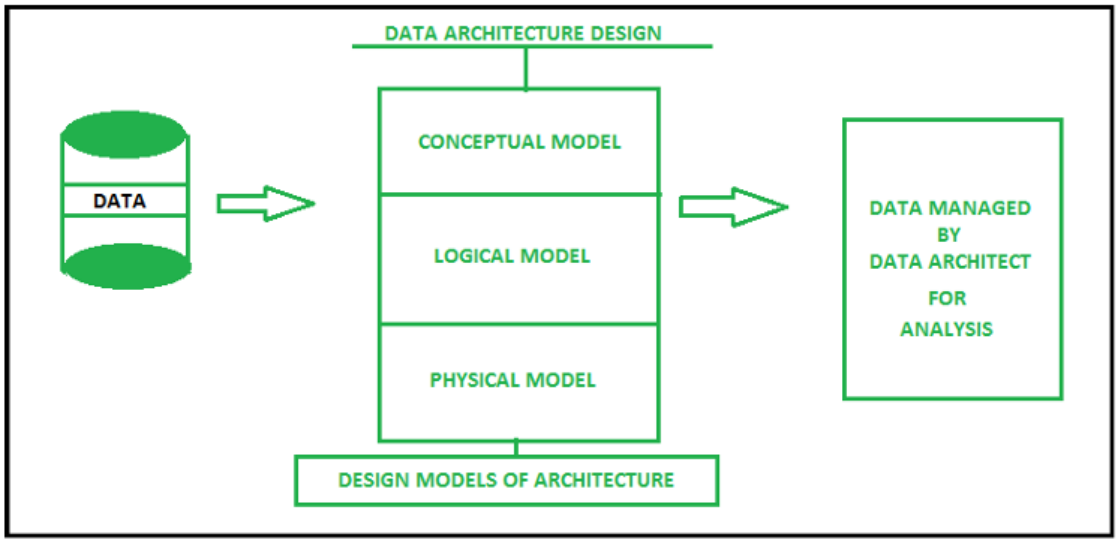
- A data architecture should set data standards for all its data systems as a vision or a model of the eventual interactions between those data systems.
- Data architectures address data in storage and data in motion; descriptions of data stores, data groups and data items; and mappings of those data artifacts to data qualities, applications, locations etc.
- Essential to realizing the target state, Data Architecture describes how data is processed, stored, and utilized in a given system. It provides criteria for data processing operations that make it possible to design data flows and also control the flow of data in the system.
- The Data Architect is typically responsible for defining the target state, aligning during development and then following up to ensure enhancements are done in the spirit of the original blueprint.
- During the definition of the target state, the Data Architecture breaks a subject down to the atomic level and then builds it back up to the desired form.
- The Data Architect breaks the subject down by going through 3 traditional architectural processes:

Conceptual model: It is a business model which uses Entity Relationship (ER) model for relation between entities and their attributes.

Logical model: It is a model where problems are represented in the form of logic such as rows and column of data, classes, xml tags and other DBMS techniques.

Physical model: Physical models hold the database design like which type of database technology will be suitable for architecture.

The data architecture is formed by dividing into three essential models and then are combined:



Factors that influence Data Architecture

Various constraints and influences will have an effect on data architecture design. These include enterprise requirements, technology drivers, economics, business policies and data processing need.

Enterprise requirements:

- These will generally include such elements as economical and effective system expansion, acceptable performance levels (especially system access speed), transaction reliability, and transparent data management.
- In addition, the conversion of raw data such as transaction records and image files into more useful information forms through such features as data warehouses is also a common organizational requirement, since this enables managerial decision making and other organizational processes.
- One of the architecture techniques is the split between managing transaction data and (master) reference data. Another one is splitting data capture systems from data retrieval systems (as done in a data warehouse).

Technology drivers:

- These are usually suggested by the completed data architecture and database architecture designs.

- In addition, some technology drivers will derive from existing organizational integration frameworks and standards, organizational economics, and existing site resources (e.g. previously purchased software licensing).

Economics:

- These are also important factors that must be considered during the data architecture phase. It is possible that some solutions, while optimal in principle, may not be potential candidates due to their cost.
- External factors such as the business cycle, interest rates, market conditions, and legal considerations could all have an effect on decisions relevant to data architecture.

Business policies:

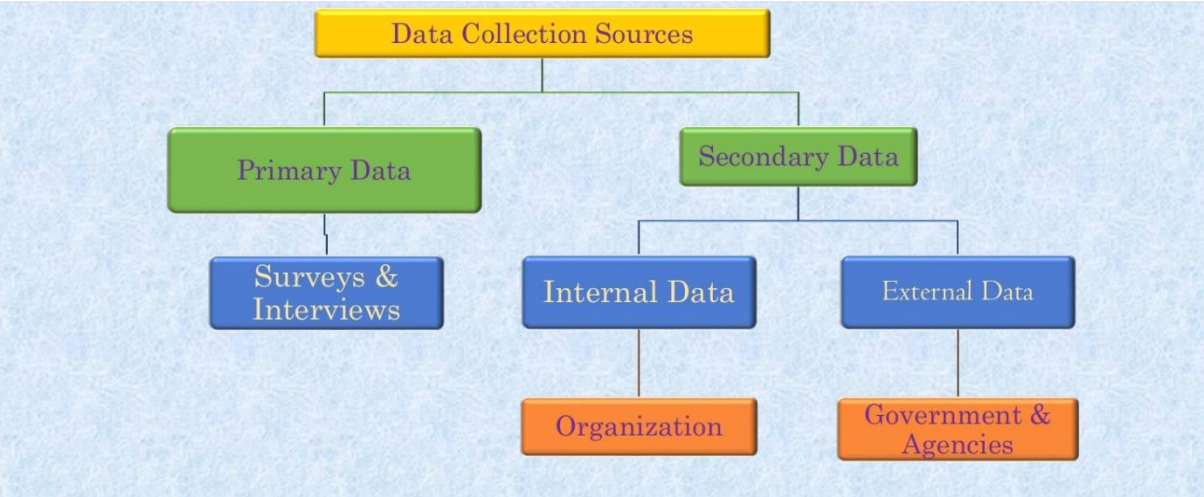
- Business policies that also drive data architecture design include internal organizational policies, rules of regulatory bodies, professional standards, and applicable governmental laws that can vary by applicable agency.
- These policies and rules will help describe the manner in which enterprise wishes to process their data.

Data processing needs

- These include accurate and reproducible transactions performed in high volumes, data warehousing for the support of management information systems (and potential data mining), repetitive periodic reporting, ad hoc reporting, and support of various organizational initiatives as required (i.e. annual budgets, new product development)
- The General Approach is based on designing the Architecture at three Levels of Specification.
  - The Logical Level
  - The Physical Level
  - The Implementation Level

1.2 Various Sources of Data Understand various primary sources of the Data.

Data can be generated from two types of sources namely Primary and Secondary Sources of Primary Data



Primary data

• The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys. The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data

1. Interview method:

- The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee.
- Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:

- The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video.
- The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analysing data. Examples are online surveys or surveys through social media polls.

3. Observation method:

- The observation method is a method of data collection in which the researcher keenly observes the behaviour and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats.
- In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behaviour towards the products. The data obtained will be sent for processing.

4. Experimental method:

- The experimental method is the process of collecting data through performing experiments, research, and investigation.
- The most frequently used experiment methods are CRD, RBD, LSD, FD.

CRD- Completely Randomized design is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.

Example 2:

A company wishes to test 4 different types of tyre. The tyres lifetime as determined from their threads are given. Where each tyre has been tried on 6 similar automobiles assigned at random to their tyres. Determine whether there is a significant difference between tyres at .05 level.

Tyres	Automobile 1	Automobile 2	Automobile 3	Automobile 4	Automobile 5	Automobile 6
A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	29	34	32	30	33	31

Solution: 33 Null Hypothesis: There is no difference between the tyres in their life time. 31 We choose a random value closest to the average of all values in the table and subtract that for each tyre in the automobile, for example by choosing 35

Tyres	Automobile 1	Automobile 2	Automobile 3	Automobile 4	Automobile 5	Automobile 6	Total
A	-2	3	1	5	-4	0	3
B	-3	5	7	3	-5	-1	6
C	-4	2	0	-2	-1	-5	-10
D	6	-1	-3	-5	-2	-4	-21
	T = Sum(X) =						-22

N = no of samples = 24 (4 rows * 6 columns)

Correction factor = $\frac{T \times T}{N}$ = 20.16

Square the values to find

Tyres	Automobile 1	Automobile 2	Automobile 3	Automobile 4	Automobile 5	Automobile 6	Total
A	4	9	1	25	16	0	55
B	9	25	9	49	25	1	118
C	16	4	0	4	1	25	50
D	36	1	9	25	4	16	91
	T = Sum(X^2) =						314

Total sum of squares

(SST) = sum(X2) – Correlation factor = 314 – 20.16 = 293.84

Sum of Squares between Treatments (SSTr) = = ((3)2/6 +(6)2/6 +(10)2/6 +(21)2/6) – Correlation factor = 77.50

Sum of Squares Error (SSE) SST – SSTr = 293.84 – 77.50 = 216.34 Now by using ANOVA (one way classification) Table, We calculate the F- Ratio.

F-Ratio: The F ratio is the ratio of two mean square values. If the null hypothesis is true, you expect F to have a value close to 1.0 most of the time. A large F ratio means that the variation among group mean is more than you'd expect to see by chance If the value of F-Ratio is closer to 1 it means that null hypothesis is true. If F-ratio is greater than then we assume that the null hypothesis is false.

Source of variation	Sum of squares	Degrees of freedom	Mean of sum of squares	F - Ratio
Between treatments	$SST_r = 77.50$	No of treatment – 1 = 4-1 =3	$MST_r = SST_r$/ Degrees of Freedom = 77.50/3 = 25.83	
				F-ratio = MST_r/MSE = 25.83/10.87 = 2.376
Within treatments	SSE = 216.34	No of values – no of treatment = 24 – 4 =20	MSE = SSE/Degrees of Freedom =216.34/20 =10.87	

In this scenario the value of F-ratio is greater than 1. This indicates there will be variation between samples. So assumed null hypothesis will be false
Level of significance = 0.05 (given in question)
Degrees of Freedom = (3, 20)
Critical value = 3.10 (calculated from 5 percentage table)
F-Ratio >critical value (i.e) 2.376> 3.10

Hence assumed null hypothesis is false. This indicates there is life time difference between tyres.

A randomized block design

The experimenter divides subjects into subgroups called blocks, such that the variability within blocks is less than the variability between blocks. Then, subjects within each block are randomly assigned to treatment conditions. Compared to a completely randomized design, this design reduces variability within treatment conditions and potential confounding, producing a better estimate of treatment effects. The table below shows a randomized block design for a hypothetical medical experiment.

Gender	Treatment	
	Placebo	Vaccine
Male	250	250
Female	250	250

Subjects are assigned to blocks, based on gender. Then, within each block, subjects are randomly assigned to treatments (either a placebo or a cold vaccine). For this design, 250 men get the placebo, 250 men get the vaccine, 250 women get the placebo, and 250 women get the vaccine. It is known that men and women are physiologically different and react differently to medication. This design ensures that each treatment condition has an equal proportion of men and women. As a result, differences between treatment conditions cannot be attributed to gender. This randomized block design removes gender as a potential source of variability and as a potential confounding variable.

LSD - Latin Square Design

A Latin square is one of the experimental designs which has a balanced two-way classification scheme say for example - 4 X 4 arrangement. In this scheme each letter from A to D occurs only once in each row and also only once in each column. The balance arrangement, it may be noted that, will not get disturbed if any row gets changed with the other.

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

The balance arrangement achieved in a Latin Square is its main strength. In this design, the comparisons among treatments, will be free from both differences between rows and columns. Thus the magnitude of error will be smaller than any other design.

FD - Factorial Designs

This design allows the experimenter to test two or more variables simultaneously. It also measures interaction effects of the variables and analyzes the impacts of each of the variables.

2. Secondary data

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

Internal source

These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

Accounting resources- This gives so much information which can be used by the marketing researcher. They give information about internal factors.

Sales Force Report- It gives information about the sales of a product. The information provided is from outside the organization.

Internal Experts- These are people who are heading the various departments. They can give an idea of how a particular thing is working.

Miscellaneous Reports- These are what information you are getting from operational reports. If the data available within the organization are unsuitable or inadequate, the marketer should extend the search to external secondary data sources.

External source:

The data which can't be found at internal organizations and can be gained through external third-party resources is external source data. The cost and time consumption are more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labour bureau, syndicate services, and other non-governmental publications.

1. Government Publications-

- Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data. These are like: Registrar General of

India- It is an office which generates demographic data. It includes details of gender, age, occupation etc.

2. Central Statistical Organization-

- This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate, and rate of major economic activities. Annual survey of Industries is also published by the CSO. It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.

3. Ministry of Commerce and Industries

- This ministry through the office of economic advisor provides information on wholesale price index. These indices may be related to a number of sectors like food, fuel, power, food grains etc.
- It also generates All India Consumer Price Index numbers for industrial workers, urban, nonmanual employees and cultural labourers.

4. Planning Commission

- It provides the basic statistics of Indian Economy.

5. Reserve Bank of India

- This provides information on Banking Savings and investment. RBI also prepares currency and finance reports.

6. Labour Bureau

- It provides information on skilled, unskilled, white collared jobs etc.

7. National Sample Survey

- This is done by the Ministry of Planning and it provides social, economic, demographic,

industrial and agricultural statistics.

8. Non-Government Publications

- These includes publications of various industrial and trade associations, such as The Indian Cotton Mill Association Various chambers of commerce.

9. The Bombay Stock Exchange

- It publishes a directory containing financial accounts, key profitability and other relevant matter) Various Associations of Press Media.
- Export Promotion Council.
- Confederation of Indian Industries (CII)
- Small Industries Development Board of India
- Different Mills like - Woollen mills, Textile mills etc

The only disadvantage of the above sources is that the data may be biased. They are likely to colour their negative points.

Survey- They conduct surveys regarding - lifestyle, sociographic, general topics.

The International Labour Organization (ILO):

- It publishes data on the total and active population, employment, unemployment, wages and consumer prices.

  1.The Organization for Economic Co-operation and development (OECD):
  It publishes data on foreign trade, industry, food, transport, and science and technology.

  2.The International Monetary Fund (IMA):
  It publishes reports on national and international foreign exchange regulations.

  Other sources:

1.3 Understanding Sources of Data from Sensor

Sensor data is the output of a device that detects and responds to some type of input from the physical environment. The output may be used to provide information or input to another system or to guide a process. Examples are as follows

• A photosensor detects the presence of visible light, infrared transmission (IR) and/or ultraviolet (UV) energy.

• Lidar, a laser-based method of detection, range finding and mapping, typically uses a

low-power, eye-safe pulsing laser working in conjunction with a camera.

• A charge-coupled device (CCD) stores and displays the data for an image in such a way that each pixel is converted into an electrical charge, the intensity of which is related to a color in the color spectrum.

• Smart grid sensors can provide real-time data about grid conditions, detecting outages, faults and load and triggering alarms.

• Wireless sensor networks combine specialized transducers with a communications infrastructure for monitoring and recording conditions at diverse locations. Commonly monitored parameters include temperature, humidity, pressure, wind direction and speed, illumination intensity, vibration intensity, sound intensity, powerline voltage, chemical concentrations, pollutant levels and vital body functions.

1.4 Understanding Sources of Data from Signal

The simplest form of signal is a direct current (DC) that is switched on and off; this is the principle by which the early telegraph worked. More complex signals consist of an alternating-current (AC) or electromagnetic carrier that contains one or more data streams. Data must be transformed into electromagnetic signals prior to transmission across a network. Data and signals can be either analog or digital. A signal is periodic if it consists of a continuously repeating pattern.

1.5 Understanding Sources of Data from GPS

The Global Positioning System (GPS) is a space based navigation system that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites. The system provides critical capabilities to military, civil, and commercial users around the world. The United States government created the system, maintains it, and makes it freely accessible to anyone with a GPS receiver.

Data Management

Data management refers to the practices, tools, and strategies used to collect, store, process, and analyse data efficiently. It ensures data quality, security, and accessibility for organizations.

Key Components of Data Management

1. **Data Collection** – Gathering data from different sources like CRM, IoT, sensors, and surveys.
2. **Data Storage** – Storing data in structured (SQL), semi-structured (JSON, XML), and unstructured formats (text, images).
3. **Data Integration** – Combining data from different systems to create a unified dataset.
4. **Data Cleaning & Processing** – Removing errors, filling missing values, and transforming data for analysis.
5. **Data Security** – Protecting data using encryption, access control, and compliance measures.
6. **Data Governance** – Establishing policies for data accuracy, privacy, and ownership.
7. **Data Access & Reporting** – Ensuring users can retrieve data through dashboards and reports.

Sources of Data Management

1. Internal Data Sources

- **Databases** – SQL and NoSQL databases store business transactions, customer details, and financial records.
- **Enterprise Systems** – CRM (Customer Relationship Management), ERP (Enterprise Resource Planning).
- **Business Intelligence Systems** – Tools like Tableau, Power BI for analytics and reporting.

2. External Data Sources

- **Government Reports** – Census data, economic reports, healthcare statistics.
- **Open Data Platforms** – Kaggle, Google Dataset Search, and government portals.
- **Research Publications** – Scientific journals and market reports.

3. Online Data Sources

- **Social Media Data** – Twitter, Facebook, and LinkedIn for sentiment analysis.
- **Web Scraping & APIs** – Extracting real-time data from websites.
- **IoT & Sensor Data** – Real-time monitoring in healthcare, agriculture, and smart cities.

4. Cloud-Based Data Sources

- **AWS, Google Cloud, Azure** – Cloud-based data lakes and warehouses.
- **Big Data Platforms** – Hadoop, Spark for large-scale data processing.

Importance of Data Management

1. **Improves Decision-Making** – Provides accurate insights for business growth.
2. **Enhances Data Security** – Protects sensitive customer information.
3. **Ensures Compliance** – Meets regulatory standards like GDPR and HIPAA.
4. **Increases Efficiency** – Reduces data errors and duplication.
5. **Supports AI & Machine Learning** – Ensures high-quality data for predictive models.

Data management plays a critical role in modern business intelligence, ensuring data-driven decisions are accurate and reliable.

1.7 Data Quality

Data quality refers to the quality of data. Data quality refers to the state of qualitative or quantitative pieces of information. There are many definitions of data quality but data is generally considered high quality if it is "fit for [its] intended uses in operations, decision making and planning

The seven characteristics that define data quality are:

1. Accuracy

2. Validity

3. Consistency

4. Timeliness

5. Completeness

6. Availability and Accessibility

7. Uniqueness

Accuracy: This characteristic refers to the exactness of the data. It cannot have any erroneous elements and must convey the correct message without being misleading. This accuracy and precision have a component that relates to its intended use. Without understanding how the data will be consumed, ensuring accuracy and precision could be off-target or more costly than necessary. For example, accuracy in healthcare might be more important than in another

industry (which is to say, inaccurate data in healthcare could have more serious consequences) and, therefore, justifiably worth higher levels of investment.

Validity: Requirements governing data set the boundaries of this characteristic. For example, on surveys, items such as gender, ethnicity, and nationality are typically limited to a set of options and open answers are not permitted. Any answers other than these would not be considered valid or legitimate based on the survey's requirement. This is the case for most data and must be carefully considered when determining its quality. The people in each department in an organization understand what data is valid or not to them, so the requirements must be leveraged when evaluating data quality.

Reliability and Consistency: Many systems in today's environments use and/or collect the same source data. Regardless of what source collected the data or where it resides, it cannot contradict a value residing in a different source or collected by a different system. There must be a stable and steady mechanism that collects and stores the data without contradiction or unwarranted variance.

Timeliness: There must be a valid reason to collect the data to justify the effort required, which also means it has to be collected at the right moment in time. Data collected too soon or too late could misrepresent a situation and drive inaccurate decisions.

Completeness: Incomplete data is as dangerous as inaccurate data. Gaps in data collection led to a partial view of the overall picture to be displayed. Without a complete picture of how operations are running, uninformed actions will occur. It's important to understand the complete set of requirements that constitute a comprehensive set of data to determine whether or not the requirements are being fulfilled.

Availability and Accessibility: This characteristic can be tricky at times due to legal and regulatory constraints. Regardless of the challenge, though, individuals need the right level of access to the data in order to perform their jobs. This presumes that the data exists and is available for access to be granted.

Uniqueness: The level of detail at which data is collected is important, because confusion and inaccurate decisions can otherwise occur. Aggregated, summarized and manipulated collections of data could offer a different meaning than the data implied at a lower level. An

appropriate level of granularity must be defined to provide sufficient uniqueness and distinctive properties to become visible. This is a requirement for operations to function effectively.

Noisy data

Noisy data is meaningless data. The term has often been used as a synonym for corrupt data. However, its meaning has expanded to include any data that cannot be understood and interpreted correctly by machines, such as unstructured text.

Noisy data

Origins of noise

• **outliers** -- values seemingly out of the normal range of data

• **Duplicate records** -- good database design should minimize this (use DISTINCT on SQL retrievals)

• **Incorrect attribute values** -- again good db design and integrity constraints should minimize this numeric only, deal with rogue strings or characters where numbers should be.null handling for attributes (nulls=missing values)

OUTLIERS:

Outlier is a point or an observation that deviates significantly from the other observations.

Outlier is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations.

Reasons for outliers: Due to experimental errors or "special circumstances".

- There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise.
- There are various methods of outlier detection. Some are graphical such as normal probability plots. Others are model-based. Box plots are a hybrid.

Types of Outliers:

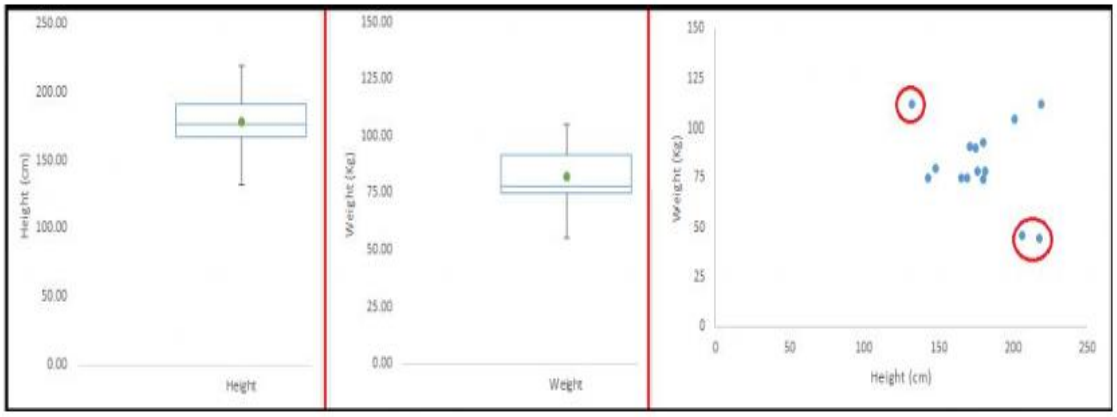
Outlier can be of two types:

Univariate: These outliers can be found when we look at distribution of a single variable

.

Multivariate: Multi-variate outliers are outliers in an n-dimensional space.

In order to find them, you have to look at distributions in multi-dimensions.



Impact of Outliers on a dataset:

Outliers can drastically change the results of the data analysis and statistical modelling. There are numerous unfavourable impacts of outliers in the data set:

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

Detect Outliers:

Most commonly used method to detect outliers is visualization. We use various visualization methods, like Box-plot, Histogram, Scatter Plot (above, we have used box plot and scatter plot for visualization).

Outlier treatments are three types:

Retention:

- There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. There are various methods of outlier detection. Some are graphical such as normal probability plots. Others are model based. Box plots are a hybrid.

Exclusion:

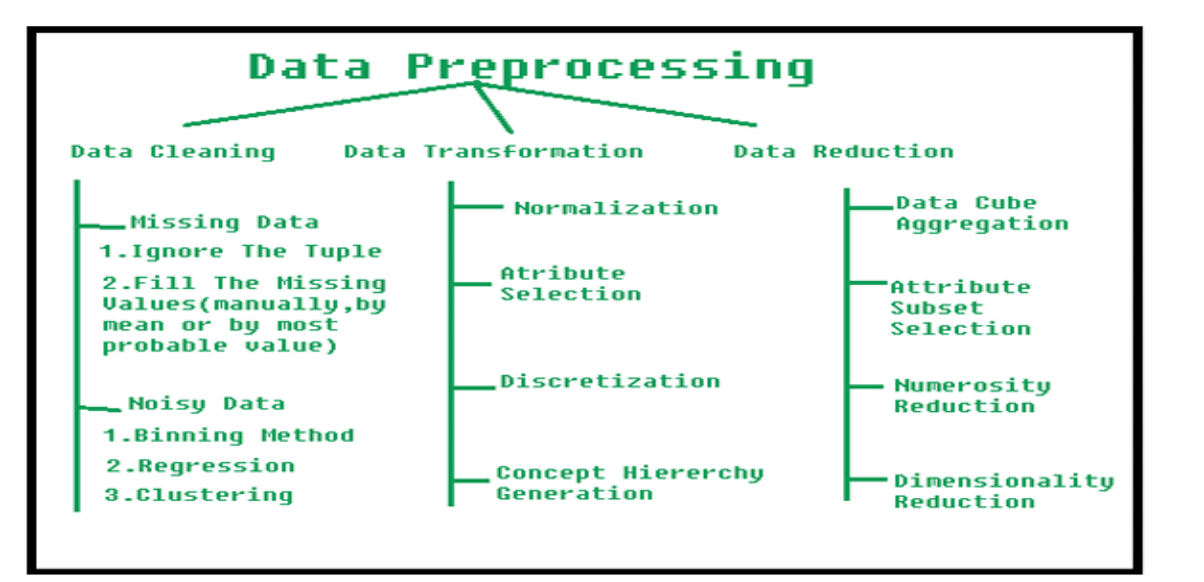
- According to a purpose of the study, it is necessary to decide, whether and which outlier will be removed/excluded from the data, since they could highly bias the final results of the analysis.

Rejection:

- Rejection of outliers is more acceptable in areas of practice where the underlying model of the process being measured and the usual distribution of measurement error are confidently known.
- An outlier resulting from an instrument reading error may be excluded but it is desirable that the reading is at least verified.

Data Pre-processing:

Preprocessing in Data Mining: Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Steps Involved in Data Preprocessing

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

1. Ignore the tuples:

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

2. 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.

(b). Noisy Data:

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:

1. Binning Method:

This method works on sorted data in order to smooth it. Binning, also called discretization, is a technique for reducing the cardinality (The total number of unique values for a dimension is known as its cardinality) of continuous and discrete data. Binning groups related values together in bins to reduce the number of distinct values

2. Regression:

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

3. Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

- Transform the data in appropriate forms suitable for mining process.
- This involves following ways:

1.Normalization:

- The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.
- It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. Attribute Selection:

- In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. Discretization:

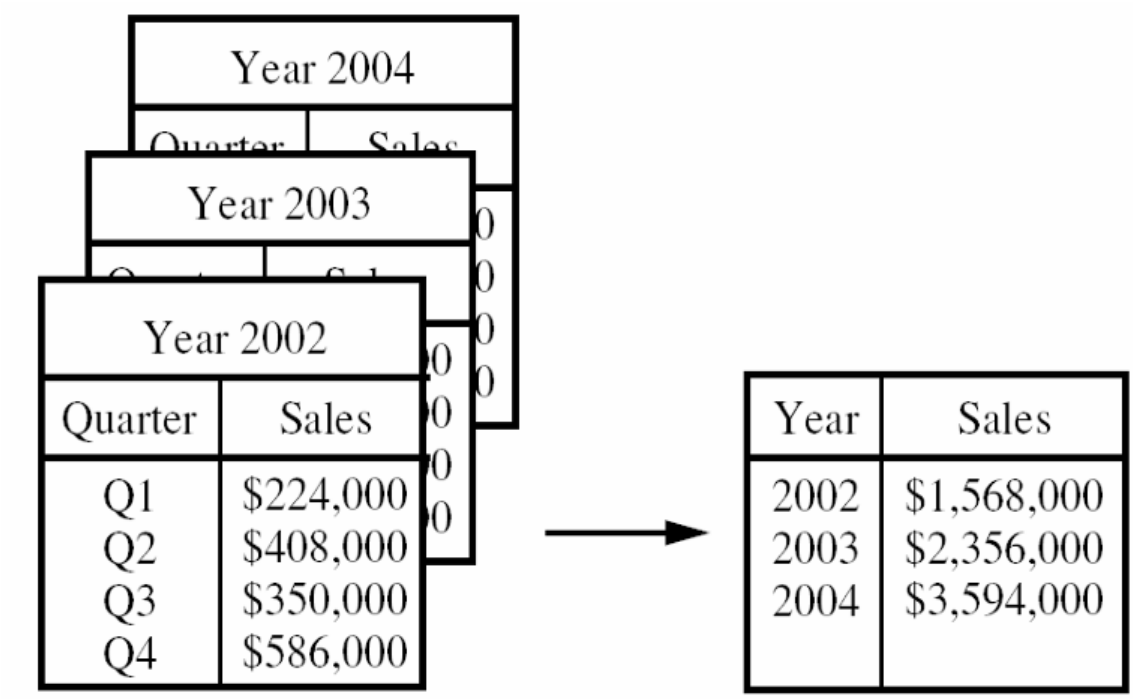
- Discretization is the process through which we can transform continuous variables, models or functions into a discrete form.
- We do this by creating a set of contiguous intervals (or bins) that go across the range of our desired variable/model/function. Continuous data is Measured, while Discrete data is Counted

3. 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.

1. Data Cube Aggregation:

- Aggregation operation is applied to data for the construction of the data cube.



2. Attribute Subset Selection:

- The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. the attribute having p-value greater than significance level can be discarded.
    - The procedure starts with an empty set of attributes as the reduced set.
    - First: The best single-feature is picked.
    - Next: At each subsequent iteration or step, the best of the remaining original attributes is added to the set

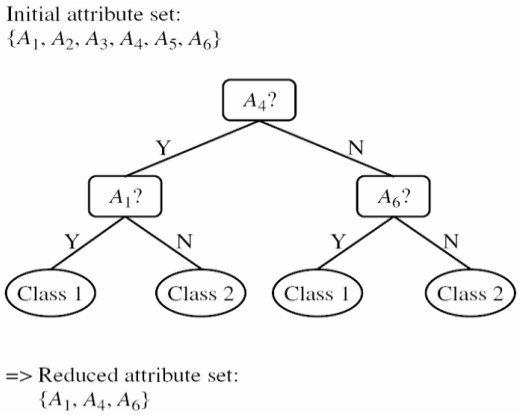
Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

Initial reduced set:
$\{\}$
$\Rightarrow \{A_1\}$
$\Rightarrow \{A_1, A_4\}$
$\Rightarrow$ Reduced attribute set:
$\{A_1, A_4, A_6\}$

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

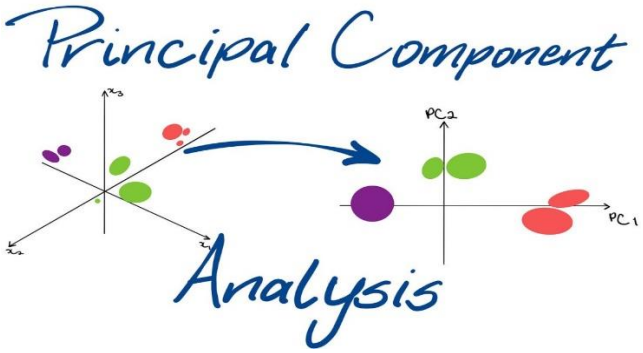
=> $\{A_1, A_3, A_4, A_5, A_6\}$
=> $\{A_1, A_4, A_5, A_6\}$
=> Reduced attribute set:
$\{A_1, A_4, A_6\}$

- Decision tree induction



3. Numerosity Reduction: (Principal Component Analysis)

- This enables to store the model of data instead of whole data



4. Dimensionality Reduction:

- This reduces the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis)

Data Processing:

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

Six stages of data processing

1. Data collection

Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.

2. Data preparation

Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as "pre-processing" is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data for the best business intelligence.

3. Data input

The clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.

4. Processing

During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed (data lakes, social networks, connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).