DATA ANALYTICS

(Professional Elective - I)

Subject Code: CS513PE

NOTES MATERIAL

UNIT1

For

B. TECH (CSE)

3rd YEAR – 1st SEM (R18)

Faculty:

B. RAVIKRISHNA

DEPARTMENT OF CSE

VIGNAN INSTITUTE OF TECHNOLOGY & SCIENCE

DESHMUKHI

Prerequisites:

1. A course on “Database Management Systems”.
2. Knowledge of probability and statistics.

Course Objectives:

1. To explore the fundamental concepts of data analytics.
2. To learn the principles and methods of statistical analysis
3. Discover interesting patterns, analyze supervised and unsupervised models and estimate the accuracy of the algorithms.
4. To understand the various search methods and visualization techniques.

Course Outcomes: After completion of this course students will be able to

1. Understand the impact of data analytics for business decisions and strategy
2. Carry out data analysis/statistical analysis
3. To carry out standard data visualization and formal inference procedures
4. Design Data Architecture
5. Understand various Data Sources

INTRODUCTION:

In the beginning times of computers and Internet, the data used was not as much of as it is today, the data then could be so easily stored and managed by all the users and business enterprises on a single computer, because the data never exceeded to the extent of 19 exabytes but now in this era, the data has increased about 2.5 quintillion per day.

Most of the data is generated from social media sites like Facebook, Instagram, Twitter, etc, and the other sources can be e-business, e-commerce transactions, hospital, school, bank data, etc. This data is impossible to manage by traditional data storing techniques. 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. But how do we do it? Well, that’s where the term ‘Data Analytics’ comes in.

Why is Data Analytics important?

Data Analytics has a key role in improving your business as it is used to gather hidden insights, Interesting Patterns in Data, 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 analyzed 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.

What are the tools used in Data Analytics?

ith 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
- Python
- Tableau Public
- QlikView
- SAS
- Microsoft Excel
- RapidMiner
- KNIME
- OpenRefine
- Apache Spark

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.

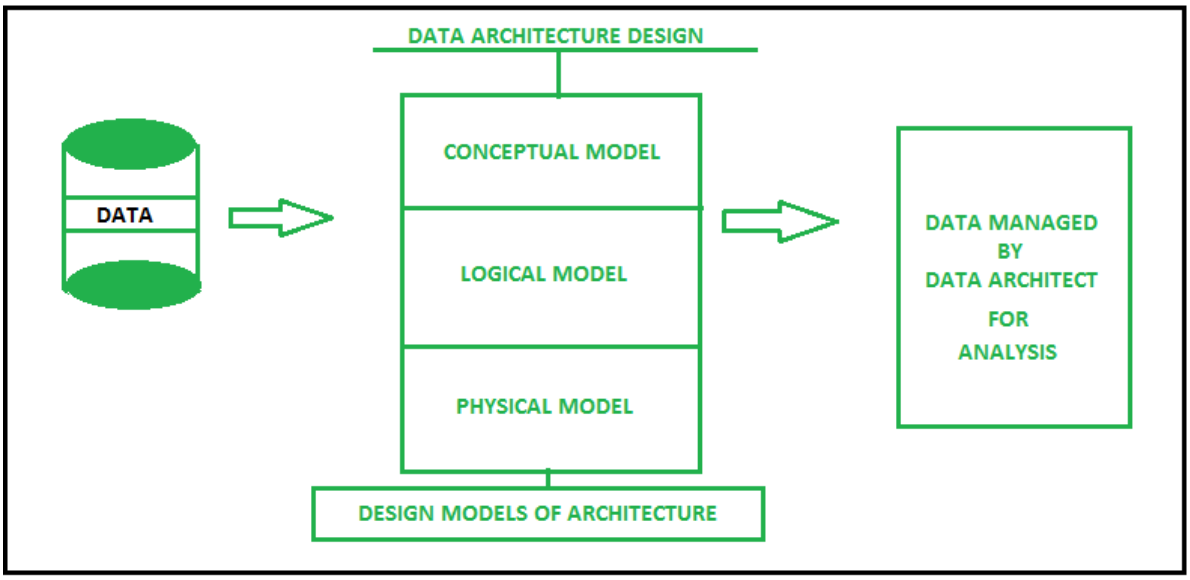
During the definition of the target state, the Data Architecture breaks a subject down to the atomic level and then builds it back up to the desired form.

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 holds the database design like which type of database technology will be suitable for architecture.

Layer	View	Data (What)	Stakeholder
1	**Scope/Contextual**	List of things and architectural standards important to the business	Planner
2	**Business Model/Conceptual**	Semantic model or Conceptual/Enterprise Data Model	Owner
3	**System Model/Logical**	Enterprise/Logical Data Model	Designer
4	**Technology Model/Physical**	Physical Data Model	Builder
5	**Detailed Representations**	Actual databases	Subcontractor

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

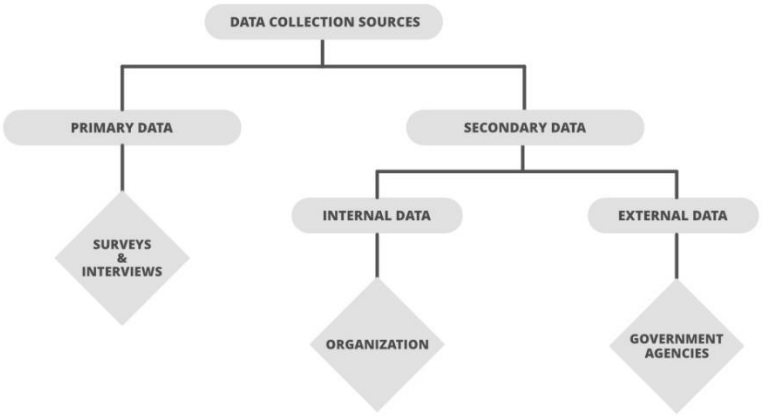
- The Physical Level
- The Implementation Level

Understand various sources of the Data:

- Data can be generated from two types of sources namely Primary and Secondary Sources of Primary Data.
- Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.
- In the process of big data analysis, "Data collection" is the initial step before starting to analyse the patterns or useful information in data. The data which is to be analysed must be collected from different valid sources.
- The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as "knowledge". Knowledge has many meanings like business knowledge or sales of enterprise products, disease treatment, etc.
- The main goal of data collection is to collect information-rich data.
- Data collection starts with asking some questions such as *what type of data is to be collected* and *what is the source of collection*.
- Most of the data collected are of two types known as **qualitative data** which is a group of non-numerical data such as words, sentences mostly focus on behaviour and actions of the group and another one is **quantitative data** which is in numerical forms and can be calculated using different scientific tools and sampling data.

The actual data is then further divided mainly into two types known as:

1. Primary data
2. Secondary data



1. 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.

RBD- Randomized Block Design is an experimental design in which the experiment is divided into small units called blocks.

- Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.

- Randomized Block Design - The Term Randomized Block Design has originated from agricultural research. In this design several treatments of variables are applied to different blocks of land to ascertain their effect on the yield of the crop.
- Blocks are formed in such a manner that each block contains as many plots as a number of treatments so that one plot from each is selected at random for each treatment. The production of each plot is measured after the treatment is given.
- These data are then interpreted and inferences are drawn by using the analysis of Variance technique so as to know the effect of various treatments like different dozes of fertilizers, different types of irrigation etc.

LSD – Latin Square Design is an experimental design that is similar to CRD and RBD blocks but contains rows and columns.

- It is an arrangement of NxN squares with an equal number of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a 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 Latin square is probably under used in most fields of research because text book examples tend to be restricted to agriculture, the area which spawned most original work on ANOVA. Agricultural examples often reflect geographical designs where rows and columns are literally two dimensions of a grid in a field.
- Rows and columns can be any two sources of variation in an experiment. In this sense a Latin square is a generalisation of a randomized block design with two different blocking systems
    - 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 design is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived. This design allows the experimenter to test two or more variables simultaneously. It also measures interaction effects of the variables and analyses the impacts of each of the variables. In a true experiment, randomization is essential so that the experimenter can infer cause and effect without any bias.

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. **Director General of Commercial Intelligence-**
   - This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.
4. **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, non-manual employees and cultural labourers.

5. **Planning Commission-**
   - It provides the basic statistics of Indian Economy.
6. **Reserve Bank of India-**
   - This provides information on Banking Savings and investment. RBI also prepares currency and finance reports.
7. **Labour Bureau-**
   - It provides information on skilled, unskilled, white collared jobs etc.
8. **National Sample Survey-**
   - This is done by the Ministry of Planning and it provides social, economic, demographic, industrial and agricultural statistics.
9. **Department of Economic Affairs**-
   - It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.
10. **State Statistical Abstract-**
    - This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.
11. **Non-Government Publications**-
    - These includes publications of various industrial and trade associations, such as The Indian Cotton Mill Association Various chambers of commerce.
12. **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.
13. **Syndicate Services-**
    - These services are provided by certain organizations which collect and tabulate the marketing information on a regular basis for a number of clients who are the subscribers to these services.
    - These services are useful in television viewing, movement of consumer goods etc.

- These syndicate services provide information data from both household as well as institution.

In collecting data from household, they use three approaches:

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

Mail Diary Panel- It may be related to 2 fields - Purchase and Media.

Electronic Scanner Services- These are used to generate data on volume.

They collect data for Institutions from

- Whole sellers
- Retailers, and
- Industrial Firms

- Various syndicate services are Operations Research Group (ORG) and The Indian Marketing Research Bureau (IMRB).

Importance of Syndicate Services:

- Syndicate services are becoming popular since the constraints of decision making are changing and we need more of specific decision-making in the light of changing environment. Also, Syndicate services are able to provide information to the industries at a low unit cost.

Disadvantages of Syndicate Services:

- The information provided is not exclusive. A number of research agencies provide customized services which suits the requirement of each individual organization.

International Organization-

These includes

- **The International Labour Organization (ILO):**
  - It publishes data on the total and active population, employment, unemployment, wages and consumer prices.
- **The Organization for Economic Co-operation and development (OECD):**
  - It publishes data on foreign trade, industry, food, transport, and science and technology.
- **The International Monetary Fund (IMA):**
  - It publishes reports on national and international foreign exchange regulations.

Other sources:

Sensor's data: With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.

Satellites data: Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.

Web traffic: Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.

Export all the Data onto the cloud like Amazon web services S3

We usually export our data to cloud for purposes like safety, multiple access and real time

simultaneous analysis.

Data Management:

Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively. The goal of data management is to help people, organizations, and connected things optimize the use of data within the bounds of policy and regulation so that they can make decisions and take actions that maximize the benefit to the organization.

Managing digital data in an organization involves a broad range of tasks, policies, procedures, and practices. The work of data management has a wide scope, covering factors such as how to:

- Create, access, and update data across a diverse data tier
- Store data across multiple clouds and on premises
- Provide high availability and disaster recovery
- Use data in a growing variety of apps, analytics, and algorithms
- Ensure data privacy and security
- Archive and destroy data in accordance with retention schedules and compliance requirements

What is Cloud Computing?

Cloud computing is a term referred to storing and accessing data over the internet. It doesn't store any data on the hard disk of your personal computer. In cloud computing, you can access data from a remote server.

Service Models of Cloud computing are the reference models on which the Cloud Computing is based. These can be categorized into

three basic service models as listed below:

1. INFRASTRUCTURE as a SERVICE (IaaS)

IaaS provides access to fundamental resources such as physical machines, virtual machines, virtual storage, etc.

2. PLATFORM as a SERVICE (PaaS)

PaaS provides the runtime environment for applications, development & deployment tools, etc.

3. SOFTWARE as a SERVICE (SAAS)

SaaS model allows to use software applications as a service to end users.

For providing the above services models AWS is one of the popular platforms. In this Amazon Cloud (Web) Services is one of the popular service platforms for Data Management

Amazon Cloud (Web) Services Tutorial

What is AWS?

The full form of AWS is Amazon Web Services. It is a platform that offers flexible, reliable, scalable, easy-to-use and, cost-effective cloud computing solutions.

AWS is a comprehensive, easy to use computing platform offered Amazon. The platform is developed with a combination of infrastructure as a service (IaaS), platform as a service (PaaS) and packaged software as a service (SaaS) offering.

History of AWS

2002- AWS services launched

2006- Launched its cloud products

2012- Holds first customer event

2015- Reveals revenues achieved of $4.6 billion

2016- Surpassed $10 billon revenue target

2016- Release snowball and snowmobile

2019- Offers nearly 100 cloud services

2021- AWS comprises over 200 products and services

Important AWS Services

Amazon Web Services offers a wide range of different business purpose global cloud-based products. The products include storage, databases, analytics, networking, mobile, development tools, enterprise applications, with a pay-as-you-go pricing model.

Amazon Web Services - Amazon S3:

- **Amazon S3** (Simple Storage Service) is a scalable, high-speed, low-cost web-based service designed for online backup and archiving of data and application programs.
- It allows to upload, store, and download any type of files up to 5 TB in size. This service allows the subscribers to access the same systems that Amazon uses to run its own web sites.
- The subscriber has control over the accessibility of data, i.e. privately/publicly accessible.

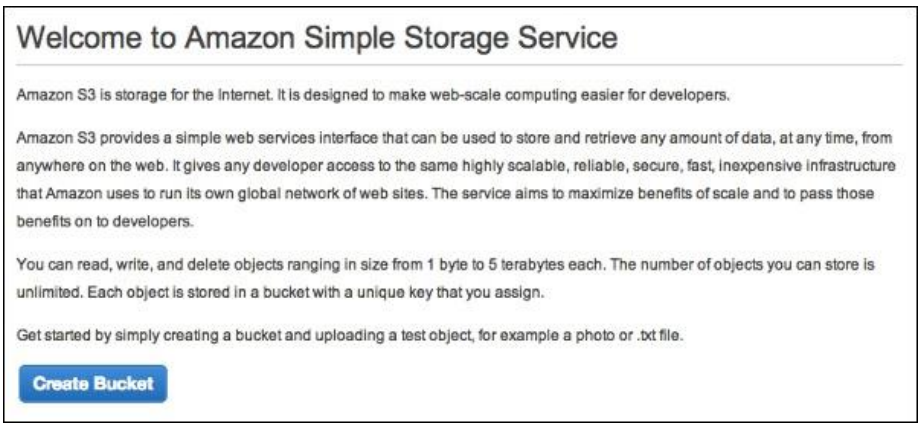
1. How to Configure S3?

Following are the steps to configure a S3 account.

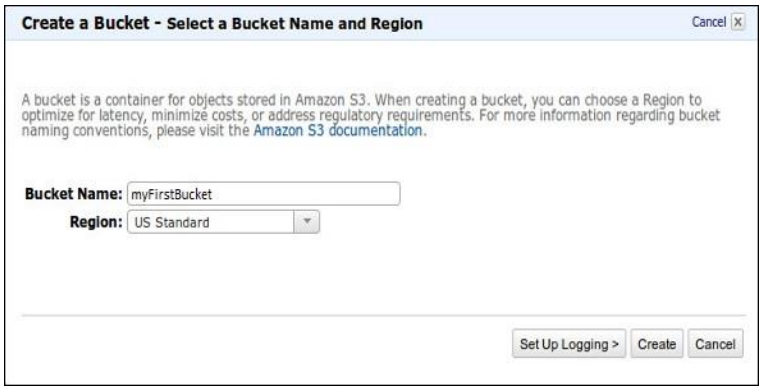
Step 1 – Open the Amazon S3 console using this link – https://console.aws.amazon.com/s3/home

Step 2 – Create a Bucket using the following steps.

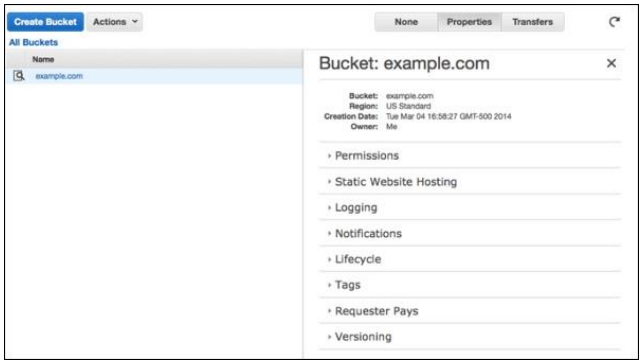
- A prompt window will open. Click the Create Bucket button at the bottom of the page.



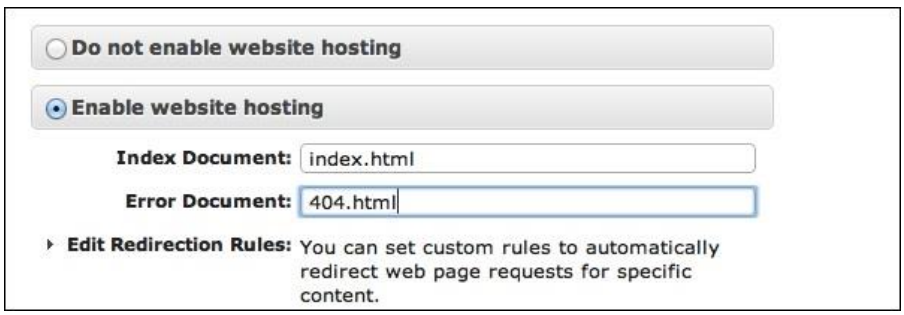
- Create a Bucket dialog box will open. Fill the required details and click the Create button.



- The bucket is created successfully in Amazon S3. The console displays the list of buckets and its properties.



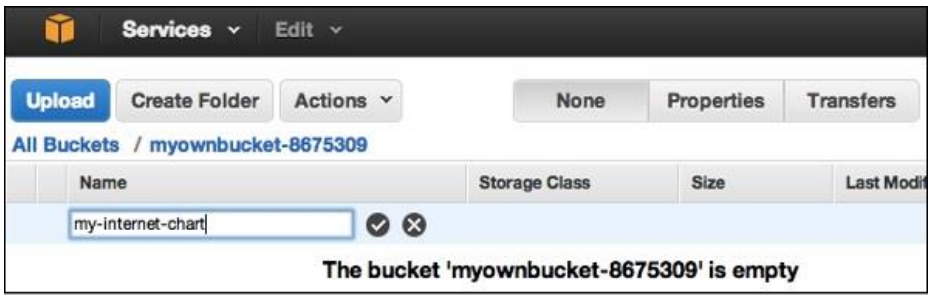
- Select the Static Website Hosting option. Click the radio button Enable website hosting and fill the required details.



Step 3 – Add an Object to a bucket using the following steps.

- Open the Amazon S3 console using the following link. https://console.aws.amazon.com/s3/home

- Click the Upload button.



- Click the Add files option. Select those files which are to be uploaded from the system and then click the Open button.

- Click the start upload button. The files will get uploaded into the bucket.
- Afterwards, we can create, edit, modify, update the objects and other files in wide formats.

Amazon S3 Features

- **Low cost and Easy to Use** – Using Amazon S3, the user can store a large amount of data at very low charges.
- **Secure** – Amazon S3 supports data transfer over SSL and the data gets encrypted automatically once it is uploaded. The user has complete control over their data by configuring bucket policies using AWS IAM.
- **Scalable** – Using Amazon S3, there need not be any worry about storage concerns. We can store as much data as we have and access it anytime.
- **Higher performance** – Amazon S3 is integrated with Amazon CloudFront, that distributes content to the end users with low latency and provides high data transfer speeds without any minimum usage commitments.
- **Integrated with AWS services** – Amazon S3 integrated with AWS services include Amazon CloudFront, Amazon CLoudWatch, Amazon Kinesis, Amazon RDS, Amazon Route 53, Amazon VPC, AWS Lambda, Amazon EBS, Amazon Dynamo DB, etc.

We are discussing Amazon S3:

https://d1.awsstatic.com/whitepapers/aws-overview.pdf

Data Quality:

What is Data Quality?

There are many definitions of data quality, in general, data quality is the assessment of how much the data is usable and fits its serving context.

Why Data Quality is Important?

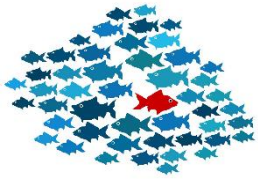
Enhancing the data quality is a critical concern as data is considered as the core of all activities within organizations, poor data quality leads to inaccurate reporting which will result inaccurate decisions and surely economic damages.

Many factors help measuring data quality such as:

- ***Data Accuracy****:* Data are accurate when data values stored in the database correspond to real-world values.
- ***Data Uniqueness***: A measure of unwanted duplication existing within or across systems for a particular field, record, or data set.
- ***Data Consistency***: Violation of semantic rules defined over the dataset.
- ***Data Completeness****:* The degree to which values are present in a data collection.
- ***Data Timeliness****:* The extent to which age of the data is appropriated for the task at hand.

Other factors can be taken into consideration such as **Availability, Ease of Manipulation, Believability.**

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. Simply speaking, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.
- **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.

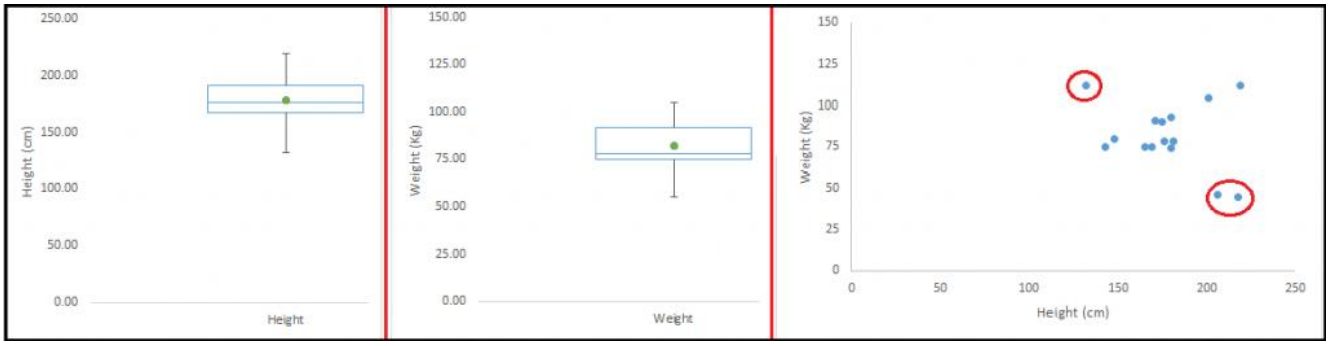
Types of Outliers:

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:

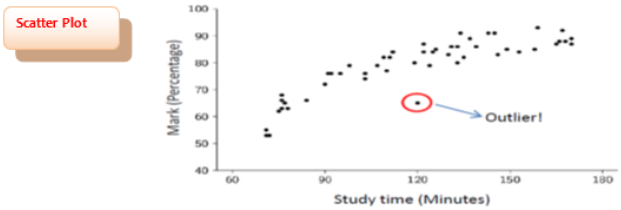
- Rejection of outliers is more acceptable in areas of practice where the underlying model of the process being measured and the usual distribution of measurement error are confidently known.
- An outlier resulting from an instrument reading error may be excluded but it is desirable that the reading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

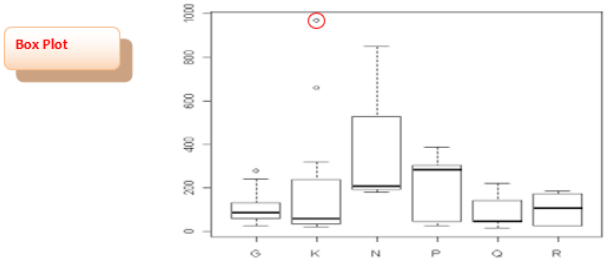
Other treatment methods

OUTLIER package in R: to detect and treat outliers in Data.

Outlier detection from graphical representation:

- Scatter plot and Box plot





- The observations out of box are treated as outliers in data

Missing Data treatment:

Missing Values

- Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

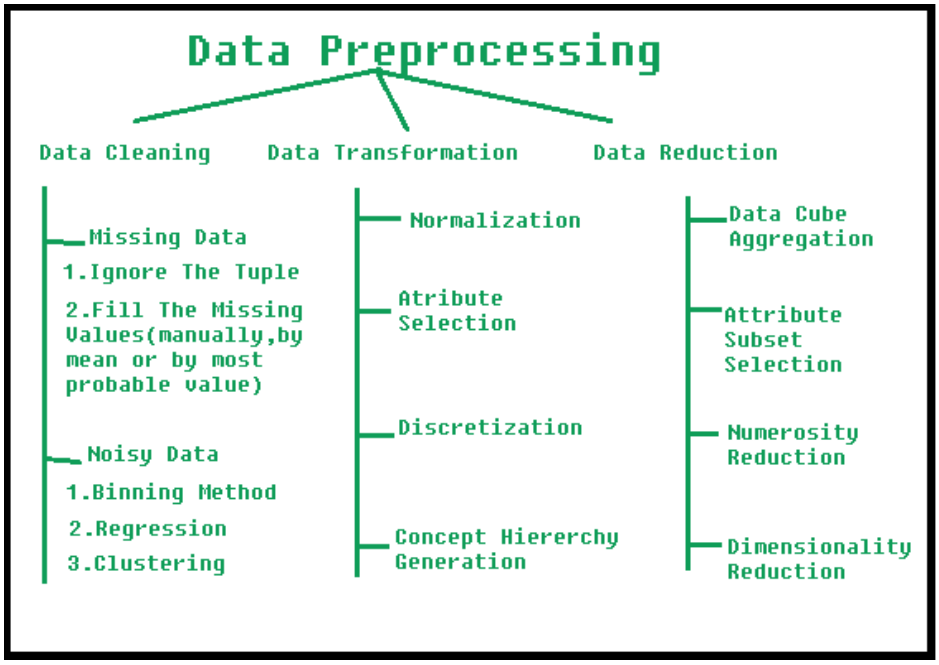
- In **R**, missing values are represented by the symbol **NA** (not available).
- Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number) and R outputs the result for dividing by zero as 'Inf'(Infinity).

PMM approach to treat missing values:

- **PMM**-> Predictive Mean Matching (PMM) is a semi-parametric imputation approach.
- It is similar to the regression method except that for each missing value, it fills in a value randomly from among the observed donor values from an observation
- whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model.

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:

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

1. **Normalization:**
   Normalization is a technique often applied as part of data preparation in Data Analytics through machine learning. 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. For machine learning, every dataset does not require normalization. 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

4. **Concept Hierarchy Generation:**
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

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 use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

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.

3. Numerosity Reduction:
   This enable to store the model of data instead of whole data, for example: Regression Models.

4. Dimensionality Reduction:
   This reduce 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.).

5. Data output/interpretation

The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.).

6. Data storage

The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

***** End of Unit-1 *****