	ABES Engineering College, Ghaziabad
	B.Tech Odd/Even Semester Sessional Test-_____ **Step-Wise Solution**
Course Code: KDS 501	
Course Name: INTRODUCTION TO DATA ANALYTICS AND VISUALIZATION	
Maximum Marks: 75	

SECTION-A

Q-1 Attempt all Parts. **(5*2=10)**

a. Differentiate the Qualitative and Quantitative Data.

Point of comparison	*Qualitative research*	*Quantitative research*
Focus of research	Quality (nature, essence)	Quantity (how many, how much)
Philosophical roots	Phenomenology, symbolic, interaction	Empiricism, logical positivism
Associated phrases	Fieldwork, ethnographic, naturalistic, grounded, subjective	Experimental, empirical, statistical
Goal of investigation	Understanding, description, discovery, hypothesis generating	Prediction, control, confirmation, hypothesis testing
Design characteristics	Flexible, evolving, emergent	Pre-determined structure
Setting	Natural, familiar	Unfamiliar, artificial
Sample	Small, non random, theoretical	Large, random, representative
Data collection	Researcher as primary instrument, interviews, observations	Inanimate instruments (scales, tests, surveys, questionnaires, computers)
Mode of analysis	Inductive (by researcher)	Deductive (by statistical methods)
Findings	Comprehensive, holistic, expansive	Precise, narrow, reductionistic

b. Define the Data? List the characteristics of Data.

Solution: Step-1

Data quality is crucial – It assesses whether information can serve its purpose in a particular context (such as data analysis, for example). So, how do you determine the quality of a given set of information? There are data quality characteristics of which you should be aware.

There are five traits that you'll find within data quality:

- Accuracy
- Completeness
- Reliability
- Relevance
- Timeliness

c. Introduce Backpropagation list their advantages.

Solution: Step-1

Backpropagation is the essence of neural network training. It is the method of fine-tuning the weights of a neural network based on the error rate obtained in the previous epoch (i.e., iteration). Proper tuning of the weights allows you to reduce error rates and make the model reliable by increasing its generalization. Backpropagation in neural network is a short form for "backward propagation of errors." It is a standard method of training artificial neural networks. This method helps calculate the gradient of a loss function with respect to all the weights in the network.

Step-2:

Most prominent advantages of Backpropagation are:

d. Backpropagation is fast, simple and easy to program

e. It has no parameters to tune apart from the numbers of input

f. It is a flexible method as it does not require prior knowledge about the network

g. It is a standard method that generally works well

h. It does not need any special mention of the features of the function to be learned**.**

d. Explain Principal Component Analysis (PCA)? List their steps.

Solution: Step-1

Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables that nonetheless retains most of the sample's information. PCA is used to reduce the problem of Overfitting.

Step-2

Step-01: Get data.

Step-02: Compute the mean vector (μ).

Step-03: Subtract mean from the given data.

Step-04: Calculate the covariance matrix.

Step-05: Calculate the eigen vectors and eigen values of the covariance matrix.

Step-06: Choosing components and forming a feature vector.

Step-07: Deriving the new data set.

e. Compare Database Management System (DBMS) with Data Stream Management System (DSMS).

Solution: Step-1

S.No	Basis	DBMS	DSMS
1	Data	Persistent relation	time windows
2	Data Access	Random	Sequential
3	Processing Model	Query-Driven	Data-Driven
4	Quaries	One-Time	Continuous
5	Query Answer	Exact	Approximate
6	Query Plans	Fixed	Adaptive

SECTION-B

Q-2 Attempt ANY ONE part from the following (5*1=5)

a. Differentiate the Structured, Unstructured, and Semi-Structured data with example.

Solution: Step-1

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table	It is based on XML/RDF(Resource Description Framework).	It is based on character and binary data
Transaction management	Matured transaction and various concurrency techniques	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as a whole
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured data	It is more flexible and there is absence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than structured data	It is more scalable.
Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

b. Introduce Big Data Anlytics. Big Data was defined by the "3Vs" but now there is "5Vs" of Big Data which are also termed as the characteristics of Big Data, explain their requirement in brief.

Solution: Step-1

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

Step-2

Volume: the size and amounts of big data that companies manage and analyse

Value: the most important "V" from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits

Variety: the diversity and range of different data types, including unstructured data, semi-structured data and raw data

Velocity: the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time

Veracity: the "truth" or accuracy of data and information assets, which often determines executive-level confidence

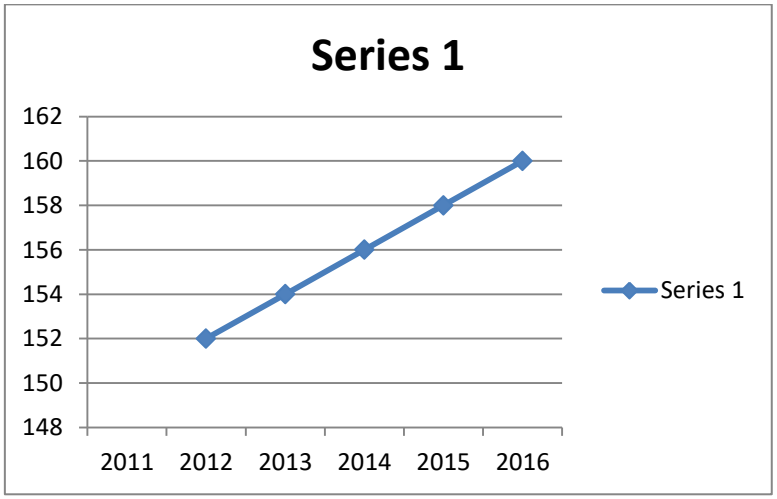
Q-3 Attempt ANY ONE part from the following **(5*1=5)**

a. Monthly sales revenue data were collected for a company. Calculate the trend of given data using Moving Average method with graphical representation.

Year	2011	2012	2013	2014	2015	2016
Sales	125	145	186	131	151	192

Solution: Step-1

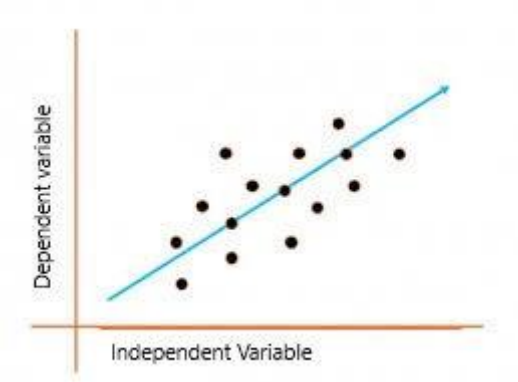
Year	Sales	Three year Moving Total	Three year Moving Average
2011	125		
2012	145	456	152
2013	186	462	154
2014	131	468	156
2015	151	474	158
2016	192	480	160



b. Explain Linear Regression in brief with example.

Solution: Step-1

- Linear regression is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression. If there is a single input variable X(dependent variable), such linear regression is called *simple linear regression*.



The above graph presents the linear relationship between the output(y) variable and predictor(X) variables. The blue line is referred to as the *best fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.

- Calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,
- $\mathbf{Y_i = \beta_0 + \beta_1 X_i}$
- Where Y_i = Dependent variable to the given value of Independent variable.
- $\boldsymbol{\beta_0}$ = constant/Intercept the predicted value of y when x is 0.
- $\boldsymbol{\beta_1}$ = Slope or regression coefficient (how much we except y to change as x increase).
- $\mathbf{X_i}$ = Independent variable (The variable we expect influencing the dependent variable y).
- This algorithm explains the linear relationship between the dependent(output) variable y and the independent(predictor) variable X using a straight line.
- But how the linear regression finds out which is the best fit line?

The goal of the linear regression algorithm is to get the **best values for B_0 and B_1** to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

Q-4 Attempt ANY ONE part from the following (5*1=5)

a. Presents the need of Stream Computing. Differentiate the Batch-Processing Streams and Real-Time streams.

Solution: Step-1

With exponential growth in data generated from sensor data streams, search engines, spam filters, medical services, online analysis of financial data streams, and so forth, there is demand for fast monitoring and storage of huge amounts of data in real-time. Traditional technologies were not aimed to such fast streams of data. Usually they required data to be stored and indexed before it could be processed.

Stream computing was created to tackle those problems that require processing and classification of continuous, high volume of data streams. It is highly used on applications such as Twitter, Facebook, High Frequency Trading and so forth. This subject will focus on the algorithms and data structures behind the analysis and management of streams. Theoretical underpinnings are emphasized, with implementation of some fundamental algorithms.

Step-2

Dimension	Batch processing	Streaming processing
Input	Data chunks	Stream of new data or updates
Data size	Known and finite	Infinite or unknown in advance
Hardware	Multiple CPUs	Typical single limited amount of memory
Storage	Store	Not store or store non-trivial portion in memory
Processing	Processed in multiple rounds	A single or few passes over data
Time	Much longer	A few seconds or even milliseconds
Applications	Widely adopted in almost every domain	Web mining, traffic monitoring, sensor networks

b. Elaborate Data Stream Management System in brief with block diagram.

Solution: Step-1

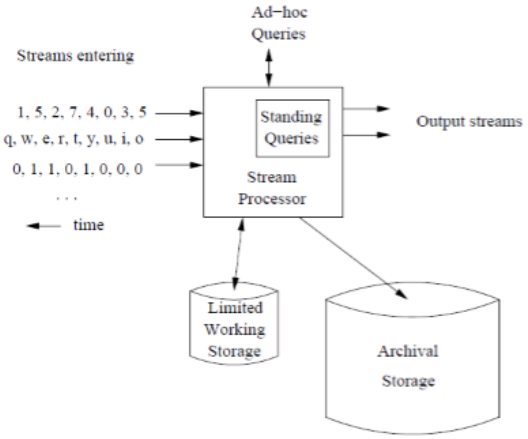


Figure: A data stream management system

A Data Stream Management System (DSMS) is a computer software system to manage continuous data streams.

A DSMS also offers a flexible query processing so that the information needed can be expressed using queries.

In DSMS, queries are executed continuously over the data passed to the system, also called continuous or standing queries. These queries are registered in the system once.

Depending on the system, a query can be formulated mainly in two ways: as a declarative expression, or as a sequence or graph of data processing operators.

A declarative query is parsed to a logical query plan, which can be optimized. This logical query s afterwards translated into a physical query execution plan (QEP).

The query execution plan contains the calls to the implementation of the operators.

Besides of the actual physical operators, query execution plans include also queues for buffering input and output for the operators.

Synopsis structures act as a support element in QEPs.

DSMS may provide specific synopsis algorithms and data structures which are required, when an operator has to store some state to produce results.

A synopsis summarizes the stream or a part of the stream.

SECTION-C

Q-5 Attempt ANY ONE part from the following (10*1=10)

a. Develop and explain the Data Analytics life cycle with appropriate block diagram.

Solution: Step-1

The Data Analytics Lifecycle is a cyclic process which explains, in six stages, how information in made, collected, processed, implemented, and analyzed for different objectives.

1. **Data Discovery**

   This is the initial phase to set your project's objectives and find ways to achieve a complete data analytics lifecycle. Start with defining your business domain and ensure you have enough resources (time, technology, data, and people) to achieve your goals.

   The biggest challenge in this phase is to accumulate enough information. You need to draft an analytic plan, which requires some serious leg work.

   Accumulate resources

   First, you have to analyze the models you have intended to develop. Then determine how much domain knowledge you need to acquire for fulfilling those models.

   The next important thing to do is assess whether you have enough skills and resources to bring your projects to fruition.

   Frame the issue

   Problems are most likely to occur while meeting your client's expectations. Therefore, you need to identify the issues related to the project and explain them to your clients. This process is called "framing." You have to prepare a problem statement explaining the current situation and challenges that can occur in the future. You also need to define the project's objective, including the success and failure criteria for the project.

   Formulate initial hypothesis

   Once you gather all the clients' requirements, you have to develop initial hypotheses after exploring the initial data.

   Data Preparation and Processing

   The Data preparation and processing phase involves collecting, processing, and conditioning data before moving to the model building process.

   Identify data sources

   You have to identify various data sources and analyze how much and what kind of data you can accumulate within a given timeframe. Evaluate the data structures, explore their attributes and acquire all the tools needed.

2. **Collection of data**

   You can collect data using three methods:

   Data acquisition: You can collect data through external sources.

   Data Entry: You can prepare data points through digital systems or manual entry as well.

   Signal reception: You can accumulate data from digital devices such as IoT devices and control systems.

3. **Model Planning**

   This is a phase where you have to analyze the quality of data and find a suitable model for your project.

   Loading Data in Analytics Sandbox

   An analytics sandbox is a part of data lake architecture that allows you to store and process large amounts of data. It can efficiently process a large range of data such as big data, transactional data, social media data, web data, and many more. It is an environment that allows your analysts to schedule and process data assets using the data tools of their choice. The best part of the analytics sandbox is its agility. It empowers analysts to process data in real-time and get essential information within a short duration.

   Data are loaded in the sandbox in three ways:

   ETL − Team specialists make the data comply with the business rules before loading it in the sandbox.

   ELT − The data is loaded in the sandbox and then transform as per business rules.

   ETLT − It comprises two levels of data transformation, including ETL and ELT both.

   The data you have collected may contain unnecessary features or null values. It may come in a form too complex to anticipate. This is where data exploration' can help you uncover the hidden trends in data.

   Steps involved in data exploration:

   Data identification

   Univariate Analysis

   Multivariate Analysis

   Filling Null values

   Feature engineering

   For model planning, data analysts often use regression techniques, decision trees, neural networks, etc. Tools mostly used for model planning and execution include Rand PL/R, WEKA, Octave, Statista, and MATLAB.

4. **Model Building**

Model building is the process where you have to deploy the planned model in a real-time environment. It allows analysts to solidify their decision-making process by gain in-depth analytical information. This is a repetitive process, as you have to add new features as required by your customers constantly.

Your aim here is to forecast business decisions and customize market strategies and develop tailor-made customer interests. This can be done by integrating the model into your existing production domain.

In some cases, a specific model perfectly aligns with the business objectives/ data, and sometimes it requires more than one try. As you start exploring the data, you need to run particular algorithms and compare the outputs with your objectives. In some cases, you may even have to run different variances of models simultaneously until you receive the desired results.

5. **Result Communication and Publication**

   This is the phase where you have to communicate the data analysis with your clients. It requires several intricate processes where you how to present information to clients in a lucid manner. Your clients don't have enough time to determine which data is essential. Therefore, you must do an impeccable job to grab the attention of your clients.

Check the data accuracy

Is the data provide information as expected? If not, then you have to run some other processes to resolve this issue. You need to ensure the data you process provides consistent information. This will help you build a convincing argument while summarizing your findings.

Highlight important findings

Well, each data holds a significant role in building an efficient project. However, some data inherits more potent information that can truly serve your audience's benefits. While summarizing your findings, try to categorize data into different key points.

Determine the most appropriate communication format

How you communicate your findings tells a lot about you as a professional. We recommend you to go for visuals presentation and animations as it helps you to convey information much faster. However, sometimes you also need to go old-school as well. For instance, your clients may have to carry the findings in physical format. They may also have to pick up certain information and share them with others.

6. **Operationalize**

   As soon you prepare a detailed report including your key findings, documents, and briefings, your data analytics life cycle almost comes close to the end. The next step remains the measure the effectiveness of your analysis before submitting the final reports to your stakeholders.

   In this process, you have to move the sandbox data and run it in a live environment. Then you have to closely monitor the results, ensuring they match with your expected goals. If the findings fit perfectly with your objective, then you can finalize the report. Otherwise, you have to take a step back in your data analytics lifecycle and make some changes.

b. Explain the technologies address in Big Data.

1. Massive Parallel Processing(MPP)
2. The Cloud
3. Grid Computing
4. Map Reduce Processing

Solution: Step-1

1. Massive Parallel Processing (MPP)

Massively parallel processing (MPP) database systems have been around for decades. While individual vendor architectures may vary, MPP is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data. An MPP database spreads data out into independent pieces managed by independent storage and central processing unit (CPU) resources. Conceptually, it is like having pieces of data loaded onto multiple network connected personal computers around a house. It removes the constraints of having one central server with only a single set CPU and disk to manage it. The data in an MPP system gets split across a variety of disks managed by a variety of CPUs spread across a number of servers.

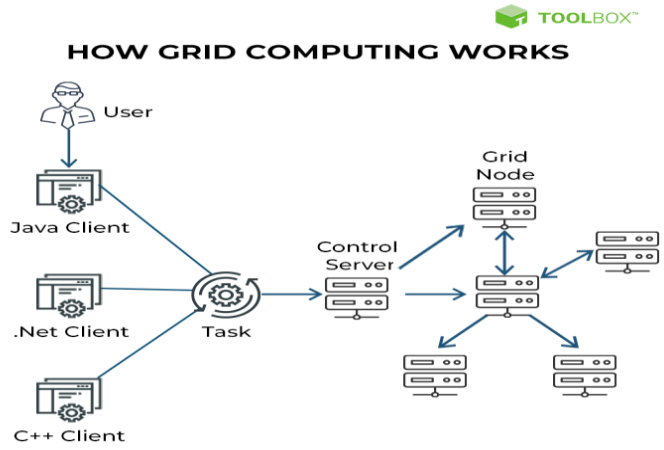
2. Cloud Computing

The concept of cloud computing is getting a lot of attention these days. As with many technologies, cloud computing is going through a hype cycle. Let's start by defining what cloud computing is all about and how it can help with advanced analytics and big data. As with any new and emerging technology, there are conflicting definitions of what cloud computing is. We'll consider two that serve as a good foundation for our discussion.

- They list five essential characteristics of a cloud environment.
- 1. On - demand self - service
- 2. Broad network access
- 3. Resource pooling
- 4. Rapid elasticity
- 5. Measured service

3. **Grid Computing**

Grid computing is a distributed architecture of multiple computers connected by networks to accomplish a joint task. These tasks are compute-intensive and difficult for a single machine to handle. Several machines on a network collaborate under a common protocol and work as a single virtual supercomputer to get complex tasks done. This offers powerful virtualization by creating a single system image that grants users and applications seamless access to IT capabilities.



4. **Map Reduce Processing**

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

Q-6 Attempt ANY ONE part from the following (10*1=10)

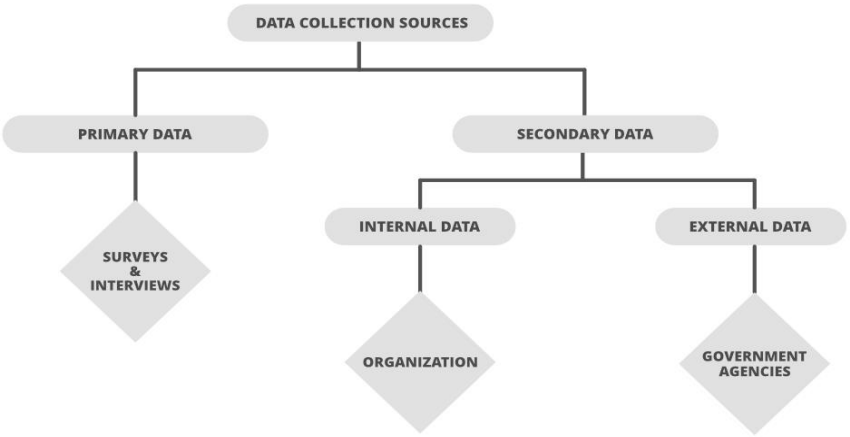
a. Data empowers to make decision informed, justify the statement and explain the various methods of primary and secondary data collection in detail.

Solution: Step-1

- 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 analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.

- Data collection starts with asking some questions such as what type of data is to be collected and what the source of collection is. 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 behavior 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.

Step-2



Primary Data Collection

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 analyzing 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 behavior 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 behavior towards the products. The data obtained will be sent for processing.

3. 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.

- **Secondary Data Collection**

1. 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.

2. 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 is 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 labor bureau, syndicate services, and other non-governmental publications.

3. Other sources:

- **Sensors 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.

b. Cloud computing is a big shift from the traditional way businesses thinking about IT resources. Explain Public Cloud and Private cloud with their major five characteristics.

Solution: Step-1

The two primary types of cloud environments: (1) public clouds and (2) private clouds.

Public Clouds

Public clouds have gotten the most hype and attention. With a public cloud users are basically loading their data onto a host system and they are then allocated resources as they need them to use that data. They will get charged according to their usage. There are definitely some advantages to such a setup:

▪ The bandwidth is as - needed and users only pay for what they use.

▪ It isn't necessary to buy a system sized to handle the maximum capacity ever required and then risk having half of the capacity sitting idle much of the time.

▪ If there are short bursts where a lot of processing is needed then it is possible to get it with no hassle. Simply pay for the extra resources.

▪ There's typically very fast ramp - up. Once granted access to the cloud environment, users load their data and start analyzing.

▪ It is easy to share data with others regardless of their location since a public cloud by definition is outside of a corporate firewall. Anyone can be given permission to log on to the environment created.

Private Clouds

▪ A private cloud has the same features of a public cloud, but it's owned exclusively by one organization and typically housed behind a corporate firewall. A private cloud is going to serve the exact same function as a public cloud, but just for the people or teams within a given organization.

▪ One huge advantage of an onsite private cloud is that the organization will have complete control over the data and system security.

▪ Data is never leaving the corporate firewall so there's absolutely no concern about where it ' s going.

Step-2

▪ **On-demand self-services:**

The Cloud computing services does not require any human administrators, user themselves are able to provision, monitor and manage computing resources as needed.

▪ **Broad network access:**

The Computing services are generally provided over standard networks and heterogeneous devices.

▪ **Rapid elasticity:**

The Computing services should have IT resources that are able to scale out and in quickly and on as needed basis. Whenever the user require services it is provided to him and it is scale out as soon as its requirement gets over.

▪ **Resource pooling:**

The IT resource (e.g., networks, servers, storage, applications, and services) present are shared across multiple applications and occupant in an uncommitted manner. Multiple clients are provided service from a same physical resource.

▪ **Measured service:**

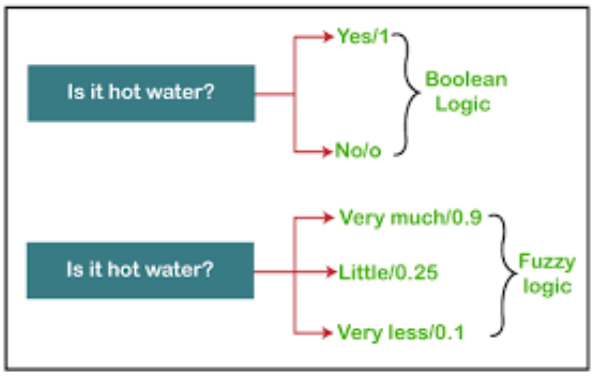
The resource utilization is tracked for each application an d occupant, it will provideboth the user and the resource provider with an account of what has been used. This is done for various reasons like monitoring billing and effective use of resource.

Q-7 Attempt ANY ONE part from the following **(10*1=10)**

a. **The word fuzzy refers to things which are not clear or are vague, justify the statement in terms of fuzzy logic. Explain three features of fuzzy logic Core, Support, and Boundary with graphical representation.**

Solution: Step-1

The word fuzzy refers to things which are not clear or are vague. Any event, process, or function that is changing continuously cannot always be defined as either true or false, which means that we need to define such activities in a Fuzzy manner. Take a look at the following diagram. It shows that in fuzzy systems, the values are indicated by a number in the range from 0 to 1. Here 1.0 represents **absolute truth** and 0.0 represents **absolute falseness**. The number which indicates the value in fuzzy systems is called the **truth value**.



Step2:

Core :

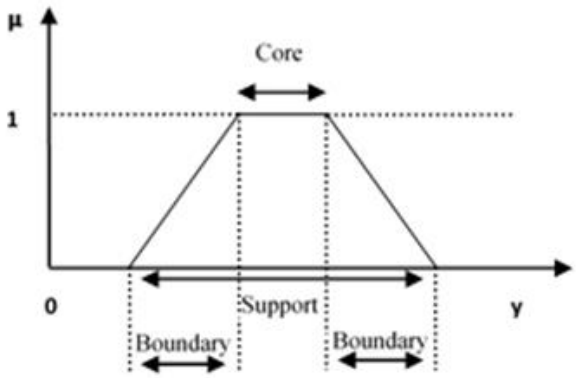
- The core of a membership function for some fuzzy set is defined as that region of the universe that is characterized by complete and full membership in the set.
- The core comprises those elements *x* of the universe such that

Support :

- The support of a membership function for some fuzzy set *A* is
- defined as that region of the universe that is characterized by nonzero membership in the set *A*.
- The support comprises those elements *x* of the universe such that

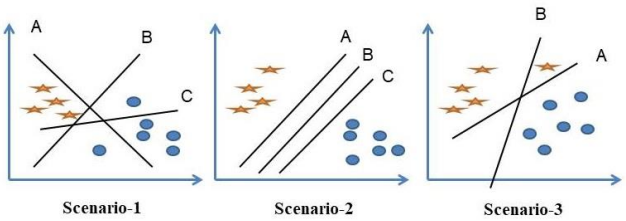
Boundaries :

- The boundaries of a membership function for some fuzzy set are defined as that region of the universe containing elements that have a non-zero membership but not complete membership.
- The boundaries comprise those elements *x* of the universe such that



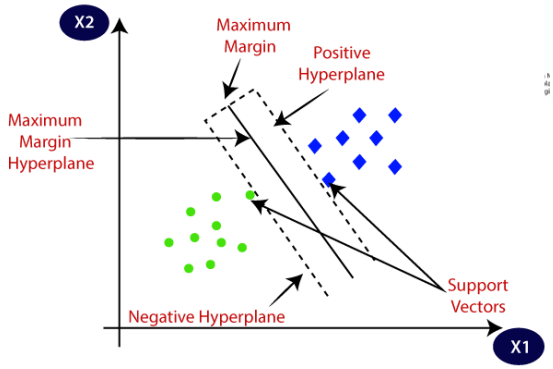
Features of Membership Function

b. Explain Support Vector Machine (SVM) and Kernel function in brief. Choose the best hyper-plane in given Scenario-1,2 and 3 with justification of the statement.



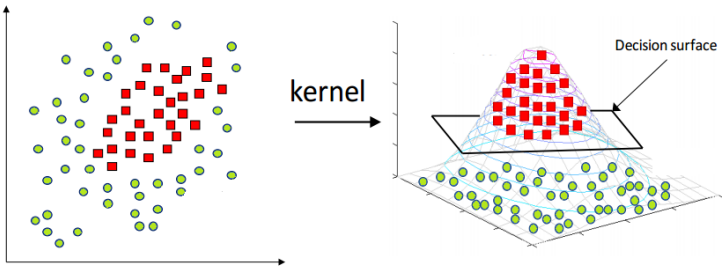
Solution: Step-1

Support Vector Machine (SVM) is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. SVM chooses the extreme points/vectors that help in creating the hyper-plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. To create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.



Kernel Function

In machine learning, a kernel refers to a method that allows us to apply linear classifiers to non-linear problems by mapping non-linear data into a higher-dimensional space without the need to visit or understand that higher-dimensional space.



- **The polynomial:** kernel is a kernel function commonly used with support vector machines and other kernelized models, that represents the similarity of vectors in a feature space over polynomials of the original variables. It is popular in image processing. Where d is the degree of polynomial. T is the transpose.

$$K(x_i, x_j) = (x_i^T . x_j + 1)^d$$

- **Radial-Basis:** It is general purpose kernel: used when their is no prior knowledge about the data.
- $$K(x, y) = \exp\left(\frac{-|x-y|^2}{2\sigma^2}\right)$$
- **Hyperbolic Tangent:** Mainly used in neural networks.
- Where k>0 and c<0

$$K(x_i, x_j) = \tanh\left(kx_i . x_j + c\right)$$

Step 2:

Scenario-1 Here we have 3 hyperplanes (A, B, and C). Now identify the right hyperplane for classify the stars and circles. You need to remember the thumb rule to identify the right hyperplane: select the hyperplane that segregates the two classes better. In this scenario hyperplane B perform their job excellently.

- **Scenario-2** Here we have 3 hyperplanes (A, B, and C)and all are segregating the classes well. Now how can we identify the right Hyperplane. Here, maximizing the distance between nearest data points and Hyperplane will help us to decide the right Hyperplane. In above picture margins of Hyperplane C is more than the others. So C has selected as the best Hyperplane.
- **Scenario-3** Here we have 2 hyperplanes (A, and B). Use the rules as discussed in previous section to identify the right Hyperplane SVM selects the Hyperplane which classifies the classes accurately prior to maximizing the margins. Hyperplane B has a classification error and A has classified all accurately. Therefore right Hyperplane is A.

Q-8 Attempt ANY ONE part from the following **(10*1=10)**

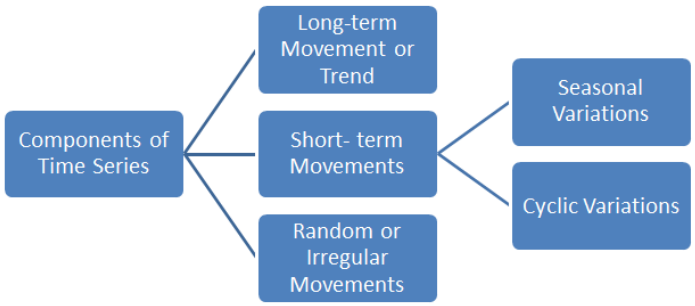
a. Explain about the Time Series analysis and their comonents. Calculate the least square method for given equation y=a+bx based on the sales values.

Year	Sales
2015	30
2016	50
2017	75
2018	80
2019	40

Solution: Step-1

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording them. Time series analysis helps organizations understand the underlying causes of trends or

systemic patterns over time. Using data visualizations, business users can see seasonal trends and dig deeper into why these trends occur data points intermittently or randomly.



Step-2:

Year	Sales	x=x-a	x^2	xy	Y=a+bx
2015	30	-2	4	-60	45
2016	50	-1	1	-50	50
2017	75	0	0	0	55
2018	80	1	1	80	60
2019	40	2	4	80	65
N=5	$\sum y=275$		$\sum x^2=10$	$\sum xy=50$	

$$y = a + bx$$
$$a = \frac{\Sigma y}{N} = \frac{275}{5} = 55$$
$$b = \frac{\Sigma xy}{x^2} = \frac{50}{10} = 5$$

b. Present the role of Activation function in Artificial Neural Network (ANN) describe in brief. Explain the most propinent activation functions.

1. Linear Function
2. ReLu Function
3. Sigmoid Function
4. Softmax Function

Solution: Step-1

Activation function decides, whether a neuron should be activated or not by calculating weighted sum and further adding bias with it. The purpose of the activation function is to introduce non-linearity into the output of a neuron. Explanation :- We know, neural network has neurons that work in correspondence of *weight, bias* and their respective activation function. In a neural network, we would update the weights and biases of the neurons on the basis of the error at the output. This process is known as *back-propagation*. Activation functions make the back-propagation possible since the gradients are supplied along with the error to update the weights and biases. Why do we need Non-linear activation functions :- A neural network without an activation function is essentially just a linear regression model. The activation function does the non-linear transformation to the input making it capable to learn and perform more complex tasks.

Step-2

1. **Linear Function**

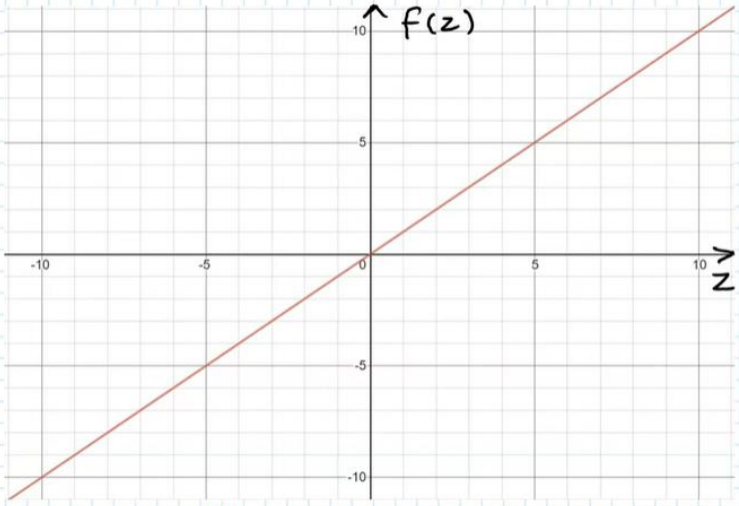
Commonly Used Activation Functions

3. Linear Function

$f(z) = z$

Range = Possible Outputs

$(-\infty, \infty)$



Used in: Hidden Layer, Output Layer for Regression

2. **Relu**

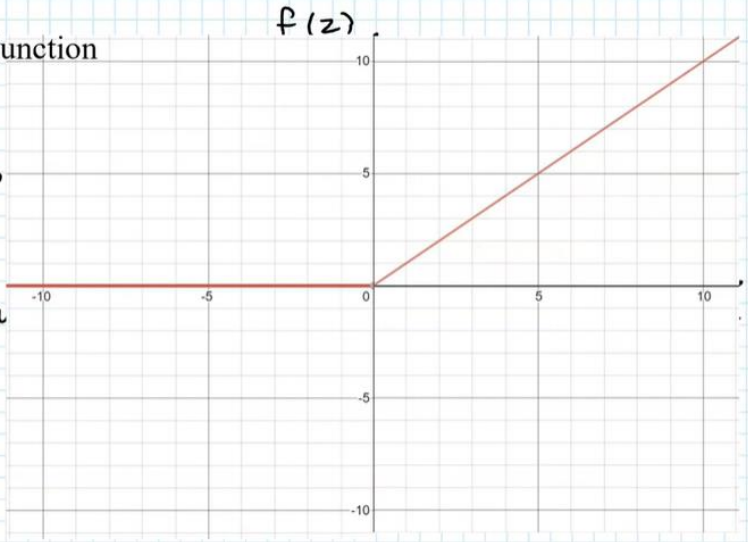
Commonly Used Activation Functions

4a. Rectified Linear Unit (ReLU) Function

$$f(z) = \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases}$$

Range = Possible Outpu

$[0, \infty)$



Used in: Hidden Layer, Output Layer for Regression (only positive output)

3. **Sigmoid**

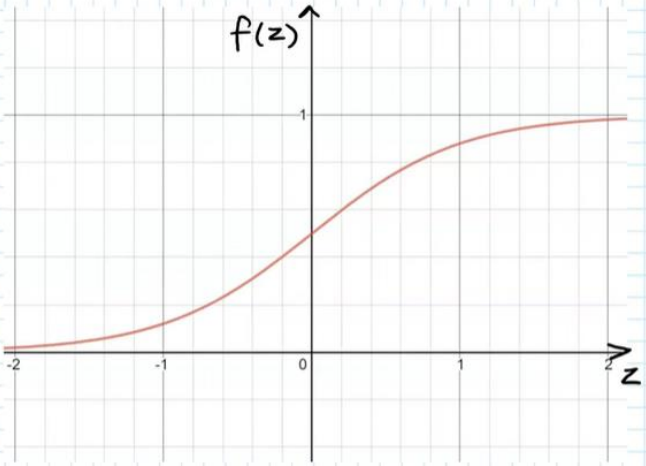
Commonly Used Activation Functions

6a. Sigmoid Function

$$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

Range = Possible Outputs

(0, 1)



Used in: Hidden Layer, Output Layer for Classification

4. Softmax

Commonly Used Activation Functions

6b. Softmax Function

$$z_1, z_2, \cdots, z_n$$

$$f(z_1) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + \cdots + e^{z_n}}$$

$$f(z_i) = \frac{e^{z_i}}{e^{z_1} + e^{z_2} + \cdots + e^{z_n}} = \frac{e^{z_i}}{\sum_{k=1}^{n} e^{z_k}}$$

Used in: Output Layer for Multiclass Classification

Q-9 Attempt ANY ONE part from the following **(10*1=10)**

a. Give the introduction of Stream Computing. Explain the different sources of stream data collection.

Solution: Step-1

A data stream is a countable infinite sequence of elements and is used to represent data elements that are made available over time. Examples are readings from sensors in an environment monitoring application, stock quotes in financial application etc. Data streams are dynamic data that are generated on the continual basis. This allows you to analyze data in real-time and gain insight on a wide range of scenarios. By using stream processing technology, data stream can be processed, stored, analyzed, and acted upon as its generated in real-time.

There are two types of data stream processing:

a. Batch Processing stream: these methods requires data to be downloaded as batches before it can be processed, stored, or analyzed.

b. Real-Time Stream: data flows in continuously, allowing that data to be processed simultaneously, in real-time the second its generated.

Step-2

Sensor Data

Sensor data are the data produced by the sensors placed at different places. Different sensors such as temperature sensor, GPS sensors and more are installed at different places for capturing the temperature, height, and many other information of that particular place. The data produced by sensor is a stream of real numbers. This data given by the sensor is stored in the main memory. These sensors sends large amount of data every 10^{th} of second.

Image Set

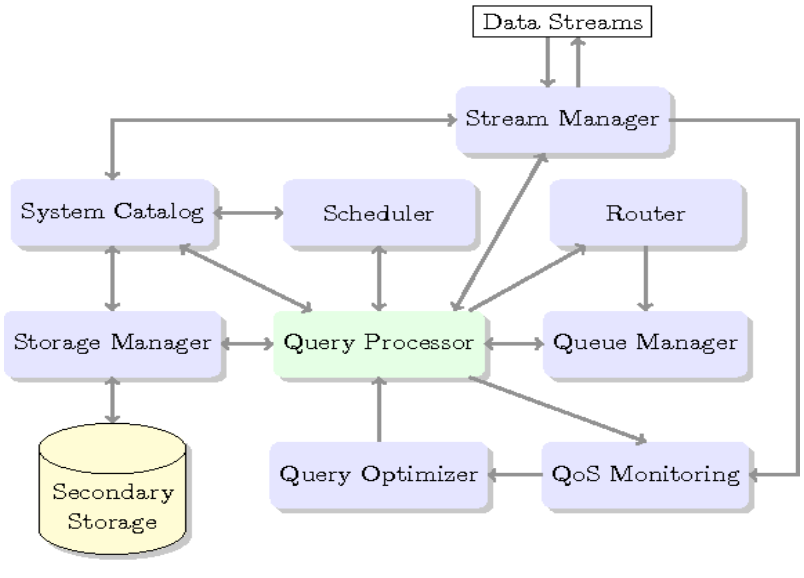
Satellites often send down to earth streams consisting of many terabytes of images per day. Surveillance cameras produce images with lower resolution than satellites, but there can be many of them, each producing a stream of images at intervals like one second. London is said to have six million such cameras, each producing a stream.

Internet and Web Traffic

Web sites receive streams of various types. For example, Google receives several hundred million search queries per day. Yahoo! accepts billions of "clicks" per day on its various sites. Many interesting things can be learned from these streams.

b. Presents a block diagram of Data Stream Management architecture with detailed component explanation.

Solution: Step-1



Solution: Step-1

1. Data stream:

a. DSMS gets data streams as input.

b. Data stream elements are represented as tuples, which adhere to a relational schema with attributes and values.

2. Stream manager :

a. Wrappers are provided which can receive raw data from its source, buffer and order it by timestamp.

b. The task of stream manager is to convert the data to the format of the data stream management system.

3. Router :

a. It helps to add tuples or data stream to the queue of the next operator according to the query execution plan.

4. Queue manager :

a. The management of queues and their corresponding buffers is handled by a queue manager.

b. The queue manager can also be used to swap data from the queues to a secondary storage, if main memory is full.

5. System catalog and storage manager:

a. To enable access to data stored on disk many systems employ a storage manager which handles access to secondary storage.

b. This is used, when persistent data is combined with data from stream sources.

c. Also it is required when loading meta-information about, queries, query plans, streams, inputs, and outputs.

d. These are held in a system catalog in secondary storage.

6. Scheduler : a. Scheduler determines which operator is executed next.

b. The Scheduler interacts closely with the query processor

7. Query processor : It helps to execute the operator by interacting with scheduler.

8. QoS monitoring :

a. Many systems also include some kind of monitor which gathers statistics about performance, operator output rate, or output delay.

b. These statistics can be used to optimize the system execution in several ways.

9. Query optimizer :

a. The throughput of a system can be increased by a load shedder which is a stream element selected by a sampling method.

b. The load shedder can be a part of a query optimizer, a single component, or part of the query execution plan.

c. The statistics can be used to re-optimize the current query execution plan and reorder the operators. For this purpose a query optimizer can be included.