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. What are the different types of Data Analytics modelling.

Solution: Step1

1. Descriptive: If you have a question about something that has already happened, then descriptive analytics can help you answer it. Descriptive analytics is often used as a means to explain something to stakeholders. For example, it can track return on investment (ROI) and other metrics of past performance.

2. Diagnostic: Like a diagnosis, diagnostic analytics provides insight as to why something happened. They work hand-in-hand with descriptive analytics to further explain critical findings.

3. Predictive: Answer questions about what could happen in the future. This analytic method leverages past data to evaluate trends and estimate the likelihood of something recurring. Statistical analysis, regression and machine learning is used to make predictive analytics function.

4. Prescriptive: If you find yourself in a critical position to make a decision about the future but feel unsure about what choice to make, prescriptive analytics can be a lifesaver.

5. Cognitive: Cognitive Analytics applies human-like intelligence to certain tasks, and brings together a number of intelligent technologies, including semantics, artificial intelligence algorithms, deep learning and machine learning.

b. Define the Qualitative data and Quantitative data.

Solution: Step-1

Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods. Qualitative data in statistics is also known as categorical data – data that can be arranged categorically based on the attributes and properties of a thing or a phenomenon.
Example: Marital status (Single, Widowed, Married) and Economic Status (High, Medium, and Low)

Step-2

Quantitative data is the value of data in the form of counts or numbers where each data set has a unique numerical value. This data is any quantifiable information that researchers can use for mathematical calculations and statistical analysis to make real-life decisions based on these mathematical derivations.
Example: Cost of a cell phone, Market share price

c. What is 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:
- Backpropagation is fast, simple and easy to program
- It has no parameters to tune apart from the numbers of input
- It is a flexible method as it does not require prior knowledge about the network
- It is a standard method that generally works well
- It does not need any special mention of the features of the function to be learned.

d. Differentiate the Linear and Logistic regression.

Solution: Step-1

Linear Regression is used to handle regression problems whereas Logistic regression is used to handle the classification problems.
Linear regression provides a continuous output but Logistic regression provides discreet output.
The purpose of Linear Regression is to find the best-fitted line while Logistic regression is one step ahead and fitting the line values to the sigmoid curve.
The method for calculating loss function in linear regression is the mean squared error whereas for logistic regression it is maximum likelihood estimation.

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)

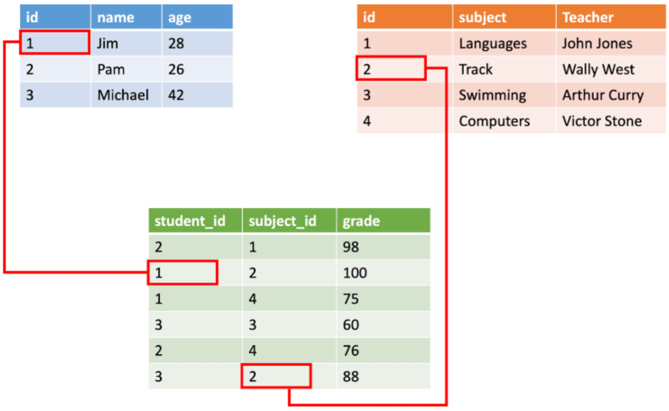
a. What are the different classifications of data? Explain Structured, Unstructured, and Semi-Structured data with example.

Solution: Step-1

We can classify data as **structured data, semi-structured data, or unstructured data**. **Structured data** resides in predefined formats and models, **unstructured data** is stored in its natural format until it's extracted for analysis, and **Semi-structured** data basically is a mix of both structured and unstructured data.

Step-2

Structured data is generally tabular data that is represented by columns and rows in a database. Databases that hold tables in this form are called relational databases. The mathematical term “relation” specify to a formed set of data held as a table. In structured data, all row in a table has the same set of columns. SQL (Structured Query Language) programming language used for structured data.

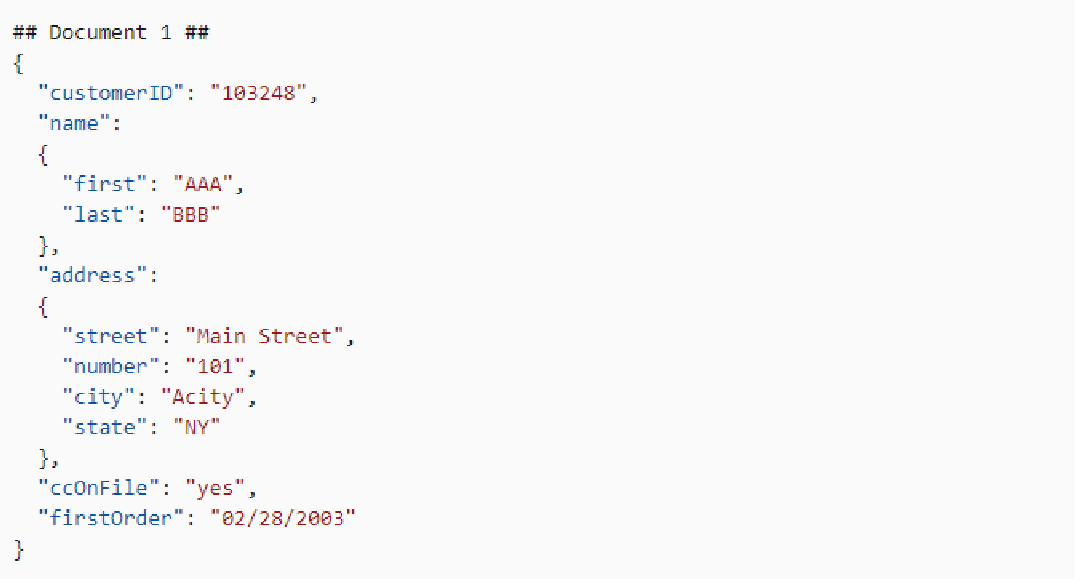


id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88

Semi-structured data is information that doesn’t consist of structured data (relational database) but still has some structure to it. Semi-structured data consist of documents held in JavaScript Object Notation (JSON) format. It also includes key-value stores and graph databases.



```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

Unstructured data is information that either does not organize in a pre-defined manner or not have a pre-defined data model. Unstructured information is a set of text-heavy but may contain data such as numbers, dates, and facts as well. **Videos, audio, and binary** data files might not have a specific structure. They’re assigned to as **unstructured** data.

Text Files and Documents
Server, Website and Application Logs
Sensor Data
Images
Video Files
Audio Files
Emails
Social Media Data

b. What is Big Data Analytics? Define the "5Vs" of Big Data which are also termed as the characteristics of Big Data.

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.

Examples of Big Data

Following are some of the Big Data examples-

The New York Stock Exchange is an example of Big Data that generates about one terabyte of new trade data per day.

Social Media: The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

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

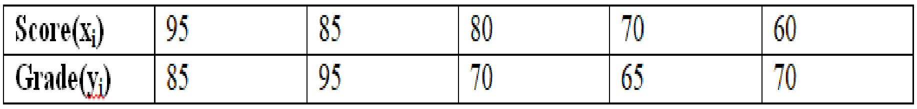
Variety: the diversity and range of different data types, including unstructured data, semi-structured data and raw data

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)

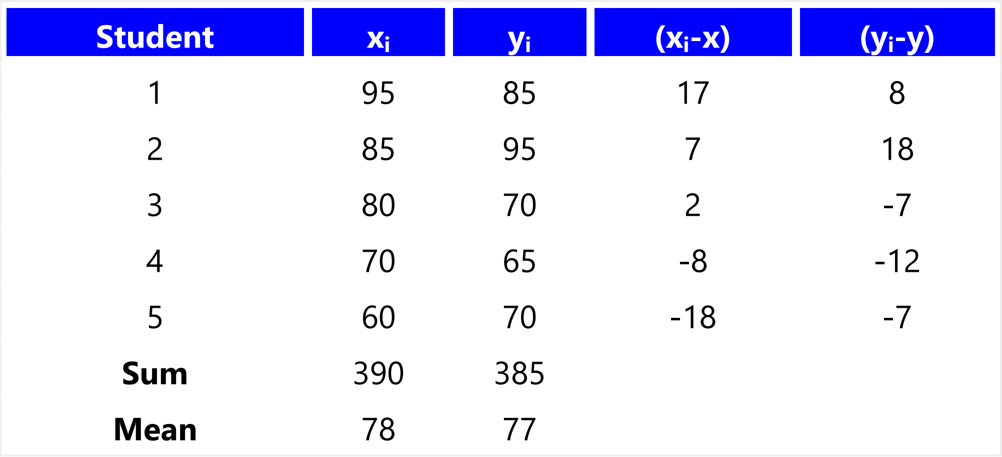
a. Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has a question. What linear regression equation best predicts statistics performance, based on math aptitude scores?



Score(x_i)	95	85	80	70	60
Grade(y_i)	85	95	70	65	70

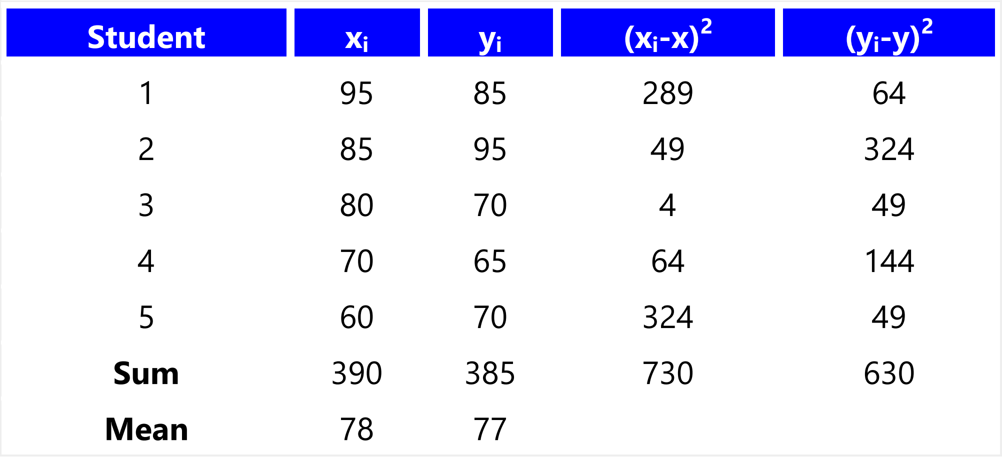
Solution: Step-1

In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two columns show deviations scores - the difference between the student's score and the average score on each measurement. The last two rows show sums and mean scores that we will use to conduct the regression analysis.



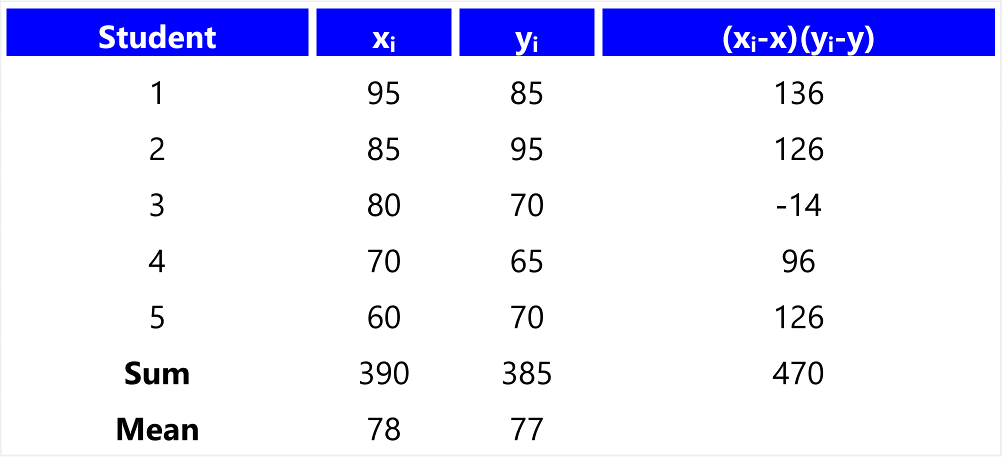
Student	x_i	y_i	(x_i-x)	(y_i-y)
1	95	85	17	8
2	85	95	7	18
3	80	70	2	-7
4	70	65	-8	-12
5	60	70	-18	-7
Sum	390	385		
Mean	78	77		

And for each student, we also need to compute the squares of the deviation scores (the last two columns in the table below).



Student	x_i	y_i	$(x_i-x)^2$	$(y_i-y)^2$
1	95	85	289	64
2	85	95	49	324
3	80	70	4	49
4	70	65	64	144
5	60	70	324	49
Sum	390	385	730	630
Mean	78	77		

And finally, for each student, we need to compute the product of the deviation scores (the last column in the table below).



Student	x_i	y_i	$(x_i-x)(y_i-y)$
1	95	85	136
2	85	95	126
3	80	70	-14
4	70	65	96
5	60	70	126
Sum	390	385	470
Mean	78	77	

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below. Notice that all of our inputs for the regression analysis come from the above three tables.

First, we solve for the regression coefficient (b_1):

$$b_1 = \Sigma \, [\, (x_i - x)(y_i - y) \,] \, / \, \Sigma \, [\, (x_i - x)^2]$$

$$b_1 = 470/730$$

$$b_1 = 0.644$$

Once we know the value of the regression coefficient (b_1), we can solve for the regression slope (b_0):

$$b_0 = y - b_1 * x$$

$$b_0 = 77 - (0.644)(78)$$

$$b_0 = 26.768$$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.

$$\hat{y} = b_0 + b_1x$$

$$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80$$

$$\hat{y} = 26.768 + 51.52 = 78.288$$

b. What is Bayesian statistics and its two basic pillars. Explain the Bayes theorem and prove.

Solution: Step-1

Classical statistics provides methods to analyze data, from simple descriptive measures to complex and sophisticated models. The available data are processed and then conclusions about a hypothetical population of which the data available are supposed to be a representative sample are drawn. Suppose, for example, we need to guess the outcome of an experiment that consists of tossing a coin. How many biased coins have we ever seen? Probably not many, and hence we are ready to believe that the coin is fair and that the outcome of the experiment can be either head or tail with the same probability. On the other hand, imagine that someone would tell us that the coin is forged so that it is more likely to land head. How can we take into account this information in the analysis of our data?

Bayesian methods provide a principled way to incorporate this external information into the data analysis process. To do so, however, Bayesian methods have to change entirely the vision of the data analysis process with respect to the classical approach. In a Bayesian approach, the data analysis process starts already with a given probability distribution. As this distribution is given *before* any data is considered, it is called *prior* distribution.

Two Pillars of Bayesian statistics are:

Conditional probability is known as the possibility of an event or outcome happening, based on the existence of a previous event or outcome. It is calculated by multiplying the probability of the preceding event by the renewed probability of the succeeding, or conditional, event.

Bayes' Theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence.

Step-2

Bayes’’theorem is a mathematical formula used to determine the conditional probability of the events.

Bayes theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

Invented – Thomas Bayes

Year- 1763

$$P(A|B) = \frac{P \cap B}{P(B)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$LHS \qquad\qquad \text{RHS}$$

$$P(A|B).P(B) = P(A \cap B)$$

$$P(B|A).P(A) = P(B \cap A)$$

$$P(A \cap B) = P(A|B).P(B) = P(B|A).P(A)$$

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

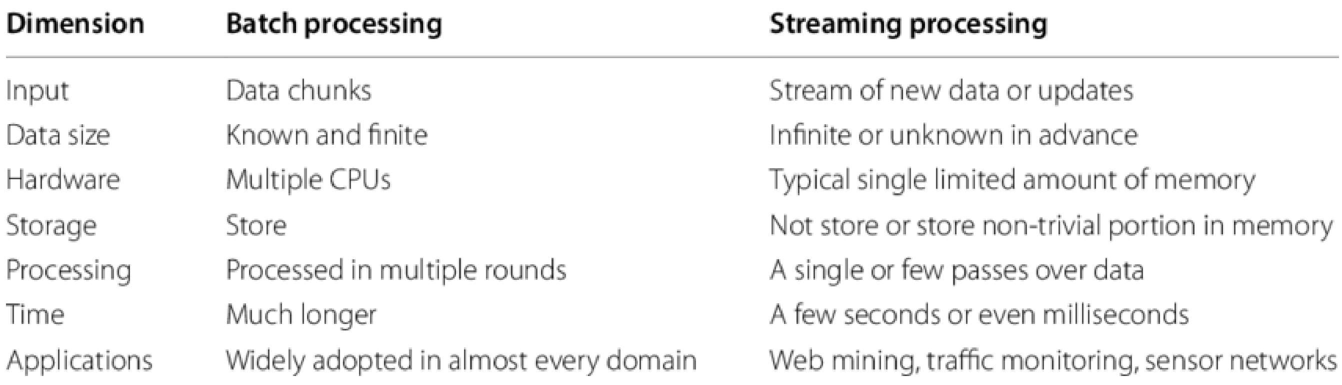
- P(A|B)- Probability of hypothesis A, given that evidence or data B.
- P(B|A)- Probability of data/evidence, given that hypothesis is true.
- P(A)- Probability of A
- P(B)- Probability of B

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

a. **What is the use 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. What is Data Stream Management System explain with block diagram.

Solution: Step-1

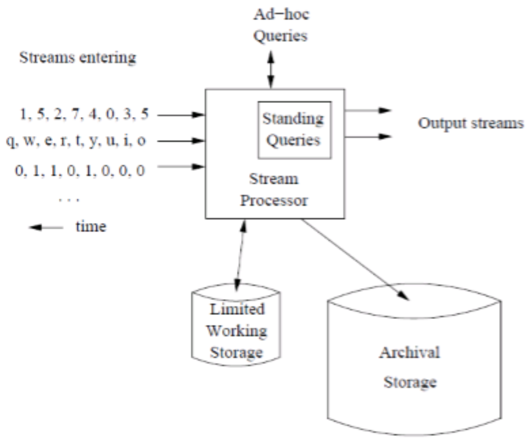


Figure: A data stream management system

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. Cloud computing is a big shift from the traditional way businesses think 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.

b. 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 is various kinds of information formatted in a particular way. Therefore, data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.

Our society is highly dependent on data, which underscores the importance of collecting it. Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity.

During data collection, the researchers must identify the data types, the sources of data, and what methods are being used. We will soon see that there are many different data collection methods. There is heavy reliance on data collection in research, commercial, and government fields.

Step-2

Primary data collection methods are different ways in which primary data can be collected. It explains the tools used in collecting primary data, some of which are highlighted below:

1. Interviews

An interview is a method of data collection that involves two groups of people, where the first group is the interviewer (the researcher(s) asking questions and collecting data) and the interviewee (the subject or respondent that is being asked questions). The questions and responses during an interview may be oral or verbal as the case may be.

2. Surveys & Questionnaires

Surveys and questionnaires are 2 similar tools used in collecting primary data. They are a group of questions typed or written down and sent to the sample of study to give responses. After giving the required responses, the survey is given back to the researcher to record. It is advisable to conduct a pilot study where the questionnaires are filled by experts and meant to assess the weakness of the questions or techniques used.

3. Observation

The observation method is mostly used in studies related to behavioral science. The researcher uses observation as a scientific tool and method of data collection. Observation as a data collection tool is usually systematically planned and subjected to checks and controls.

4. Experiments

An experiment is a structured study where the researchers attempt to understand the causes, effects, and processes involved in a particular process. This data collection methods is usually controlled by the researcher, who determines which subject is used, how they are grouped, and the treatment they receive.

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.

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 labour bureau, syndicate services, and other non-governmental publications.

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

a. **Data analytics is important because it helps businesses optimize their performances. Explain various popular applications of Data Analytics.**

Solution: Step-1

1. **Transportation**

▪ Data analytics can be applied to help in improving Transportation Systems and the intelligence around them. The predictive method of the analysis helps find transport problems like Traffic or network congestion. It helps synchronize the vast amount of data and uses them to build and design plans and strategies to plan alternative routes and reduce congestion and traffic, which in turn reduces the number of accidents and mishappenings.

2.Web Search or InternetWeb Results

▪ The web search engines like Yahoo, Bing, Duckduckgo, and Google use a set of data to give you when you search a data. Whenever you hit on the search button, the search engines use algorithms of data analytics to deliver the best-searched results within a limited time frame. The set of data that appears whenever we search for any information is obtained through data analytics.

2. **Manufacturing**

▪ Data analytics helps the manufacturing industries maintain their overall work through certain tools like prediction analysis, regression analysis, budgeting, etc. The unit can figure out the number of products needed to be manufactured according to the data collected and analyzed from the demand samples and likewise in many other operations increasing the operating capacity as well as the profitability.

3. **Security**

▪ Data analyst provides utmost security to the organization, Security Analytics is a way to deal with online protection zeroed in on the examination of information to deliver proactive safety efforts. No business can foresee the future, particularly where security dangers are concerned, yet by sending security investigation apparatuses that can dissect security occasions it is conceivable to identify danger before it gets an opportunity to affect your framework and main concern.

5. **Education**

Data analytics applications in education are the most needed data analyst in the current scenario. It is mostly used in adaptive learning, new innovations, adaptive content, etc. Is the estimation, assortment, investigation, and detailing of information about students and their specific circumstances, for reasons for comprehension and streamlining learning and conditions in which it happens.

6. Healthcare

▪ Applications of data analytics in healthcare can be utilized to channel enormous measures of information in seconds to discover treatment choices or answers for various illnesses. This won't just give precise arrangements dependent on recorded data yet may likewise give accurate answers for exceptional worries for specific patients.

8. Insurance ▪ There is a lot of data analysis taking place during the insurance process. Several data, such as actuarial data and claims data, help insurance companies realize the risk involved in insuring the person. Analytical software can be used to identify risky claims and bring them before the authorities for further investigation.

b. 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.

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

a. Write the steps involved in Principal Component Analysis (PCA). Consider the two dimensional patterns: (2, 1), (3, 5), (4, 3), (5, 6), (6, 7), (7, 8). Compute the principal component using PCA Algorithm.

Solution: Step-1

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.

Step-2

Solution-

We use the above discussed PCA Algorithm-

Step-01:

Get data.

The given feature vectors are-

$x_1 = (2, 1)$

$x_2 = (3, 5)$

$x_3 = (4, 3)$

$x_4 = (5, 6)$

$x_5 = (6, 7)$

$x_6 = (7, 8)$

Step-02:

Calculate the mean vector (μ).

Mean vector $(\mu) = ((2 + 3 + 4 + 5 + 6 + 7) / 6, (1 + 5 + 3 + 6 + 7 + 8) / 6)$

$= (4.5, 5)$

Step-03:

Subtract mean vector (μ) from the given feature vectors.

$x_1 – \mu = (2 – 4.5, 1 – 5) = (-2.5, -4)$

$x_2 – \mu = (3 – 4.5, 5 – 5) = (-1.5, 0)$

$x_3 – \mu = (4 – 4.5, 3 – 5) = (-0.5, -2)$

$x_4 – \mu = (5 – 4.5, 6 – 5) = (0.5, 1)$

$x_5 – \mu = (6 – 4.5, 7 – 5) = (1.5, 2)$

$x_6 – \mu = (7 – 4.5, 8 – 5) = (2.5, 3)$

Step-04:

Calculate the covariance matrix.

Covariance matrix is given by-

Now,

Covariance matrix

= (m_1 + m_2 + m_3 + m_4 + m_5 + m_6) / 6

On adding the above matrices and dividing by 6, we get-

Step-05:

Calculate the eigen values and eigen vectors of the covariance matrix.

λ is an eigen value for a matrix M if it is a solution of the characteristic equation $|M – \lambda I| = 0$.

From here,

$(2.92 – \lambda)(5.67 – \lambda) – (3.67 \times 3.67) = 0$

$16.56 – 2.92\lambda – 5.67\lambda + \lambda^2 – 13.47 = 0$

$\lambda^2 – 8.59\lambda + 3.09 = 0$

Solving this quadratic equation, we get $\lambda = 8.22, 0.38$

Thus, two eigen values are $\lambda_1 = 8.22$ and $\lambda_2 = 0.38$.

Clearly, the second eigen value is very small compared to the first eigen value.

So, the second eigen vector can be left out.

Eigen vector corresponding to the greatest eigen value is the principal component for the given data set.

So. we find the eigen vector corresponding to eigen value λ_1.We use the following equation to find the eigen vector-

$$MX = \lambda X$$

where-

M = Covariance Matrix
X = Eigen vector
λ = Eigen value

Solving these, we get-
$2.92X_1 + 3.67X_2 = 8.22X_1$
$3.67X_1 + 5.67X_2 = 8.22X_2$

On simplification, we get-
$5.3X_1 = 3.67X_2$ ………(1)
$3.67X_1 = 2.55X_2$ ………(2)

b. What is Support Vector Machine (SVM) and Kernel functions. Explain their important concepts by visualizing on appropriate diagram.
c. 1. Support Vectors
d. 2. Hyper-Plane
e. 3. Margin

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

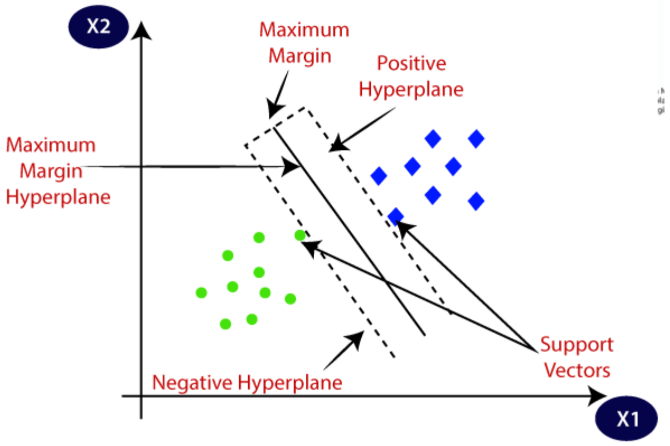
In machine learning, a kernel refers to a method that allows us to apply linear classifiers to non-linear problems by mapping non-linear data into a higher-dimensional space without the need to visit or understand that higher-dimensional space.

Step-2

The followings are important concepts in SVM −

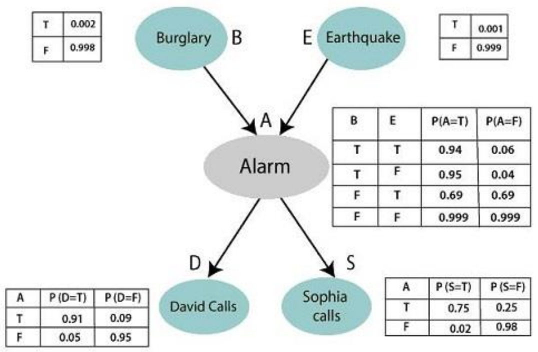
- Support Vectors
- Hyper-plane
- Margin
- **Support Vectors** − Data points that are closest to the hyper-plane is called support vectors. Separating line will be defined with the help of these data points.
- **Hyperplane** − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.
- **Margin** − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.
    - The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyper-plane (MMH) and it can be done in the following two steps –
    - First, SVM will generate hyper-planes iteratively that segregates the classes in best way.

Then, it will choose the hyper-plane that separates the classes correctly



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

a. Explain Baysian Belief Network. Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and David and Sophia both called the Harry. Solve this equation P(S, D, A, ¬B, ¬E) = P (S|A) *P (D|A)*P (A|¬B ^ ¬E) *P (¬B) *P (¬E) by probabilities given in below tables.



T	0.002
F	0.998

T	0.001
F	0.999

B	E	P(A=T)	P(A=F)
T	T	0.94	0.06
T	F	0.95	0.04
F	T	0.69	0.69
F	F	0.999	0.999

A	P (D=T)	P (D=F)
T	0.91	0.09
F	0.05	0.95

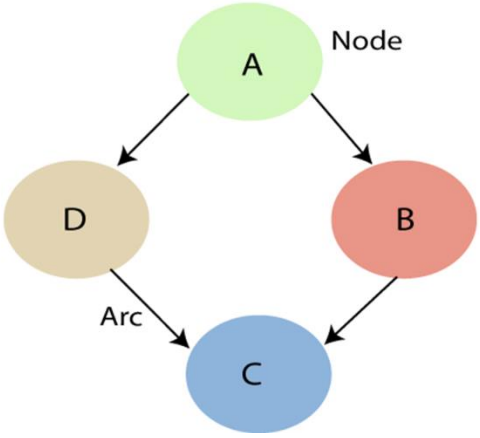
A	P (S=T)	P (S=F)
T	0.75	0.25
F	0.02	0.98

Solution: Step-1

Bayesian Belief Network in artificial intelligence. Bayesian belief network is key computer technology for dealing with probabilistic events and to solve a problem which has uncertainty. We can define a Bayesian network as:

"A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph." It is also called a Bayes network, belief network, decision network, or Bayesian model. Bayesian networks are probabilistic, because these networks are built from a probability distribution, and also use probability theory for prediction and anomaly detection.

Note: The Bayesian network graph does not contain any cyclic graph. Hence, it is known as a directed acyclic graph or DAG.



A directed graph G with four vertices A,B,C, and D. If p(xA,xB,xC,xD) factorizes with respect to G, then we must have p(xA,xB,xC,xD) = p(xA)p(xB|xA)p(xC|xB)p(xD|xC).

Step-2

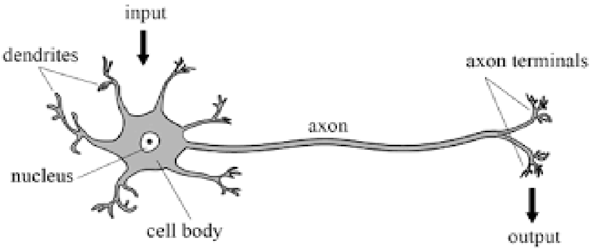
P(S, D, A, ¬B, ¬E) = P (S|A) *P (D|A)*P (A|¬B ^ ¬E) *P (¬B) *P (¬E).
= 0.75* 0.91* 0.001* 0.998*0.999
= 0.00068045.

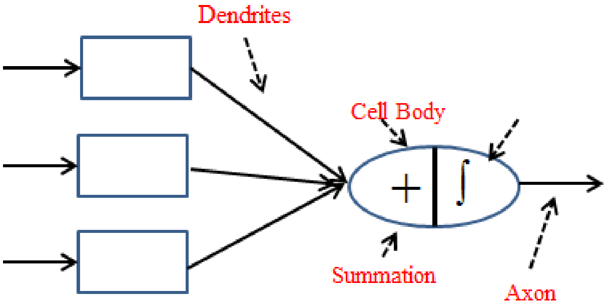
b. Describe the concept of Artificial Neural Network (ANN) relating with biological neurons. Illustrate the different types of neural networks with their procedural figure.
1. Single Layer Perceptron Model
2. Radial Basis Function Neural Network
3. Multi-Layer Perceptron Neural Network
4. Recurrent Neural Network

Solution: Step-1

Artificial Neural Networks (ANN) are algorithms based on brain function and are used to model complicated patterns and forecast issues. The Artificial Neural Network (ANN) is a deep learning method that arose from the concept of the human brain Biological Neural Networks. The development of ANN was the result of an attempt to replicate the workings of the human brain. The workings of ANN are extremely similar to those of biological neural networks, although they are not identical. ANN algorithm accepts only numeric and structured data.

ANN are biological inspired simulations performed on computer to perform certain tasks like- clustering, classification, pattern recognition.





- Dendrites: Receives signal from other neurons.
- Soma/Cell Body: It seems all the “incoming” signals to generate outputs.
- Axon Structure: When the sum reaches to the threshold value, neuron fires and the signal travel to axon to other neuron.
- Semapses/Axon Terminals: The point of intersection of one neuron with other neuron. The amount of signal transmitted depends on the strength (weights) of the connections.

Step-2:

Single Layer Perceptron

Perceptron model, proposed by Minsky-Papert is one of the simplest and oldest models of Neuron. It is the smallest unit of neural network that does certain computations to detect features or business intelligence in the input data. It accepts weighted inputs, and apply the activation function to obtain the output as the final result. Perceptron is also known as TLU(threshold logic unit). Perceptron is a supervised learning algorithm that classifies the data into two categories, thus it is a binary classifier. A perceptron separates the input space into two categories by a hyperplane represented by the following equation:

Advantages of Perceptron:
Perceptrons can implement Logic Gates like AND, OR, or NAND.

Disadvantages of Perceptron:
Perceptrons can only learn linearly separable problems such as boolean AND problem. For non-linear problems such as the boolean XOR problem, it does not work.

Radial Basis Function

A **Radial Basis Function** Network, or RBFN for short, is a form of neural network that relies on the integration of the Radial Basis Function and is specialized for tasks involving non-linear classification.

The RBFN approach is more intuitive than the MLP. The RBFN performs classification by measuring the input’s similarity to examples from the training set. Each RBFN neuron stores a “prototype”, which is just one of the examples from the training set. When we want to classify a new input, each neuron computes the

Euclidean distance between the input and its prototype. Roughly speaking, if the input more closely resembles the class A prototypes than the class B prototypes, it is classified as class A.

Multi-Layer Perceptron

An entry point towards complex neural nets where input data travels through various layers of artificial neurons. Every single node is connected to all neurons in the next layer which makes it a fully connected neural network. Input and output layers are present having multiple hidden Layers i.e. at least three or more layers in total. It has a bi-directional propagation i.e. forward propagation and backward propagation. Inputs are multiplied with weights and fed to the activation function and in backpropagation, they are modified to reduce the loss. In simple words, weights are machine learnt values from Neural Networks. They self-adjust depending on the difference between predicted outputs vs training inputs. Nonlinear activation functions are used followed by softmax as an output layer activation function.

Advantages on Multi-Layer Perceptron

Used for deep learning [due to the presence of dense fully connected layers and back propagation]

Disadvantages on Multi-Layer Perceptron:

Comparatively complex to design and maintain

Comparatively slow (depends on number of hidden layers)

Recurrent Neural Network (RNN)

A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation, natural language processing (nlp), speech recognition, and image captioning; they are incorporated into popular applications such as Siri, voice search, and Google Translate. Like feedforward and convolutional neural networks (CNNs), recurrent neural networks utilize training data to learn.

They are distinguished by their "memory" as they take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depend on the prior elements within the sequence. While future events would also be helpful in determining the output of a given sequence, unidirectional recurrent neural networks cannot account for these events in their predictions.

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 10th of second.

Image Set

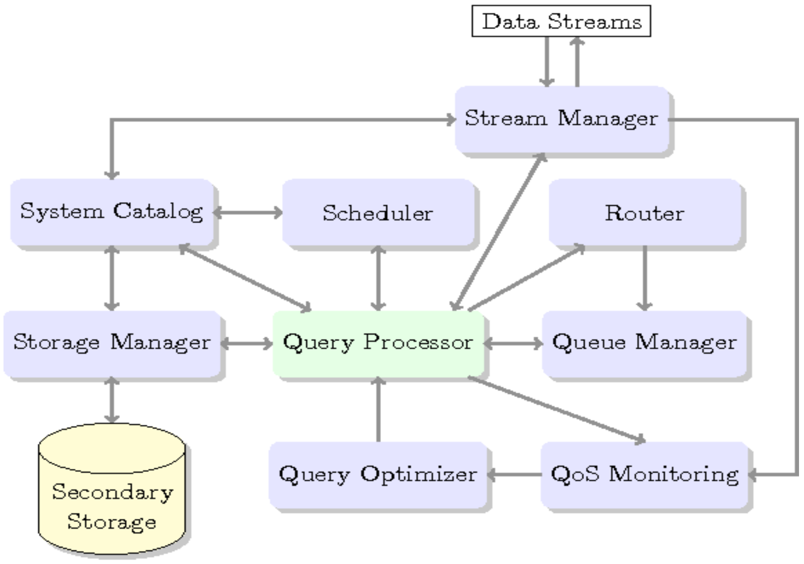
Satellites often send down to earth streams consisting of many terabytes of images per day. Surveillance cameras produce images with lower resolution than satellites, but there can be many of them, each producing a stream of images at intervals like one second. London is said to have six million such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