VTU Examination June 2019

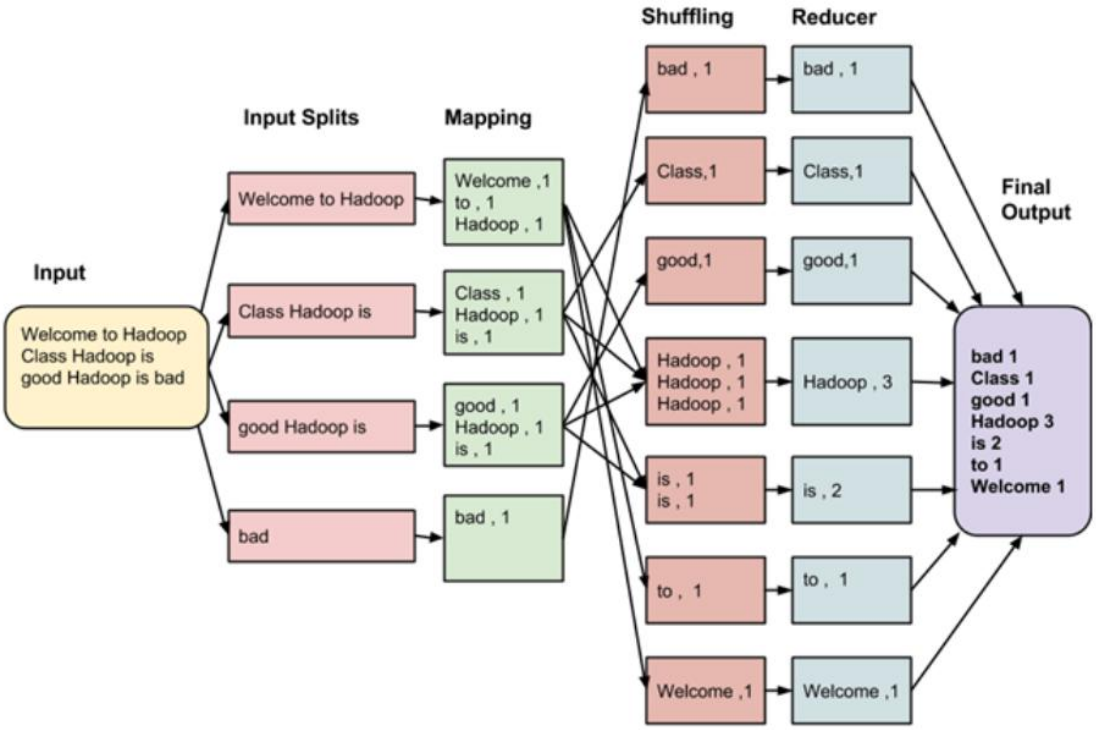
Solution for Big Data Analytics(15CS82)

Q.1 a) How does the Hadoop Map reduce data flow work for a word count program? Give an example?

What is MapReduce in Hadoop?
MapReduce is a programming model suitable for processing of huge data. Hadoop is capable of running MapReduce programs written in various languages: Java, Ruby, Python, and C++. MapReduce programs are parallel in nature, thus are very useful for performing large-scale data analysis using multiple machines in the cluster.
MapReduce programs work in two phases:
1. Map phase
2. Reduce phase.
An input to each phase is **key-value** pairs. In addition, every programmer needs to specify two functions: **map function** and **reduce function**.

How MapReduce Works:-
The whole process goes through four phases of execution namely, splitting, mapping, shuffling, and reducing.
Let's understand this with an example –
Consider you have following input data for your Map Reduce Program

Welcome to Hadoop Class
Hadoop is good
Hadoop is bad



The final output of the MapReduce task is

bad	1
Class	1
good	1
Hadoop	3
is	2
to	1
Welcome	1

The data goes through the following phases

Input Splits:

An input to a MapReduce job is divided into fixed-size pieces called **input splits** Input split is a chunk of the input that is consumed by a single map

Mapping

This is the very first phase in the execution of map-reduce program. In this phase data in each split is passed to a mapping function to produce output values. In our example, a job of mapping phase is to count a number of occurrences of each word from input splits (more details about input-split is given below) and prepare a list in the form of <word, frequency>

Shuffling

This phase consumes the output of Mapping phase. Its task is to consolidate the relevant records from Mapping phase output. In our example, the same words are clubed together along with their

respective frequency.

Reducing

In this phase, output values from the Shuffling phase are aggregated. This phase combines values from Shuffling phase and returns a single output value. In short, this phase summarizes the complete dataset.

In our example, this phase aggregates the values from Shuffling phase i.e., calculates total occurrences of each word.

b) Briefly explain HDFS Name node federation, NFS Gateway, Snapshots, Checkpoints, and Backups?

HDFS Name node Federation: It enhances an existing HDFS architecture. In prior HDFS architecture for entire cluster allows only single namespace. In that configuration, Single NameNode manages namespace. If NameNode fails, the cluster as a whole would be out of services. The cluster will be unavailable until the NameNode restarts or brought on a separate machine. Hadoop Federation overcomes this limitation by adding support for many NameNode/Namespaces to HDFS.

The NFS Gateway: It supports NFSv3 and allows HDFS to be mounted as part of the client's local file system. Currently NFS Gateway supports and enables the following usage patterns:

- Users can browse the HDFS file system through their local file system on NFSv3 client compatible operating systems.
- Users can download files from the the HDFS file system on to their local file system.
- Users can upload files from their local file system directly to the HDFS file system.
- Users can stream data directly to HDFS through the mount point. File append is supported but random write is not supported.

The NFS gateway machine needs the same thing to run an HDFS client like Hadoop JAR files, HADOOP_CONF directory. The NFS gateway can be on the same host as DataNode, NameNode, or any HDFS client.

Checkpoint:

The Checkpoint node periodically creates checkpoints of the namespace. It downloads fsimage and edits from the active NameNode, merges them locally, and uploads the new image back to the active NameNode. The Checkpoint node usually runs on a different machine than the NameNode since its memory requirements are on the same order as the NameNode. The Checkpoint node is started by bin/hdfs namenode -checkpoint on the node specified in the configuration file.

The location of the Checkpoint (or Backup) node and its accompanying web interface are configured via the dfs.namenode.backup.address and dfs.namenode.backup.http-address configuration variables.The start of the checkpoint process on the Checkpoint node is controlled by two configuration parameters.

- dfs.namenode.checkpoint.period, set to 1 hour by default, specifies the maximum delay between two consecutive checkpoints

- dfs.namenode.checkpoint.txns, set to 1 million by default, defines the number of uncheckpointed transactions on the NameNode which will force an urgent checkpoint, even if the checkpoint period has not been reached.

The Checkpoint node stores the latest checkpoint in a directory that is structured the same as the NameNode's directory. This allows the checkpointed image to be always available for reading by the NameNode if necessary. See Import checkpoint.

Multiple checkpoint nodes may be specified in the cluster configuration file.

Back up node:

The Backup node provides the same checkpointing functionality as the Checkpoint node, as well as maintaining an in-memory, up-to-date copy of the file system namespace that is always synchronized with the active NameNode state. Along with accepting a journal stream of file system edits from the NameNode and persisting this to disk, the Backup node also applies those edits into its own copy of the namespace in memory, thus creating a backup of the namespace.

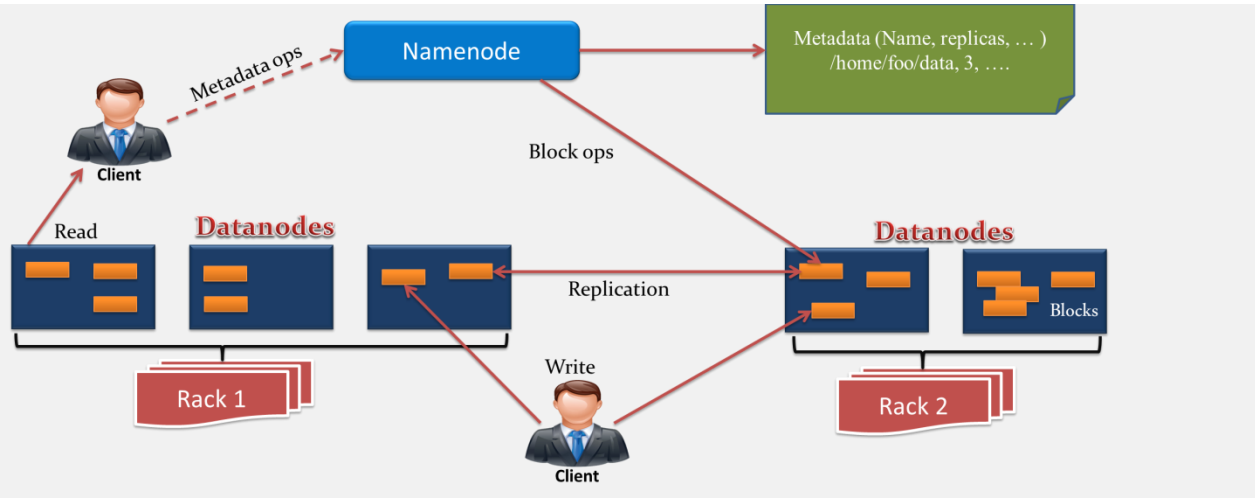
The Backup node does not need to download fsimage and edits files from the active NameNode in order to create a checkpoint, as would be required with a Checkpoint node or Secondary NameNode, since it already has an up-to-date state of the namespace state in memory. The Backup node checkpoint process is more efficient as it only needs to save the namespace into the local fsimage file and reset edits.

Use of a Backup node provides the option of running the NameNode with no persistent storage, delegating all responsibility for persisting the state of the namespace to the Backup node. To do this, start the NameNode with the -importCheckpoint option, along with specifying no persistent storage directories of type edits dfs.namenode.edits.dir for the NameNode configuration.

Q2 a) What do you understand by HDFS? Explain its component with neat diagram?

HDFS Architecture

This architecture gives you a complete picture of Hadoop Distributed File System. There is a single namenode which stores metadata and there are multiple datanodes which do actual storage work. Nodes are arranged in racks and Replicas of data blocks are stored on different racks in the cluster to provide fault tolerance. In remaining section of this tutorial we will see, how read and write operations are performed in HDFS? To read or write a file in HDFS, the client needs to interact with Namenode. HDFS applications need a *write-once-read-many* access model for files. A file once created and written cannot be edited.



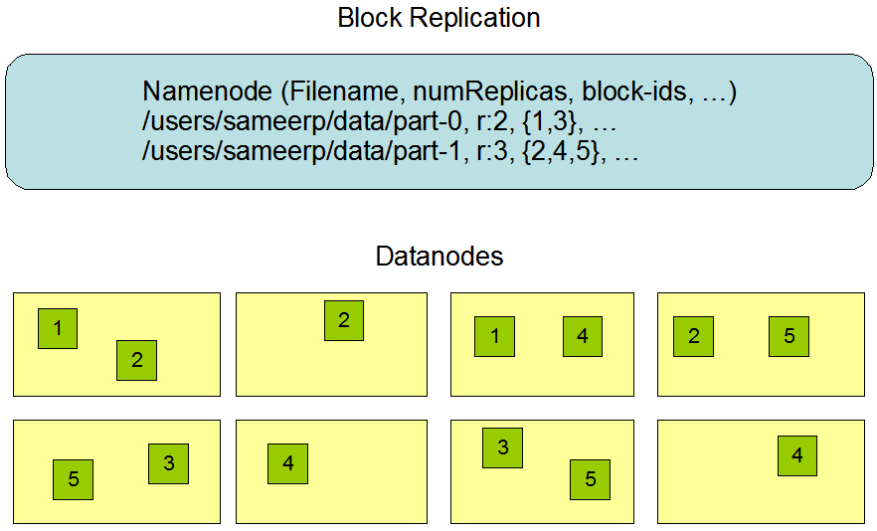
Namenode stores metadata and datanode which stores actual data. The client interacts with namenode for any task to be performed as namenode is the centerpiece in the cluster. There are several datanodes in the cluster which store HDFS data in the local disk. Datanode sends a heartbeat message to namenode periodically to indicate that it is alive. Also, it replicates data to other datanode as per the replication factor.

b) Bring out the concepts of HDFS block replication, with example.

Data Replication

HDFS is designed to reliably store very large files across machines in a large cluster. It stores each file as a sequence of blocks; all blocks in a file except the last block are the same size. The blocks of a file are replicated for fault tolerance. The block size and replication factor are configurable per file. An application can specify the number of replicas of a file. The replication factor can be specified at file creation time and can be changed later. Files in HDFS are write-once and have strictly one writer at any time.

The NameNode makes all decisions regarding replication of blocks. It periodically receives a Heartbeat and a Blockreport from each of the DataNodes in the cluster. Receipt of a Heartbeat implies that the DataNode is functioning properly. A Blockreport contains a list of all blocks on a DataNode.

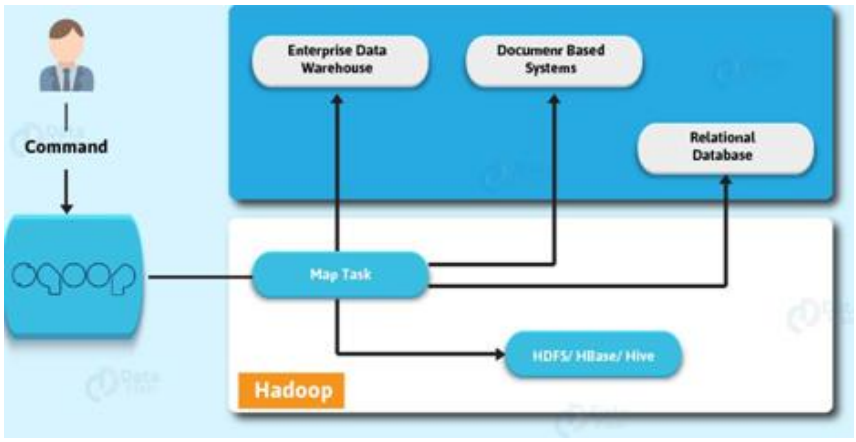


Replica Placement:

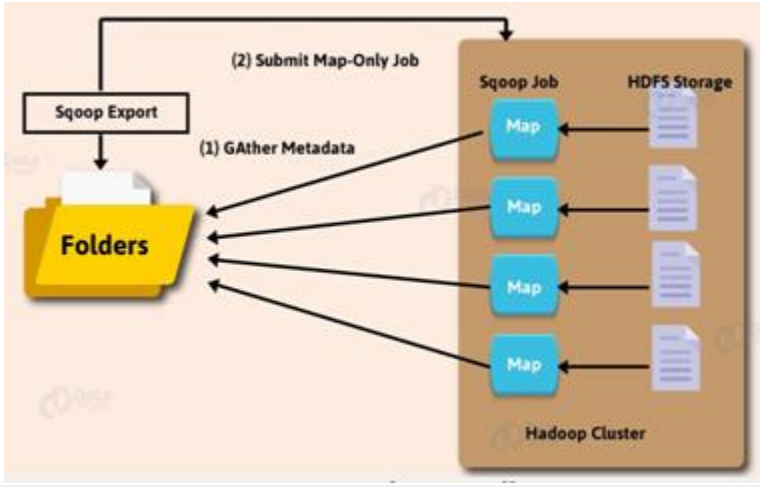
The placement of replicas is critical to HDFS reliability and performance. Optimizing replica placement distinguishes HDFS from most other distributed file systems. This is a feature that needs lots of tuning and experience. The purpose of a rack-aware replica placement policy is to improve data reliability, availability, and network bandwidth utilization. The current implementation for the replica placement policy is a first effort in this direction. The short-term goals of implementing this policy are to validate it on production systems, learn more about its

Q3 a) Explain apache Scoop Import and Export method with neat diagram.

Sqoop Architecture and Working



Basically, a tool which imports individual tables from RDBMS to HDFS is what we call **Sqoop import tool**. However, in HDFS we treat each row in a table as a record. Moreover, our main task gets divided into subtasks, while we submit Sqoop command. However, map task individually handles it internally. On defining map task, it is the subtask that imports part of data to the Hadoop Ecosystem. Likewise, we can say all map tasks import the whole data collectively.

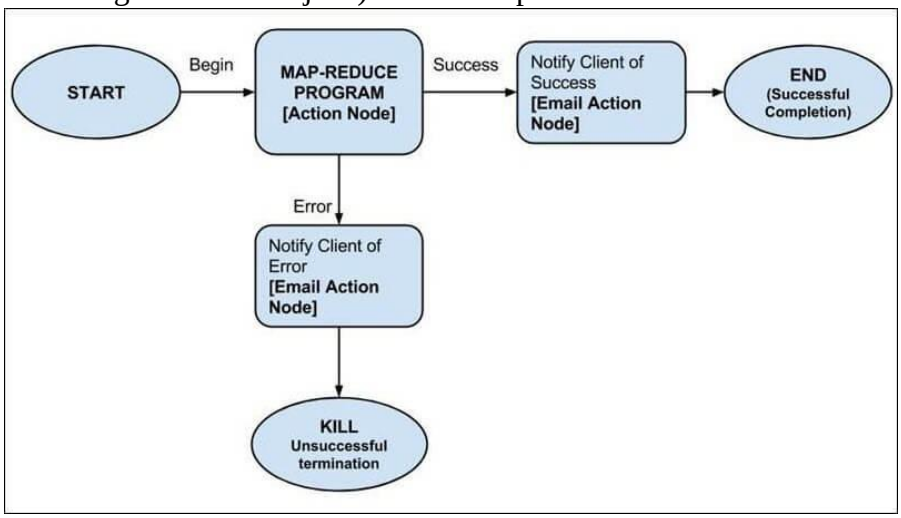


However, Export also works in the same way.
A tool which exports a set of files from HDFS back to an RDBMS is a **Sqoop Export tool**. Moreover, there are files which behave as input to Sqoop which also contain records. Those files what we call as rows in the table.
Moreover, the job is mapped into map tasks, while we submit our job, that brings the chunk of data from **HDFS**. Then we export these chunks to a structured data destination. Likewise, we receive the whole data at the destination by combining all these exported chunks of data. However, in most of the cases, it is an RDBMS (MYSQL/Oracle/SQL Server).
In addition, in case of aggregations, we require reducing phase. However, Sqoop does not perform any aggregations it just imports and exports the data. Also, on the basis of the number defined by the user, map job launch multiple mappers.
In addition, each mapper task will be assigned with a part of data to be imported for Sqoop import. Also, to get high-performance Sqoop distributes the input data among the mappers equally. Afterwards, by using JDBC each mapper creates the connection with the database. Also fetches the part of data assigned by Sqoop. Moreover, it writes it into HDFS or **Hive** or **HBase** on the basis of arguments provided in the CLI.

b) Explain with neat diagram, the Apache Oozie work flow for Hadoop architecture.
Apache Oozie is a workflow scheduler for Hadoop. It is a system which runs the workflow of dependent jobs. Here, users are permitted to create **Directed Acyclic Graphs** of workflows, which can be run in parallel and sequentially in Hadoop.
It consists of two parts:

- Workflow engine: Responsibility of a workflow engine is to store and run workflows composed of Hadoop jobs e.g., MapReduce, Pig, Hive.
- **Coordinator engine**: It runs workflow jobs based on predefined schedules and availability of data.

Oozie is scalable and can manage the timely execution of thousands of workflows (each consisting of dozens of jobs) in a Hadoop cluster.



Oozie is very much flexible, as well. One can easily start, stop, suspend and rerun jobs. Oozie makes it very easy to rerun failed workflows. One can easily understand how difficult it can be to

catch up missed or failed jobs due to downtime or failure. It is even possible to skip a specific failed node.

Q.4 a) How do you run MapReduce and Message Passing interface (MPI) on YARN architecture.

YARN, for those just arriving at this particular party, stands for Yet Another Resource Negotiator, a tool that enables other data processing frameworks to run on Hadoop. The glory of YARN is that it presents Hadoop with an elegant solution to a number of longstanding challenges.

YARN is meant to provide a more efficient and flexible workload scheduling as well as a resource management facility, both of which will ultimately enable Hadoop to run more than just MapReduce jobs.

Distributed storage: Nothing has changed here with the shift from MapReduce to YARN — HDFS is still the storage layer for Hadoop.

- **Resource management:** The key underlying concept in the shift to YARN from Hadoop 1 is decoupling resource management from data processing. This enables YARN to provide resources to any processing framework written for Hadoop, including MapReduce.
- **Processing framework:** Because YARN is a general-purpose resource management facility, it can allocate cluster resources to any data processing framework written for Hadoop. The processing framework then handles application runtime issues.

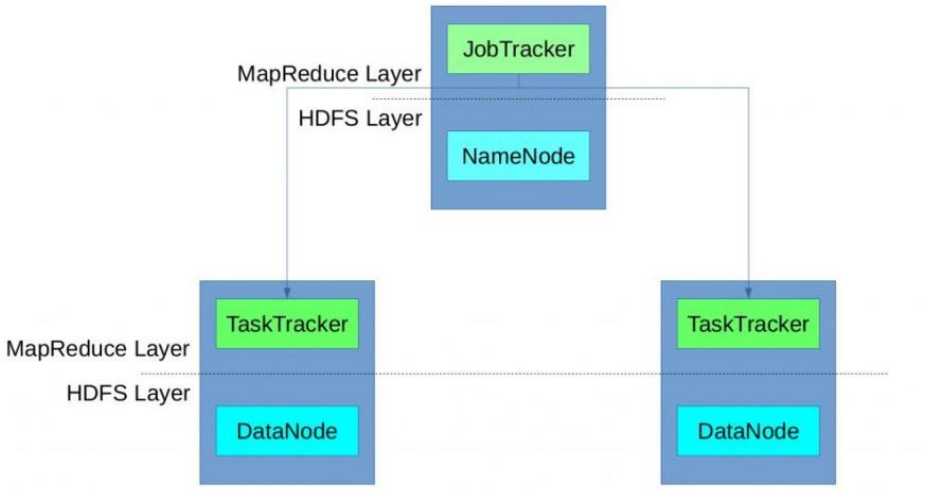
To maintain compatibility for all the code that was developed for Hadoop 1, MapReduce serves as the first framework available for use on YARN. At the time of this writing, the Apache Tez project was an incubator project in development as an alternative framework for the execution of Pig and Hive applications. Tez will likely emerge as a standard Hadoop configuration.

- **Application Programming Interface (API):** With the support for additional processing frameworks, support for additional APIs will come. At the time of this writing, Hoya (for running HBase on YARN), Apache Giraph (for graph processing), Open MPI (for message passing in parallel systems), Apache Storm (for data stream processing) are in active development.

b) What do you understand by YARN Distributed Shell?

YARN (Yet Another Resource Negotiator) has been introduced to Hadoop with version 2.0 and solves a few issues with the resources scheduling of MapReduce in version 1.0. In order to understand the benefits of YARN, we have to review how resource scheduling worked in version 1.0.

A MapReduce job is split by the framework into tasks (Map tasks, Reducer tasks) and each task is run on of the DataNode machines on the cluster. For the execution of tasks, each DataNode machine provided a predefined number of slots (map slots, reducers slots). The JobTracker was responsible for the reservation of execution slots for the different tasks of a job and monitored their execution. If the execution failed, it reserved another slot and re-started the task. It also cleaned up temporary resources and make the reserved slot available to other tasks.



The fact that there was only one JobTracker instance in Hadoop 1.0 led to the problem that the whole MapReduce execution could fail, if the the JobTracker fails (single point of failure). Beyond that, having only one instance of the JobTracker limits scalability (for very large clusters with thousands of nodes).

The concept of predefined map and reduce slots also caused resource problems in case all map slots are used while reduce slots are still available and vice versa. In general it was not possible to reuse the MapReduce infrastructure for other types of computation like real-time jobs. While MapReduce is a batch framework, applications that want to process large data sets stored in HDFS and immediately inform the user about results cannot be implemented with it. Beneath the fact that MapReduce 1.0 did not offer realtime provision of computation results, all other types of applications that want to perform computations on the HDFS data had to be implemented as Map and Reduce jobs, which was not always possible.

Hence Hadoop 2.0 introduced YARN as resource manager, which no longer uses slots to manage resources. Instead nodes have "resources" (like memory and CPU cores) which can be allocated by applications on a per request basis. This way MapReduce jobs can run together with non-MapReduce jobs in the same cluster.

Q5a) Write any four basic Business Intelligence applications for various sectors?

1. Customer Relationship Management

A business exists to serve a customer. A happy customer becomes a repeat customer. A business should understand the needs and sentiments of the customer, sell more of its offerings to the existing customers, and also, expand the pool of customers it serves. BI applications can impact many aspects of marketing.

1. Maximize the return on marketing campaigns: Understanding the customer’s pain points from data-based analysis can ensure that the marketing messages are fine-tuned to better resonate with customers.

2. Healthcare and Wellness:

Health care is one of the biggest sectors in advanced economies. Evidence-based medicine is the newest trend in data-based health care management. BIapplications can help apply the most effective diagnoses and prescriptions for various ailments. They can also help manage public health issues, and reduce waste and fraud.1. Diagnose disease in patients: 2. Treatment effectiveness: 3. Wellness management: 4. Manage fraud and abuse

3.Education: As higher education becomes more expensive and competitive, it becomes a great user of data-based decision-making. There is a strong need for efficiency, increasing revenue, and improving the quality of student experience at all levels of education.

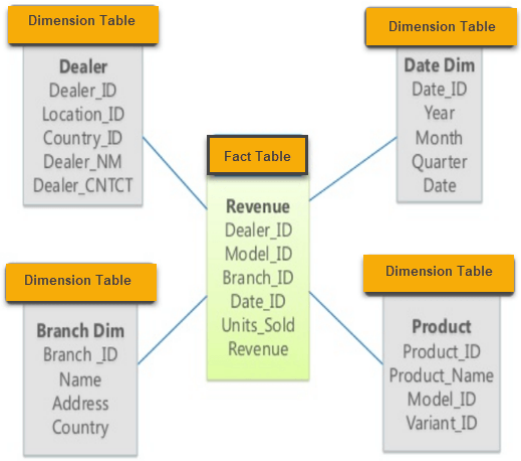
1. Student Enrollment (Recruitment and Retention): 2. Course offerings: 3. Fund-raising from Alumni and other donors:

4. Retail: Retail organizations grow by meeting customer needs with quality products, in a convenient, timely, and cost-effective manner. Understanding emerging customer shopping patterns can help retailers organize their products, inventory, store layout, and web presence in order to delight their customers, which in turn would help increase revenue and profits. Retailers generate a lot of transaction and logistics data that can be used to diagnose and solve problems.

1. Optimize inventory levels at different locations: 2. Improve store layout and sales promotions
3. Optimize logistics for seasonal effects: 4. Minimize losses due to limited shelf life:

b) Explain Star schema of data warehousing.

The star schema is the simplest type of Data Warehouse schema. It is known as star schema as its structure resembles a star. In the Star schema, the center of the star can have one fact tables and numbers of associated dimension tables. It is also known as Star Join Schema and is optimized for querying large data sets.



For example, as you can see in the above-given image that fact table is at the center which contains keys to every dimension table like Deal_ID, Model ID, Date_ID, Product_ID, Branch_ID & other attributes like Units sold and revenue.

Characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure

- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are **not normalized**. For instance, in the above figure, Country_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools

c) What is confusion matrix?

There are two primary kinds of data mining processes: supervised learning and unsupervised learning. In supervised learning, a decision model can be created using past data, and the model can then be used to predict the correct answer for future data instances.

Classification is the main category of supervised learning activity. There are many techniques for classification, decision trees being the most popular one. Each of these techniques can be implemented with many algorithms. A common metric for all of classificationtechniques is predictive accuracy.

Predictive Accuracy = (Correct Predictions) / Total Predictions

Suppose a data mining project has been initiated to develop a predictivemodel for cancer patients using a decision tree. Using a relevant set ofvariables and data instances, a decision tree model has been created. Themodel is then used to predict other data instances. When a true positive data point is positive, that is a correct prediction, called a true positive (TP). Similarly, when a true negative data point is classified as negative, that is a true negative (TN). On the otherhand, when a true-positive data point isclassified by the model as negative, that is an incorrect prediction, called afalse negative (FN). Similarly, when a true-negative data point is classified as positive, that is classified as a false positive (FP). This is represented using the confusion matrix.

ConfusionMatrix		True Class	
		Positive	*Negative*
Predicted Class	Positive	**True Positive (TP)**	False Positive (FP)
Predicted class	Negative	False Negative (FN)	**True Negative (TN)**

Thus the predictive accuracy can be specified by the following formula. Predictive Accuracy = (TP +TN) / (TP + TN + FP + FN).

Q.6 a) Explain Crisp-DM cycle with a neat diagram.

Business is the act of doing something productive to serve someone's needs, and thus earn a living and make the world a better place. Business activities are recorded on paper or using

electronic media, and then these records become data. There is more data from customers' responses and on the industry as a whole. All this data can be analyzed and mined using special tools and techniques to generate patterns and intelligence, which reflect how the business is functioning. These ideas can then be fed back into the business so that it can evolve to become more effective and efficient in serving customer needs. And the cycle continues (Figure 1.1).

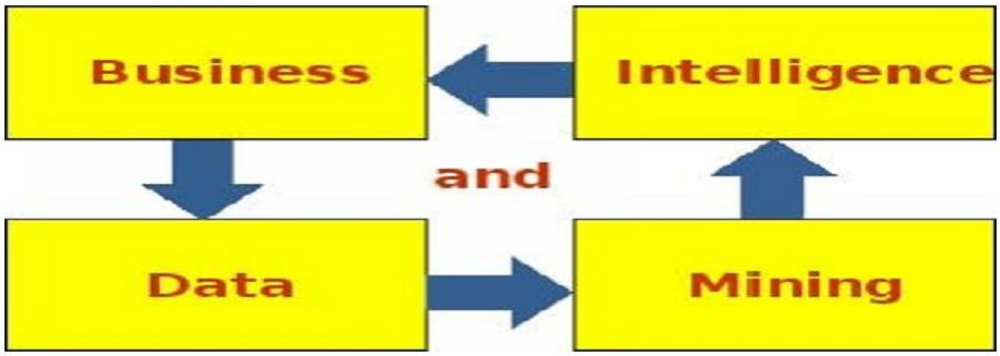


Figure 1.1: Business Intelligence and Data Mining Cycle

Business Intelligence

Any business organization needs to continually monitor its business environment and its own performance, and then rapidly adjust its future plans. This includes monitoring the industry, the competitors, the suppliers, and the customers. The organization needs to also develop a balanced scorecard to track its own health and vitality. Executives typically determine what they want to track based on their key performance Indexes (KPIs) or key result areas (KRAs). Customized reports need to be designed to deliver the require information to every executive. These reports can be converted into customized dashboards that deliver the information rapidly and in easy-to grasp formats.

b) What do you understand by the term data visualization? How is it important in big data analytics?

1. Line graph. This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables.
2. Scatter plot: This is another very basic and useful graphic form. It helps several the relationship between two variables. In the above case let, it shows two dimensions: Life Expectancy and Fertility Rate. Unlike in a line graph, there are no line segments connecting the points.
3. Bar graph: A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented. The bars can be plotted vertically or horizontally. The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate.
4. Stacked Bar graphs: These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story. Bars can also be

normalized such as the total height of every bar is equal, so it can show the relative composition of each bar.

5. Histograms: These are like bar graphs, except that they are useful in showing data frequencies or data values on classes (or ranges) of a numerical variable.

6. Pie charts: These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value.

7. Box charts: These are special form of charts to show the distribution of variables. The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction.

8. Bubble Graph: This is an interesting way of displaying multiple dimensions in one chart. It is a variant of a scatter plot with many data points marked on two dimensions. Now imagine that each data point on the graph is a bubble (or a circle) ... the size of the circle and the color fill in the circle could represent two additional dimensions.

9. Dials: These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low range, medium range, or high range. These ranges could be colored red, yellow and green to give an instant view of the data.

10. Geographical Data maps are particularly useful maps to denote statistics.

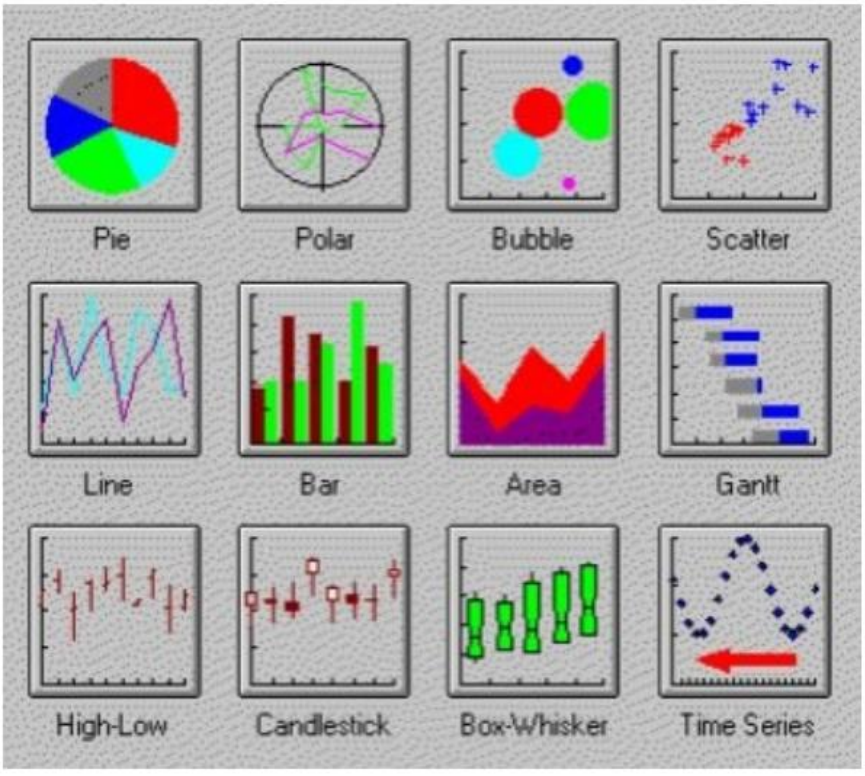


Fig.1 Different types of graphs

c) Differentiate between data Mining and Data warehousing.

Data Mining Vs Data Warehouse: Key Differences

Data Mining	Data Warehouse
Data mining is the process of analyzing unknown patterns of data.	A data warehouse is database system which is designed for analytical instead of transactional work.
Data mining is a method of comparing large amounts of data to finding right patterns.	Data warehousing is a method of centralizing data from different sources into one common repository.
Data mining is usually done by business users with the assistance of engineers.	Data warehousing is a process which needs to occur before any data mining can take place.
Data mining is the considered as a process of extracting data from large data sets.	On the other hand, Data warehousing is the process of pooling all relevant data together.
One of the most important benefits of data mining techniques is the detection and identification of errors in the system.	One of the pros of Data Warehouse is its ability to update consistently. That's why it is ideal for the business owner who wants the best and latest features.
Data mining helps to create suggestive patterns of important factors. Like the buying habits of customers, products, sales. So that, companies can make the necessary adjustments in operation and production.	Data Warehouse adds an extra value to operational business systems like CRM systems when the warehouse is integrated.

Q7a) What is a splitting variable? Describe three criteria for choosing splitting variable.
Splitting criteria

1. Which variable to use for the first split? How should one determine the most important variable for the first branch, and subsequently, for each sub-tree? There are many measures like least errors, information gain, gini's coefficient, etc.
2. What values to use for the split? If the variables have continuous values such as for age or blood pressure, what value-ranges should be used to make bins?

3. How many branches should be allowed for each node? There could be binary trees, with just two branches at each node. Or there could be more branches allowed.

b) List the advantages and disadvantages of Regression Models
Regression Models are very popular because they offer many advantages.
1. Regression models are easy to understand as they are built upon basicstatistical principles such as correlation and least square error.
2. Regression models provide simple algebraic equations that are easy tounderstand and use.
3. The strength (or the goodness of fit) of the regression model is measuredin terms of thecorrelation coefficients, and other related statisticalparameters that are well understood.
4. Regression models can match and beat the predictive power of othermodelling techniques.
5. Regression models can include all the variables that one wants to include in the model.
6. Regression modelling tools are pervasive. They are found in statistical packages as well as data mining packages. MS Excel spreadsheets can provide simple regression modeling

capabilities.

Regression models can however prove inadequate under many circumstances.

1. Regression models can not cover for poor data quality issues. If the data is not prepared well to remove missing values or is not well-behaved in terms of a normal distribution, the validity of the model suffers.
2. Regression models suffer from collinearity problems (meaning strong linear correlations among some independent variables). If the independent variables have strong correlations among themselves, then they will eat into each other's predictive power and the regression coefficients will lose their ruggedness. Regression models will not automatically choose between highly collinear variables, although some packages attempt to do that.
3. Regression models can be unwieldy and unreliable if a large number of variables are included in the model. All variables entered into the model will be reflected in the regression equation, irrespective of their contribution to the predictive power of the model. There is no concept of automatic pruning of the regression model.
4. Regression models do not automatically take care of non-linearity. The user needs to imagine the kind of additional terms that might be needed to be added to the regression model to improve its fit.
5. Regression models work only with numeric data and not with categorical variables. There are ways to deal with categorical variables though by creating multiple new variables with a yes/no value.

c) Create a decision tree for the following data set. The objective is to predict the class category. (Loan approved or not)

Age	**Job**	**House**	**Credit**	**LoanApproved**
Young	False	No	Fair	*No*
Young	False	No	Good	*No*
Young	True	No	Good	*Yes*
Young	True	Yes	Fair	*Yes*
Young	False	No	Fair	*No*
Middle	False	No	Fair	*No*
Middle	False	No	Good	*No*
Middle	True	Yes	Good	*Yes*
Middle	False	Yes	Excellent	*Yes*
Middle	False	Yes	Excellent	*Yes*
Old	False	Yes	Excellent	*Yes*
Old	False	Yes	Good	*Yes*
Old	True	No	Good	*Yes*
Old	True	No	Excellent	*Yes*
Old	False	No	Fair	*No*

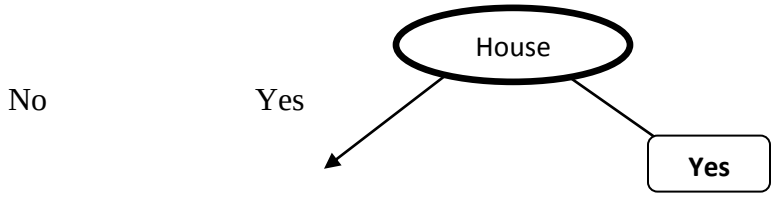
Then solve the following problem using the model.

Age	Job	House	Credit	LoanApproved
Young	False	No	Good	*??*

Solution :

Attributes	Rules	Error	Total Error
Age	Young→No	2/5	5/15
	Middle→Yes	2/5	
	Old→Yes	1/5	
Job	False→No	4/10	4/15
	True→Yes	0/5	
House	No→No	3/9	3/15
	Yes→Yes	0/6	
Credit	Fair→	1/5	3/15
	Good→	2/6	
	Excellent→	0/4	

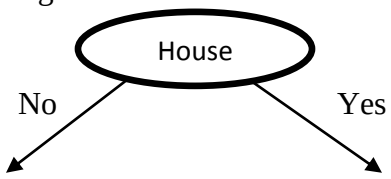
- To select the root node, we find the attribute which is having least number of error. But there is a tie between two attributes, House and Credit.
- Select attribute House as a root node as it has two branches as compared with Credit attribute.

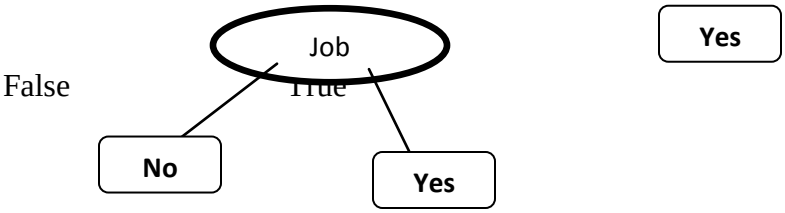


- Now grow the tree for House=No

Attributes	Rules	Error	Total Error
Age	Young→No	1/4	2/9
	Middle→No	0/2	
	Old→Yes	1/3	
Job	False→No	0/6	0/9
	True→Yes	0/3	
Credit	Fair→No	0/4	2/9
	Good→Yes	2/4	
	Excellent→Yes	0/1	

- Job attribute is having least error than others





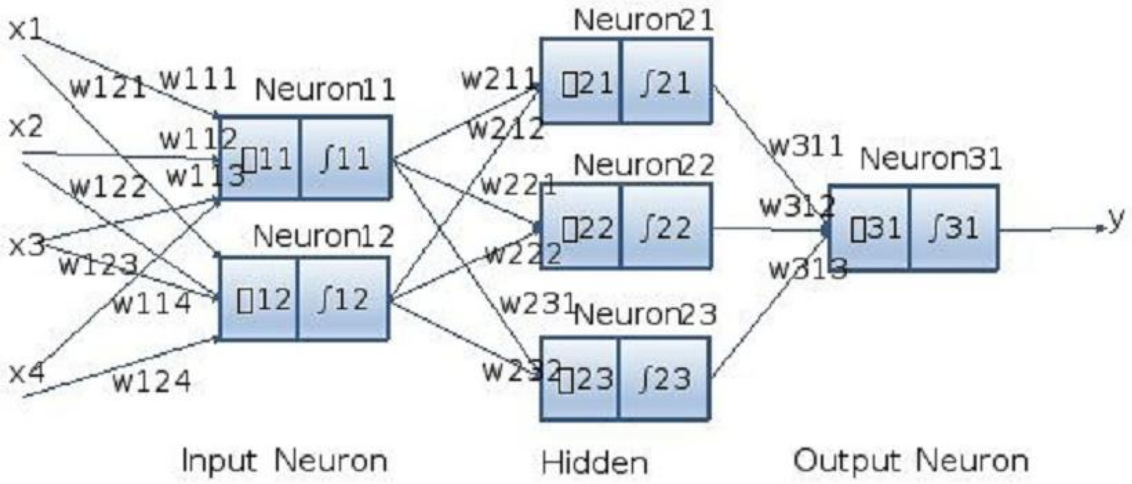
- For the following test data A**nswer is No**

Age	**Job**	**House**	**Credit**	**LoanApproved**
Young	False	No	Good	***No***

Q8a) Design Principles of an Artificial Neural Network

1. A neuron is the basic processing unit of the network. The neuron (orprocessing element)receives inputs from its preceding neurons (or PEs),does some nonlinear weightedcomputation on the basis of those inputs,transforms the result into its output value, and then passes on the outputto the next neuron in the network. X's are the inputs, w's arethe weights for each input, and y is the output.

2. A Neural network is a multi-layered model. There is at least one inputneuron, one output neuron, and at least one processing neuron. An ANNwith just this basic structure would be a simple, single-stagecomputational unit. A simple task may be processed by just that oneneuron and the result may be communicated soon. ANNs however, mayhave multiple layers of processing elements in sequence. There could bemany neurons involved in a sequence depending upon the complexity ofthe predictive action. The layers of PEs could work in sequence, or theycould work in parallel



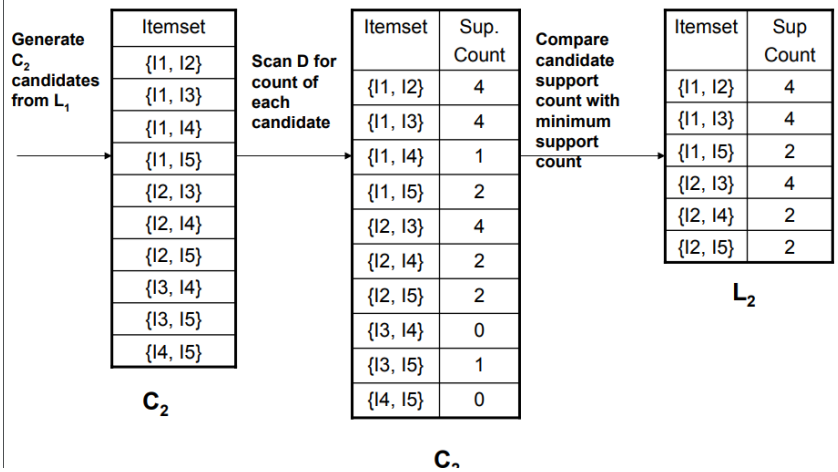
3. The processing logic of each neuron may assign different weights to thevarious incoming input streams. The processing logic may also use nonlinear transformation,such as a sigmoid function, from the processedvalues to the output value. This processing logic and the intermediate weight and processing functions are just what works for the system as a whole, in its objective of solving a problem collectively. Thus, neural networks are considered to be an opaque and a black-box system.

4. The neural network can be trained by making similar decisions over and over again with many training cases. It will continue to learn by adjusting its internal computation and communication based on feedback about its previous decisions. Thus, the neural networks become better at making a decision as they handle more and more decisions. Depending upon the nature of the problem and the availability of good training data, at some point the neural network will learn enough and begin to match the predictive accuracy of a human expert. In many practical situations, the predictions of ANN, trained over a long period of time with a large number of training data, have begun to decisively become more accurate than human experts. At that point ANN can begin to be seriously considered for deployment in real situations in real time.

b) How does Apriori Algorithm work? Apply the same for the following example

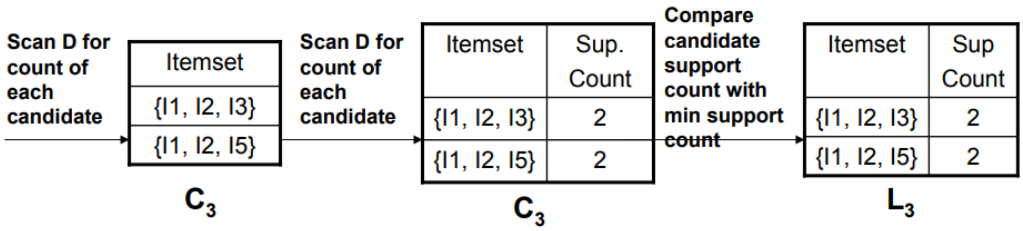
TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Step 2: Generating 2-itemset Frequent Pattern



Step 2: Generating 2-itemset Frequent Pattern

- To discover the set of frequent 2-itemsets, L_2 , the algorithm uses L_1 *Join* L_1 to generate a candidate set of 2-itemsets, C_2.
- Next, the transactions in D are scanned and the support count for each candidate itemset in C_2 is accumulated (as shown in the middle table).
- The set of frequent 2-itemsets, L_2 , is then determined, consisting of those candidate 2-itemsets in C_2 having minimum support.



- The generation of the set of candidate 3-itemsets, C_3 , involves use of the Apriori Property.
- In order to find C_3, we compute L_2 *Join* L_2.
- C_3 = L2 *Join* L2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}.
- Now, Join step is complete and Prune step will be used to reduce the size of C_3. Prune step helps to avoid heavy computation due to large C_k.

We had L = {{I1}, {I2}, {I3}, {I4}, {I5}, {I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}, {I1,I2,I3}, {I1,I2,I5}}.

- Lets take ***l*** = {I1,I2,I5}.
- Its all nonempty subsets are {I1,I2}, {I1,I5}, {I2,I5}, {I1}, {I2}, {I5}.

Let minimum confidence threshold is , say 70%.

The resulting association rules are shown below each listed with its confidence.

- R1: I1 ^ I2 → I5
  - Confidence = sc{I1,I2,I5}/sc{I1,I2} = 2/4 = 50%
  - R1 is Rejected.
- R2: I1 ^ I5 → I2
  - Confidence = sc{I1,I2,I5}/sc{I1,I5} = 2/2 = 100%
  - R2 is Selected.
- R3: I2 ^ I5 → I1
  - Confidence = sc{I1,I2,I5}/sc{I2,I5} = 2/2 = 100%
  - R3 is Selected

- R4: I1 → I2 ^ I5
  - Confidence = sc{I1,I2,I5}/sc{I1} = 2/6 = 33%
  - R4 is Rejected.
- R5: I2 → I1 ^ I5
  - Confidence = sc{I1,I2,I5}/{I2} = 2/7 = 29%
  - R5 is Rejected.
- R6: I5 → I1 ^ I2
  - Confidence = sc{I1,I2,I5}/ {I5} = 2/2 = 100%
  - R6 is Selected.

  In this way, We have found three strong association rules.

Q9 a) what is Naïve Bayes Technique? Explain its Model.

In machine learning we are often interested in selecting the best hypothesis (h) given data (d). In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d). One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge.
Bayes' Theorem is stated as: P(h|d) = (P(d|h) * P(h)) / P(d)
Where

- **P(h|d)** is the probability of hypothesis h given the data d. This is called the posterior probability.
- **P(d|h)** is the probability of data d given that the hypothesis h was true.
- **P(h)** is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- **P(d)** is the probability of the data (regardless of the hypothesis).

You can see that we are interested in calculating the posterior probability of P(h|d) from the prior probability p(h) with P(D) and P(d|h).
After calculating the posterior probability for a number of different hypotheses, you can select the hypothesis with the highest probability. This is the maximum probable hypothesis and may formally be called the maximum a posteriori (MAP) hypothesis.
This can be written as: MAP(h) = max(P(h|d))

or

MAP(h) = max((P(d|h) * P(h)) / P(d))

or

MAP(h) = max(P(d|h) * P(h))

The P(d) is a normalizing term which allows us to calculate the probability. We can drop it when we are interested in the most probable hypothesis as it is constant and only used to normalize.
Back to classification, if we have an even number of instances in each class in our training data,

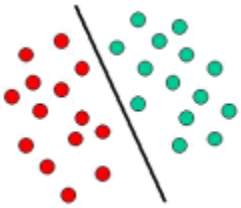
then the probability of each class (e.g. P(h)) will be equal. Again, this would be a constant term in our equation and we could drop it so that we end up with:

$$MAP(h) = \max(P(d|h))$$

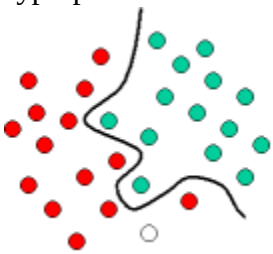
This is a useful exercise, because when reading up further on Naive Bayes you may see all of these forms of the theorem.

b) What is Support Vector Machine? Explain its Model.

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).

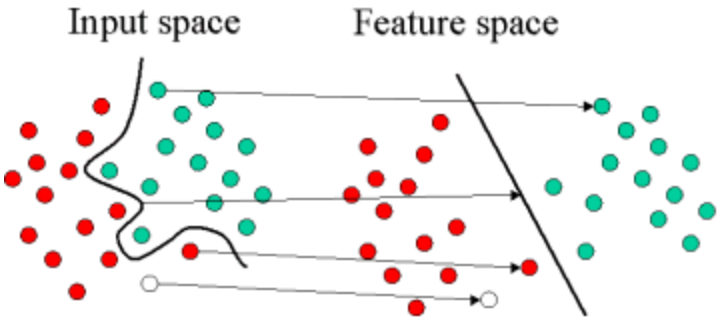


The above is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective groups (GREEN and RED in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). This situation is depicted in the illustration below. Compared to the previous schematic, it is clear that a full separation of the GREEN and RED objects would require a curve (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suited to handle such tasks.



The illustration below shows the basic idea behind Support Vector Machines. Here we see the original objects (left side of the schematic) mapped, i.e., rearranged, using a set of mathematical functions, known as kernels. The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) is linearly separable and, thus, instead of constructing the complex curve (left schematic), all we have to do is to find an optimal line that can separate the GREEN and the RED objects.

c)



Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. For categorical variables a dummy variable is created with case values as either 0 or 1. Thus, a categorical dependent variable consisting of three levels, say (A, B, C), is represented by a set of three dummy variables:
A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

c) Mention the 3-step process of Text Mining.
Text Mining is a rapidly evolving area of research. As the amount of social media and other text data grows, there is need for efficient abstraction and categorization of meaningful information from the text.

The first level of analysis is identifying frequent words. This creates a bag of important words. Texts – documents or smaller messages – can then be ranked on how they match to a particular bag-of-words. However, there are challenges with this approach. For example, the words may be spelled a little differently. Or there may be different words with similar meanings.

The next level is at the level of identifying meaningful phrases from words. Thus ‘ice’ and ‘cream’ will be two different key words that often come together. However, there is a more meaningful phrase by combining the two words into ‘ice cream’. There might be similarly meaningful phrases like ‘Apple Pie’.

The next higher level is that of Topics. Multiple phrases could be combined into Topic area. Thus the two phrases above could be put into a common basket, and this bucket could be called ‘Desserts’. Text mining is a semi-automated process. Text data needs to be gathered, structured, and then mined, in a 3-step process (Figure 1)

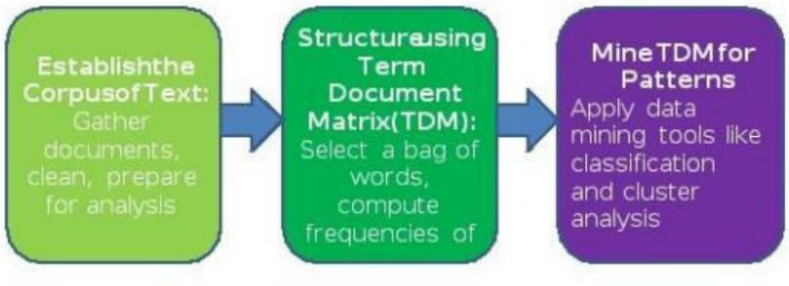


Figure 1 Text Mining Architecture

1. The text and documents are first gathered into a corpus, and organized.
2. The corpus is then analysed for structure. The result is a matrix mapping important terms to source documents.
3. The structured data is then analysed for word structures, sequences, and frequency.

Q10 a) Explain briefly the three different types of Web Mining?

Web mining is the art and science of discovering patterns and insights from the World-wide web so as to improve it. The world-wide web is at the heart of the digital revolution. More data is posted on the web every day than was there on the whole web just 20 years ago. Billions of users are using it every day for a variety of purposes. The web is used for electronic commerce, business communication, and many other applications. Web mining analyzes data from the web and helps find insights that could optimize the web content and improve the user experience. Data for web mining is collected via Web crawlers, web logs, and other means.

Here are some characteristics of optimized websites:

1. Appearance: Aesthetic design. Well-formatted content, easy to scan and navigate. Good colorcontrasts.

2. Content: Well-planned information architecture with useful content. Fresh content. Searchengine optimized. Links to other good sites.

3. Functionality: Accessible to all authorized users. Fast loading times. Usable forms. Mobileenabled. This type of content and its structure is of interest to ensure the web is easy to use. Theanalysis of web usage provides feedback on the web content, and also the consumer's browsinghabits. This data can be of immense use for commercial advertising, and even for socialengineering. The web could be analyzed for its structure as well as content. The usage pattern ofweb pages could also be analyzed.

Depending upon objectives, web mining can be divided into three different types:

1. Web usage mining

As a user clicks anywhere on a webpage or application, the action is recorded by many entities inmany locations. The browser at the client machine will record the click, and the web serverproviding the content would also make a record of the pages served and the user activity on thosepages. The entities between the client and the server, such as the router, proxy server, or adserver, too would record that click

2. Web content mining

A website is designed in the form of pages with a distinct URL (universal resource locator). Alarge website may contain thousands of pages. These pages and their content is managed usingspecialized software systems called Content Management Systems. Every page can have text,graphics, audio, video, forms, applications, and more kinds of content including usergenerated content.

3. Web structure mining

The Web works through a system of hyperlinks using the hypertext protocol (http). Any page cancreate a hyperlink to any other page, it can be linked to by another page. The intertwined or self-referral nature of web lends itself to some uniquenetwork analytical algorithms. The structure ofWeb pages could also be analyzed toexamine the pattern of hyperlinks among pages. Thereare two basic strategic models for successful websites: Hubs and Authorities.

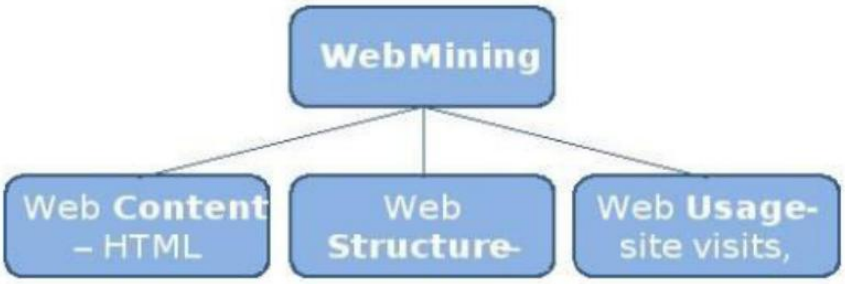
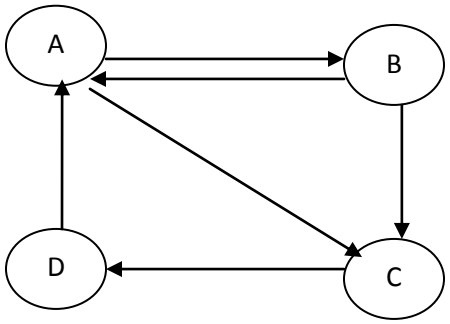


Figure: 1 Web Mining structure

b) Compute the Rank values for the nodes for the following network. Which the highest rank node after computation?



Solution :

a) Compute the Influence matrix (rank matrix)

- Assign the variables for influence value for each node, as Ra, Rb, Rc, Rd.
- There are two bound links from node A to nodes B and C. Thus, both B and C receives half of node A's influence. Similarly, there are two outbound links from node B to nodes C and A, So both C and A received half of node B's influence.

$$Ra = 0.5*Rb + Rd$$
$$Rb = 0.5*Ra$$
$$Rc = 0.5*Ra + 0.5*Rb$$
$$Rd = Rc$$

b) Set the initial set of rank values such as 1/n (n is number of nodes). As 4 nodes are there, initial rank values for all nodes are ¼ i.e 0.25

Variables	Initial Values
Ra	0.25
Rb	0.25
Rc	0.25
Rd	0.25

c) Compute the rank values for 1st iteration and theniteratively compute new rank values till they stabilized.

Variables	Initial Values	Iteration 1
Ra	0.25	0.375
Rb	0.25	0.125
Rc	0.25	0.250
Rd	0.25	0.250

Variables	Initial Values	Iteration 1	Iteration 2
Ra	0.25	0.375	0.3125
Rb	0.25	0.125	0.1875

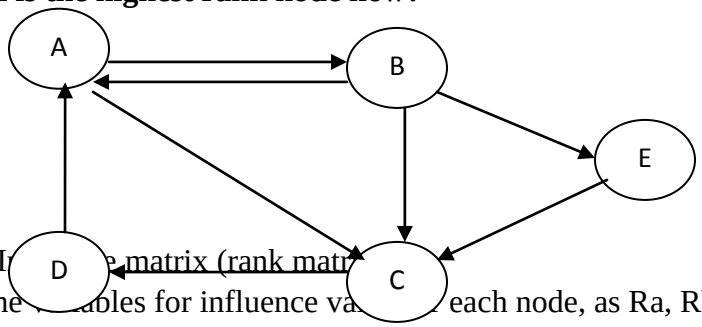
	Ra	Rb	Rc	Rd
Ra	0	0.5	0	1.0
Rb	0.5	0	0	0
Rc	0.5	0.5	0	0
Rd	0	0	1.0	0

Rc	0.25	0.250	0.250
Rd	0.25	0.250	0.250

Variables	Initial Values	Iteration 1	Iteration 2	--------	Iteration 8
Ra	0.25	0.375	0.3125	…….	0.333
Rb	0.25	0.125	0.1875	……	0.167
Rc	0.25	0.250	0.250	……	0.250
Rd	0.25	0.250	0.250	…….	0.250

The Final rank shows of node A is highest at 0.333

Exercise:Which is the highest rank node now?



a) Compute the I matrix (rank mat

- Assign the ables for influence va each node, as Ra, Rb, Rc, Rd, Re

Ra = 1/3 *Rb + Rd

Rb = ½ *Ra

Rc =1/2 *Ra + 1/3*Rb +Re

Rd = Rc

Re = 1/3*Rb

b) Set the initial set of rank values such as 1/n (n is number of nodes). As 4 nodes are there, initial rank values for all nodes are 1/5i.e 0.2

Variables	**Initial Values**
Ra	0.2
Rb	0.2
Rc	0.2
Rd	0.2
Re	0.2

c) Compute the rank values for 1st iteration and then iteratively compute new rank values till they stabilized.

Variables	**Initial Values**	**Iteration 1**
Ra	0.2	0.267
Rb	0.2	0.1
Rc	0.2	0.367
Rd	0.2	0.367

Re	0.2	0.06

Variables	**Initial Values**	**Iteration 1**	**Iteration 2**
Ra	0.2	0.267	0.400
Rb	0.2	0.1	0.134
Rc	0.2	0.367	0.234
Rd	0.2	0.367	0.234
Re	0.2	0.067	0.033

Variables	**Initial Values**	**Iteration 1**	**Iteration 2**	**Iteration 3**
Ra	0.2	0.267	0.400	0.279
Rb	0.2	0.1	0.134	0.200
Rc	0.2	0.367	0.234	0.278
Rd	0.2	0.367	0.234	0.278
Re	0.2	0.067	0.033	0.045

Continue the iterations …. The Final rank shows of node A is highest