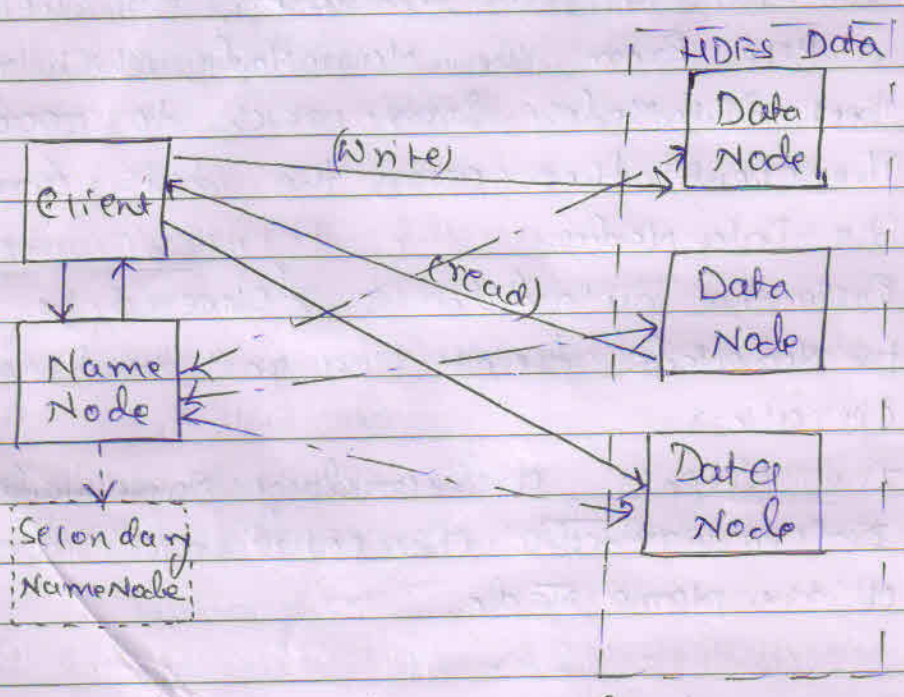
Course: Big Data Analytics
Course Code:- 17CS82

MODULE - 1

Q1 a) With a neat diagram, explain components of HDFS (Hadoop Distributed File System).

Soln



* There are two types nodes:-
① Name Node ② Data Node.

* The design of HDFS follows Master/slave relationship. in which master (Name Node) manages the file system namespace & regulates access to files by clients. The Slaves (Data Nodes) are responsible for serving read and write requests from the file system to the clients.

* When a client writes data, it first communicates with NameNode & requests to create a file. the NameNode determines how many blocks are needed and provides client with Data Nodes that will store data.

* Depending on how many nodes are in the cluster, the NameNode attempt to write replicas of data blocks on the nodes that are in other rack.
* After DataNode acknowledges that file block replication complete, the client closes the file & informs the NameNode that the operation is complete.
* In case of read operation, the client requests a file from the NameNode, which returns the best DataNodes from which to read the data. The client then access the data directly from the DataNodes.
* Once the metadata is delivered to the client, the NameNode doesn't take part in Read or write Operation.
* The purpose of Secondary NameNode is to perform periodic checkpoints that evaluate status of the NameNode.

* NameNode contains two disk files:-
  ① fsimage_* :- image of the file system state.
  ② edit_* :- Stores the series of modification done to file system.

Various roles in HDFS:-
① HDFS uses Master/slave model designed for large file reading/streaming.
② The NameNode is a metadata server or "traffic cop".
③ HDFS provides a single Namespace that is managed by the NameNode.

④ Data is redundantly stored on Data Nodes; there is no data on the NameNode.

⑤ The Secondary NameNode performs checkpoints of NameNode file system's state but is not a Failover node.

Q1

(b) Write and explain the mapper & reducer scripts for the mapreduce model.

Solution:-

Mapper Script

```
#!/bin/bash
While read line ; do
  for token in $line : do
  if [ "$token" = "kutuzov" ] ;then
   echo "kutuzov, 1"
  elif [ "$token" = "Petersburg" ] ; then
    echo "petersburg, 1"
  fi
  done
done
```

The mapper inputs a text file and then outputs data in a (key, value) pair (token-name, count). In our case, input is text file & keys are "kutuzov & Petersburg".

Reducer Scipt

```
#!/bin/bash
kcount =0
pcount =0
```

```
while read line ; do
  if [ "$line" = "Kutuzov," ] ; then
  let kcount = kcount + 1
  elif [ "$line" = "Petersburg," ] ; then
  let pcount = pcount + 1
  fi
done
echo "kutuzov, $kcount"
echo "Petersburg, $pcount"
```

The reducer script takes the key-value pairs output from mappers & combines the similar tokens & counts the total number of instances. The result is a new-key value pair (token-name, sum).

Output

```
$ cat war-and-peace.txt | ./mapper.sh | ./reducer.sh
Kutuzov, 315
Petersburg, 128
```

Q2

(a) With a neat diagram, describe the steps of MapReduce parallel data flow.

Solution: Parallel execution of MapReduce involves following steps :-

① Input Splits ② Map step ③ Combiner step
④ Shuffle step ⑤ Reduce step

① Input Splits :- the input splits are logical boundaries based on the input data. Splits are usually smaller than HDFS block size. The number of splits corresponds to the

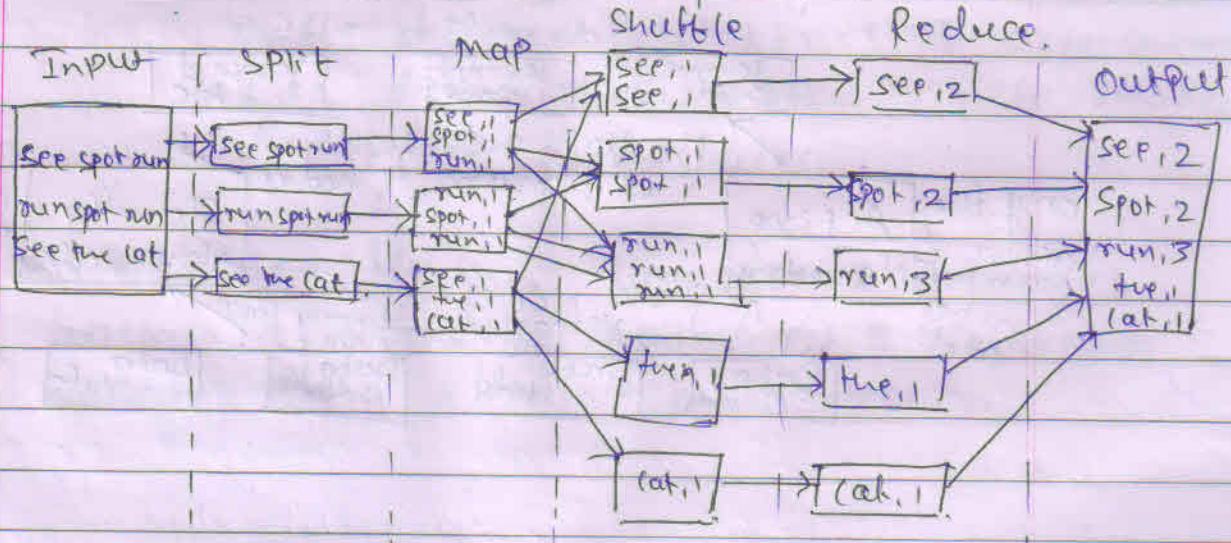
number of mapping processes used in the map stage.

(2) Map step :- In case of large amounts of data, several mappers can be operating at the same time. The user provides the specific mapping process. MapReduce will execute the mapper on machine where block is available & is least busy.

(3) Combiner step :- It provides an optimization as part of map stage where key-value pairs are combined prior to next stage. It is optional.

(4) Shuffle step :- All similar keys are combined and counted by same reducer process. Hence results of the map stage must be collected by key-value pairs & shuffled to the same reducer process. If only one reducer available, then shuffle stage is not needed.

(5) Reduce step :- The data reduction is performed in this step. The results are written to HDFS. Each reducer will write an output file.



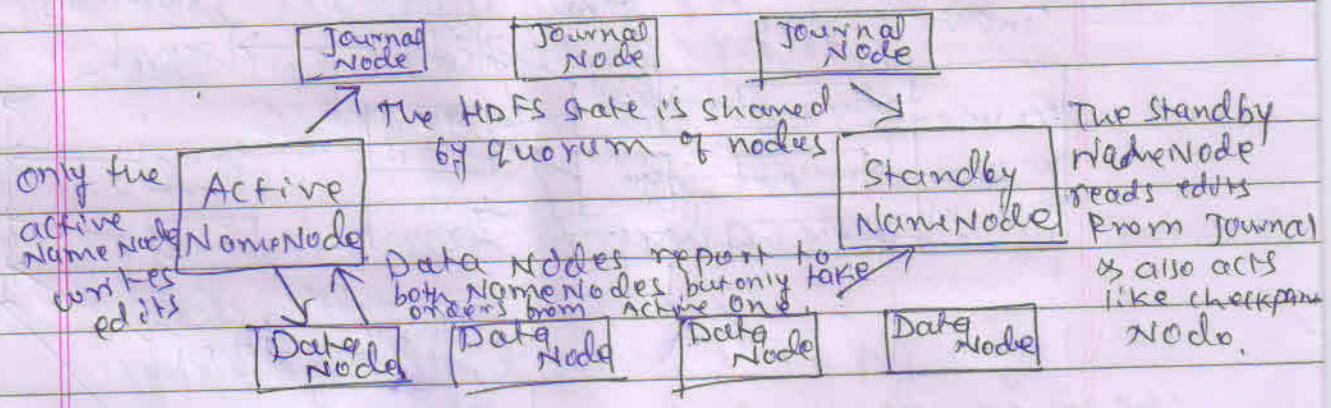
Q2

(b) Explain the following roles in HDFS deployment with a diagram (i) High Availability (ii) Name Node Federation

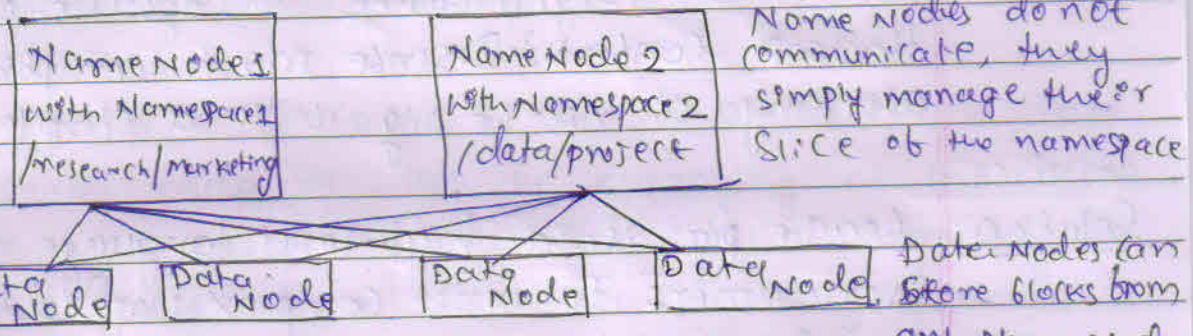
Solution

(i) High Availability (HA)

* Name Node High Availability provides true Failover Service.
* An HA Hadoop cluster has two Name Node Machines :-
  1. Active Name Node
  2. Standby Name Node
* Active Name Node is responsible for all client HDFS operation in the cluster. The standby Name Node maintains enough state to provide a fast failover when needed.
* Both Active & Standby NameNodes receive blocks reports from Data Nodes.
* The Journal Nodes provide proper Synchronization for the Standby Node.



ii) Name Node Federation



* Instead of having single namespace for the entire cluster, the Federation provides multiple NameNodes/namespaces to the HDFS file system.

* The benefits of multiple Namenodes/namespaces are:

i) Namespace scalability - HDFS cluster storage scales horizontally without placing a burden on the Name Node.

ii) Better Performance:- Adding more Namenodes to the cluster scales the file system read/write operations throughput by separating total namespace.

iii) System Isolation:- Multiple NameNodes enable different categories of applications to be distinguished & users can be isolated to different namespaces.

* In the above figure NameNode 1 manages /research & /marketing namespaces & NameNode 2 manages /data & /project namespaces.

Module-2

Q3

a) What is significance of Apache pig in Hadoop context? Describe main components and the working of Apache pig with a simple example.

Solution Apache pig is a high-level language that enables programmers to write complex MapReduce transformations using a simple scripting language.

Apache Pig Components

i) Parser:- Initially Pig scripts are handled by the parser. It checks the syntax of the script, does type checking, etc. The output of the parser will be Directed Acyclic Graph (DAG).

ii) Optimizer :- The DAG is passed to the logical optimizer, which carries out the logical optimizations such as projections & push docon.

iii) Compiler:- The compiler compiles the optimized logical plan into a series of MapReduce jobs.

iv) Execution Engine:- Finally the MapReduce jobs are submitted to Hadoop in a sorted order. Finally these MapReduce jobs are executed on Hadoop producing the desired results.

Example:-

Pig starts with a grunt> prompt. pig commands ends with a semicolon(;)

```
grunt> A = load 'passwd' using pigStorage(':');
grunt> B = foreach A generate $0 as id;
grunt> dump B;
```

The output is list of user Names printed on screen.

To exit, enter quit.

```
$grunt>quit
```

Pig can also be run from script.

```
/* id.pig */
A= load 'passwd' using PigStorage (':');
B= foreach A generate $0 as id;
dump B;
Store B into 'id.out'
```

Comments are written in pig with /*...*/

To execute id.pig use below command

```
$pig -x local id.pig
```

Q3

(b) Explain the features and the benefits of Apache HIVE in Hadoop.

Solution: Apache pig & Hive is a datawarehouse infrastructure built on top of Hadoop for data summarization, adhoc queries, & the analysis of large data sets using SQL-like language called Hive-QL.

* Hive is the defacto standard for interactive SQL queries over petabytes of data using Hadoop.

* Features of Hive :-

→ Provides tools to enable easy data extraction transformation & loading (ETL)

→ Provides a mechanism to impose structure on a variety of data formats.

→ Access to files stored either directly in HDFS or in other database system like HBASE

→ provides Query execution via MapReduce and Tez.

*Hive provides users with SQL capability to query the data on Hadoop clusters.

Benefits of Hive.
→ keeps queries running fast.
→ Takes very less time to write Hive Query in comparison to MapReduce code.
→ Hive QL is a declarative language like SQL.
→ Provides the structure on an array of data formats.
→ Multiple users can query the data with the help of HiveQL.
→ Simple to learn & use.

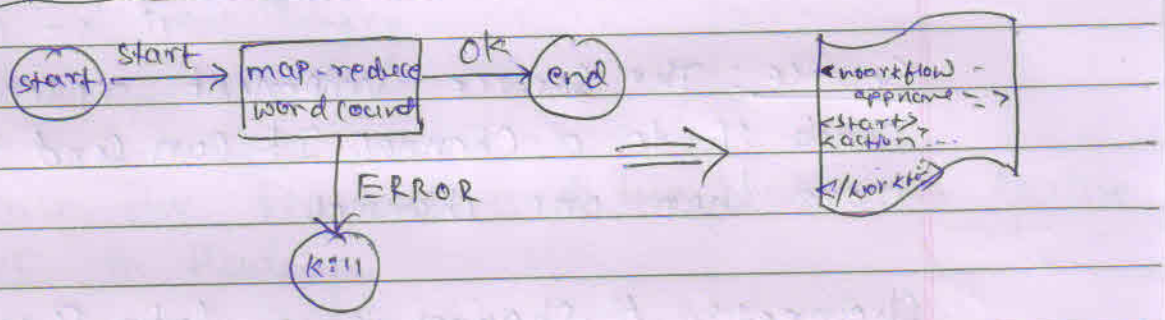
Q 4
a) With neat diagrams, explain the oozie DAG workflow and the types of nodes in workflow.

Solution) oozie is a workflow director system designed to run and manage multiple related Apache Hadoop jobs.
Types of nodes :-
i) Control flow nodes - define the beginning and the end of a workflow. They include start, end and optional fail nodes.
ii) Action nodes - These are where the actual processing tasks are defined. When an action node finishes, the remote system notify oozie & the next node in the workflow is executed.

→ Fork/join nodes:- These enable parallel execution of tasks in the workflow. The fork node enables two or more tasks to run at the same time. A join node represents a rendezvous point that must wait until all forked tasks complete.

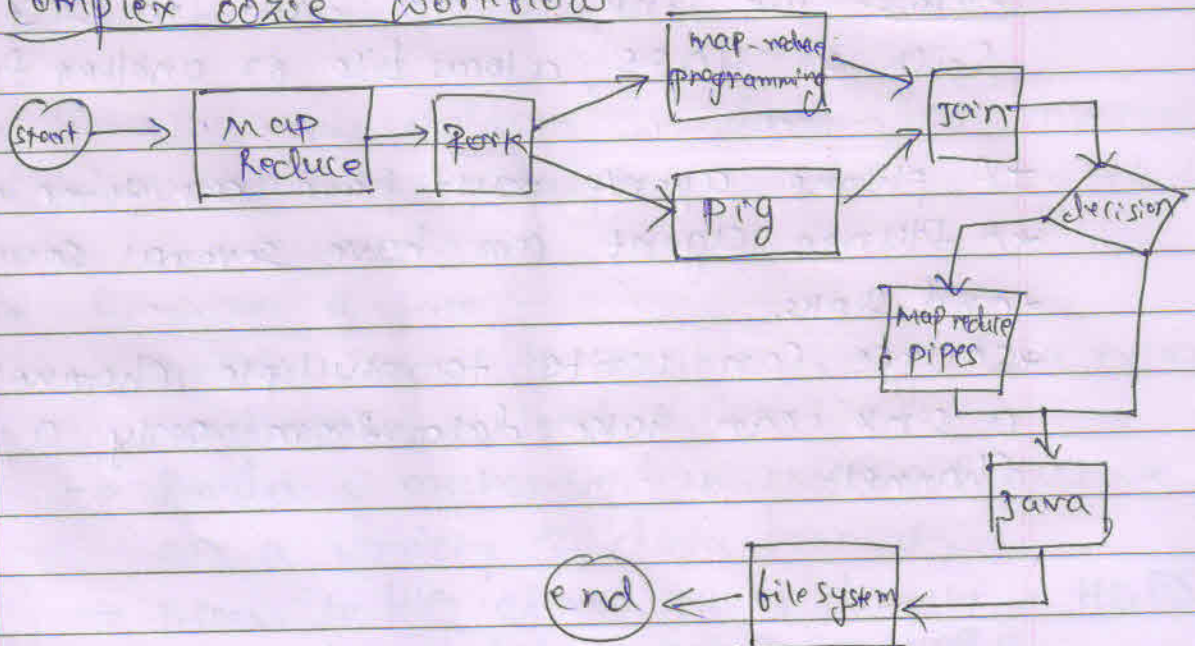
→ Control flow nodes. These enable decisions to be made about the previous task. Control decisions are made based on the results of previous actions.

Simple oozie workflow



Here oozie runs a basic MapReduce operation. If the application was successful, the job ends; if an error occured, the job is killed.

Complex oozie workflow



This workflow uses all node types.

b) What is Apache Flume? Describe the features, components and the working of Apache Flume.

Solution * Apache Flume is an independent agent designed to collect, transport & store data into HDFS.

*Flume is often used for log files, social media generated data, email messages & just about any continuous data source.

Components of Flume agent.

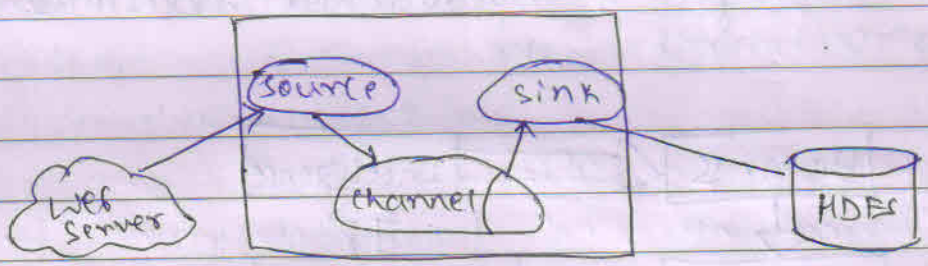
Source: The source component receives data & sends it to a channel. It can send the data to more than one channel.

Channel:- A channel is a data queue that forwards the source data to the sink destination. It can be thought of as a buffer that manages input and output flow rates.

Sink:- The sink delivers data to destination such as HDFS, a local file or another flume agent.

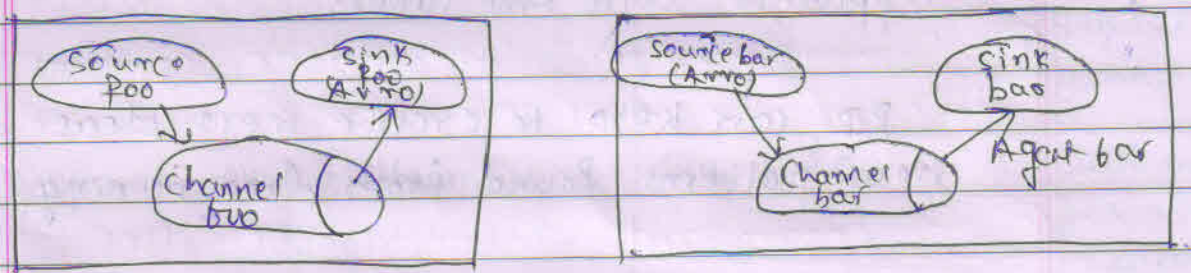
*A flume agent must have all three components.
*A flume agent can have several sources, channels and sinks.
*Source can write to multiple channels, but a sink can take data from only a single channel.

Flume Agents

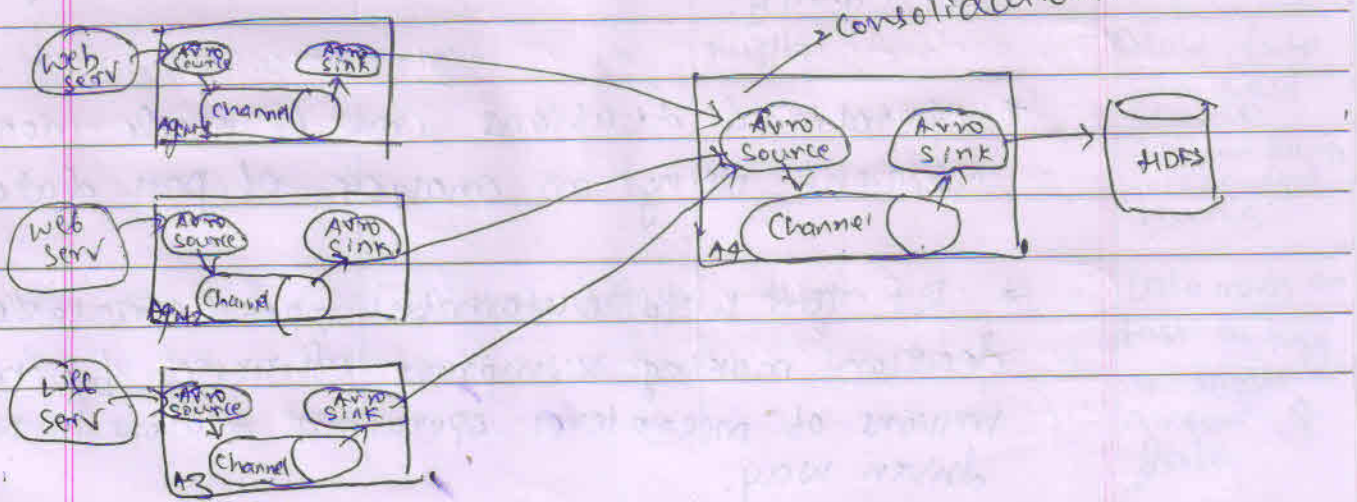


Flume agents may be placed in a pipeline, to traverse several machines or domains. This configuration is normally used when data are collected on one machine & sent to another machine that has access to HDFS.

In a Flume pipeline, the Sink from one agent is connected to the source of another. The data transfer is done by Avro, Avro uses RPC to send data.



Flume can consolidate several data sources before committing them to HDFS

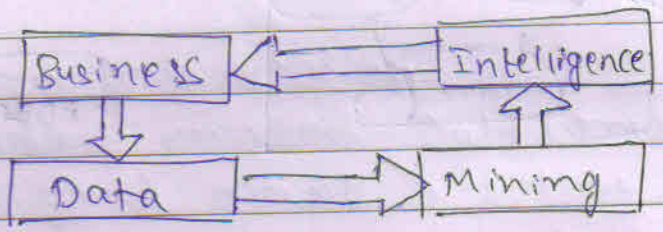


Module – 3

Q5

(a) Draw the flow of BIDM cycle. Explain the strategic & operational decisions.

Solution



Strategic decisions

* Strategic decisions are those that impact the direction of company.

* In Strategic decision-making, the goal itself may or may not be clear & the same is true for the path to reach the goal.

* The consequences of the decision would be apparent some time later.

* BI can help to create new ideas based on new patterns found from data mining.

Operational decisions

* operational decisions are more routine & tactical decisions focused on developing greater efficiency.

* Operational decisions can be made more efficient using an analysis of past data.

* BI can help automate operation level decision making & improve efficiency by making millions of micro level operational decisions in model-driven way.

Q3

(b) Differentiate between datamart & data warehouse based on the following with justifications:
i) Scope ii) Target organization iii) Cost iv) Approach
v) Complexity vi) Time.

	Datamart	Datawarehouse	Justification
i) Scope	one subject/functional area	Complex enterprise data needs.	Smaller data marts aligen to deliver Datawarehouse capabilities.
~~Virtue~~	~~Functional area reporting & insights~~	~~Centralized management~~	
ii) Target organization	Decentralized management	Centralized management	Datamart-one line business Datawarehouse-multiple areas
iii) Cost	Low	High	Datamart has less size than Datawarehouse
iv) Approach	Bottom-up	Top-down	Data mart focuses on single subject, warehouse takes data from multiple functional units
v) Complexity	Low	High	Data mart draws data from few sources, warehouse collects from varied sources.
vi) Time	Low to medium	High	Data marts are fast as they use small amount of data.

Q6

a) Describe any 8 considerations for a data warehouse and explain the key elements with a diagrammatic representation.

Solution Design Consideration

(1) Subject oriented - Data warehouse should be designed around a subject domain.

(2) Integrated - Data warehouse should include data from many functions that can shed light on a particular subject area.

(3) Time variant (time series): Data in a data warehouse should grow at daily or other chosen intervals.

(4) Non volatile :- Data warehouse should be persistent, it should not be created on the fly from operational databases.

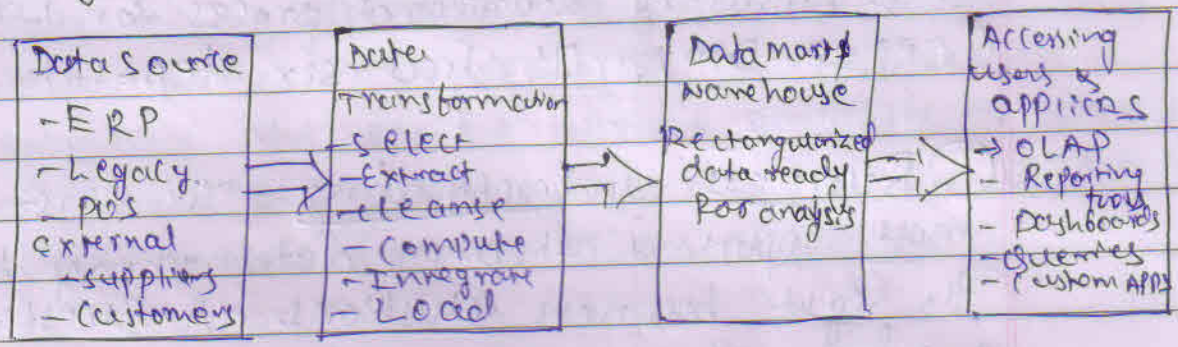
(5) Summarized :- Data warehouse contains rolled-up data at the right level for queries and analysis.

(6) Not normalized :- Data warehouse uses star schema, which is a rectangular central table surrounded by some look up tables the single table view significantly enhances speed of queries.

(7) Metadata :- Many of the variables in the database are computed from other variables in the operational database.

(8) Near Real time / Right time / Active): Data warehouses should be updated in near real time in many high transaction volumes.

Key elements



i) Data source :- Provies the raw data.

ii) Data transformation :- Transforms the data to meet the decision needs.

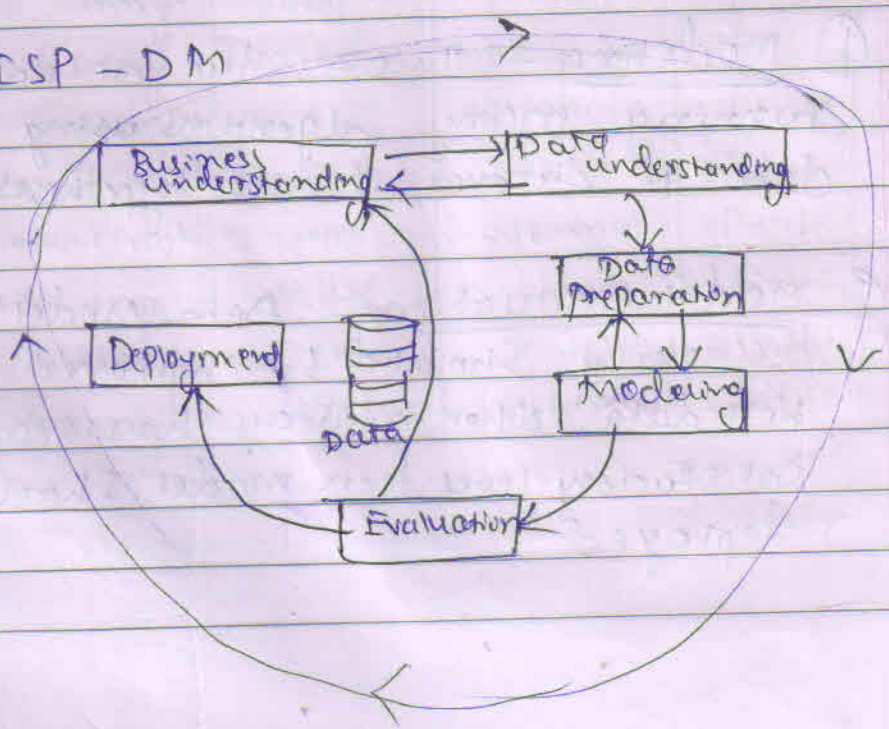
iii) Data Mart/Warehouse :- Method of regularly and accurately loading of data in Data warehouse/Mart.

iv) Accessing users & applications :- The devices & applications use data from Data warehouse to deliver insights and other benefits to users.

Q 6 b) Explain CRISP DM Cycle with a diagram.

Solution

CRISP - DM



The Data Mining Industry has proposed a Cross-Industry standard process for Data Mining (CRISP-DM). It has six steps:-

(1) Business understanding :- The first and the most important step in data mining is asking the right business questions. A question is good if answering it would lead to large payoffs for the organization, financially. A relative important step is to be creative and open in proposing imaginative hypotheses for the solution.

(2) Data Understanding :- One needs to be imaginative in scouring many elements of data through many sources in helping address the hypotheses to solve a problem.

(3) Data Preparation - The data should be relevant, clean & of high quality. Data cleaning can take 60-70% of the time in data mining project. New elements can be added from external sources of data that can improve predictive accuracy.

(4) Modeling - This is the actual task of running many algorithms using the available data to discover if the hypotheses are supported.

(5) Model evaluation - ~~Data~~ Model's predictive accuracy should be improved with more test data. When accuracy has reached some satisfactory level, then model should be deployed.

6. Dissemination & Roll out - It is important that data mining solution is presented to the key stakeholders & is deployed in organization. Otherwise the project will be a waste of time & a setback for establishing and supporting a data-based decision-process culture in the organization.

Module - 4

Q 7

a) Explain the steps & three differentiating criteria of a decision tree algorithm. Construct a decision tree for the following dataset and predict the outcome for the given question.

OUTLOOK	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	NO
overcast	Hot	High	False	Yes
Rainy	mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	~~Yes~~ No
overcast	Cool	Normal	True	Yes
Sunny	mild	High	False	NO
Sunny	Cool	Normal	False	Yes
Rainy	mild	Normal	False	Yes
Sunny	mild	Normal	True	Yes
overcast	Mild	High	True	Yes
overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No.

OUTLOOK	Temp	Humidity	Windy	Play
Sunny	Hot	Normal	True	?

Solution Steps in decision tree algorithm:

1. Create a root node and assign all of the training data to it.
2. Select the best splitting attribute according to certain criteria.
3. Add a branch to the root node for each value of the split.
4. Split the data into mutually exclusive subsets along the lines of the specific split.
5. Repeat steps 2 & 3 for each and every leaf node until a stopping criteria is reached.

Decision trees algorithms differ on three key elements

i) Splitting Criteria:-

a) Deciding Splitting Variable:- Algorithms use different measures like least errors, information gain, Gini's coefficient, etc. to compute splitting variable that provides the most benefit

b) Values to be splitted:- If the variables are continuous then what value-range should be used to make bins?

c) Branch creation for nodes:- There could be binary trees, with just two branches at each node. Or there could be more branches allowed.

ii) Stopping Criteria - Tree building can be stopped when a certain depth of the branches has been reached & trees become unreadable after that.

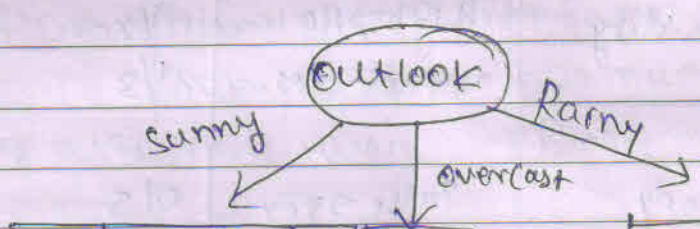
ii) Pruning: It is the act of reducing the size of decision tree by removing sections of the tree that provide little value.

pre pruning - Halt the tree construction early, when certain criteria are met.

Post pruning - Removing branches or subtrees from a fully grown tree.

Attribute	Rules	Error	Total Error
Outlook	Sunny → NO	2/5	4/14
	Overcast → yes	0/4	
	Rainy → yes	2/5	
Temperature	Hot → NO	2/4	5/14
	Mild → yes	2/6	
	Cool → yes	1/4	
Humidity	High → NO	3/7	4/14
	Normal → yes	1/7	
windy	False → yes	2/8	5/14
	True → NO	3/6	

The variable that leads to least number of errors should be chosen as the first node.



Temp	Humidity	Windy	Play
Hot	High	False	No
Hot	High	True	No
Mild	High	False	No
Cool	Normal	False	No
Mild	Normal	True	Yes

Temp	Humidity	Wind	Play
Mild	High	False	Yes
Cool	Normal	False	Yes
Cool	Normal	True	No
Mild	Normal	False	Yes
Mild	High	True	No

Determine the next node at left branch.

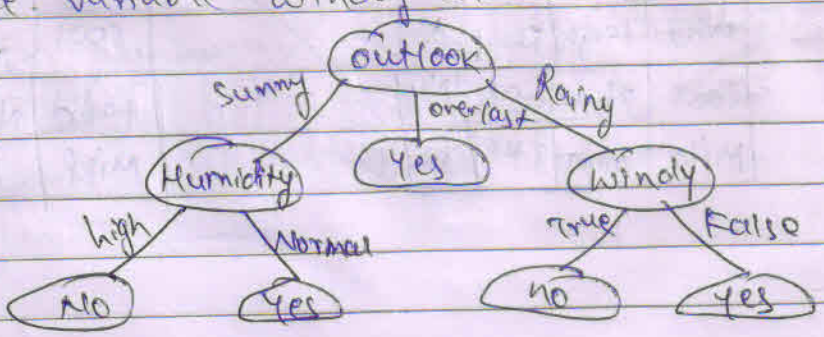
Attribute	Rules	Error	Total Error
Temperature	Hot → No	0/2	1/5
	Mild → No	1/2	
	Cool → Yes	0/1	
Humidity	High → No	0/3	0/5
	Normal → Yes	0/2	
	False → No	1/3	
Windy	False → No	1/3	2/5
	True → Yes	1/2	

The variable Humidity shows least error.

Right branch -

Attribute	Rules	Error	Total Error
Temperature	Mild → Yes	1/3	2/5
	Cool → Yes	1/2	
Humidity	High → No	1/2	2/5
	Normal → Yes	1/3	
Windy	False → Yes	0/3	0/5
	True → No	0/2	

The variable windy shows least error.



In the given problem outlook is sunny & Humidity is normal, which leades to answer - Yes.

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	Normal	True	Yes.

Q7

(b) Differentiate between C4.5, CART & CHAID decision tree algorithms.

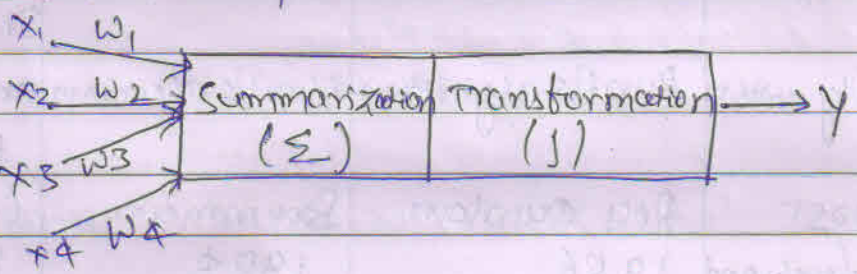
Solution	Decision Tree	C4.5	CART	CHAID
①	Full Name	ID3 (Iterative Dichotomizer)	Classification & Regression Trees	Chi-square Automatic Interaction detector
②	Basic Algorithm	Hunt's algorithm	Hunt's algorithm	Adjusted significance testing.
③	Developer	Ross Quinlan	Breiman	Gordon Kass
④	When developed	1986	1984	1980
⑤	Type of trees	classification	Classification & Regression trees	Classification & Regression trees
⑥	Serial implementn	Tree growth and tree pruning	Tree growth & Tree pruning	Tree growth & Tree pruning
⑦	Type of Data	Discrete & continuous Incomplete data	Discrete & continuous	Non normal data also accepted
⑧	Type of splits	Multi-way splits	Binary splits only	Multiway splits
⑨	Splitting criteria	Information gain	Gini's coefficient & others	Chi-square test
⑩	Pruning criteria	Clever bottom-up technique avoid over-fitting	Remove weakest link first	Trees can become very large
⑪	Implementation	publicly available	Publicly available in most packages	Popular in market research for segmnts

Q 8

a) Explain the design principles of ANN by constructing a model representation for a single and multilayer ANN. Describe the steps to build an ANN.

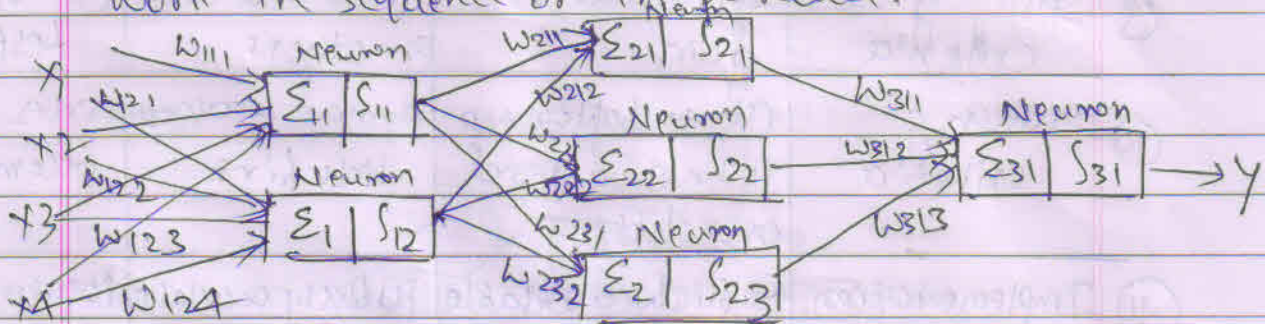
Solution: Design principles of ANN.

① A neuron is the basic processing unit of the network. The neuron receives inputs from its preceding neurons, does some nonlinear weighted computation on the basis of those inputs, transforms the result into its output value, & then passes on the output to the next neuron in the network.



⇒ x's are inputs, w's are the weights for each input, & y is output.

② A neural network is a multilayered model. ANN's can have multiple layers of processing elements in sequence. There could be many neurons involved in a sequence depending upon the complexity of the predictive action. The layers of PEs could work in sequence or in parallel.



(3) The processing logic of each neuron may assign different weights to the various incoming input patternstreams. The processing logic may also use non linear transformation.

(4) The neural network can be trained by making similar decisions over & over again with many training cases. It will continue to learn by adjusting its internal computation & communication based on feedback about its previous decisions.

Depending upon the nature of the problem and the availability of good training data, at some point, the neural network will learn enough & begin to match the predictive accuracy of a human expert.

Steps to build an ANN

(1) Gather data & divide into training data & test data. The training data needs to be further divided into training data & validation data.

(2) Select the network architecture such as Feedforward network.

(3) Select the algorithm, such as multilayer perceptron.

(4) Set network parameters.

(5) Train the ANN with training data.

(6) Validate the model with validation data.

(7) Freeze the weights & other parameters.

(8) Test the trained network with test data.

(9) Deploy the ANN when it achieves good predictive accuracy.

Q 8 (b) For the dataset, find the affinities of the product - product which sell together. Consider s = 33%, c = 50% and 3-itemset level only.

Transaction List				
1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	ketchup
3	Bread	Butter	ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	ketchup

Solution

~~Transaction~~

1 itemsets	Frequency
Milk	9
Bread	10
Butter	10
Egg	3
ketchup	3
Cookies	5

If itemsets that occur 4 or more times out of 12 are selected, that corresponds to meeting a minimum

Support level of 33 percent (4 out of 12). only 4 items make the cut. The frequent items that meet the support level of 33% are:

Frequent 1-item sets	Frequency
Milk	9
Bread	10
Butter	10
Cookies	5

The next step is to go for next level of itemsets using selected earlier

2-item sets	Frequency
Milk, Bread	7
Milk, Butter	7
Milk, Cookies	3
Bread, Butter	9
Butter, Cookies	3
Bread, Cookies	4

only four transactions meet the minimum support level of 33%.

2-Itemsets	Frequency
Milk, Bread	7
Milk, Butter	7
Bread, Butter	9
Bread, Cookies	4

The next step is to list the next higher level of Itemsets i.e. 3-item itemsets.

3-itemsets	Frequency
Milk, Bread, Butter	6
Milk, Bread, Cookies	1
Bread, Butter, Cookies	3

only one 3-item itemset meets the minimum support requirements

3-item sets	Frequency
Milk, Bread, Butter	6

Module-5

Q9

a Compare Text mining & data mining.

Solution

Dimension	Text mining	Data mining
Nature of Data	unstructured data: words, phrases, sentences	Numbers, alphabetical, & logical values.
Language used	Many languages & dialects used in world;	similar numerical systems across the world.
Clarity & precision	Sentences can be ambiguous. Sentiment may contradict the words	Numbers are precise.
Consistency	Different parts of the text can contradict each other	Different parts of the data can be inconsistent, thus requiring statistical significance analysis.
Sentiment	Text may present a clear & consistent or mixed sentiment across continuum	Not applicable
Quality	Spelling errors. Differing values of proper nouns	Issues with missing values, outliers, etc.

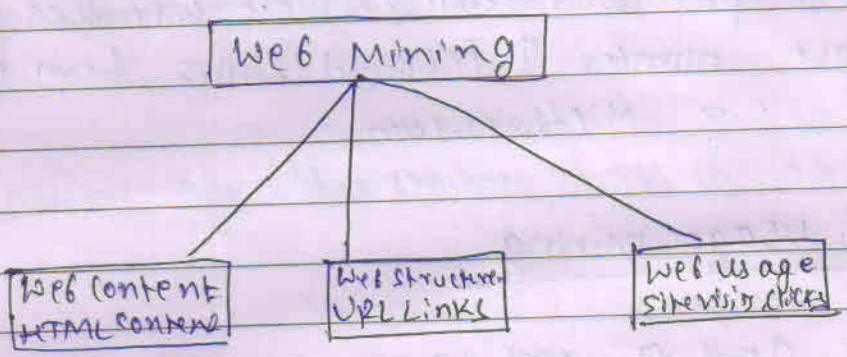
Dimension	Text mining	Data Mining
Nature of Analysis	Keyword based search; sentiment mining	A true wide range of statistical & machine learning analysis for relationships & differences

Q 9

b) Explain the 3 types of Web mining. Use appropriate flow diagrams to represent the same.

Solution

Three types of web mining



Web content mining : A Website is designed in the form of pages with a distinct URL. A large website contain thousands of pages. These pages are managed by content management system.

The Websites keep a record of all requests received for its page/URLs. The log of these requests could be analyzed to guage the popularity of those pages among different segments of population. The text & application content on the pages could be analyzed for its usage by visit counts. The pages on a website themselves could be analyzed for quality of content that attract most users.

Web Structure mining :- The structure of web pages could be analyzed to examine the pattern

of hyperlinks among the pages. There are two basic strategic models for successful websites -
i) Hubs ii) Authorities.

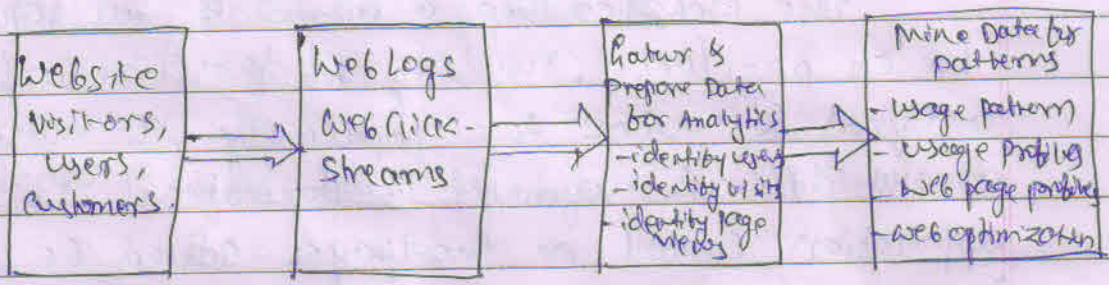
Hubs:- These are the pages with a large number of interesting links. They serve as a hub or a gathering point, where people visit to access a variety of information.
e.g. yahoo.com

Authorities:- People gravitate towards page that provide the most complete and authoritative information on a particular subject. These websites have the most number of inbound links from other websites.
e.g. NYtimes.com

Web usage mining

The goal of web usage mining is to extract useful information and patterns from data generated through web page visits & transactions.

Multiple level analysis of web content.



* The server side analysis would show the relative popularity of the web pages accessed.

* The client side analysis focus on usage pattern of the actual content consumed by created by users.

a) Usage pattern could be analyzed using 'clickstream' analysis i.e. analyzing web activity for patterns of sequence of clicks & location & duration of visits on websites. Clickstream analysis can be useful for web activity analysis, software testing, etc.

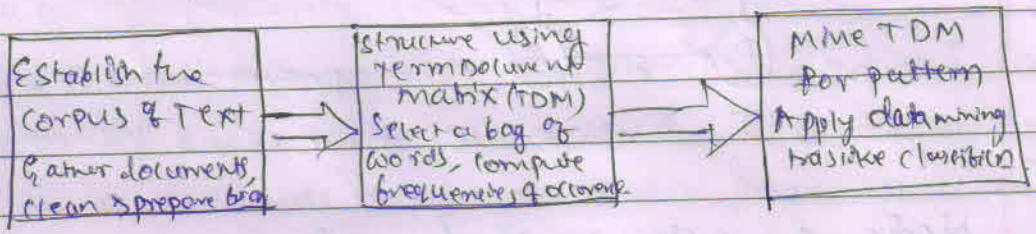
b) Textual information accessed on the pages retrieved by users could be analyzed using text mining techniques.

Web usage mining helps to predict user behavior based on previously learned rules & user's profiles and can help determine lifetime value of clients.

Q10

(a) Explain the text mining process & the architecture.

Solution



* The first level of analysis is identifying frequent words. This creates a bag of important words. Text can then be ranked on how they match to a particular bag-of-words.

* The next level is to identify meaningful phrases from words.

* The next level is that of topics. Multiple phrases can be combined into Topic area.

* Text mining is a semi automated process. Text data needs to be gathered, structured & then mined in a 3-step process:-

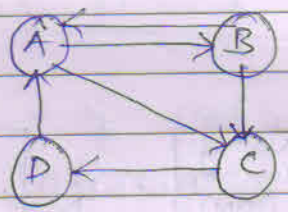
(1) The text & documents are first gathered into a corpus and organized.

(2) The corpus is then analyzed for structure. The result is a matrix mapping important terms to source documents.

(3) The structure data is then analyzed for word structures, sequences & frequency.

Q10

(b) Compute the rank values for the below network.



Solution

Node A gets all of the influence of node D and half the influence of node B

$$R_a = 0.5 * R_b + R_d$$

Node B gets half the influence of node A.

$$R_b = 0.5 * R_a$$

Node C gets half the influence of node A & half the influence of node B.

$$R_c = 0.5 \times R_a + 0.5 \times R_b$$

Node D gets all of the influence of node C & half influence of node B.

$$R_d = R_c .$$

Influence matrix

	R_a	R_b	R_c	R_d
R_a	0	0.5	0	1.00
R_b	0.5	0	0	0
R_c	0.5	0.5	0	0
R_d	0	0	1.00	0

Assume initial rank values as below.

Variable	Initial value
R_a	0.250
R_b	0.250
R_c	0.250
R_d	0.250

Computing revised values using earlier equations. we get,

Variable	Initial value	Iteration 1
R_a	0.250	0.375
R_b	0.250	0.125
R_c	0.250	0.250
R_d	0.250	0.250

$R_a = R_b \times 0.5 + R_d$
$= 0.25 \times 0.5 + 0.25 = 0.375$

$R_b = 0.5 \times R_a$
$= 0.5 \times 0.25$
$= 0.125$

$R_c = 0.5 \times R_a + 0.5 \times R_b$
$= 0.5 \times 0.25 + 0.5 \times 0.25$
$= 0.25$

$R_d = R_c$
$= 0.25$

III'y Iteration 2

Variable	Initial value	Iteration 1	Iteration 2
R_a	0.25	0.375	0.3125
R_b	0.25	0.125	0.1875
R_c	0.25	0.250	0.25
R_d	0.25	0.250	0.25

We do few more iterations till the values stabilize.

Variable	Initial value	Iteration 1	Iteration 2	....	Iteration 8
R_a	0.25	0.375	0.313	....	0.333
R_b	0.25	0.125	0.188	....	0.167
R_c	0.25	0.25	0.250	....	0.250
R_d	0.25	0.25	0.250	....	0.250

The final rank shows that rank of node A is highest at 0.33. Hence important node is 'A'.