CBCS SCHEME

USN | 2 | V | D | 1 | 6 | C | S | 0 | 2 | 8 | 17CS82

Eighth Semester B.E. Degree Examination, July/August 2021
Big Data Analytics

Time: 3 hrs. Max. Marks: 100

Note: *Answer any FIVE full questions.*

1 a. What is HDFS? Explain its components with a neat diagram. **(10 Marks)**
b. Explain HDFS safemode and rack awareness, with neat diagram. **(10 Marks)**

2 a. What is MapReduce Program? Explain MapReduce parallel data flow, with neat functional diagram. **(10 Marks)**
b. What is Nano node federation? Explain NaneNode high availability design with diagram. **(10 Marks)**

3 a. Explain Apache Sqoop Import and Export methods, with neat diagram. **(10 Marks)**
b. How do you run MapReduce and message passing interface on Yarn architecture? **(10 Marks)**

4 a. Explain with a neat diagram, the Apache Oozie work flow for Hadoop architecture. **(10 Marks)**
b. What is YARN? Explain Yarn application frame work. **(10 Marks)**

5 a. What is Business Intelligence? List the different BI applications and explain in detail any four applications. **(10 Marks)**
b. Draw and explain flow of BIDM cycle. Explain the Strategic and Operational decisions. **(10 Marks)**

6 a. Explain CRISP DM cycle, with neat diagram. **(10 Marks)**
b. Define Data warehouse and illustrate design considerations for data warehouse. **(10 Marks)**

7 a. What is Association Rule? Explain below given rules with suitable examples :
i) Support ii) Confidence iii) Lift. **(10 Marks)**
b. What is Unsupervised Machine Learning concept? Explain K – Means clustering techniques, with suitable example. **(10 Marks)**

8 a. Write and explain Apriori Algorithm with example. **(10 Marks)**
b. List and explain the steps for developing an ANN (Artificial Neural Network). **(10 Marks)**

9 a. Discuss the application and practical consideration of Social network Analysis. **(10 Marks)**
b. Explain the 3 types of Web mining. Use appropriate flow diagrams to represent the same. **(10 Marks)**

10 a. Explain the Text Mining process and the architecture. **(10 Marks)**
b. Briefly explain the Data Mining. Compare text mining and data mining. **(10 Marks)**

* * * * *

BIG DATA ANALYTICS (17CS82)

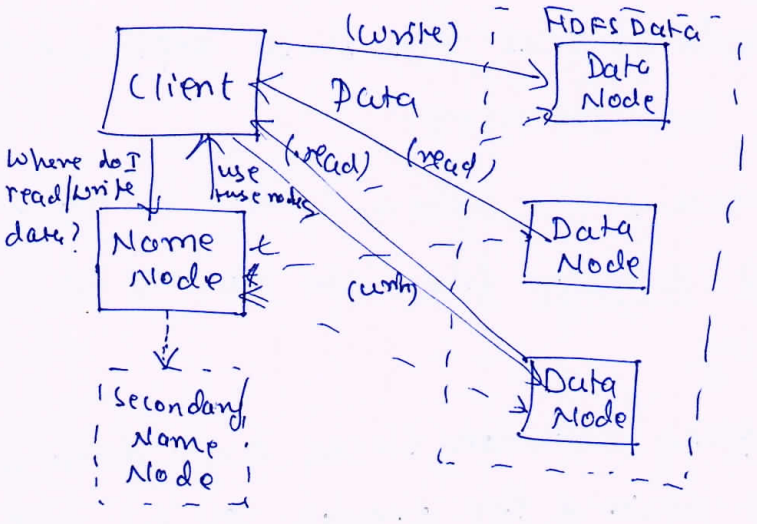
Eighth Semester B.E. Degree Examinations July/August 2021

Note: Answer Any FIVE Full questions

Q1 a) What is HDFS? Explain its components with a neat diagram.

Solution Hadoop Distributed File System (HDFS) is used for storing data in Hadoop cluster. HDFS is a redundant and highly reliable distributed file system.

HDFS Components



-5M

There are two types of nodes:

1) NameNode
2) multiple Data Node.

Single NameNode manages all the metadata needed to store and retrieve the actual data from Data Node. No data is actually stored on the NameNode. For a minimal Hadoop installation, there needs to be a single NameNode daemon & single DataNode daemon running on at least one ML

SM

The DataNodes are responsible for serving read & 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 the client with the DataNodes that will store the data.

- The purpose of Secondary Name Node is to perform periodic checkpoints that evaluate the status of the NameNode.

Q1(b) Explain HDFS Safemode and rack awareness, with neat diagram.

Solution

HDFS Safemode

SM

- In safemode, blocks cannot be replicated or deleted.
- Performs two processes:

1) The previous file system state is reconstructed by loading fsimage file into memory and replaying the edit log.

2) The mapping between blocks and data nodes is created by waiting for enough of the DataNodes to register so that atleast one copy of the data is available.

Rack Awareness

SM

- Rack awareness deals with data locality. Hadoop Cluster will exhibit three levels of data locality:

1. Data resides on the local machine
2. Data resides in the same rack
3. Data resides in a different rack.

when the YARN scheduler is assigning MapReduce containers to work as mappers, it will try to place the container first on the local machine, then on the same rack and finally on another rack.

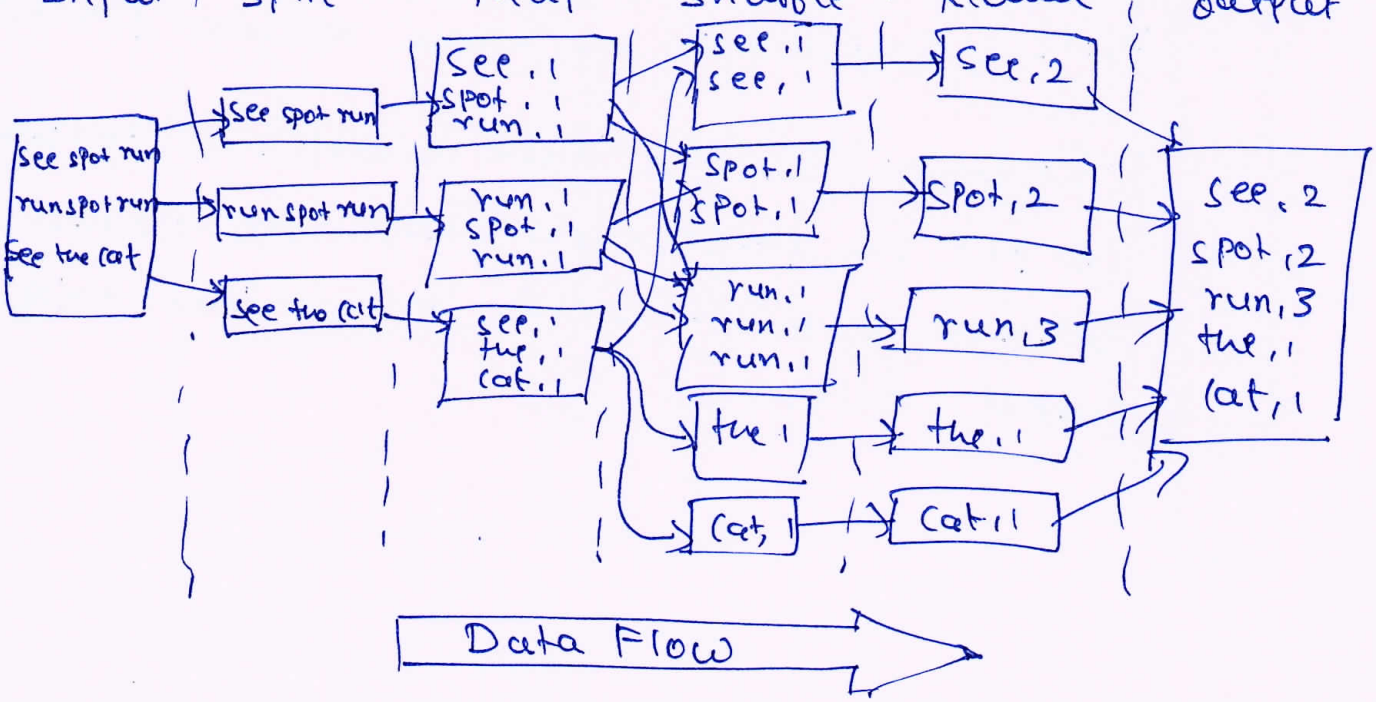
The NameNode tries to place replicated data blocks on multiple racks for improved fault tolerance.

HDFS can be made rack aware by using a user-derived script that enables the master node to map the network topology of the cluster. A default Hadoop installation assumes all the nodes belong to the same rack.

Q2(a) What is mapReduce program? Explain Map Reduce Parallel data flow, with neat functional diagram.

Solution Apache Hadoop mapReduce can be scaled from one to thousands of processors. 1 M

MapReduce Parallel Data Flow. 4 M



Steps

5M

1. Input splits: These are logical boundaries based on the input data. Splits are smaller than the HDFS block size. The number of splits corresponds to the number of mapping processes used in map stage.

2. Map step: Many mappers can be operating at the same time. MapReduce will try to execute the mapper on the machines where the block resides.

3. Combiner step: It is possible to provide an optimization as part of the map stage where key-value pairs are combined prior to the next stage. The combiner stage is optional.

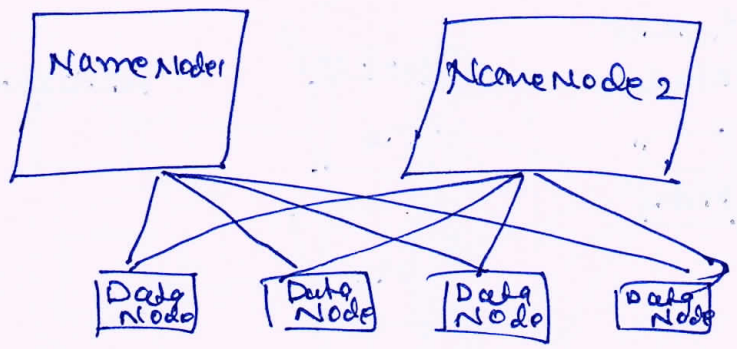
4. Shuffle Step: All similar keys must be combined and counted by the same reducer process. Therefore, results of map stage must be collected by key-value pairs & shuffled to the same reducer process.

5. Reduce step: The data reduction is performed as per the programmers design. The results are written to HDFS. Each reducer will write an output file.

Q2
(b) What is Name Node Federation? Explain Name Node high availability design with diagram.

3 M

Solution Name Node Federation addresses the problem of single namespace by adding support for multiple namenodes/ namespaces to the HDFS file System.

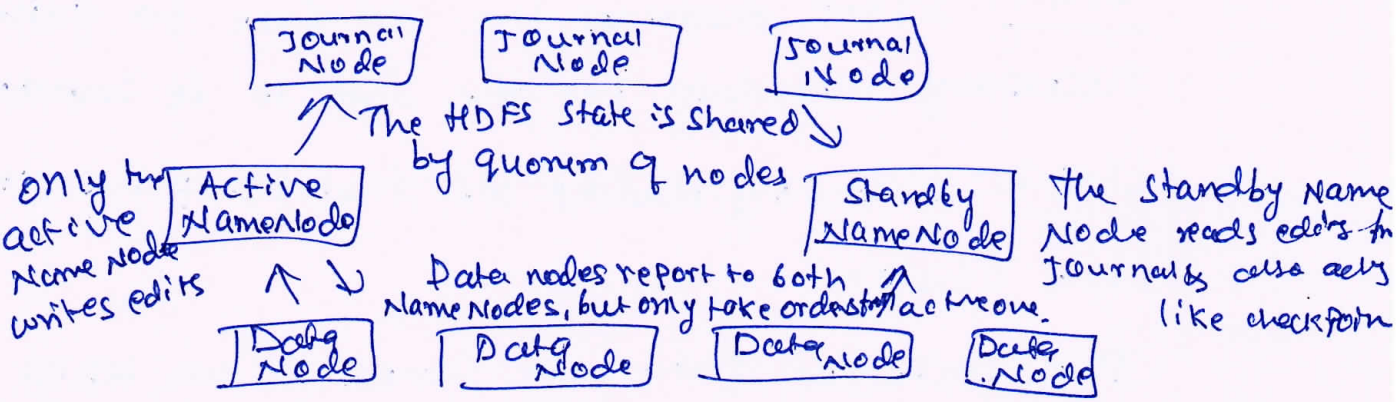


Benefits:

- Namespace Scalability
- Better performance
- System isolation.

Name Node High Availability

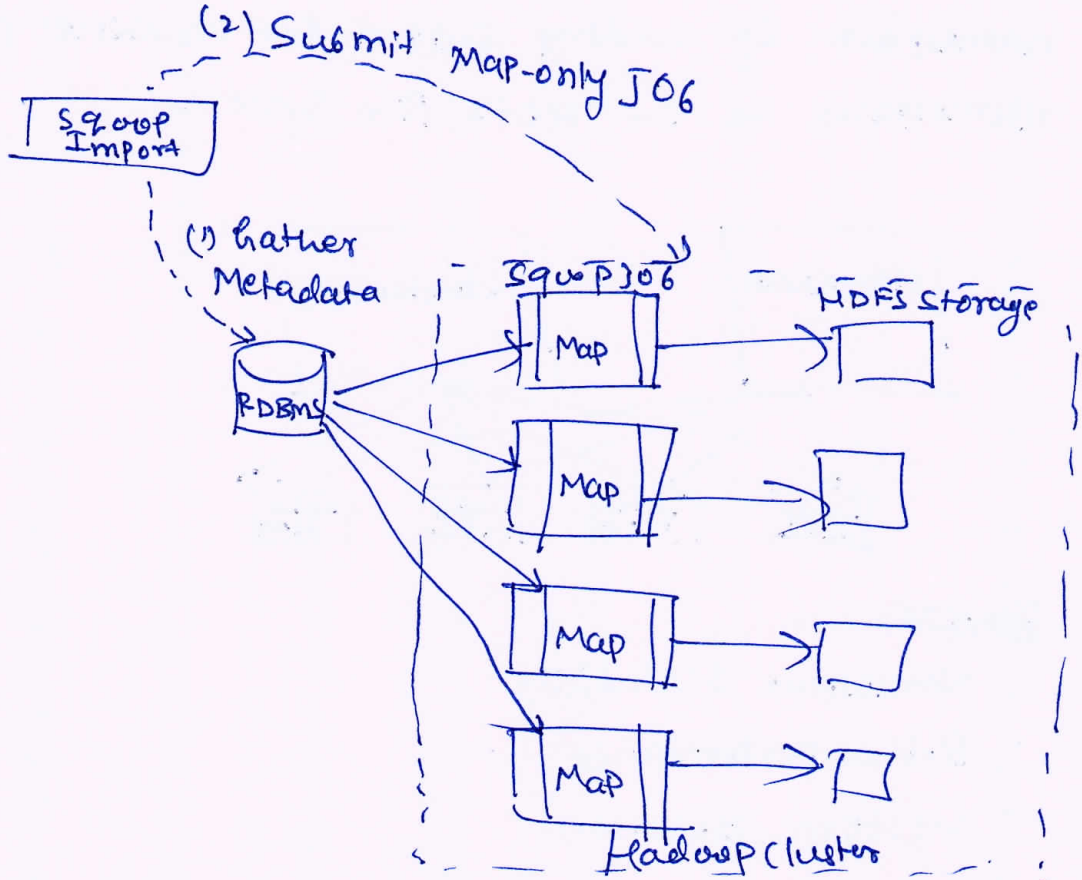
7 M



- Name Node (HA) provides true failover service.
- There are two Namenodes one is active & other is in Standby State.
- Active NameNode is responsible for all HDFS operation in the cluster.

Q3 (a) Apache Sqoop Import & Export methods, with neat diagram.

Solution **Sqoop Import** 3M



Process 2M

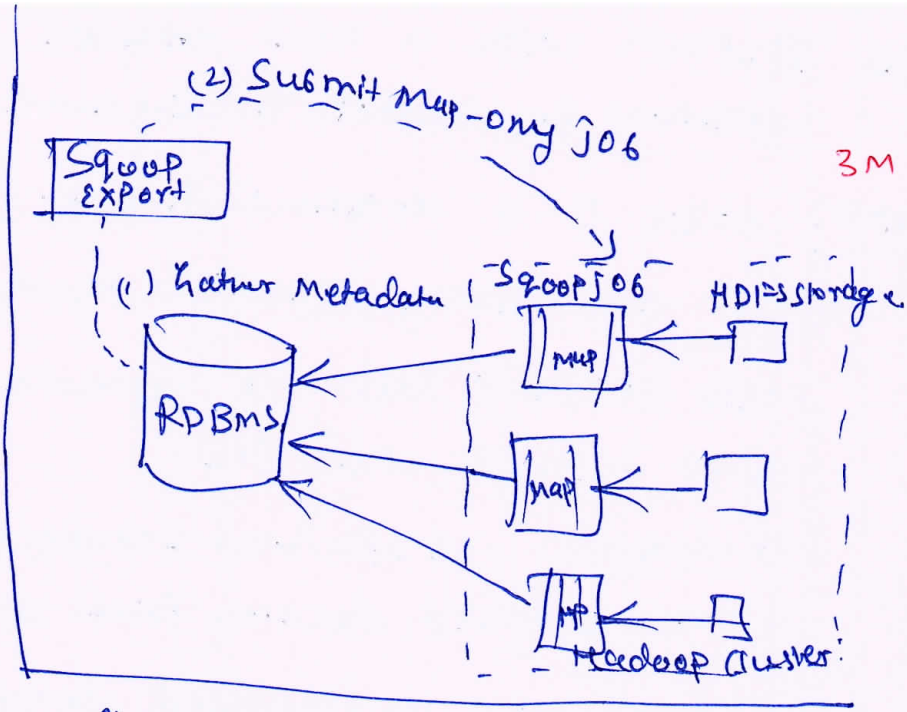
Step 1: Sqoop examines the database to gather the necessary metadata for the data to be imported.

Step 2: Map-only Hadoop job that Sqoop submits to the cluster.

The imported data are saved in an HDFS directory. Sqoop will use the databasename for the directory, or the user can specify any alternative directory where files should be populated.

Once placed in HDFS, the data is ready for processing.

Sqoop export



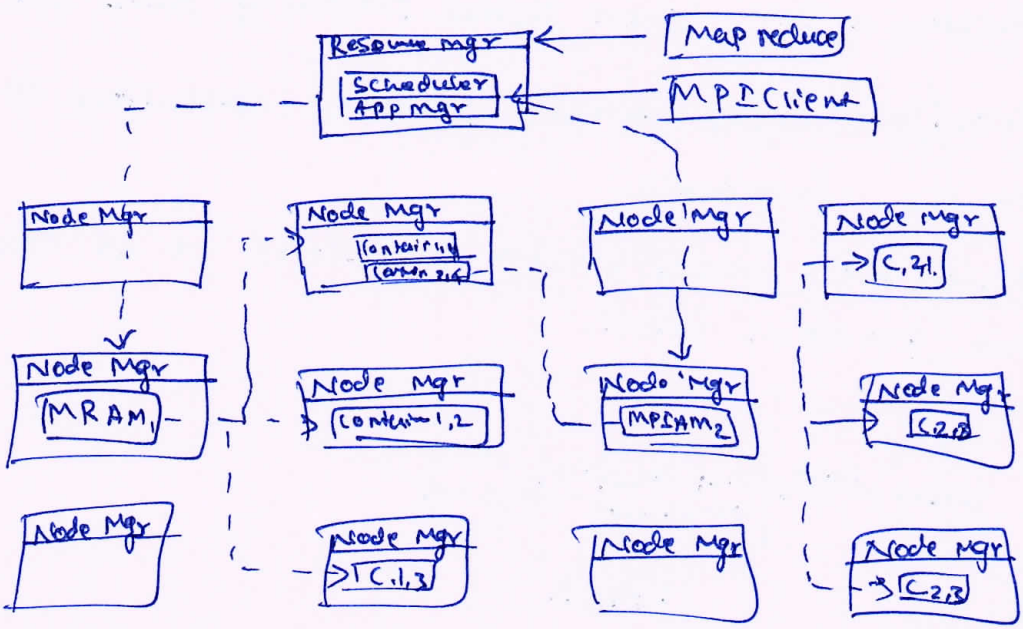
Data Export: It is a two step process:

Step1: Examine the database for metadata.

Step2: The export step again uses map-only job to write the data to the database. 2M

Sqoop divides the input data set into splits, then uses individual map tasks to push the splits to the database.

Q3 (b) How do you run MapReduce and message passing interface on YARN architecture?



YARN presents a resource management layer of Hadoop and plated into Platform, which provides services such as scheduling, fault monitoring, data locality & more to MapReduce & other frameworks.

Q4 (a) Explain with a neat diagram, the Apache oozie workflow for Hadoop architecture. 10M

Solution: Oozie is a workflow director system designed to run and manage multiple related Apache Hadoop jobs. Oozie workflow jobs are represented as DAGs. There are three types of oozie jobs

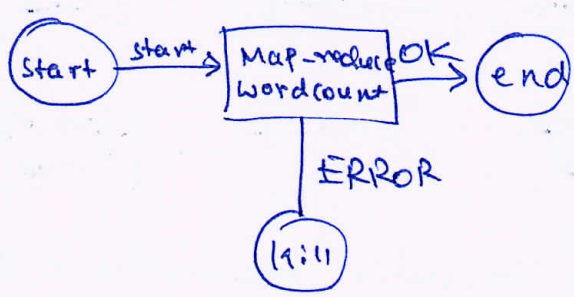
① Workflow - a specified sequence of Hadoop jobs with outcome based decision points and control dependency.

② Coordinator - a scheduled workflow job that can run at various time intervals or when data is available.

③ Bundle - a higher-level oozie abstraction that will batch a set of coordinator jobs.

Oozie workflow definitions are written in hPDL. Such workflows have several nodes:-

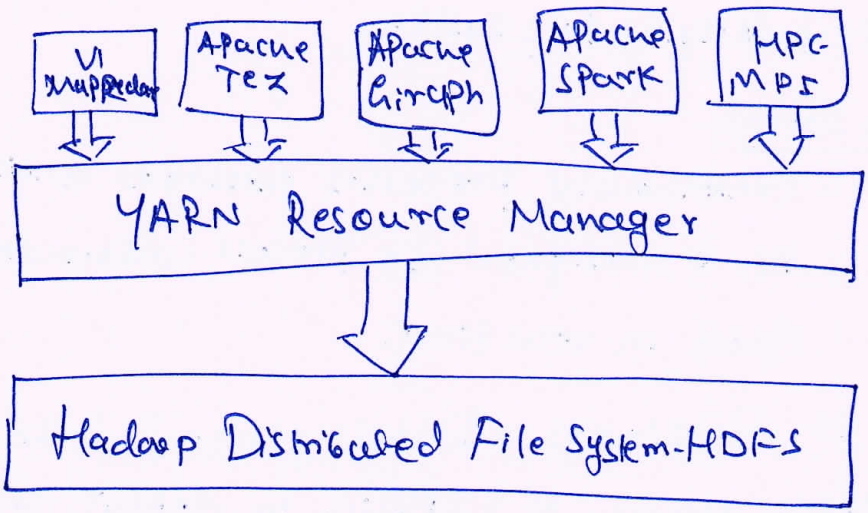
- Control flow nodes - define the beginning and end of a workflow
- Action nodes - where actual processing tasks are defined.
- Fork/Join nodes - Enable parallel execution of tasks in the workflow.
- Control flow nodes: Enable decisions to be made about the previous task.



4 b) What is YARN? Explain YARN Application Framework.

Solution YARN uses separate Resource Manager to schedule and manage all jobs on the cluster.

Application Framework 2M



→ Hadoop MapReduce 8M

- It was first YARN framework and drove many of YARN's requirements. It is integrated tightly with the rest of the Hadoop ecosystem projects.

Apache Tez -

Apache Tez generalizes the execution of complex DAGs and enables these tasks to be spread across stages so that they can run as single.

Apache - Giraph

- It is an iterative graph processing system built for high Scalability.

Hoya: HBase on YARN

- Creates dynamic & elastic Apache HBase clusters on top of YARN.

Dryad on YARN

- Dryad provides a DAG as the abstraction of execution flow.

Apache Spark

- used for in memory data processing. The advantage of Spark is common resource management and single underlying file system.

Apache Storm

- continuously processes messages until it is stopped.
- It is designed to process unbounded streams of data in real time.

Q5(a) What is Business Intelligence? List the different BI applications & explain in detail any four Applications.

Solution BI includes variety of IT applications that are used to analyze an organisation's data & communicate the information to relevant users. 1M

Applications

1. Customer Relationship Management 9M
2. Healthcare & wellness
3. Education
4. Retail
5. Banking
6. Financial services
7. Insurance
8. Manufacturing

⑨ Telecom

⑩ Public Sector.

① Customer Relationship Management:

- Maximize the return on marketing campaigns
- Improve customer Retention.
- Maximize customer value
- Identify and Delight Highly-valued customers
- Manage Brand Image.

② Healthcare & wellness

- Diagnose Disease in patients
- Treatment Effectiveness
- Wellness Management
- Manage Fraud and Abuse
- Public Health management

③ Education

- Student Enrollment
- Course offerings
- Fund-Raising from Alumni & other Donors.

④ Retail

- Optimize Inventory Levels at Different locations
- Improve Store Layout and Sales promotions.
- Optimize logistics for seasonal Effects
- Minimize Losses due to limited shelf life

5(b) Draw and explain flow of BIDM cycle. Explain the Strategic and operational decisions.

10 M

Solution BIDM cycle - Business Intelligence and Data Mining cycle.

* Business activities are recorded on paper or using electronic media, & then these records become data.

* All this data can be analyzed and mined using special tools and techniques to generate patterns & 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;

Strategic decisions:

- These decisions impact the direction of company.
- The goal here is may or may not be clear & the same is true for the path to reach the goal.
- The consequences of the decision could be apparent some time later.
- BI can help with what-if analysis of many possible scenarios.

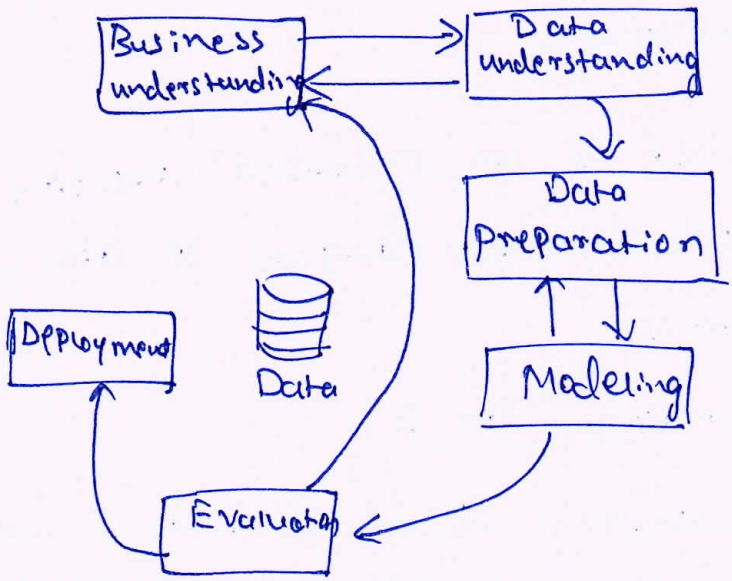
Operational Decisions

- These are more routine and tactical decisions, focused on developing greater efficiency.
- These can be made more efficient using an analysis of past data.
- BI can help automate operation level decision-making and improve efficiency by making millions of microlevel operational decisions in a model-driven way.

Q6 a) Explain CRISP DM cycle, with neat diagram.

Solution CRISP DM - Cross-Industry Standard Process for Data Mining.

3M



7M

① Business understanding : Ask the right business questions. A question is good if answer lead to large payoffs, for organization, financially & otherwise. An important step is to be creative & open in proposing imaginative hypotheses for the solution.

② Data understanding : one needs to be imaginative in scouring for many elements of data through many sources in helping address the hypotheses to solve a problem.

③ Data Preparation : Data should be relevant, clean and of high quality. Data cleaning can take 60-70% of the time in a data mining project.

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

⑤ Model Evaluation - It is better to triangulate the analysis by applying multiple data mining techniques.

⑥ Dissemination & Rollout - The data mining solution is presented to the key stakeholders & is deployed in the

6(b) Define Data warehouse & illustrate design considerations for Data warehouse.

Solution Data warehouse is an organised collection of integrated, subject oriented, databases designed to aid decision support functions. 5M

Design considerations for DW 9M

① Subject oriented - To be effective, a Data warehouse should be designed around a subject domain i.e. to help solve a certain category of problems.

② Integrated - The DW should include data from many functions that can shed light on a particular subject area.

③ Time-variant (time series) - The data in a DW should grow at daily or other chosen intervals.

④ Non volatile - DW should be persistent, it should not be created on the fly from the operations databases.

⑤ Summarized - DW contains rolled-up data at the right level for queries and analysis

⑥ Not normalized - DW uses star schema, which is a rectangular central table, surrounded by some look up tables.

⑦ Metadata - Many of the variables in the database are computed from other variables in the operational databases.

⑧ Near Real-time / Right time - DWs should be updated in near real-time in many high transaction volume industries

Q7a) What is Association Rule? Explain below given rules with suitable Examples:

i) Support ii) Confidence iii) Lift.

Solution Association Rule: It is a popular, unsupervised learning technique, used in businesses to help identify shopping patterns. 1M

A generic association rule is represented between a set X & Y: $X \Rightarrow Y [S\%, C\%]$ 9M

Ex: {Hotel booking, Flight booking} ⇒ {Rental car} {30%, 60%

i) Support: The no. of transactions that include items in the {X} & {Y} parts of the rule as a percentage of total no. of transaction.

e.g. Consider 1000 transactions in a dataset, 300 occurences of X & 150 occurences of (X, Y).

Support S for $X \Rightarrow Y = P(X \cup Y) = 150/1000 = 15\%$.

ii) Confidence: It is the ratio of transactions that includes all items in {B} as well as the no. of transactions that include all items in {A} to the no. of transactions that includes all items in {A}.

e.g.

confidence for $X \Rightarrow Y$ will be $P(Y/X)$ or

$P(X \cup Y) / P(X) = 150/300 = 50\%$.

iii) Lift: the lift of the rule $X \Rightarrow Y$ is the confidence of the rule divided by the expected confidence, assuming itemset X & Y are independent of each other.

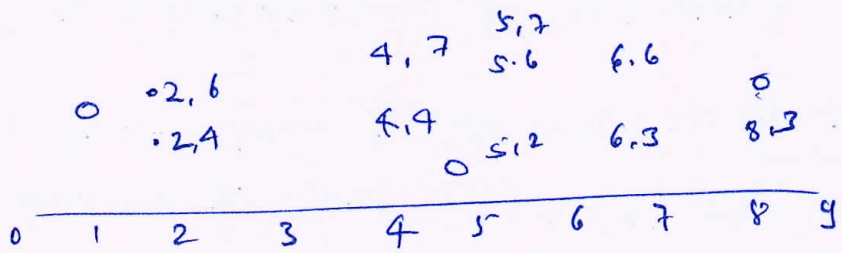
Q 7(b) ~~Explain~~ What is Unsupervised Machine Learning concept? Explain K-means clustering with suitable example.

Solution: Unsupervised learning is a type of Machine learning in which the algorithm is not provided with any pre-assigned labels or scores for training data. 1M

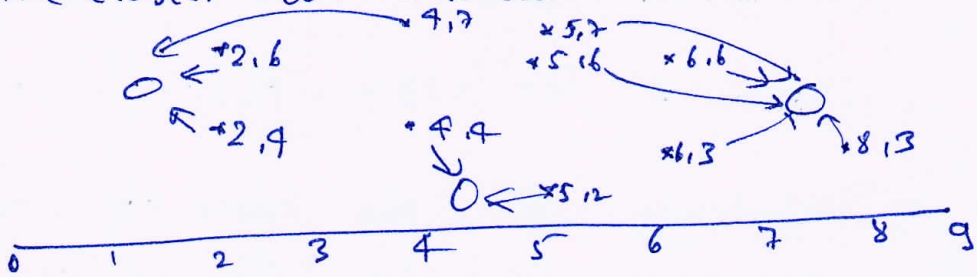
K Means Clustering 9M

- K-means is the most popular clustering algorithm.
- It iteratively computes the clusters & their centroids.

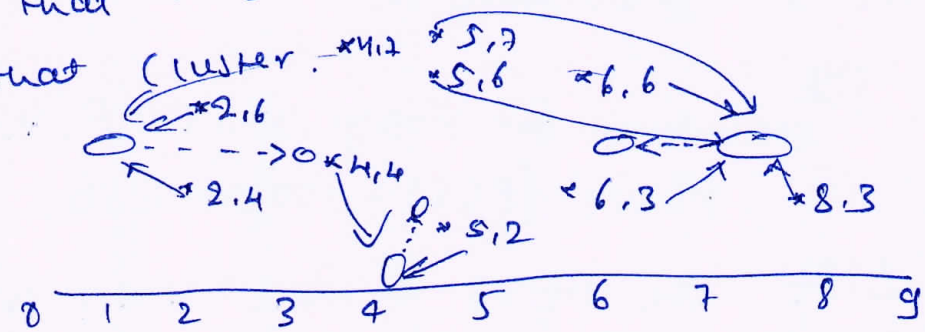
eg.



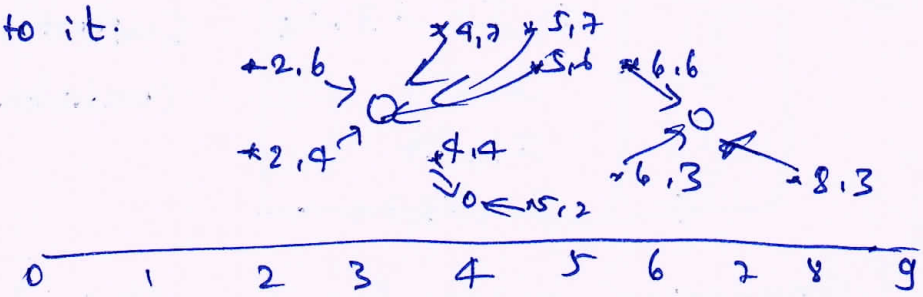
Step 1: For a data point, distance values will be from each of the 3 centroids. the data point will be assigned to the cluster with shortest distance to the centroid.



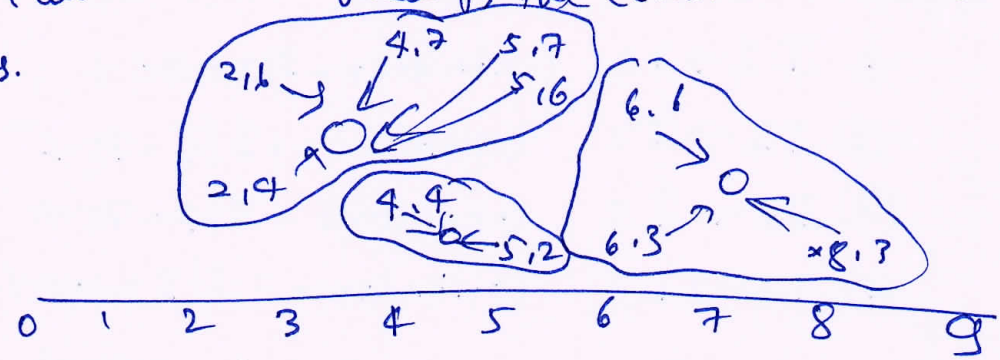
Step 2: The centroid for each cluster is recalculated, such that it is closest to all the data points allocated to that cluster.



Step 3: once again, data points are assigned to 3 centroids closest to it:



The new centroid will be computed from the data points in the cluster until finally, the centroids stabilize in their locations.



The 3-clusters shown are- 3-data points cluster with centroid (6.5, 4-5), a 2-data point cluster with centroid (4.5, 3) & a 5 data point cluster with centroid (3.5, 3)

Q 8
(a) write and Explain Apriori Algorithm with an example.

Solution Algorithm

3M

Step 1: Calculate the support of itemsets in the transactional database.

Step 2: Prune the candidate set by eliminating items with a support less than the given threshold.

Step 3: Join the frequent itemsets to form sets of size k+1 & repeat the above sets until no more itemsets can be formed.

e.g.

7M

T1	I1, I2, I3, I4
T2	I2, I3
T3	I3, I4
T4	I2, I3, I4

Support = 3

Confidence = 80%.

Following rules can be obtained from the size of 2-frequent itemsets

① I2 → I3 Confidence = 3/3 = 100%

② I3 → I2 Confidence = 3/4 = 75%

③ I3 → I4 Confidence = 3/4 = 75%

④ I4 → I3 Confidence = 3/3 = 100%

Rule ① & ④ are included in result.

8b List & explain steps in developing an ANN.

10 M

Steps:

① Gather data & divide into training data & test data. Training data is further divided into training data & Validation data.

② Select the network architecture, such as Feedforward network.

③ Select the algorithm, such as MLP.

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 ANN when it achieves good predictive accuracy.

9(a) Discuss Practical considerations of SNA and applications of SNA

Solution

Practical Considerations 3M

1) Gathering Data: Electronic communication records can be harnessed to gather social network data more easily. Capturing & cleansing & organizing data can take lot of time & effort.

2) computation & visualization – Modeling large networks can be computationally challenging & visualizing them also would require special skills.

3) Dynamic Networks : Relationships between nodes in a social network can be fluid. They can change in strength & functional nature.

Applications 7M

1) Self Awareness – visualizing his/her social network can help a person organize their relationships and support network.

2) Communities – SNA can help identification, construction, and strengthening of networks within communities to build wellness, comfort & resilience.

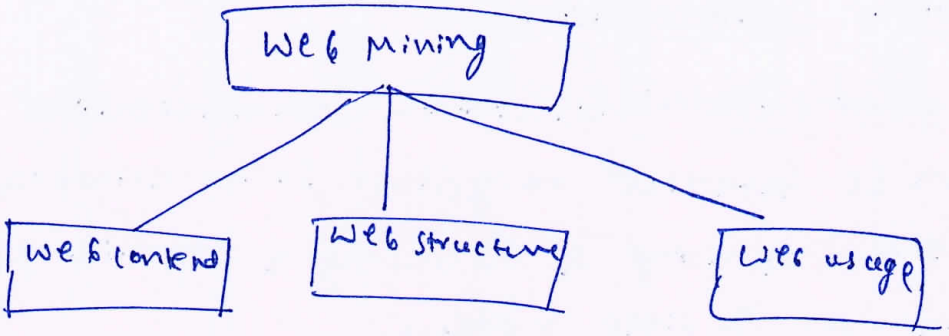
3) Marketing – Any two people are related to one another through atmost 7 degrees of ~~Freedom~~ links. This can be used to reach out organizers with their messages to large number of people.

4) Public Health – Awareness of networks can help identify the paths that certain diseases take to spread.

9 q(b) Explain 3-types of web mining. Use appropriate block diagrams to represent the same.

3M

xution



Web content Mining - A website is designed in the form of pages with a distinct URL. The website keeps a record of all requests received for its page/URLs, including the requestor information using cookies. The log of these requests could be analyzed to gauge the popularity of those pages.

Web Structure Mining

3M

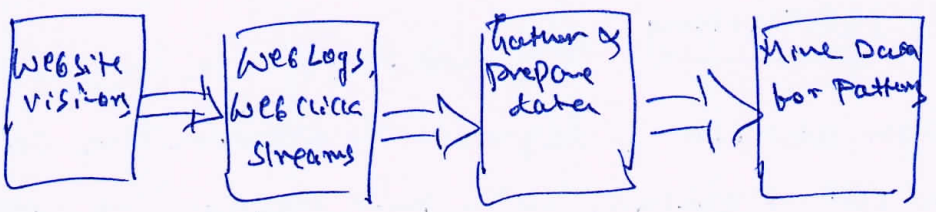
The web works through a system of hyperlinks using HTTPS. There are two basic models for successful websites

1) Hubs: These are pages with a large no. of interesting links. They serve as a hub or a gathering point.

2) Authorities: These websites have lot of inbound links.

Web usage Mining

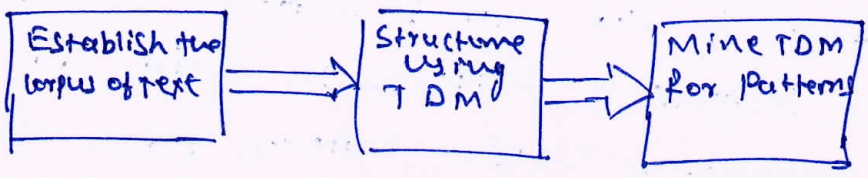
4M



- the goal of web usage mining is to extract useful information & patterns from data generated through web page visits & transactions

Q10 (a) Explain Text mining process & Architecture.

Solution



10 M

- The first level of analysis is identifying frequent words. This creates a bag of important words.
- The next level is identifying meaningful phrases from words.
- The next higher level is that of TOPICS. Multiple Phrases can be combined into Topic area.
- Text mining is a semi automated process. Text needs to be gathered, structured and then mined in a 3-step process.

Step 1: The text & documents are first gathered into a corpus and organized.

Step 2: The corpus is then analyzed for structure. The result is a matrix mapping important terms to source document.

Step 3: The structured data is then analyzed for word structures, sequences and frequency.

10(b) Briefly Explain the Data Mining. Compare text mining & data mining.

2M

Solution Data mining is the art & science of discovering knowledge, insights & patterns in data. It is the act of extracting useful patterns from an organized collection of data.

8M

	Text Mining	Datamining.
Nature of Data:	unstructured data: words Phrases, Sentences	Numbers, alphabetical & logical values.
Language used:	Many languages & dialects used in the world;	Similar numerical systems across the world.
Clarity & precision	Sentences can be ambiguous; sentiments may contradict the words	Numbers are precise.
Consistency	Different parts of the text can contradict each other	Different parts of data can be inconsistent, thus need statistical significance analysis
Sentiment	Text may present a clear & consistent or mixed sentiment a/c a continuum	Not applicable
Quality	Spelling errors. Differing values of proper nouns such as names	Issues with missing values, outliers, etc.
Nature of Analysis	Keyword based search; Sentiment mining	A full wide range of statistical & ML analysis for relationships & differences.

Rajkal