Q.1 Characteristics of big data (5V's)

→ 1) Volume :- The large amount of data generated, measured in petabytes or exabytes.

2) Velocity :- The speed at which data is generated & processed in real time.

3) Variety :- The different types of data

Q.2 (A)

① Structured - SQL DB

② Unstructured - predefined fmt, text

③ semistructured - JSON, XML files

4) Veracity - The accuracy & trustworthiness of data.

5) Value - The meaningful insight Derived from process data.

Q.3. Features of HDFS (Hadoop distributed file system)

→ ① Scalability - Easily scalable by adding more nodes to the cluster.

⑤ Opensource ② Fault tolerance - ensure reliability.

② Cost effective ③ High throughput - optimize for large datasets enabling high data access.

③ Easy to use ④ Streaming data access :- Designed batch process allowing streaming access to dat

④ ML Capab ⑤ Large Data storage - Capable of storing very large files across multiple machines.

⑤ Secure

Q.4. Challenges of Big data :-

① storage ④ Transfer

② Capture ⑤ visulization.

③ Analysis ⑥ privacy.

Q.5 Applications of Big data Analytics

→ ① Smarter healthcare
② Manufacturing
③ Search quality.
④ Trading Analytics
⑤ Traffic Control
⑥ Telecom
⑦ Homeland security

Q.6 Hadoop – i) Hadoop is an open source framework designed for distributed storage & distributed processing of large datasets using a cluster of computers.

ii) It enables scalable, fault tolerant & efficient data processing through its Core Components :- ① HDFS
② map reduce
③ YARN
④ Common Hadoop.

① Massive data storage
② Faster processing

• Why it matters :- ① Scalability
② Cost effective
③ Flexibility
④ Reliability
⑤ Wide Adoption

(features)

Q.7 Advantages & limitation of hadoop.

→ ① Cost reduction
② Flexible
③ Enhaced scalibility
④ Minimal Traffic

- Limitions of hadoop -

① Complex programming model - Java

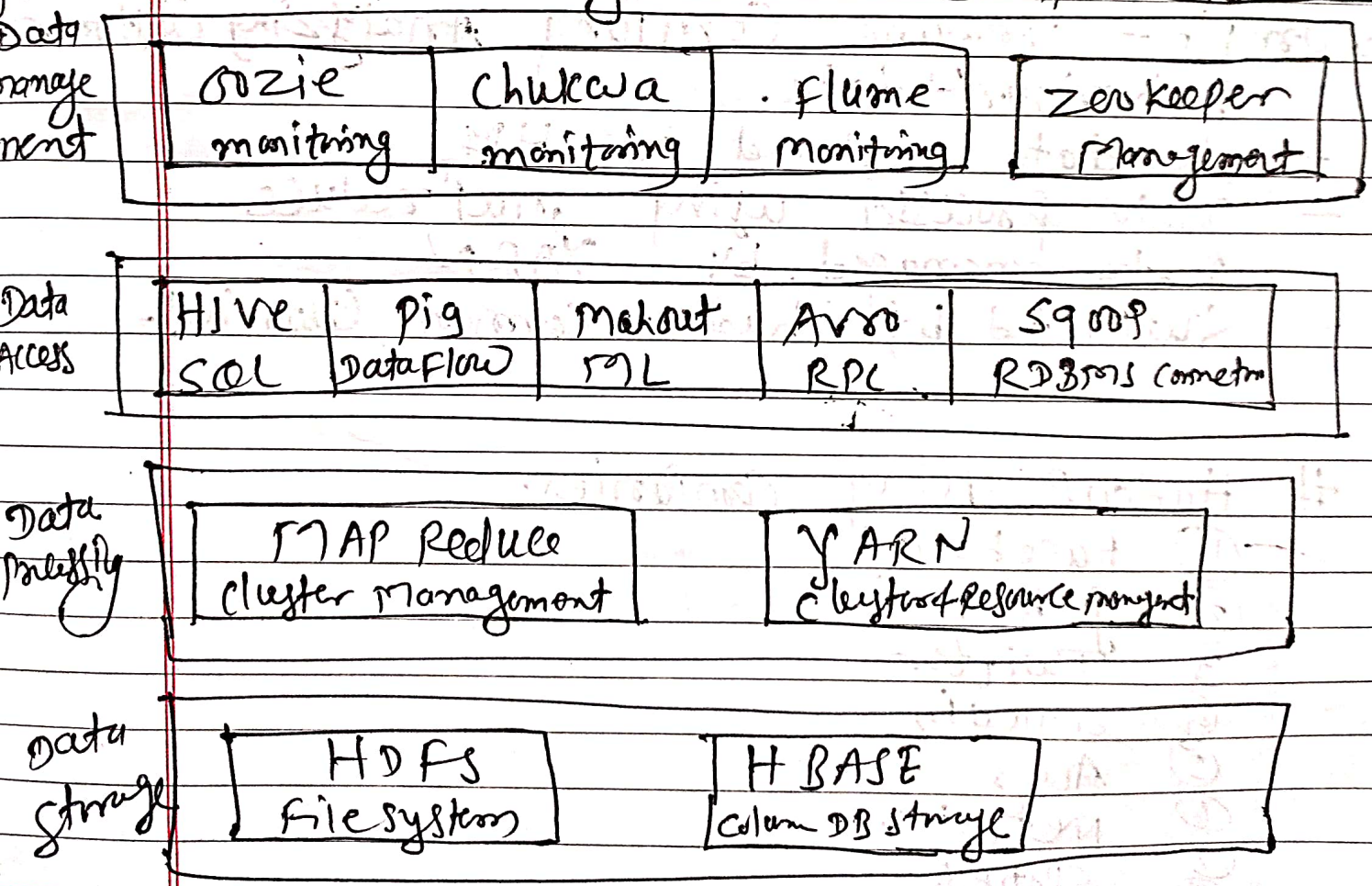
② Not suitable for Real time processing - Batch processing design.

③ Small File problem :- inefficient at handling small files. due to excessive metidata strage in namenode.

④ Latency issues - High latency due to batch oriented nature of mapreduce, making it less suitable for time sensitive data processing.

⑤ Resource intensive - require significant hardware resources for setup & maintenance.

8# Hadoop Ecosystem & its Components.



(90)

Components of Hadoop

i) Hadoop Ecosystem holds variety of tools & frameworks to support big data.

① Data management — zookeeper
② oozie — job scheduling & monitoring
③ chukwa & flume — monitoring
④ Data ~~Processing~~ Access — Pig, Hive — Query SQL, dataflow
Sqoop RDBMS Connector
⑤ Mahout, Spark — ML
⑥ Data Processing — Mapreduce — Programming based
— YARN — Yet another Resource Negotiator
⑦ HDFS — Hadoop Distributed file system.
⑧ HBASE — NO SQL Database
⑨ Spark — In memory data Processing

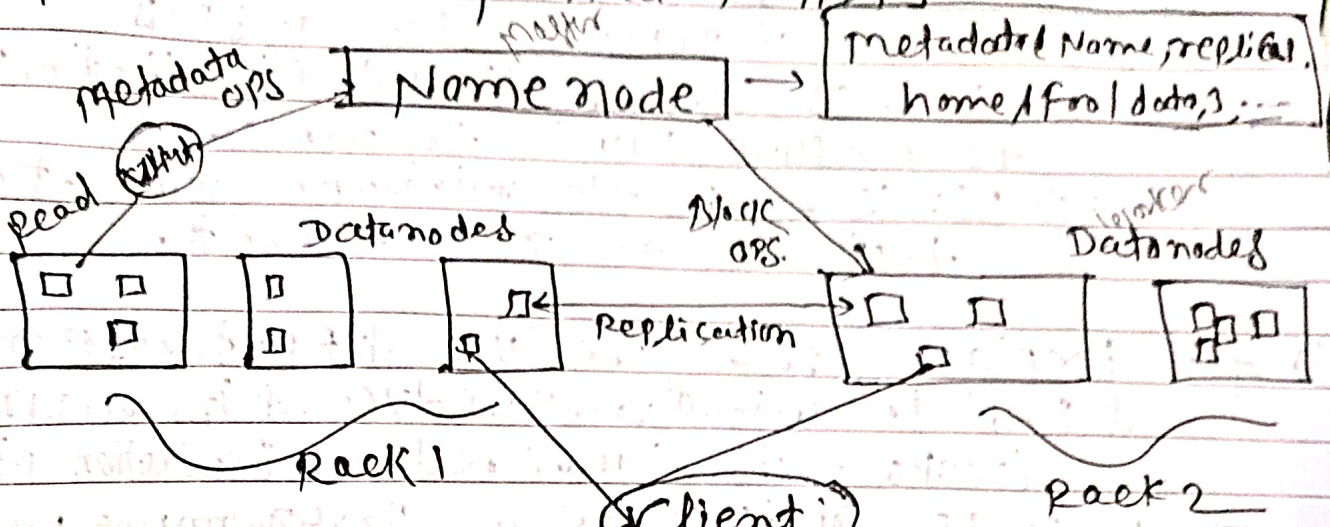
Example — Imagine Company Analyzing Customer behaviour data.
- The data is stored in HDFS
- Data processed using mapreduce
- Data managed by YARN
- Supported by hadoop common libraries.

Hadoop Using Companies.

① Facebook
② Yahoo
③ Twitter
④ LinkedIn
⑤ AWS
⑥ Netflix.
⑦ Spotify
⑧ Bank of America

① Name node
② Data nodes
③ secondary data nodes
④ Block storage

9(#) Architectures of Hadoop HDFS.



① Name node - i) Manage the metadata of the file system [Master node]
ii) Manager & Controller of HDFS
③ keep track of the location of data blocks across data nodes.
④ manage file system namespace & control access to files.

② Datanodes - i) [Worker node] that store actual data blocks.
ii) perform read & write operations as instructed by the name node.
iii) periodically send heartbeat signal to the name-node to confirm their status.

③ Secondary name node - i) It is seperate physical machine which act as [helper of namenode]
ii) It perform periodic check points. It cominicate with namenode and take snapshot of meta data which help to minimize the down time.

④ Block Storage — i) Data in HDFS is stored in blocks
default size 128MB
ii) Blocks are replicated (replication factor is 3)
to ensure fault tolerance.
iii) Block — is the minimum amount of data
that it can read or write.

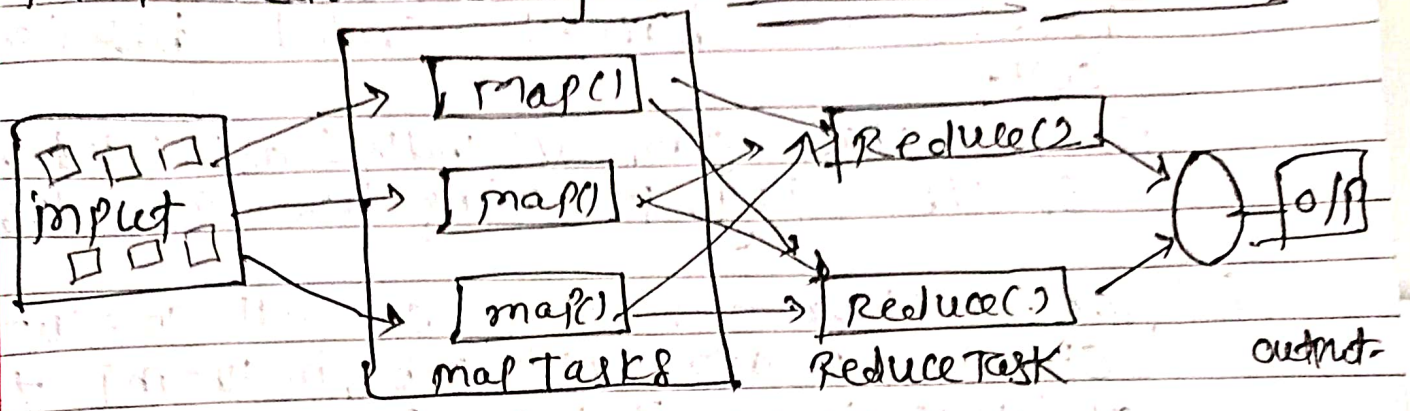
10. # Explain map reduce & features of map reduce
& Explain with diag map reduce. & its main
components of map reduce. Execution pipeline.

→ i) map reduce is a programming framework
that allows us to perform distributed
& parallel processing on large data sets
in a distributed environment.
ii) Data processing Breaking down tasks into
two main functions. ① map ② Reduce.
iii) map — [Sorting & filtering the data]
- organizing them into the form of group
- key-value pair based result which is later
on processed by the reduce() method.
iv) Reduce — (Summarization by aggregating
the mapped data.)
Reduce — takes o/p of the map as input.
& Combine them into smaller tuples.

Features of map reduce

① Scalability
② Fault tolerance - Reliability.
③ Parallel Processing
④ Data Locality
⑤ Simplicity.
⑥ extensibility

Map reduce Diagram & Execution Pipeline



Components

① Input Splits ② map function ③ Reduce function ④ output format ⑤ Shuffle & Sort ⑥ Job & task Tracker.

① Input Splits - i/p data is split into small chunks called i/p splits. Each split processed independently

ii) spliting helps in parallel processing by distributing the splits across different nodes.

② map ~~fun~~ function - map function process each i/p split & generate a set key-value pairs.

ii) It perform data filtering, sorting & xer based on the logic defined in the map function.

③ Reduce function - Takes the grouped key-value pairs & perform aggregation or summerization.

ii) o/p is final set key value pairs, which represent the result of the map reduce job.

④ output format - HDFS, output format defines how the key-value pairs are stored & represented.

(5) Shuffle & Sort - After map phase key value pairs are shuffled & sorted based on the key.
ii) This phase group all values associated with specific key, preparing them to reduce phase

(6) Job Tracker - manage overall execution of the mapreduce job, including task assignment, monitoring & job completion.

(7) Task Tracker - Runs on each node in the cluster & responsible for executing the task assigned by the job tracker.

(2)
NoSQL

(1) Types of NoSQL DB.
(1) Document store - json, xml, BSON, eg. MongoDB
(2) key-value store - key-value pair, eg, Redis, Dynamo
(3) Column family store - eg. Apache Cassandra, HBase
(4) Graph Database - nodes, edge, eg. NEO4j, Amazon Neptune.

(2) Features of NoSQL DB.
(1) Schema Flexibility - Dynamic schemas.
(2) Scalability.
(3) High performance.
(4) Distributed Architecture.
(5) High Availability
(6) Support for unstructured data.
(7) Eventual Consistency
(8) Fault tolerance.

Advantages of NOSQL	Disadvantages of NOSQL
① Scalability.	① Complexity
② Flexibility	② Lack of Standardization
③ Availability	③ Limited Acid Support
④ performance	④ Less mature tools
⑤ cost effective	⑤ Eventual consistency
⑥ Suitability for Big data	

Q.4	SQL	NOSQL
→	① Relational DB (Rows & Column Tables)	i) Non Relational (document key-value, graph, column)
ii)	fixed schema	ii) Dynamic schema.
iii)	vertical Scalability	iii) Horizontal scalability
iv)	structured query langues	iv) JSON query.
v)	Full support ACID properties	v) Limited support ACID
vi)	Example. - MySQL, Postgresql	vi) mongoDB, Redis, Neo4j

Q.5	DBMS	DSMS
→	i) Database management system	i) Data stream managment system
ii)	Handle structured data.	ii) Handle streaming data.
iii)	Store static data in tables	iii) process Continuous data
iv)	Support SQL	iv) support Continuous queries
v)	Batch processing higher latency	v) Real time processing, Low latency
vi)	Ex- MySQL, Postgresql	vi) Apache flink, Apache storm.
vii)	Support CRUD operation	viii) support read operation

Q.1 DGIM Algorithm [- Datar - Gionis - Indyk - motwani Algorithm]

→ Used for Counting the number of 1's in a binary stream over a sliding window.

We modify the buckets after a new 1's arrives.

- It uses $O(\log_2 N)$ bits to represent a window of N bits.
- Error rate is no more than 50%.

Elements

① Timestamp

② Bucket (0 & 1)

Rules for Bucket

① Every bucket should contain at least a single 1 in it.

② Right side of the bucket should strickly start from 1.

③ Length of the bucket is equal to the number of 1's in it.

④ Every bucket length should be in Power of 2.

$2^0 = 1, 2^1 = 2, 2^2 = 4, 2^3 = 8, 2^4 = 16 \ldots$

⑤ As we move to left, the bucket size should not decreases.

⑥ No more than two bucket can have same size.

for Data stream mining.

(2) (A) Bloom Filter – is a probabilistic Data Structure used to test whether an element is part of a set.

ii) It uses multiple hash function and bits arrays, allowing space-efficient storage but with a possibility of false positives.

Application:-
Used in stream Data mining to quickly check membership of elements without storing the entire dataset.

(3) (A) Explain issues in Data stream query processing.

→ ① High throughput - stream generate large amount of data rapidly, require efficient processing techniques.

② Approximate Answers - Full accuracy is often sacrificed for performance, require approximate query answers.

③ memory Constraints:- Storing & processing an entire data stream in memory is Impractical.

④ Data skew :- Un-even data distribution in streams can impact performance and accuracy.

2 Given — Bloom filter

$n = 15$

$k = 3$

$S = \{25, 36, 49\}$

Check membership of 16 & 81

- $h_1(x) = x \% 15$
- $h_2(x) = (x+2) \% 15$
- $h_3(x) = (2x+1) \% 15$

→ Step ① Initialize bit array / vector

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

① $S = 25$ Insert it.

- $h_1(x) = 25 \% 15 = 10$
- $h_2(x) = (25+2) \% 15 = 12$
- $h_3(x) = (25 \times 2 + 1) \% 15 = 6$

$15 \overline{)25}$ (1, remainder 10) Remainder

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	1	0	0	0	1	0	1	0	0	0

Bit vector.

② Let insert $S = 36$

- $h_1(x) = 36 \% 15 = 6$
- $h_2(x) = (36+2) \% 15 = 8$
- $h_3(x) = (2 \times 36 + 1) \% 15 = 13$

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	1	0	1	0	1	0	1	1	0	0

Bit vector

③ Let insert $S = 49$

- $h_1(x) = 49 \% 15 = 4$
- $h_2(x) = (49+2) \% 15 = 6$
- $h_3(x) = (2 \times 49 + 1) \% 15 = 9$

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
			1		1		1				1	1		

Bit vector.

(1) Lets Lookup 16
- $h_1(x) = (16) \% 15 = 1$
- $h_2(x) = (16+2) \% 15 = 3$
- $h_3(x) = (16 \times 2 + 1) \% 15 = 3$

Since 1 & 3 are 0's 16 is not present in the set

(2) Lets Lookup 81
- $h_1(x) = 81 \% 15 = 6$
- $h_2(x) = 81+2 \% 15 = 8$
- $h_3(x) = 81 \times 2 + 1 \% 15 = 13$

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
			1		1		1				1	1		

Since 6, 8 & 13 are 1's, 81 is present in the set (False +ve)

(*) Applications of Bloom Filters
1. Quora
2. google Chrome
3. medium
4. Big Table / Hbase / Cassandra / Postgresql.

(P-3) $n = 12$ (Practice)
$k = 3$
$S = \{17, 23\}$ check membership of 13 & 29

(4) Flajolet-Martin Algorithm – estimates the number of distinct element in a data stream using hash functions & bit patterns.

P-1 Given : Stream – 4, 2, 5, 9, 1, 6, 3, 7

Hash Function

$h(x) = ax+b \bmod 32$

$h(x) = 3x+1 \bmod 32$

$h(x) = x+6 \bmod 32$

(1) $h(x) = 3x+1 \bmod 32$

16 8 4 2 1

$= 3(4)+1 \bmod 32 = 12+1 \bmod 32 = 13 = 01101$

$= 3(2)+1 \bmod 32 = 6+1 \bmod 32 = 7 = 00111$

$= 3(5)+1 \bmod 32 = 15+1 \bmod 32 = 16 = 10000$

$= 3(9)+1 \bmod 32 = 27+1 \bmod 32 = 28 = 11100$

$= 3(1)+1 \bmod 32 = 3+1 \bmod 32 = 4 = 00100$

$= 3(6)+1 \bmod 32 = 18+1 \bmod 32 = 19 = 10011$

$= 3(3)+1 \bmod 32 = 9+1 \bmod 32 = 10 = 01010$

$= 3(7)+1 \bmod 32 = 21+1 \bmod 32 = 22 = 10110$

$r(a)$ = Trailing zeros = {0, 0, 4, 2, 2, 0, 1, 1}

$R = \max[\text{Trailing zero}] = 4$

Estimated output = $N = 2^R = 2^4 = 16$

(2) $h(x) = x+6 \bmod 32$

16 8 4 2 1

$= 4+6 \bmod 32 = 10 \bmod 32 = 10 =$ 01010

$= 2+6 \bmod 32 = 8 \bmod 32 = 8 =$ 01000

$= 5+6 \bmod 32 = 11 \bmod 32 = 11 =$ 01011

$= 9+6 \bmod 32 = 15 \bmod 32 = 15 =$ 01111

$= 1+6 \bmod 32 = 7 \bmod 32 = 7 =$ 00111

$= 6+6 \bmod 32 = 12 \bmod 32 = 12 =$ 01100

Trailing zero = The zero's at the right most end
Counting right side zero

16	8	4	2	1
0	1	0	0	1
0	1	1	0	1

3+6 mod 32 = 9 mod 32 = 9
7+6 mod 32 = 13 mod 32 = 13

r(a) = Trailing zero's = {1, 1, 0, 0, 2, 0, 0}
R = max [Trailing zero] = 2
Estimated output $N = 2^R = 2^2 = 4$

(P2) Practice - Stream = 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1
Hash function $h(x) = 6x + 1 \bmod 5$
Flajolet martin Algorithm.

(5) Finding Similar Items & Clustering

① CURE Algorithm - (Clustering Using Representa
It is a hierarchical clustering algorithm that handles large datasets & outliers better than traditional methods.

Steps - ① Sample selection - select random sample of data points.
② cluster the sample - perform an initial clustering on the sample using a traditional Algorithm [e.g K-means]
③ Select Representatives Points :- For each cluster, select representatives points.
④ shrink Representative points → toward the centroid using shrinking factor.
⑤ merge clusters - based on the proximity of their representative points.
⑥ label the Remaining Data - Assign the remaining Data Points to the nearest clusters.

Example — In a dataset with scattered clusters, CURE selects a few representatives points from each clusters & moves from closer to the center, effectively handling non-spherical clusters & outliers.

④ Distance measures — ① Hamming Distance
② Cosine Angle
③ Jaccard Distance
[Jaccard similarity]
④ Euclidean Distance.

P-1). Explain with example TWO major classes of Distance measures?

① Euclidean Distance — measures the straight line distance between two points

$$d(x,y) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

Ex: (1, 2) & (4, 6)
$(x_1\ y_1)$ $(x_2\ y_2)$

$$= \sqrt{(4-1)^2 + (6-2)^2}$$

$$= \sqrt{9+16}$$

$$= \sqrt{25} = 5$$

② Cosine angle similarity — measure cosine angle between two vectors, indicating their similarity in terms of direction

$$\text{Cosine similarity} = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

E.g.
A = (1, 2)
B = (2, 4)

$$= \frac{(1\times2)\cdot(2\times4)}{(\sqrt{1^2+2^2}) \times (\sqrt{2^2+4^2})} = \frac{16}{\sqrt{5} \times \sqrt{20}} = 1.6$$

$\cos^{-1}(-0.6) = 126.^0$

Cosine distance = 1 − Cosine similarity

= 1 − 1.6 = −0.6

(2) 1 indicate → vector A & B pointing in exactly same direction.

0 indicate → they considered "close" to each other in terms of direction

(3) Jaccard distance — measure the dissimilarity between two datasets & is calculated

Jaccard distance = 1 − Jaccard similarity.

example = JS = 80%

JD = 1 − 0.8 = 0.2 or 20%

(4) Jaccard similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

P-1 Example = A = {0, 1, 2, 5, 6, 8, 9}

B = {0, 2, 3, 4, 5, 7, 9}

→

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$= \frac{\{0,2,5,9\}}{\{0,1,2,3,4,5,6,7,8,9\}} = \frac{4}{10} = 0.4$$

$J(A,B) = 0.4 \simeq 40\%$

④ Harming distance - Difference between Corresponding bits.

Q-1 e.g- ① 110011 & 010101

② 11010 & 01011

→ ① → Compare bit by bit 110011 & 010101

```
  1 1 0 0 1 1
  0 1 0 1 0 1
  ↓ ↓ ↓ ↓ ↓ ↓
  1 0 0 1 1 0
```

Position → 1 4 5 Difference position 1, 4, 5 = 3

Harming distance = 3

②

```
  1 1 0 1 0
  0 1 0 1 1
  ↓ ↓ ↓ ↓ ↓
  1 0 0 0 1
```

Position → 1 5 Difference Position = 1, 5 = 2

Harming distance = 2

Q-2 cosine angles between

1. (3, -1, 2) & (-2, 3, 1)

$$\frac{A \cdot B}{\|A\| \cdot \|B\|}$$

$$= \frac{(3 \times -2) + (-1 \times 3) + (2 \times 1)}{\left(\sqrt{(3)^2 + (-1)^2 + (2)^2}\right) \times \left(\sqrt{(-2)^2 + (3)^2 + (1)^2}\right)}$$

$$= \frac{-7}{\sqrt{14} \times \sqrt{14}} = -0.5$$

$$\cos^{-1}(-0.5) = 120^\circ$$