BDA.

PAGE :
DATE : / /

* What is Big data? -

Data which is so large, fast or complex that its difficult to process using traditional methods is Big data.

Generally - TB's to 10^5 of PB's.

Data ~~th~~ which is large, fast or complex that is difficult to process using traditional methods is Big data.

Data which is large, fast or complex that is difficult to process using traditional methods is Big data.

Data which is large, fast and complex that is difficult to process ~~traditionally~~ using traditionally method is Big data

Data which is large, fast and complex which is ~~difficult~~ difficult to process using Traditional methods is a big data.

* Veracity :-

Data and algorithm should be same. Then only we will get the accurate result.

~~a~~ How it will work in Amazon.

① User click on particular photo.

② then amazon will take that data ~~in~~ that is photo and ~~run~~ run the algorithm with that only and with this help it will recommend the user similar type of the product. with the Help of algorithm and data.

* So Amazon makes more profit - by suggesting this product

PAGE:
DATE: / /

* There are 310 million active users on Amazon.
↓
1%
↓
3-1 million → 31 lakh users.

* Value:-
Whenever user search about the product is the raw data.
ii) Amazon takes that Raw data and suggest the product to the user and user buy that product and amazon makes money with this.

Hadoop.

Hadoop is opensource framework of tools designed for storage and processing of large scale data (big data).

Hadoop is opersource framework of ~~of~~ Tool designed for storage and processing of large-scale data (big data)

Free to use to all.

Hadoop is opensource framwork of tools designed ~~to~~ for storage and processing of large scale data (big data).

Hadoop is open source framwork of tools which is designed to storage and processing the big amount of data (that is Big data).

I.M.P. points

i] It is created by Doug cutting & Mike cafarella in the year ~~2006~~ April 1, ~~2006~~ 2005.

~~ii~~ ii] In ~~2006~~ 2005 only yahoo! funded this Hadoop.

Why its symbol is elephant?-

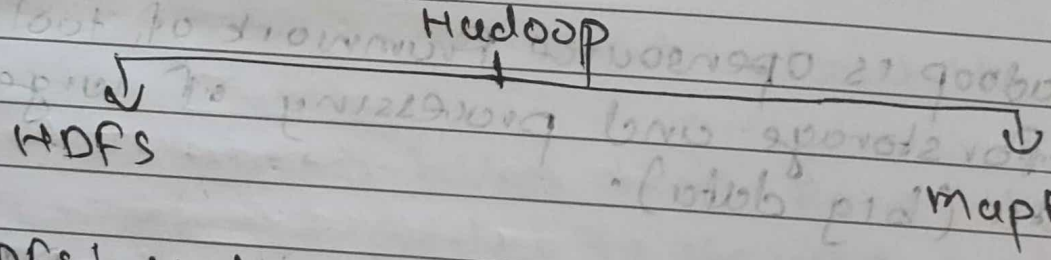
i) because the main Author Doug cutting His favourite toy was the elephant Hence he given the symbol ~~as~~ of elephant.

ii] and In 2006 Apache software found-ation taken the Hadoop. from yahoo!

PAGE:
DATE: / /

* Who are the major users of Hadoop
  i] Netflix.
  ii] AWS
  iii] Microsoft
  iv] IBM

* Two major components of Hadoop:



HDFS :- Hadoop Distributed file system

* HDFS :-
  i] Designed to store and manage huge data in efficient manner.
  ↳ Distributed storage.
  Designed to store and manage huge data in efficient manner.
  i] HDFS is the Hadoop distributed file system. Designed to storage and manage the large ~~data~~ data in efficient manner.
  ii] For the large data only we uses the HDFS system. It is not suitable for small data

centralized storage → storage on the one place
distributed storage → storage on multiple nodes.
- this can original, duplicate, or replicate etc.

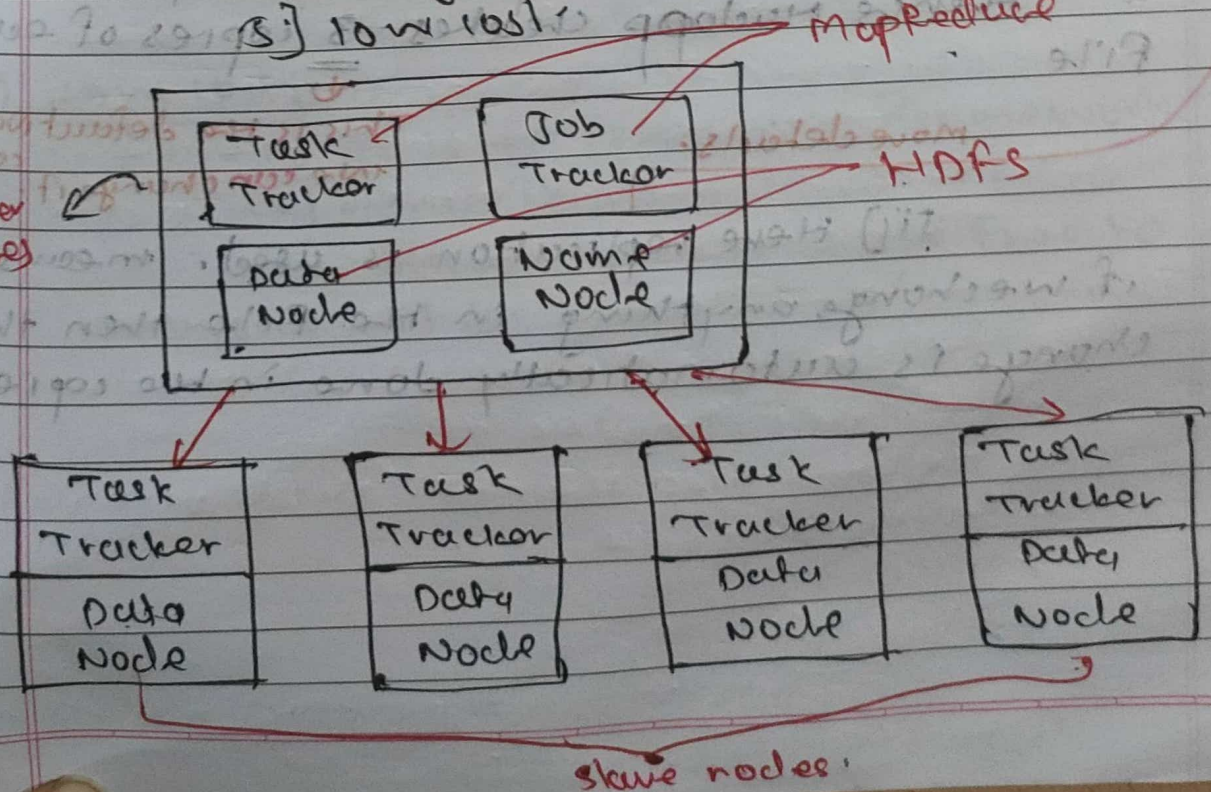
* MapReduce:-

i) MapReduce is the massive parallel processing technique used to processing data.
MapReduce is the massive parallel processing technique which is used for processing large amount of data parallely. so we get the output quickly.

Input:- It takes input in the form of list
output:- It gives the output in the form of list.
ii) Most of Java language is used.

* Features of Hadoop:-

i) Fault Tolerance.
ii) Highly scalable.
iii) Easy programming
iv) Huge and flexible storage.
v) low cost.



Replication	duplication
* If we change anything in file then it will automatic get updated on each node where it is present	* If we change ~~any file~~ anything in file then then it will not get change or update on each node where the same file is present on node

i) Fault Tolerance:

i) All the nodes in Hadoop are no Hardware if any node get damage because of any reason.

ii) If suppose one node get damage then the information inside that ~~data~~ data will get loss. Hence Hadoop what do,

Hence Hadoop creates 3 copies of each file

move defaults.

This is the default No of copies. We can change it.

iii) Here replication is used. means if we change anything in the file then that change is automatically done in the copies.

PAGE :
DATE : / /

ii) Highly scalable :

i] In the Hadoop all the slave nodes are highly scalable means we can scale that slave nodes in the thousands of number.

iii) Easy programming :

i] Here as a programmer we have to know the Java programming language, and programmer should know the logic means solve to create the logic (implementation will become easy)

iv) Huge and flexible storage :

i] the data we are storing that can be structured or un-structured

ii] Before storing the data there is no ~~need~~ need to preprocess the data.

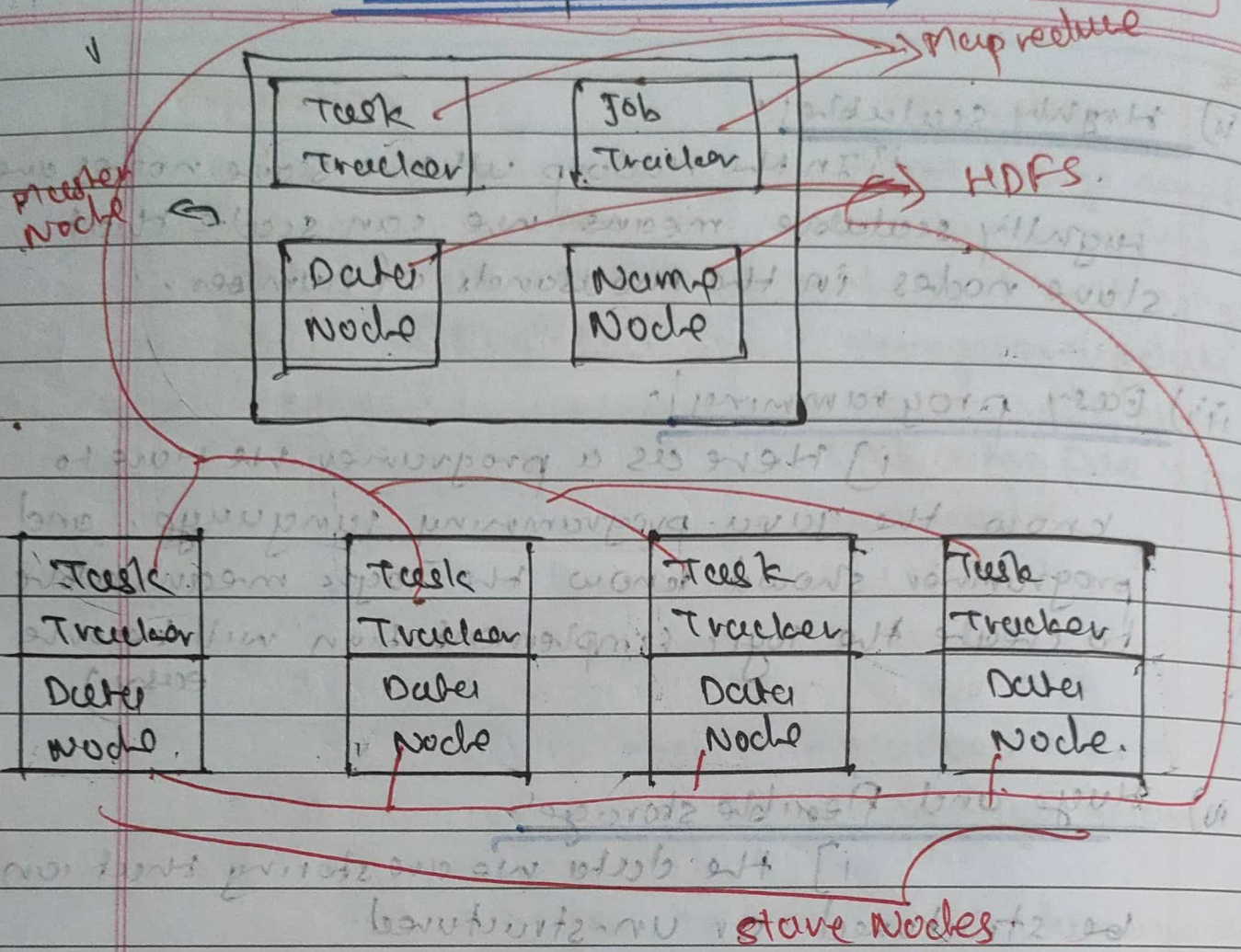
v) Low cost :

i] It is ~~fr~~ open source framework

ii] Hence it is free to use all

iii) ~~because~~ because of the free to use everyone it is low cost.

Hadoop Architecture



(1) Task Tracker :-

It process the small piece of data given to that particular node.

It process the small piece of data given to that particular node.

(2) Data node :-

It manages the data that reserved on the particular node.

It manages the data that reserved on the particular node.

(3) MapReduce :-

It is massive parallel processing technique to process the data.

PAGE :
DATE : / /

* MapReduce :-
It is massive parralel processing technique to process the data.

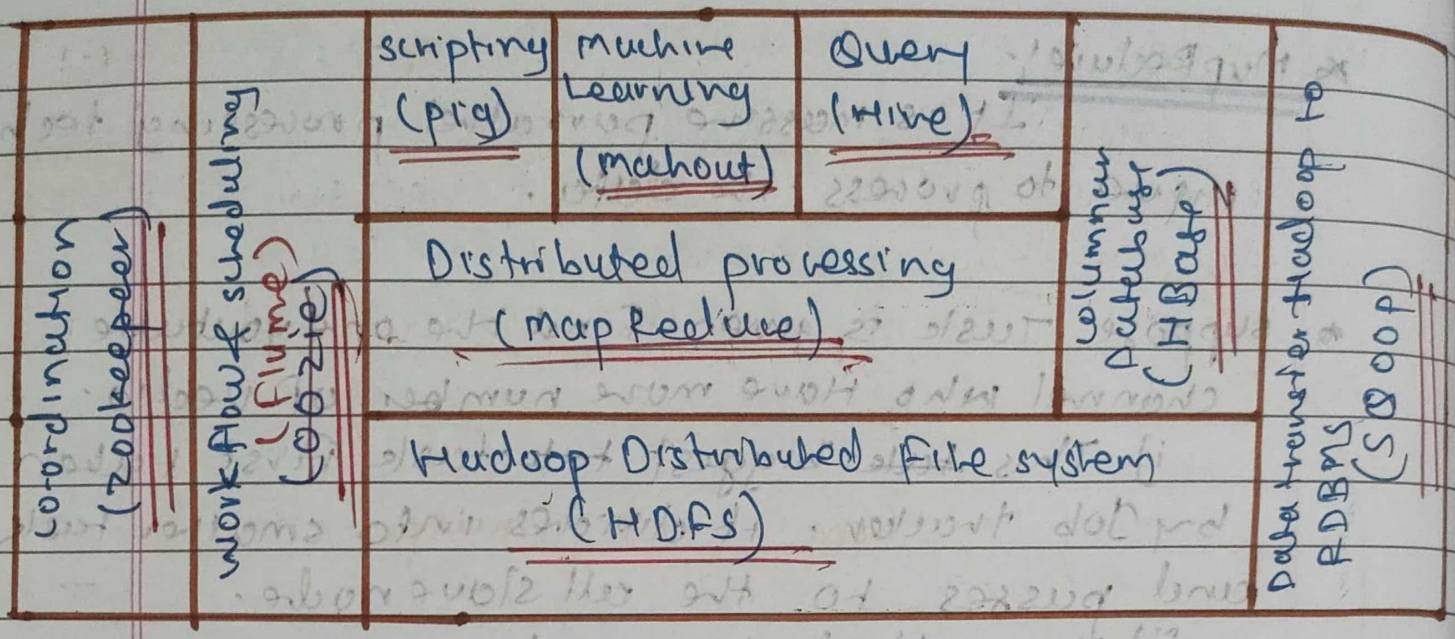
* suppose Task is given find the one youtube channel who have more number of vedio.
i) This ~~task~~ biggest ~~task~~ task first taken by Job tracker. it divides into smaller task and passes to the all slave node.
ii) then that sleve nodes process process that data parallel and give the output.

* HDFS :- (Hadoop distributed File system).
Designed to store and manage the lavge amount of data in effrient way.

* Job Tracker :-
i) Breaks the bigger task into smaller task and forward it to the Task Tracker

* Name Node :-
it keeps information about the which part of the data is having with which node

Hadoop ecosystem



1. HDFS
2. MapReduce
3. Flume
4. Hive
5. HBase
6. mahout
7. pig
8. SQOOP
9. zookeeper.

(1) **HDFS** :- (Hadoop distributed filesystem).

It have majorly two components

i) Data node.

ii) Name Node.

"HDFS is designed to store the large amount of data in efficient way"

i) **Data Node**.

i) It manage all the data that received on particular node.

ii)

ii) Name Node:- (each file have 3 copies it track that also).

i) It keeps all the information about the which part of the data is given to the which slave node.

ii) It maintain above record or information.

② MapReduce:- (Input - List, output - List)

MapReduce is a massive parallel processing technique used to process data parallel.

MapReduce is a massive parallel processing technique used to process data parallel.

MapReduce it's a massive parallel processing technique used to process data parallel.

It have major two component

i) Job Tracker

ii) Task Tracker.

i) Job Tracker:- ii) It also set the priority of time to nodes

i) Job tracker breakes the bigger task into smaller task and forword it to the task Tracker.

ii) Task Tracker:-

i) Job Task Tracker process the the small pieces of task to generate the output

4 Bx+

PAGE :
DATE : / /

③ **Flume!-**

i] It is the data injection tool in HDFS. means we can inject the data with the help of flume, that data can be, Unstructure, structured, or semi-structured data.

ii] It is also more reliable and fault tolerant.

④ **Apache Hive!-**

i] It is an open-source data warehouse system. For i) querying data
ii] Analyzing data.
iii] summarizing data.
This is large dataset stored in H-files.

ii] It uses HQL = Hive + SQL
iii] It is SQL like Query language
iv) Feature :- Highly scalable.

⑤ **HBase!-**

i) It is distributed column oriented database which is built on top of Hadoop file system.

ii] It is design to provide the Quick Random access to huge amount of data.

4 Bx+

PAGE :
DATE : / /

③ **Flume!-**

i] It is the data injection tool in HDFS. means we can inject the data with the Help of flume ; that data can be, Unstructure, structured, or semi-structured data.

ii] It is also more reliable and fault tolerant.

④ **Apache Hive!-**

i] It is an open-source data warehouse system. For i) querry data
ii] Analyzing data.
iii] summarizing data.
This is large dataset stored in H-files.

ii] It uses HQL = Hive + SQL
iii] It is SQL like Query language
iv) Feature :- Highly scalable.

⑤ **HBase!-**

i] It is distributed column oriented database which is built on top of Hadoop file system.

ii] It is design to provide the Quick Random access to huge amount of data.

6. Mahout :-

i] With the help of mahout it become a easy to implement the machine learning algorithms on the Hadoop ecosystem.

ii] It Have three techniques :-

i] Recommendation → movie, series.

ii] Classification ← once classify as spam it go in spam folder

iii] clustering. → one type of things come together and form their group is known as clustering.

7] Pig :-

i] This is ~~also~~ data processing tool.

ii] It Have its own language that is Pig-latien it is scripting language

iii] It also Have its own component that is Apache pig Engine. → It accepts the pig-latien scripts.

iv] 1 ~~in~~ line in pig-latien = 100 line in Java

8] Sqoop :-

i] It import structured data from RDBMS to HDFS in Hadoop

ii] It exports structured data from HDFS to RDBMS in Hadoop

9] Zookeeper :-

i] With the Help of zookeeper each component in the Hadoop ecosystem they can communicate with each other. ~~on behalf the problem.~~

ii] And perform or complete the task in less time.

iii] It also take care that no one component can get interuption in between the task processing.

* MAP-Reduce

mapping → (key, value) pair

Reducing.

Shuffling → similar word group

Reducing → Add the final result.

PAGE :
DATE : / /

* Relational database.

student ID	Name	Location ~~Geo~~	Gender	college.
18	Lallu	Nerul	M	Don-Bosco
27	Gullu	CST	M	K.J. Somaiya
36	Puppu	Thane	M	R.A.I.T
61	Pappa ki Pari	Seawoods	M	Don. Bosco.

* NO-SQL Database.

Types of database in NO-SQL - i] key value store
ii] Wide column store iii] Document DB
iv] Graph DB.

```
{ "student_id : "gg"
  "name"      : "Babbu"
  "Hobby"     : "Drinking"
  "Branch"    : "Computer Engineering".
}

{ student_id : "69",
  "Name"     : "Puppu",
  "Location" : "Dadar"
  "Hobby"    : "~~King~~ killing kids For Fun"
}
```

RDBMS VS NO-SQL.

PAGE:
DATE: / /

RDBMS	NO-SQL.
i] Have fixed or static predefined schema	i] Have dynamic schema.
ii] Not suited for hierarchical data storage	ii] Best suited for Hierarchical data storage.
iii] vertically scalable	iii] Horizontally scalable.
iv] follow Acid property	iv] follow CAP (consistency, availability, partition tolerance) cap theorem.
v] RDBMS supports Transactions. (also complex transaction with joins).	v] NOSQL databases don't support transactions (support only simple transactions)
vi] RDBMS manage only structured data.	vi] NO-SQL databases can manage structured, Un-structured, semi-structured data.
vii] Relational databases have a single point of failure with failover	vii] NO-SQL database have no single point failover.
viii] ~~ix~~ ~~Rel~~ support powerfull Query language	viii] support a very simple Query language.
ix] support complex transaction	ix] support simple transaction.

Example:- MySQL, Oracle, MS-SQL etc.	example:- MongoDB, BigTable, Redis, RavenDB. etc.

Now to make the input file. We will make the four columns :-

i/p file.

Matrix	i/k	j	Val.
A	0	0	1
A	0	1	2
A	1	0	3
A	1	1	4
B	0	0	5
B	0	1	6
B	1	0	7
B	1	1	8

Note:-
A → i
B → k

Now we have to make the map function. so we will now consider above table.

P.T.O.

Most Imp question.

Matrix Multiplication with 1 Map Reduce step

DATE: / /

1 step 2 step.

i] Map

ii] Reduce.

* there is two type of matrix multiplication.
  - i) one step → 1 time map, 1 time Reduce
  - ii] two step → 2 time map, 2 time Reduce

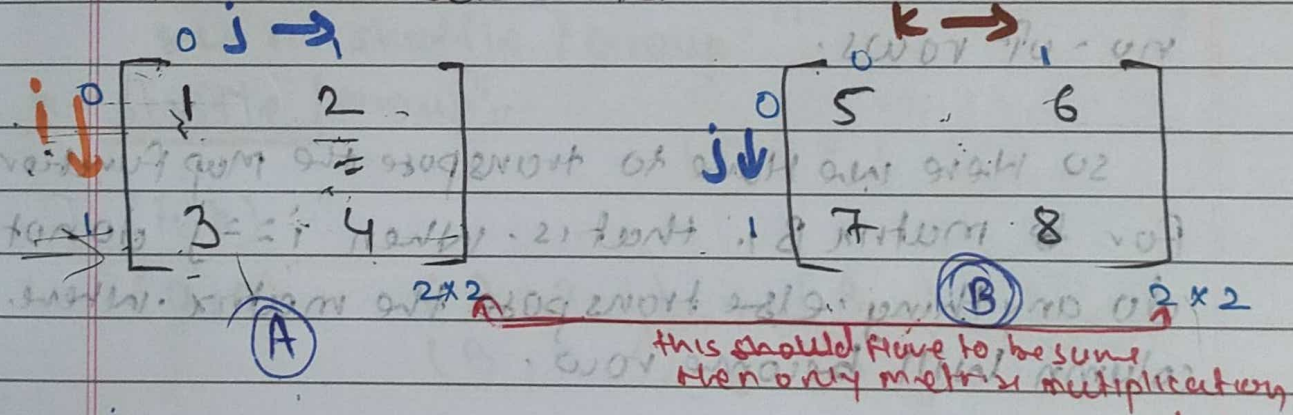
i) one step Matrix multiplication:

soln: Assume there is A matrix of size 2×2
& also Assume there is B matrix of size 2×2.

1st condition:

we can do matrix multiplication when only, the columns of first matrix is equal to the rows of second matrix;

* Assume the below matrix.



i → No. of ~~Rows~~ Rows in matrix A.

j → No. of column of matrix B

j → No. of Rows ~~in~~ of matrix B

k → No. of columns of matrix B.

2×2 2×2 =

↓ ↓

2×3 3×2 = ?

Example:- MySQL, Oracle, MS-SQL etc.	example:- MongoDB, BigTable, Redis, RavenDB. etc.

Now to make the input file. We will make the four columns :-

i/p file.

Matrix	i/k	j	val
A	0	0	1
A	0	1	2
A	1	0	3
A	1	1	4
B	0	0	5
B	0	1	6
B	1	0	7
B	1	1	8

Note:-
A → i
B → k

Now we have to make the map function so we will now convert above table.

Most Imp question.

Matrix Multiplication with 1 MapReduce step

1 step 2 step.

i] Map

ii] Reduce.

* there is two type of matrix multiplication.
  - i) one step → 1 time map, 1 time Reduce
  - ii) two step. → 2 time map, 2 time Reduce.

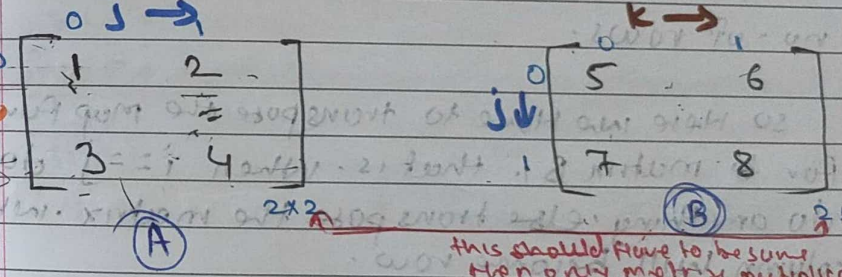
i) one step Matrix multiplication:

sol:- Assume there is A matrix of size 2×2 also Assume there is B matrix of size 2×2.

1st condition:-

we can do matrix multiplication when only the columns of first matrix is equal to the rows of second matrix.

* Assume the below matrix..



this should have to be same then only matrix multiplication is possible.

i → No. of ~~Rows~~ Rows in matrix A.

j → No. of column of matrix B

j → No. of Rows ~~in~~ of matrix B

k → No. of columns of matrix B.

2×2 2×2 =

PAGE:
DATE: / /

* Map function :-

i/k.	value
0	(A, 0, 1)
0	(A, 1, 2)
1	(A, 0, 3)
1	(A, 1, 4)
0	(B, 0, 5)
0	(B, 1, 6)
1	(B, 0, 7)
1	(B, 1, 8).

wrong

Note :-
To fill the value format is
(matrix, i, j, value)

Now while making the map function we forgot that the j for matrix A is no. of columns and for matrix B it is no. of rows.

So Here we have to transpose the map function for matrix B i.e. that is. When i == j do not do anything. else transpose the matrix. where column will become row.

Hence above map function is wrong.
correct one is :-

map function correct one, ~~after~~ :-

i/k	value
0	(A, 0, 1)
0	(A, 0, 2)
1	(A, 0, 3)
1	(A, 1, 4)
0	(B, 0, 5)
0	(B, 0, 7)
1	(B, 1, 6)
1	(B, 1, 8)

correct MAP-Function.

This above one is correct map function.

"Now before Reduce function we have to see the shuffle / Group!"

* Shuffle / Group :-

(i, k)

(0, 0) (A, 0, 1) ; (A, 1, 2)
(B, 0, 5) ; (B, 0, 7)

(0, 1) (A, 0, 1) ; (A, 1, 2)
(B, 1, 6) (B, 1, 8)

(1, 0) (A, 0, 3) (A, 1, 4)
(B, 0, 5) (B, 0, 7)

(1, 1) (A, 0, 3) (A, 1, 4)
(B, 1, 6) (B, 1, 8)

PAGE :
DATE : / /

Now Reduce :-

(0,0) $((1\times5)+(7\times2)) = 19$

(0,1) $((1\times6)+(2\times8)) = 22$

(1,0) $((3\times5)+(4\times7)) = 43$

(1,1) $((3\times6)+(4\times8)) = 50$

Note :- Vertical "multiplication" Horizontal "Addition", we have to perform

Hence Resultent matrix = $\begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$

In this way we do matrix multiplication using MapReduce.