Big Data Analytics-IAT 2

Enter your USN *

1CR17CS048

Enter Your name *

HARSHIT KUMAR

Enter your Section *

A

The first level of Analysis in text mining is (CO5,L2) 1 point

○ Identifying phrases

● Identifying frequent words

○ Ientifying topic area

○ None of the above

The Naïve-Bayes algorithm is special in that it takes into consideration the ----------- of an instance belonging to a class(CO4,L2) 1 point

● prior probability

○ Posterior probability

○ Either a or b

○ None

Naïve-Bayes assumes that all the features are -------- for most instances that work fine (CO4,L2) 1 point

○ dependent

● independent

In SVM the Kernel methods operate using what is called the ------------- (CO4,L2) 1 point

○ Seed trick

○ Core trick

● kernel trick

○ None

Training the SVMs is an inefficient and time consuming process (C04,L2) 1 point

● True

○ False

The pages with a large number of interesting links are called ----- (CO4,L2)

1 point

○ Authorities

● Hubs

The----------- model used in computing the importance of a node is social network . (CO4,L2)

1 point

○ Influence model

○ Data flow model

● Influence flow model

○ None

In HDFS block size if of (CO1,L2)

1 point

○ 128 MB

○ 256 MB

● 64 MB

○ 1 GB

Which of the following depicts the data flow in Map reduce Model (CO1,L2) 1 point

○ Split, Map , Reduce, output

○ Map, split, reduce, shuffle

○ Input split map shuffle

● Split, Map , Reduce, Shuffle

Debugging on a distributed system is easy and should be encouraged at all costs. (CO1,L2) 1 point

○ True

● False

Which of the following scenario may not be a good fit for HDFS? (CO1,L2) 2 points

● HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file

○ b) HDFS is suitable for storing data related to applications requiring low latency data access

○ c) HDFS is suitable for storing data related to applications requiring low latency data access

○ d) None of the mentioned

HDFS Snapshots are similar to Backups created by Administrators (CO1,L2) 2 points

● True

○ False

A 500 MB file is broken down to ----- blocks on HDFS (CO1,L2)

2 points

○ 16

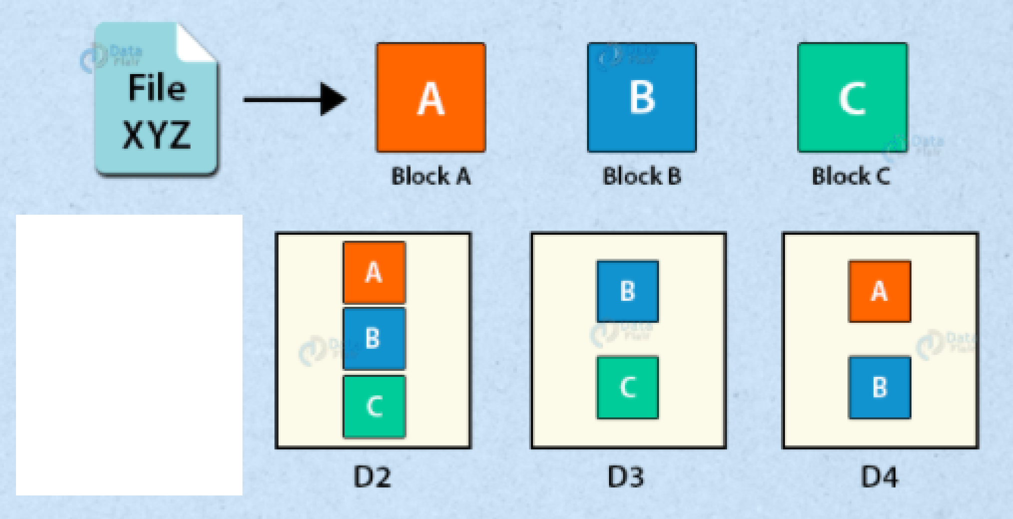
○ 64

● 8

○ 12

What is the replication factor of block C in the figure (CO1,L3)

2 points



○ 1

○ 4

○ 3

● 2

Point out the wrong statement (CO1,L2) 2 points

○ Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level

○ Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode

○ User data is stored on the local file system of DataNodes

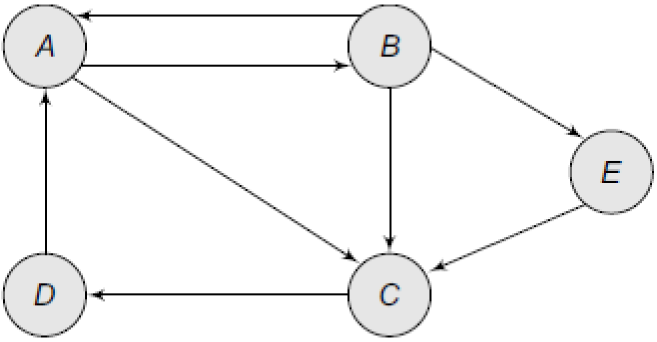
● DataNode is aware of the files to which the blocks stored on it belong to

With a neat diagram explain the components of HDFS (Scan and upload link:https://drive.google.com/drive/folders/1RwgmP_aqOzyNZoyR3ZRL1I_yk5LMdJUU?usp=sharing) (CO1,L2) 10 points

Explain the Map reduce programming model with a neat diagram((Scan and upload link:https://drive.google.com/drive/folders/1RwgmP_aqOzyNZoyR3ZRL1I_yk5LMdJUU?usp=sharing) (CO1,L2) 10 points

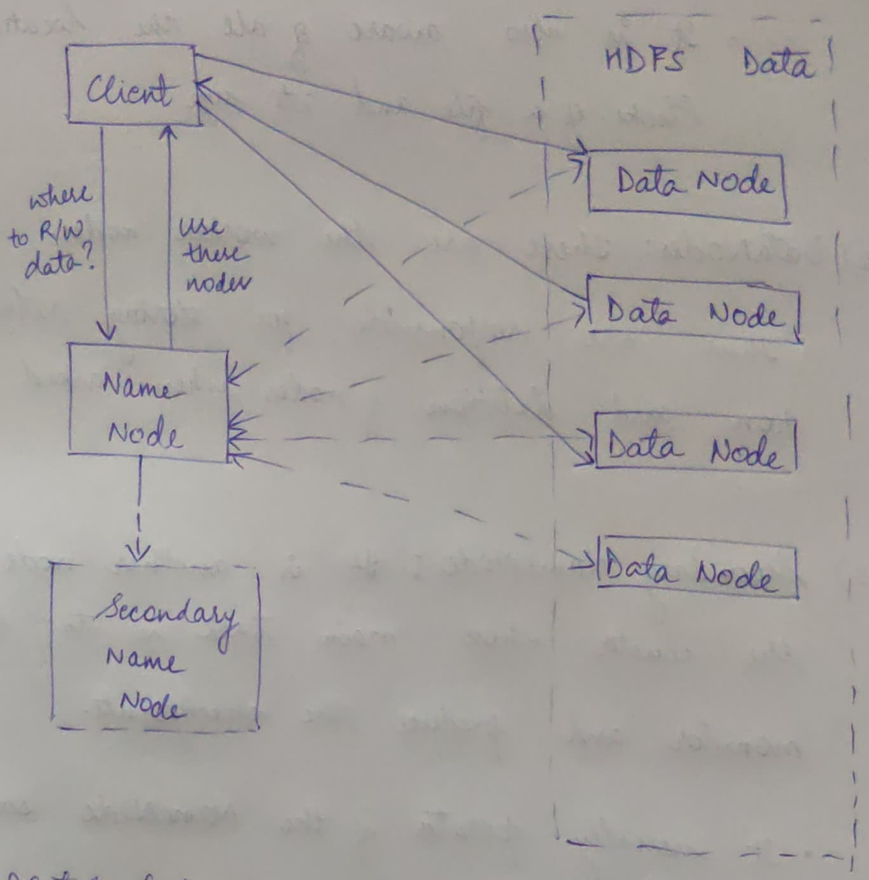
Compute the rank values for the nodes for the following network . Which is the highest ranked node now?(Scan and upload link:https://drive.google.com/drive/folders/1RwgmP_aqOzyNZoyR3ZRL1I_yk5LMdJUU?usp=sharing) (CO4,L3)

10 points



Shruti Singh
1CR17CS144

* HDFS Components



- Hadoop Distributed File System
- HDFS uses a master/slave model to maintain huge volumes of data.

- HDFS is defined by three properties:
  - Huge Volumes
  - Data Access
  - Cost Effective

- HDFS Components

1) <u>Namenodes</u>: It is the master node that runs on a separate node in the cluster.

   - It manages the filesystem napspace which is the filesystem tree or hierarchy of the files

- Stores information like owners of files and permissio
- It is also aware of all the locations of all th blocks of a file and its size.

2) DataNodes: These are the worker nodes.

- These are responsible for storing retrieving, replication and delition of nodes when asked by NameNode

3) Secondary NameNode: It is another node present in the cluster whose main task is to regularly monitor and produce the checkpoints.

- It usually operates if the NameNode somehow fails in the operation.

4, Client: It asks to the NameNode where it should read and write the data from?

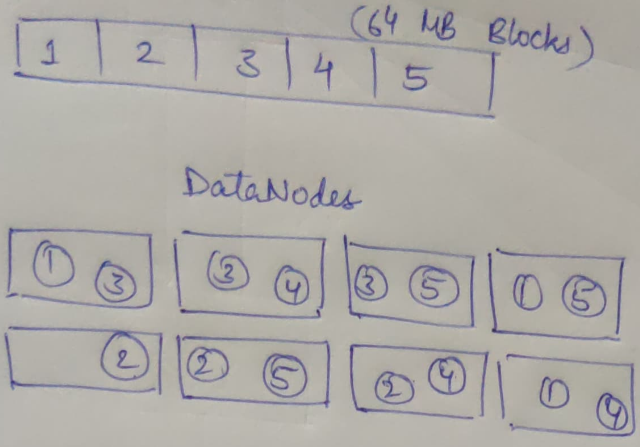
- It Reads/Write the data into one/more datanodes depending upon the no. of blocks and amount of replication.
- It closes the file and informs the namenode after the operation is complete.

→ Replication of blocks:

- every block stored in the filesystem is replicated on different Data Nodes.
- The default replication factor of HDFS is 3.

- The default block size is 64MB

Example.



* HDFS Safe Mode

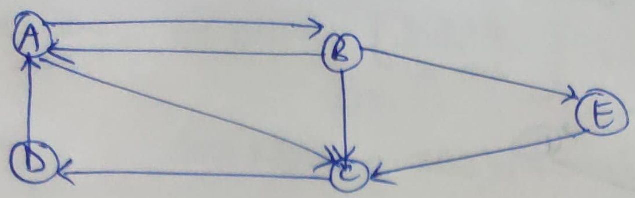
- Name Node starts as a read-only safe mode, where blocks cannot be replicated or deleted.
- It enables the namenode to perform 2 important processes.

* Rack Awareness :

- Hadoop runs on a cluster of computers which are commonly spread accross many racks.
- NameNode places replicated data block on multiple racks.
- Purpose of Rack Awareness is to improve data reliability, availability and n/w bandwidth utilization.

Q.

Shruti Singh
1CR17CS144



$L(A) = 3$; $L(B) = 3$; $L(C) = 1$; $L(D) = 1$; $L(E) = 1$.

Influence Matrix

i/i	R_a	R_b	R_c	R_d	R_e
R_a	0	0.33	0	1	0
R_b	0.5	0	0	0	0
R_c	0.5	0.33	0	0	1
R_d	0	0	1	0	0
R_e	0	0.33	0	0	0

$R_a = 0.33R_b + R_d$

$R_b = 0.5R_a$

$R_c = 0.5R_A + 0.33R_b + R_e$

$R_d = R_c$

$R_e = 0.33R_b$.

Var.	initial values	itr 1	itr 2	itr 3	itr 4	~~itr 5~~
R_a	0.2	0.266	0.233	0.233	0.233	
R_b	0.2	0.1	0.15	0.15	0.15	
R_c	0.2	0.366	0.283	0.283	0.283	
R_d	0.2	0.2	0.2	0.2	0.2	
R_e	0.2	0.066	0.133	0.133	0.133	

Highest ranked node : Node C

$R_c > R_a > R_d > R_b > R_e$.

Shruti Singh
TCR17CS144

☆ Map-Reducing Programming Model.

↳ It helps to process BigData in paralal on multiple nodes.

- Map Reduce algorithm contains two important tasks

① Map → takes a set of data and converts it into another set of data, where individual elements are broken down into tuples.

② Reduce → Takes the output from the Map as an input and combines those data tuples into a smaller set.

Input → split → Map → Shuffle → Reduce → Output.

Steps:

① Input split

- HDFS distributes and replicates data over multiple servers

- 500 MB file broken into 8 blocks.

② Map step

- It takes the key-value pair and processes each one of them

- They operate in parallel.

③ Combiner Step

- An optional stage

- Combines similar data into single sets.

④ shuffle step :

- All similar keys combined and counted by the same reducer process before the parallel reduction stage complete.

- Results of map stage must be collected by key-value pairs and shuffled to the same reducer process.

⑤ Reduce step

- It runs a reducer function on each grouped key-value paired data.

- The data is processed in this layer. Once the execution is over it gives zero or more key-value pairs to the final step.

⑥ Output

- Output formatter translates the final key-value pair from the reducer function and writes them onto a file using a record writer.

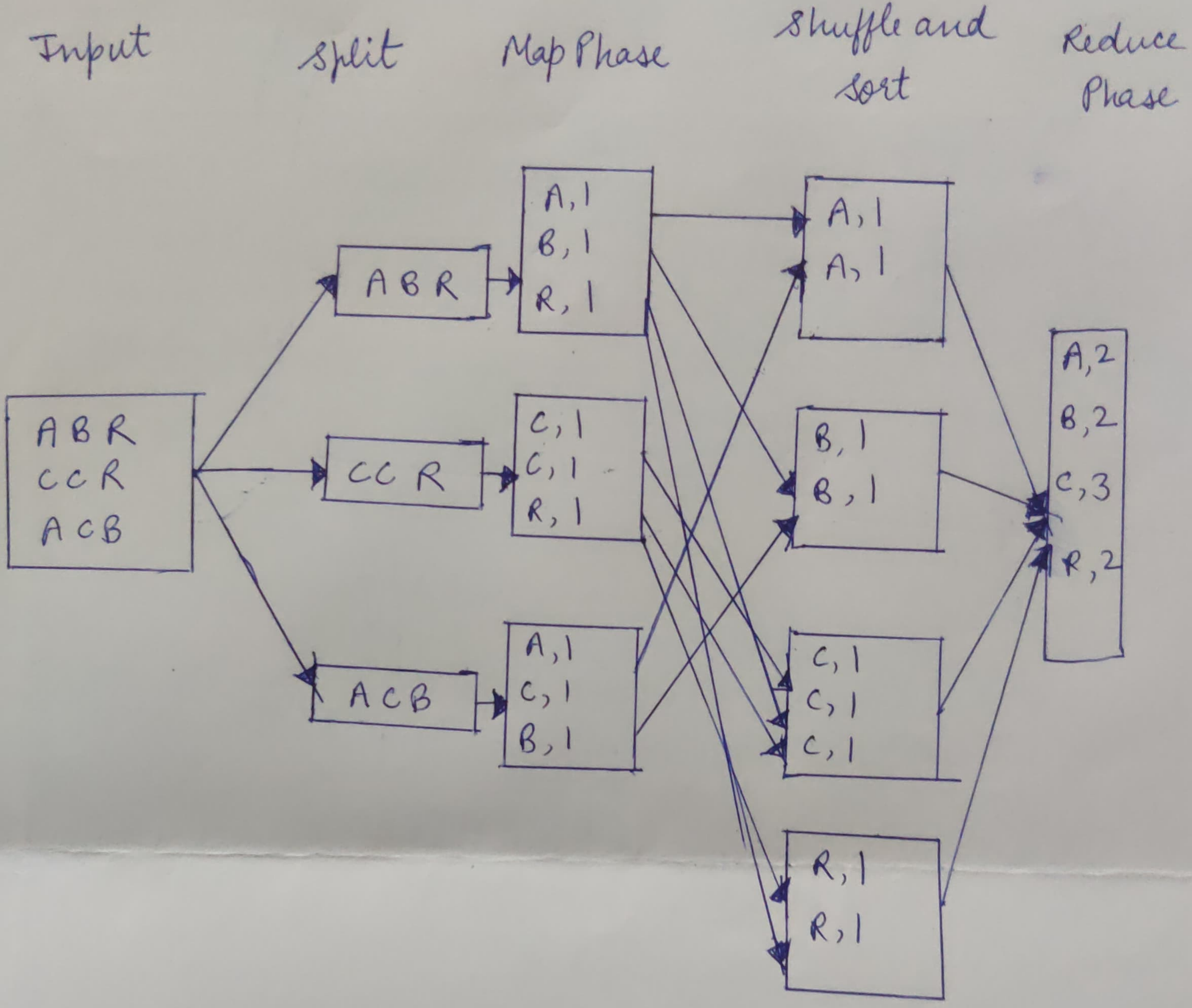
Mapper & Reducer.

Mapper.py

```
import sys
~~inp = input.txt~~

for line in ~~inp:~~ sys.stdin:
    line = line.strip()
    words = line.split()

    for word in words:
        print (word, 1)
```

Reducer.py

```
current_word = None
current_count = 0
word = None
 for line in sys.stdin:
     line = line.strip()
     word, count = line.split('\t', 1)
     try:
         count = int(count)
     except ValueError:
         continue

 if current_word == word:
     current_word += count
 else:
     if current_word:
         print (current_word, current_count)
     current_count = count
     current_word = word
to output last word
 if current_word == word:
     print (current_word, current_count)
```


Input
Split
Map Phase
Shuffle and sort
Reduce Phase
ABR
CCR
ACB
ABR
CCR
ACB
A,1
B,1
R,1
C,1
C,1
R,1
A,1
C,1
B,1
A,1
A,1
B,1
B,1
C,1
C,1
C,1
R,1
R,1
A,2
B,2
C,3
R,2