THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

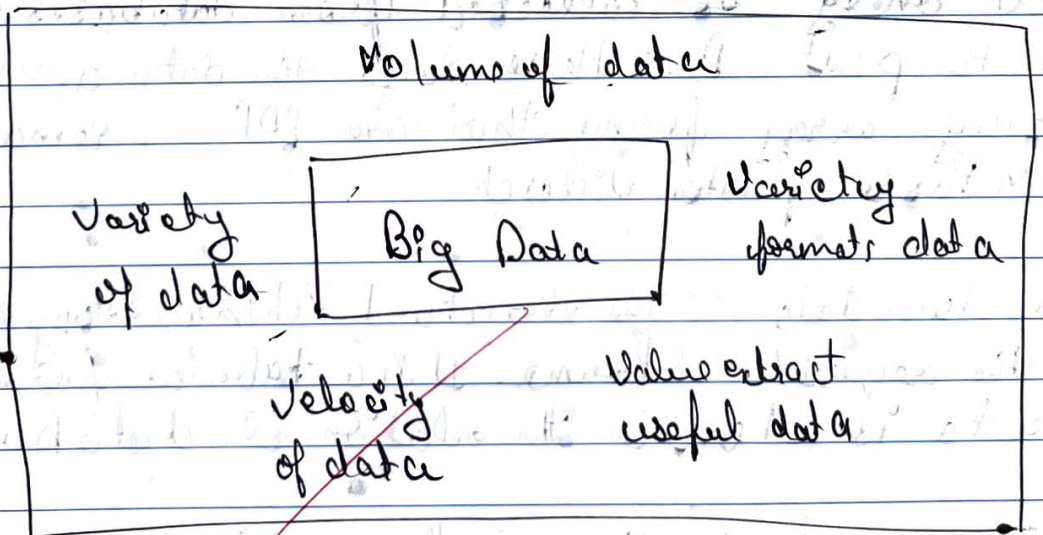
Assignment No. 1

Q Write Briefly on Big data characteristics Hadoop architecture Hadoop ecosystem

⇒ Big data Characteristics

Big data contain a large amount of data that is not being processed by traditional data storage or the processing unit It is used by many mathematical companies to process the data and business of many organization the data flow could be exceed 150 Exabyte per bytes before replication

there are five V's of Big Data that explains the Characteristics



the 5 V's of Big data are as follows:

1) Volume

The name of Big data itself is related to an enormous size. Big data is vast volume of data generated from many sources daily such as business process, machines, social media platforms, networks human interatices and many more

eg :- Facebook can generate approximately a billion message 4.5 B the that the 'like' button is recorded & more than 350 million new post uploaded each day

2) Variety

Big data can be structured, unstructured & semi structured that are being collected from different sources. data will only be collected from databases & sheets in the past, But these days the data comes in the array array forms that are PDF's, emails, audios, SM Posts, photos, videos etc

1) structured data :- The structured schema, along with all the required columns. It is a tabular form. structured data is stored in the relational databases

2) Semi structured :- the semi structured the schema is not appropriately defined eg: JSON, XML, S.

OTP (online transaction processing) systems are built to work with semi-structured data. It is stored in realtime ie table

Unstructured data : All the unstructured files log files, audio files & image files are included in the unstructured data. Some organization have such data available but they did not know how to derive the value of the data since the dat is raw

Quasi structured data : the data formats contains data with inconsistent data formats that are formated with that and time with some tools

eg: Web server logs i the log file is created & maintained by some series that contain a list of activities.

Veracity

Veracity means how much the data is reliable. It has many ways filter or translate the data Veracity is the process of being able to handle & manage data efficiently Big data in also essential in business development.

eg :- facebook posts with hastags

Value :-

Value is an essential characteristic of Big data. It is not the data that we process or store. It is valuable and reliable data that we store process & also analyze

5) Velocity:

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in real time. It contains the linking of incoming data sets speeds, rate of damage, activity Bursts. The primary aspects of Big data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like application, logs Business process network & social media site, sensors mobile devices, etc.

Hadoop Architecture:

Hadoop is an open source framework from apache and is used to store process & analyze the data which are very huge in volume. Hadoop is written in Java & is not OLAP (online analytical processing). It is used for batch/offline processing. It is used by Facebook, Yahoo, Google.

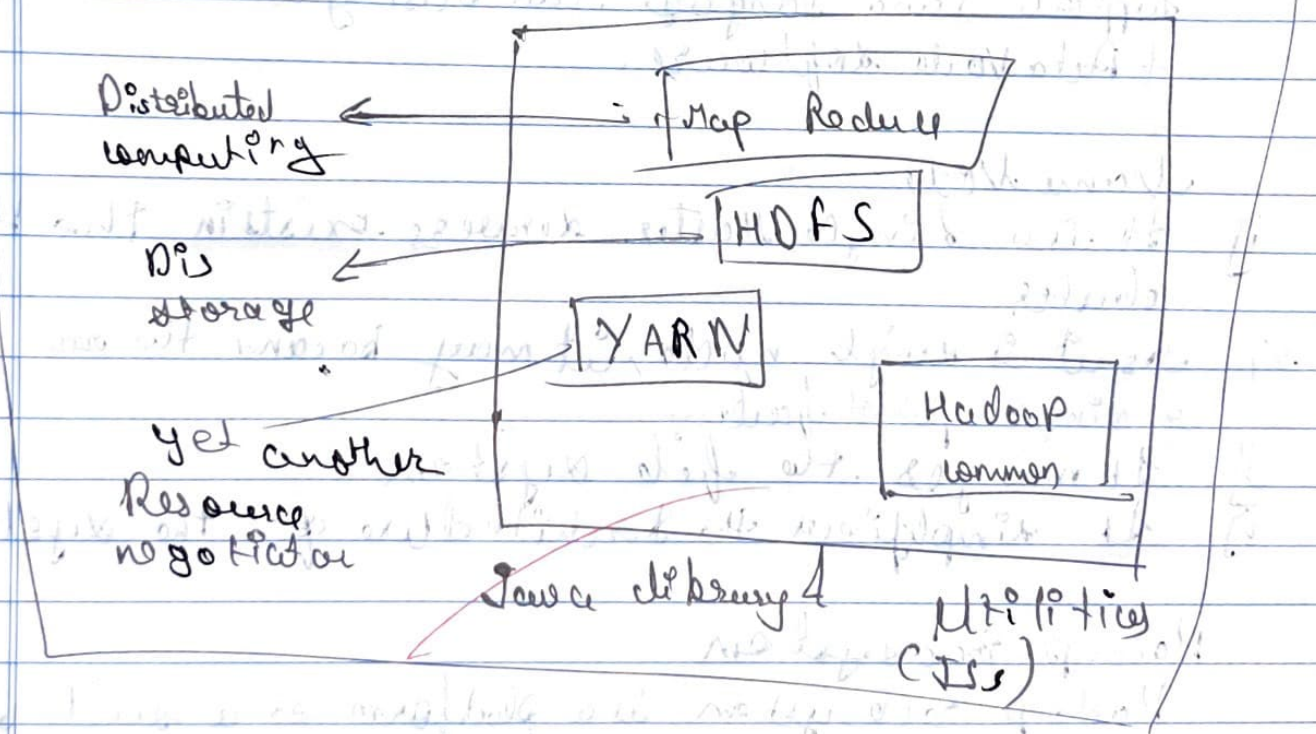
Modules of Hadoop

1) HDFS:- Hadoop distributed file system google published its paper GFS & on the basic of that HDFS was developed. It states that the file will be broken into blocks & stored in nodes. The distributed architecture

2) Yarn :- Yet another Resource Negotiator is used for job scheduling & manage the cluster

3) Map Reduce :- This is a framework which helps with java programs to do in the parallel computation on data using key value pair. the Map task takes input data and converts it into data set Map task is consumed by reduce task & then the out of reduces gives the desired results

4) Hadoop Common :- These Java developer are used to start Hadoop and are used by other Hadoop modules



Hadoop architecture is a package of the file system, Map Reduce engine & the HDFS (Hadoop Distributed file system

* Hadoop cluster consists of a single Master & multiple slave Nodes. The Master Node - Includes Job tracker, task tracker, Nodes.

Hadoop distributed file system
the Hadoop Distributed file system (HDFS) is a distributed file system for Hadoop. It contains a Master / slave architecture. this Java language is used to develop HDFS. So any Machine that supports Java language can easily run the Name & Data Node software.

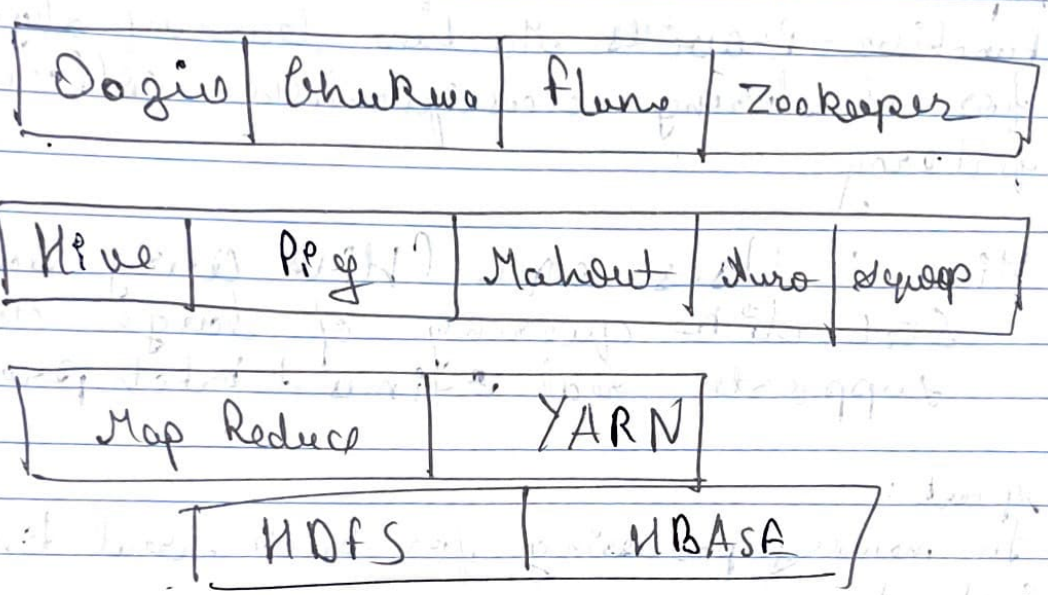
Name Node
1) It is a single Master servers exist in the HDFS cluster
2) As it is single node, it may become the reason of single point failure
3) It manages the file system
4) It simplifies the architecture of the system

Hadoop Ecosystem
Hadoop Ecosystem is a platform or a suit which provides various services to solve big data problems It include apache projects and various components tools & so on

THADOMAL SHAHANI TSEC ENGINEERING COLLEGE

components of Hadoop

1] HDFS :- File system
2] Map Reduce :- Data Processing
3] Yarn :- Yet another Resource negotiator
4] Spark - In Memory data processing
5] PigHive :- Query Based
6] HBase - No-SQL Database



1] HDFS components Name Node (meta data) & Node (Actual Node) Purpose function - stores large datasets across multiples nodes, ensuring fault tolerance

2] YARN
Components Resource manager ; Node Manager ; App manager
function : Process large datasets using distributed parallel algorithms

3] MapReduce
Component - Map Node Hangover. Reduce (aggregate & summarize data).

4] Pig :- Uses Pig latin SQL like language to process & analyse data. Execute commands. MapReduce & store results in HDFS

5] Apache Hive Mahout
function - Provides Machine learning algorithms for clustering, classification collaborative filtering

6] Hive :- Uses SQL (Hive Query language) for SQL like querying of large datasets supports real time & batch processing

7] Spark :
In memory processing for batch, real time & interact

8] HBase :- NoSQL database for real time & interactive data efficient for quick lookups in large data types

9] Zookeeper :- Coordinates & synchronize distributed system, ensuring consistency

10] Oozie :- Workflow scheduler. Manages job execution in sequential & event Based (coordinates job manner

15/1/2024

THADOMAL SHAHANI TSEC ENGINEERING COLLEGE

Assignment 2

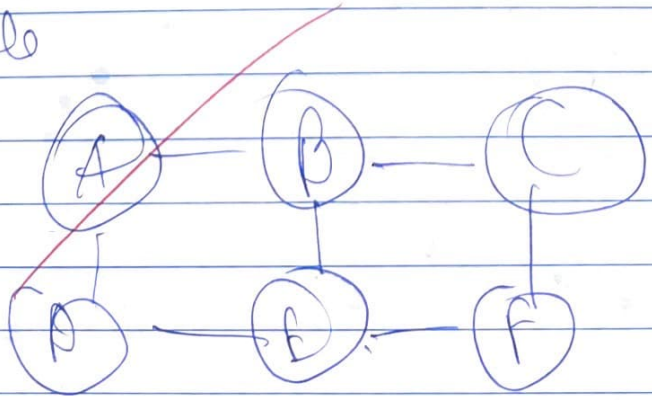
1)

⇒ In Girvan-Newman algorithm we are giving to divide the nodes of the graph into 2 or more cluster. this algorithm removes the edges with the highest betweens until there is no edges remain or any specific no. of edges is present Between is the number of the shortest paths between pairs of nodes that are run through it

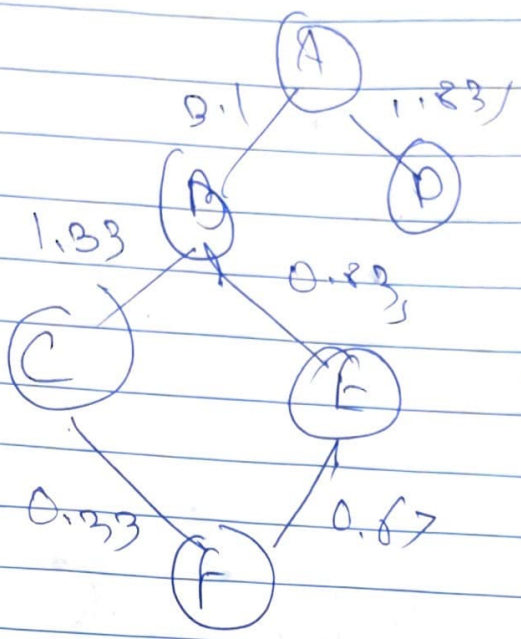
Algorithm.

- Create a graph of N nodes & its edges or take an inbuild graph like karate graph
- Calculate the betweens of all existed graphs in the graph.
- Now remove all the edges with the highest betweens
- Recalculate the betweens of all the edge that got affected by the removed of edges
- Now repeat the step 3 & 4 until no. edges remain

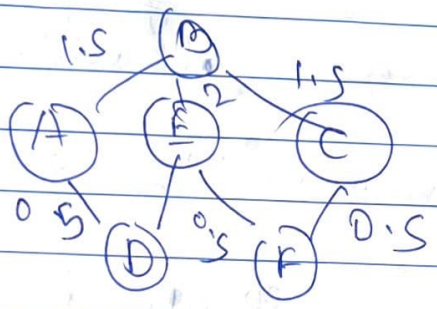
Example



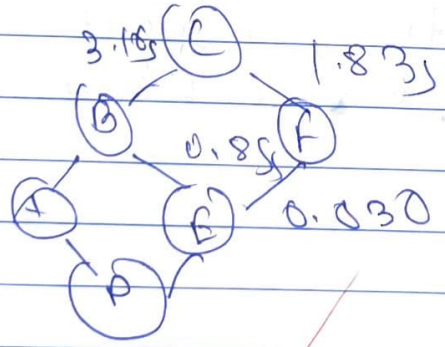
Starting with A



for B

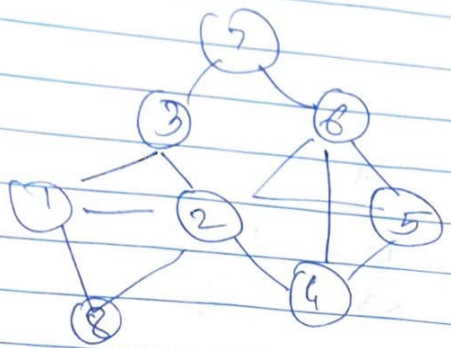


for C



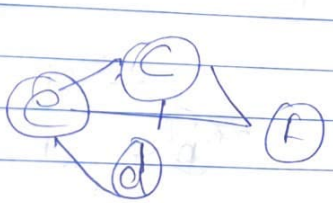
2) Clique percolation method builds up the communities from k-cliques which correspond to complete subgraphs of k-nodes

Ex

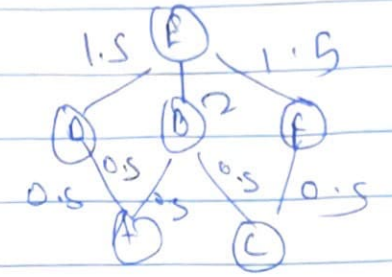


there are six k = 3 cliques

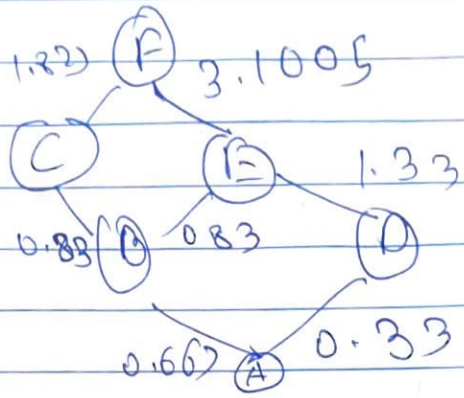
a : 1, 2, 3 > community 1
b : 1, 2, 8
c : 2, 4, 5 > community 2
d : 2, 4, 6
e : 2 5, 6 > community 3
f : 4, 5, 6



for E



for F



Edges	Edge Between E
AB	
AD	5.33
BC	8
BE	7.34
CF	5.33
DE	8
EF	8