AJAY KUMAR GARG ENGINEERING College, GHAZIABAD.

Dept of MCA

Solution ST-2

Course: MCA

Session: 2017-18

Subject: Big Data

Max Marks: 50

Semester: 5

Section: MCA1&2

Sub code: NMCA E-44

Time: 2 hour.

Section A

Q1) what is the purpose of Hadoop Pipe.

Ans 1. Hadoop Pipes: Apache Hadoop provides an Adapter layer called pipes which allows c++ application code to be used in Map Reduce programs.

It allows c++ code to use Hadoop DFS and Map/Reduce.

Q2) what is the advantage of using materialized view.

Ans 2 A Materialized view is a database object that contains the result of a query.

For Example, it may be a local copy of data located remotely, or may be sub set of rows and/or columns of a table or join result, or may be a summary using an aggregate function.

It is a database query that contains the

result of pre computed query.

Q3 Differentiate between Scale up and Scale out.

Ans3 In vertical scaling: the data resides on a single node and scaling is done through Multicore i.e spreading the load between the CPU and RAM resources of that Machine. It is often limited to the capacity of a single machine, scaling beyond a capacity becomes too difficult.

Horizontal scaling: is often based on Partition of the data i.e each node contains only a part. It allows you to Combine the power of Multiple M/c into a Unified Single M/c unit Combined power of all of them together.

Before we Conclude that Horizontal Scaling is achieved through partitioning and Vertical Scaling is achieved through multicore support.

Q4) What is Codec? Name the interface for its Implementation.

Ans4. Codec is the implementation of a Compression — Decompression Algo. In Hadoop a Codec is represented by an implementation of the Compression Codec interface.

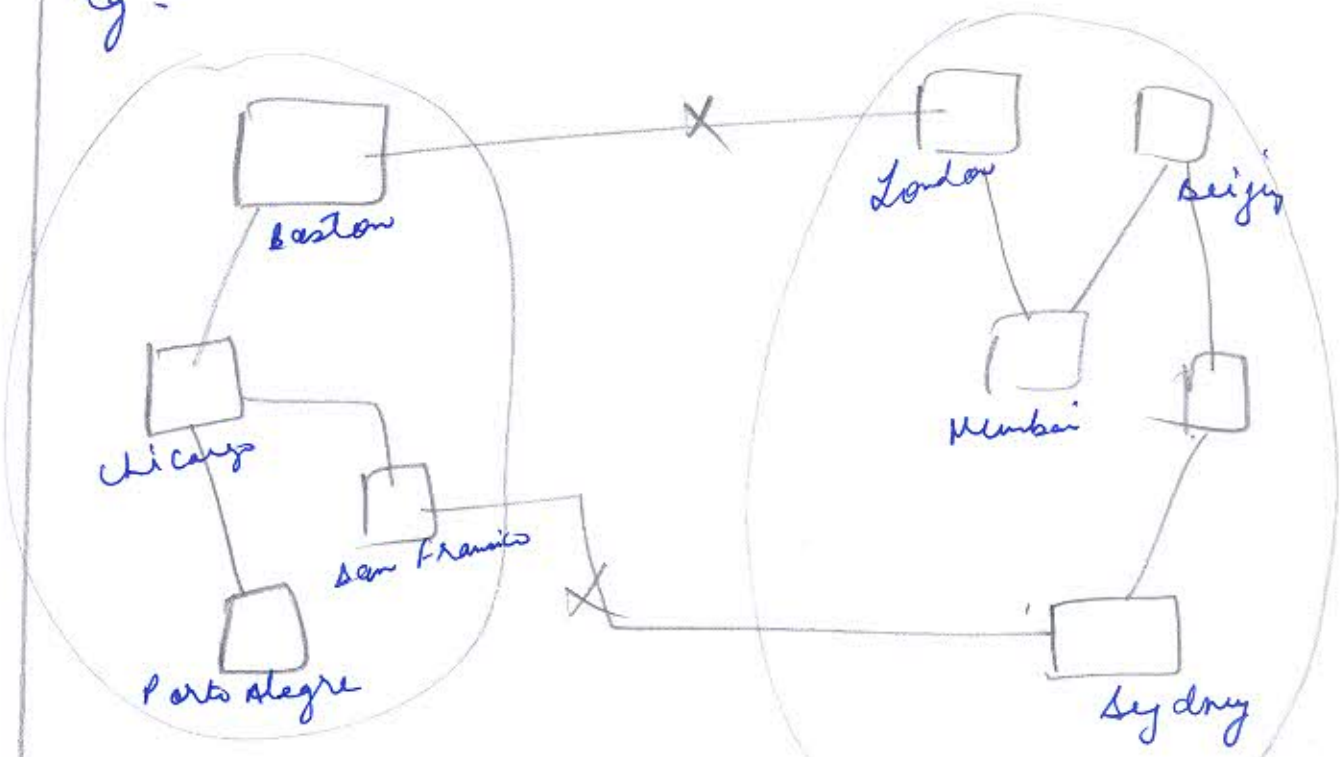
For Eg: GZIPCodec encapsulates the Compression & decompression algorithm for gzip.

Q5. Explain Partition Tolerance through an Example

Ans 5 - Partition Tolerance means that the cluster can survive communication breakages in the cluster that seperate the cluster into multiple partitions unable to communicate with each other.

This situation is also called Split brain.

Eg -



With two breaks in the communication lines, the network partitions into two groups.

∴ clusters have to be tolerant of network partitions.

Section B

Q6 Compare and Contrast various Compression Algorithms in Hadoop.

Ans 6 Compression brings two major benefit:

- It reduces the space needed to store file
- It speeds up data transfer across the network or to or from disk.

Various Compression formats & respective features is as following:

Compression Format	Tool	Algo	Filename Extension	Splittable
Deflate	N/A	Deflate	.deflate	NO
gzip	gzip	Deflate	.gz	NO
bzip2	bzip2	bzip2	.bz2	Yes
lzo	lzop	LZO	.lzo	NO
Snappy	N/A	Snappy	Snappy	NO

All compression Algo exhibit a space/time trade off: Faster compression and decompression speed usually comes at the expense of smaller space savings.

- GZIP is general purpose compressor, and in the middle of speed/time trade off.
- BZIP2 compresses more effectively than GZIP, but is slower.
- BZIP2's decompression speed is faster than its compression speed, but it is slower than other formats.
- LZO & Snappy, both optimize for speed and are around an order of magnitude faster than gzip, but compress less effectively.
- Snappy is slightly faster than LZO for decompression

Q7) What is the use of Serilization? What are Methods of serilization in Hadoop.

Ans. 7. Serilization is the process of turning structured objects into a byte stream for transmission over a network or for writing to persistent storage.

Serialization is used for:

(1) IPC (Interprocess Communication)

(2) Persistent storage.

Methods of Serialization in Hadoop are.

- Writables
- AVRO.

Writables is compact & fast but not so easily to extend or use from language other than Java.

Avro: supports many language other than Java including for serialization.

Q8 Differentiate between Schemaless and Schema based data Base.

Ans- NoSQL is another name for Schemaless database while SQL is Schema based Data Base. Following are the difference between the two based on following parameters.

Entity	SQL	NoSQL
① Scaling	Vertical scaling	Horizontal scaling
② Schema	Fix	Dynamic
③ Development Model	Mix	opensource
④ Consistency	Follow ACID (Atomicity - Consistency Isolation Durability)	Follow BASE (Basically Available, Soft State Eventually consistent)
⑤ Model	Relational	Non-Relational. Simple
6 Development	Since 1970	

Q9 Discuss various Distribution Model with their advantages and disadvantages.

Ans 9 Various Distribution Models can be:

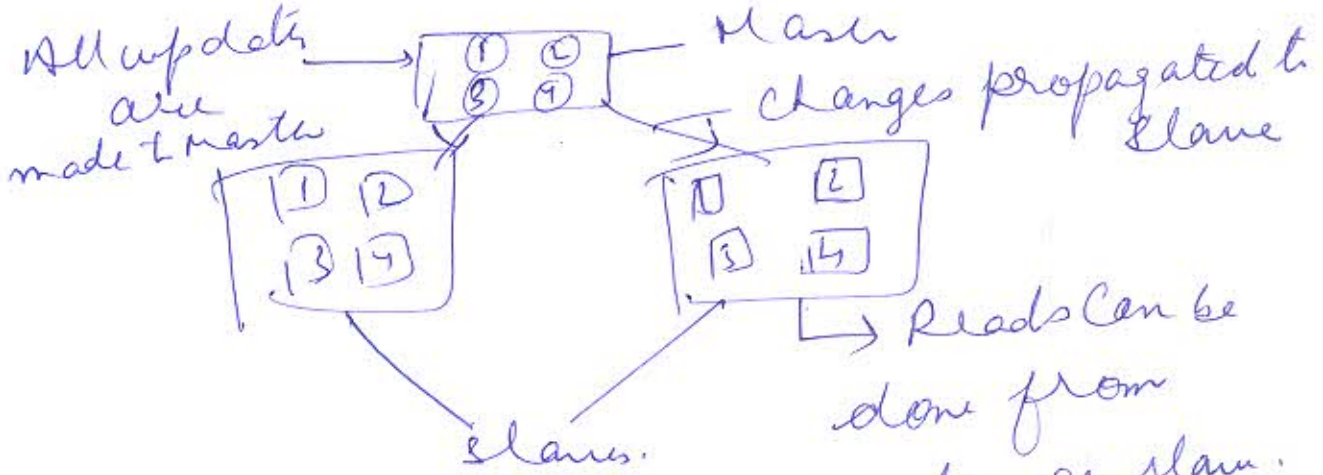
Basic Replication & Sharding:

Replication puts takes same data and copies it over multiple nodes while Sharding puts different data on different nodes.

They can broadly be.

1. Master slave
2. Peer - to - Peer Replication
3. Sharding
4. Combining sharding & Replications

In master slave: One node is designated as master other nodes are slave

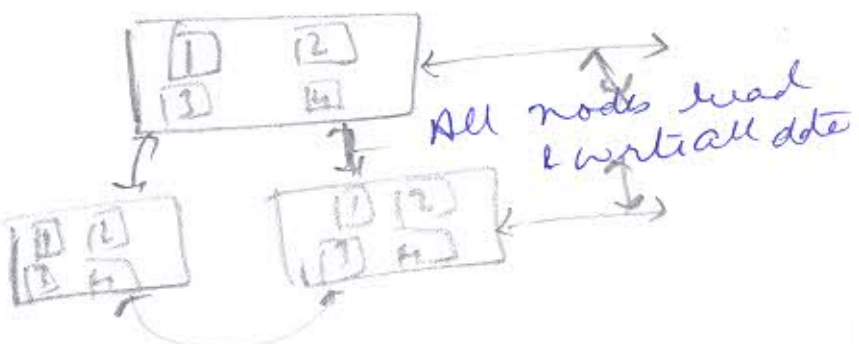


Master slave.

Advantage

1. It is read Intensive
2. Read Resilience.

Peer to Peer.



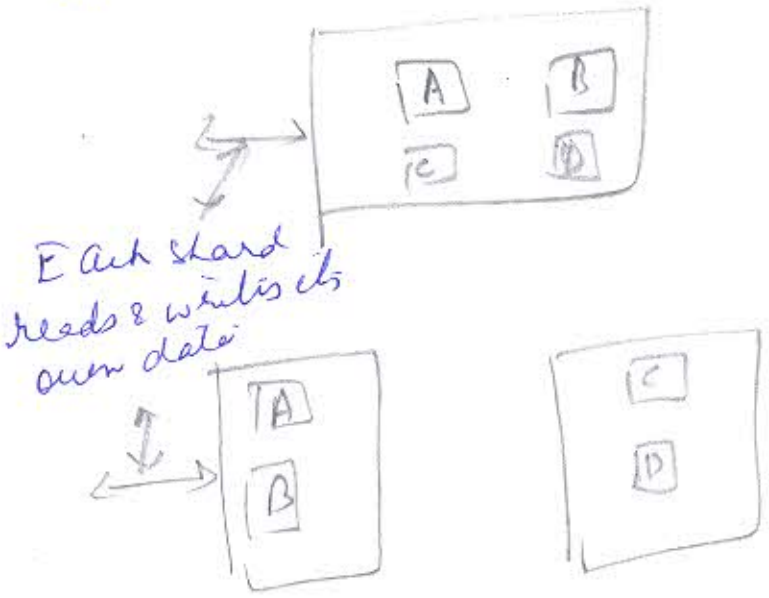
Peer to Peer

Peer to Peer Disadvatage is :

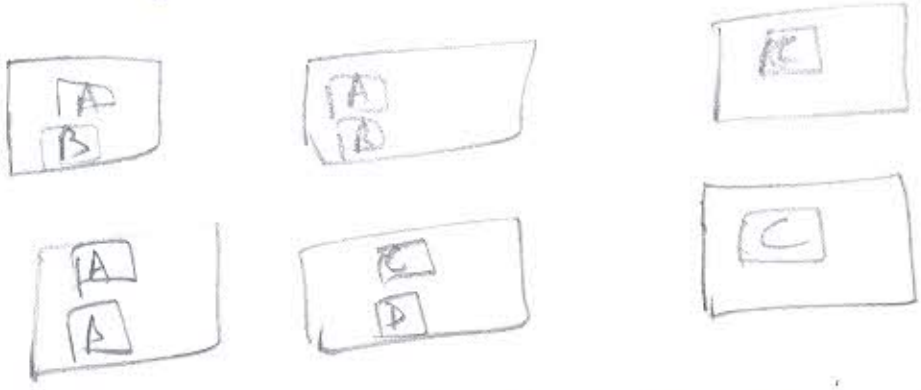
1). write - write conflict.

In sharding different people are accessing different part of datasets.

different parts of data are put onto different servers :



Combining sharding & Replication



Thus Sharding & Replication are orthogonal strategies that can be.

Q10) What is the advantage of AVRO File Based data structures over sequence file format?

Ans10 Sequence file is a file format consisting of binary key/value pair

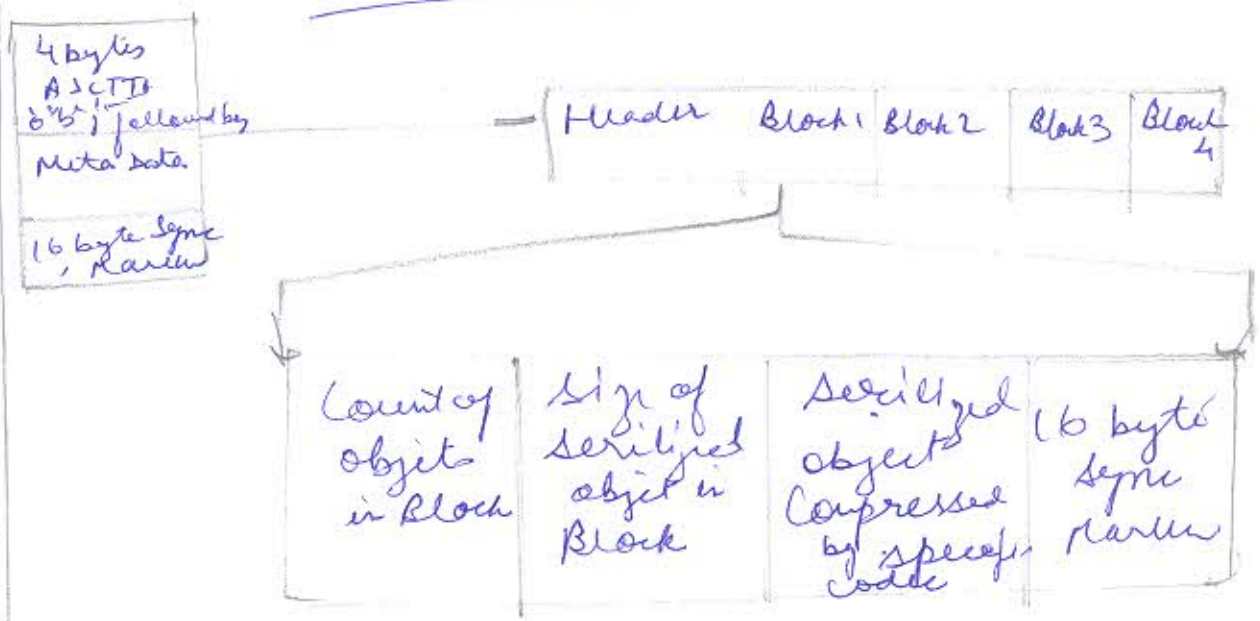
It provides writer, Reader & Sorter classes

There are three different formats for sequence file depending on the compression types specified:

- Uncompressed
- Record compressed
- Block compressed format

Sequence File Header
3 Bytes (SEQ) + 1 Byte (version)
Text - key class Name
Text - value class Name
Boolean - Is compressed
Boolean - Is Block Compressed
Compression codec class name
Metadata
Sync Mark

Avro File :-



~~Big data Avro~~

Avro schema are usually written in JSON

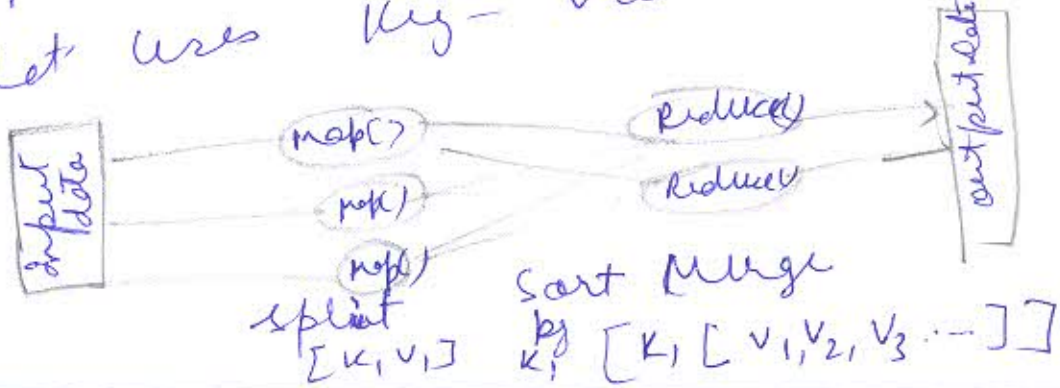
Supports splitability

Its Row Based, Compact binary form.

Section C

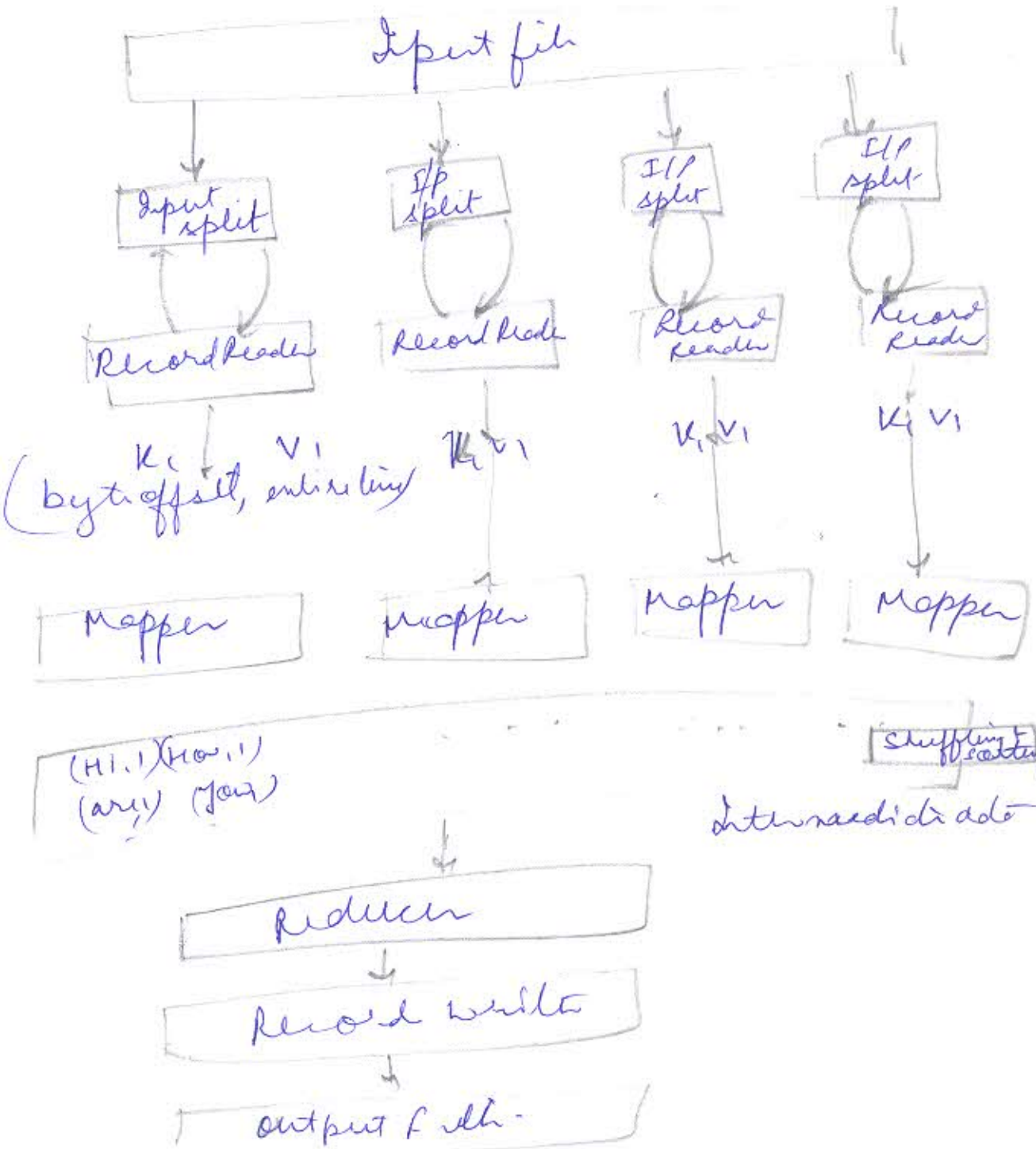
Q11 ~~map Reduce~~ what are the stages involved in map-Reduce? Explain giving an example

Ans 11 Map Reduce is a programing Paradigm that uses key-value pair.



F.h. input (200 MB)

Hi How are you How is your job How is ur family How is ur Bro what is time now How old are you what is the strength of India?	64MB 64MB 64MB 64MB



Q12) Explain key value and graph data model in detail giving examples?

Key value database:

A 12

Basic data structure is may, where we can store value a string, a JSON structure along with key to reference that value

eg are Riak, Redis, Memcached,

eg:

Car	
key	Attributes
1	Make: model: Color: Year:
2	Make: model Color: color: year: Transmission:

Various key value pair Can be:

(Business) key → value

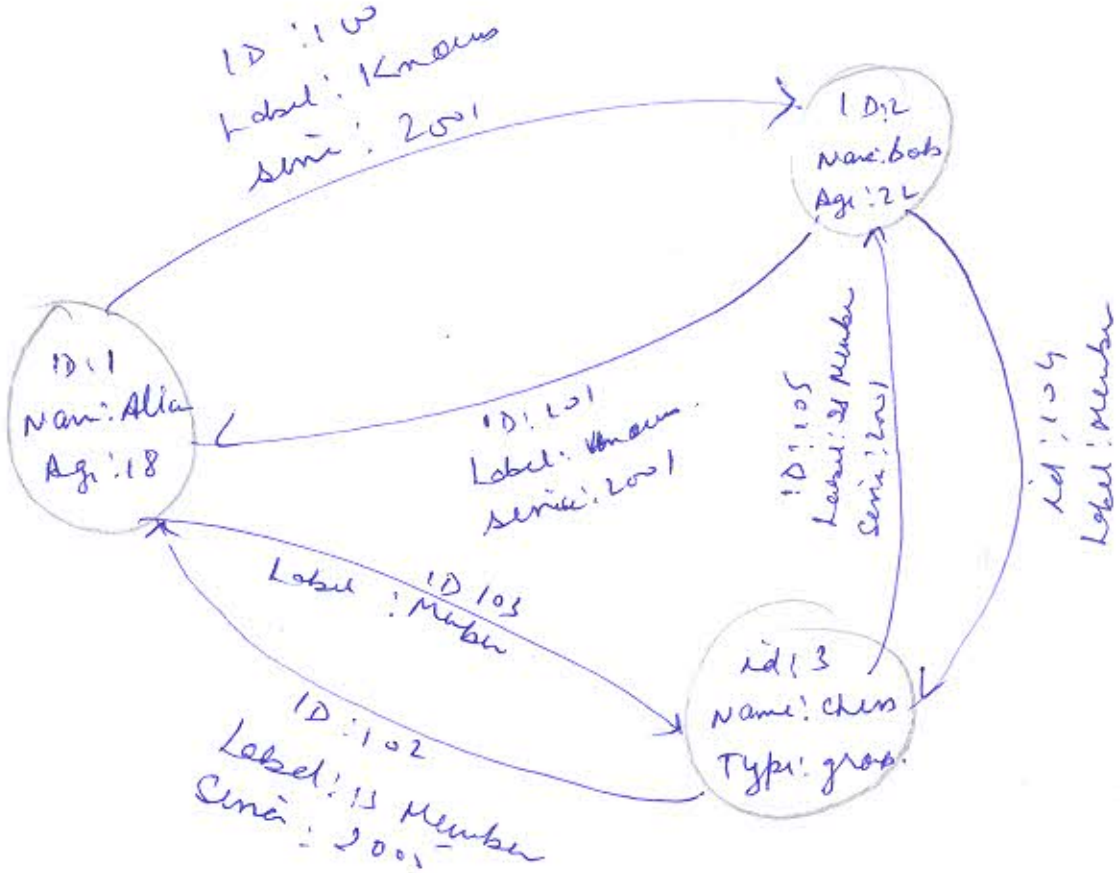
(twitter ID) key → Information about that value

Graph model: uses graph data structure to store data.

It is used when traversing relationships are core to the application like social

Social network connection

eg Neo 4j & ~~Graph~~ Giraph are its commercial product.



The Node user: Bob is a vertex with a property Bob.

We also see Relations which are edges:

- Boss of
- Friend of
- Known to
- married to.