Big data.

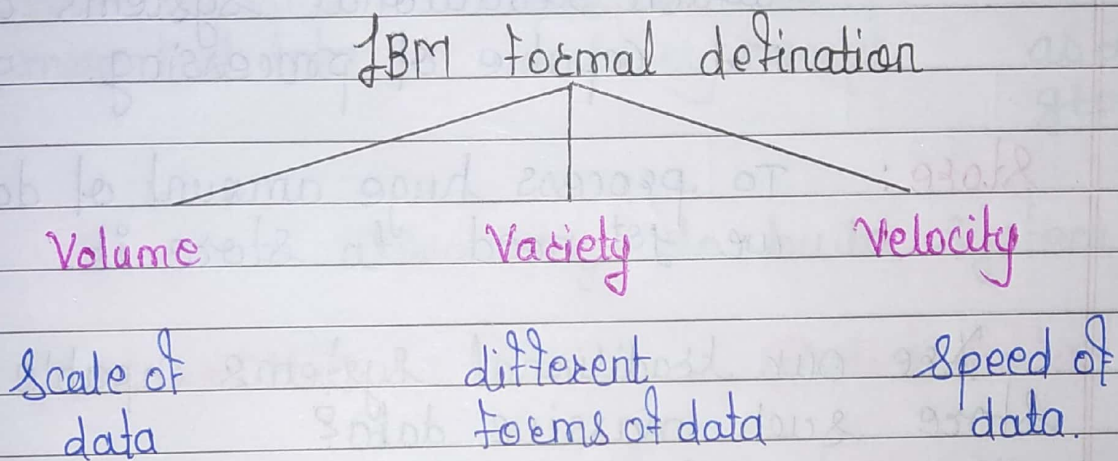## What is big data?

"big data is a term that describes the large volume of data."

How to classify big data?

3v's of big data by IBM.

IBM formal defination

Volume	Variety	Velocity
Scale of data	different forms of data	Speed of data.

- Volume   - 2.5 quitillion
- Variety :
  - ① Structured data: RDBMS databases (Oracle & maySQL)
  - ② Semi-Structured data: CSV, XML, JSON.
  - ③ Unstructured data: Audio, video, Image log files

- velocity: 900 million photos on facebook.
  600 million tweets on twitter.

4th v of data:

Veracity : uncertainity of data
i.e poor quality data
unclean data.

Why big data?

process: To process huge amount of data
which tradilional systems are
not capable of processing.

Store: To process huge amount of data
we 1st need to store it.

Are our tradilional systems capable to
Store such massive data?

Big data System Requirements?

* Store huge amount of data :
  • Teaditional Systems are NOT fit
  to store such huge amount of data.

  • So Store massive amount of
  data.

* process huge amount of data in a
  efficient and timely manner.

X tradilional system are not capable to handle.

* Scale easily to accomodate growing requirement

• tradilional systems have serious limitalions

store	Process	Scale
Store massive amount of data	process it is a timely manner	Scale easily as data grows.

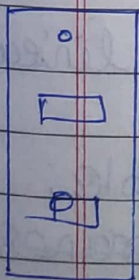Two ways to build system.

① Monolithic   ② distributed

Monolithic: One powerful system with lot of resources.

distributed: Many smaller systems come together.

Q. Monolithic or distributed?

© Monolithic is a single powerful server. Hard to do add resources after a certain limit.

Monolithic

② Resources: ① RAM — 8 GB (Memory)
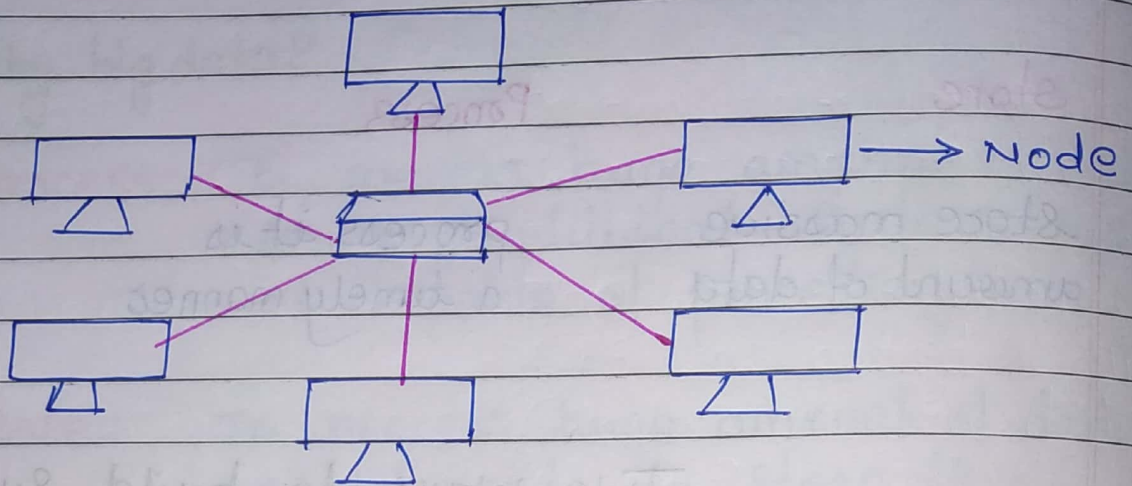② Hard disk - 1 TB (storage)
③ CPU quad core (compute)

Is monolithic scalable?

No!

2x resources ≠ 2x performance

Distributed:

6 Node cluster.

① The set of all nodes is known as cluster.

② Many small and cheap computers come together. to act as a single entity

Is distributed system scalable?

Yes, distributed systems are linearly scalable.

Distributed systems are scalable.

2x resources = 2x performance speed.

Monolithic Aechitecture

Disleibuted Aechitecture

Vertically Scaling
(Not true Scaling)

Horizontal scaling
(True scaling)

So, Monolithic or disteibuted?

✗ Monolithic   distributed ✓

'That's Why all good big data Systems are based on distributed architecture!

Hadoop

What is hadoop?
        Hadoop is a feamework to solve bigdata problems.

Hadoop Evolution:

2003	Google released a paper to desceibes how to stoce large datasets.
2003	This papes was called as GFS (google file system).

2004		Google released another paper to described how to process large dataset
2004		This paper was called as MapReduce.
2006		Yahoo took these papers & implemented it.
2006		The implementation of GFS was named as HDFS (Hadoop distributed File System).
		The implementation of Mapreduce was named as Mapreduce (unchange)

Hadoop Version 1.0

HDFS - for distributed storage
MapReduce - for distributed processing.

2009		Hadoop comes under Apache Software Foundation & become openc. Source.
2013		Apache released Hadoop 2.0 to provide major performance enhancement.

2003	2004	2006	2009	2013
google GFS	google MR	Yahoo Implementation	Hadoop under Apache	Hadoop 2.0 released

Hadoop 1.0

MapReduce
HDFS

Hadoop 2.0

MapReduce   YARN
          HDFS.