Big data.

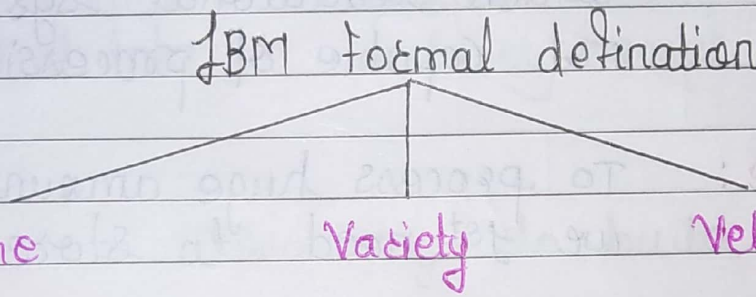
Page No.
Date

What is big data?

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

How to classify big data?

3V's of bigdata by IBM.



Volume — Scale of data

Variety — different forms of data

Velocity — Speed of data.

- Volume - 2.5 quintillion
- Variety:
  1. Structured data: RDBMS databases (oracle & mySQL)
  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.

Page No.
Date

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.

Page No.
Date

X traditional System are not capable to handle.

* Scale easily to accomodate growing requirement

• traditional Systems have serious limitations

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

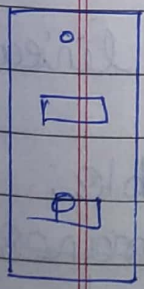
Two ways to build System.

① Monolithic ② distributed

Monolithic: one powerful system with lot of resources.

distributed: many smaller systems come together.

Q. Monolithic or distributed?



② Monolithic is a single powerful server. Hard to do add resources after a certain limit.

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

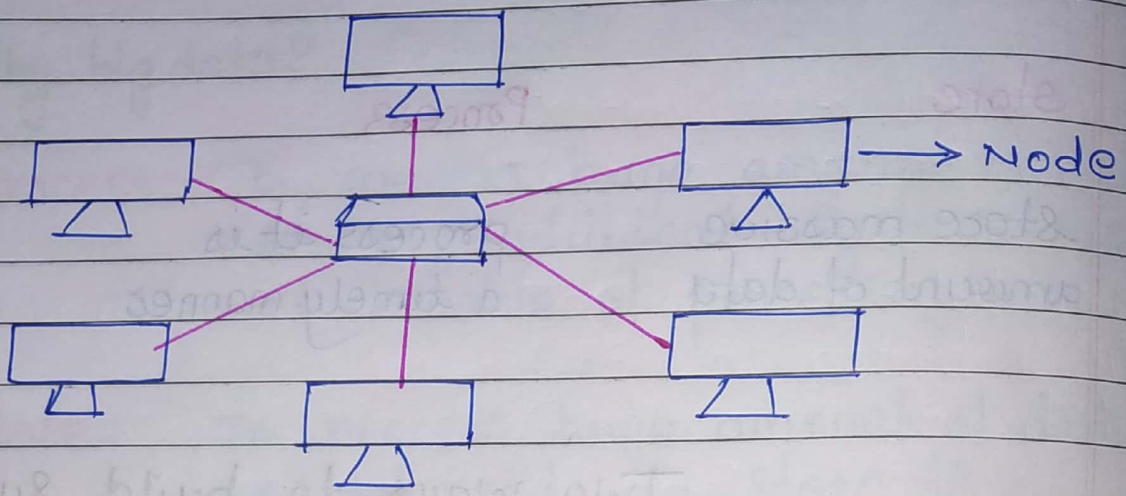
Monolithic

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.

Page No.
Date

Monolithic Architecture → Vertically Scaling (Not true Scaling)

Distributed Architecture → Horizontal Scaling (True scaling)

So, Monolithic or distributed?

✗ Monolithic distributed ✓

That's why all good big data systems are based on distributed architecture.

Hadoop

What is hadoop?

Hadoop is a framework to solve bigdata problems.

Hadoop Evolution:

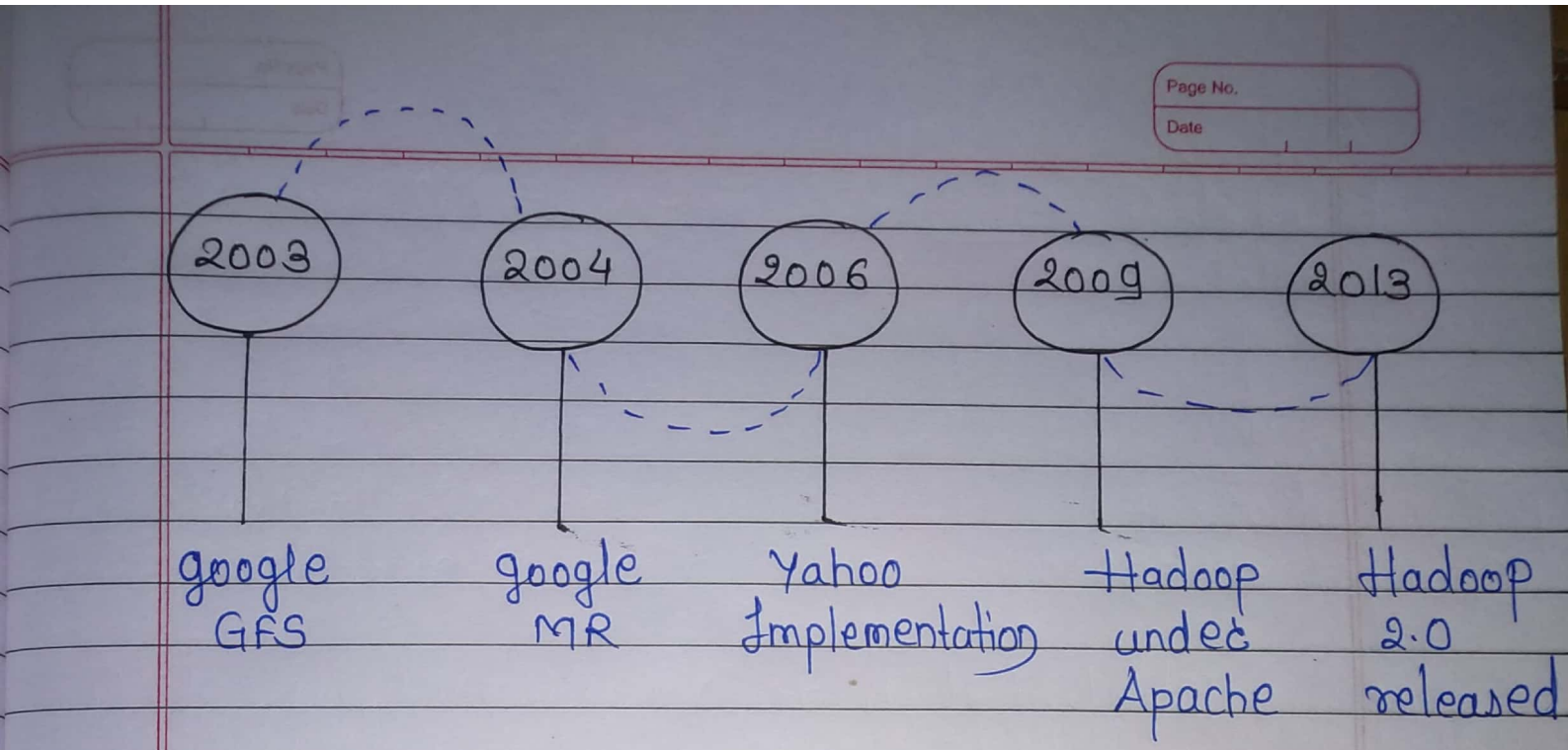
2003	Google released a paper to describes how to store large datasets.
2003	This paper was called as GFS (google file system).

2004	Google released another paper to described how to process large datas
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 ~~of~~ 1.0

HDFS - for distributed storage
MapReduce - for distributed processing.

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



Hadoop 1.0

MapReduce
HDFS

Hadoop 2.0

MapReduce YARN
HDFS