IS BIG DATA A TECHNOLOGY?

As a matter of fact, Facebook handles 105 petabytes of data every hour, which is roughly 100 million gigabytes.
So, how this huge bulk of data is getting managed. Is it because there is massive hard disk available to store data?
The above problem is known as **#BIGDATA**.
So, bigdata is not a technology. It's a problem we have to overcome with technology.

And some sub-problems of BIGDATA are:
VOLUME and VELOCITY:

Let's discuss a scenario:

You know there are companies like dell EMC, who have the technology to make storage drives in respect of exabytes. Although it will be costly but if neglect the matter of cost. Will it be viable to use?

The answer is a big NO. Generally speaking, if it takes almost 1 min to copy 1 GB data in a typical SATA hard drive. So, storing petabytes of data in a single storage device will be a calamity. It will take several days.

Why not use SSD?

The transfer of data will be fast. Yes, indeed. But the concept here is not about deciding which type of storage to be used. It's about the how efficiently we can use the storage!

The business of google, Facebook runs on the instant access of data. And we people i.e. including me too, don't want to wait even a second. Forget about a day.

VOLUME represent the problem of how to store the data in an instant.
And **VELOCITY** represent the problem of how to read that data it an instant and successfully transfer to the requested user.

The solution is DISTRIBUTED STORAGE: DISCOVERY OF SUPER COMPUTERS.

We split the incoming data into many smaller parts and parallelly store them in many storage units. This configuration is based on the topology of MASTER-SLAVE MODEL. The setup in which multiple computers are working together is known as CLUSTER. In this design, one

module is the master unit, which handles the splitting of incoming and outgoing data from the server and all other units are slaves as storage units.

And how are they connected?

It is none other resource but networking. In a cluster, every computer is a node. The master is termed as name node and the slaves as a data node and is completely known as a multi-node cluster. One of the software used is Hadoop to implement these clusters.

How is the data distributed among the data node?

Transferring data from one to another involves the use of the protocol. E.g., a famously known protocol is HTTP. Here, HDFS (Hadoop Distributive File System) is used.

A very interesting fact:
Facebook use approximately 15,000-50,000 units in a cluster. E.g. if one storage unit can store 1,000 gb of data in 1000 min. then a cluster of 1000 units can save it in 1min.
That's the power of Distributive storage.

By:

AKSHAY ANIIL