Let's assume that, you have 100 TB of data to store and process with Hadoop. The configuration of each available DataNode is as follows:

* 8 GB RAM
* 10 TB HDD
* 100 MB/s read-write speed

You have a Hadoop Cluster with replication factor = 3 and block size = 64 MB.

In this case, the number of DataNodes required to store would be:

* Total amount of Data \* Replication Factor / Disk Space available on each DataNode
* 100 \* 3 / 10
* 30 DataNodes

Now, let's assume you need to process this 100 TB of data using MapReduce.

And, reading 100 TB data at a speed of 100 MB/s using only 1 node would take:

* Total data / Read-write speed
* **100 \* 1000000/ 100 (Please Note: 1 TB is 1000000 MB)**
* **1000000 seconds**
* **277.78 hours or 16666.67 minutes**

So, with 30 DataNodes you would be able to finish this MapReduce job in:

* **277.78 / 30**
* **9.26 hours**

1. Problem Statement

How many such Data Nodes you would need to read 100TB data in 5 minutes in your

Hadoop Cluster?

277.78 \* 60 /5 ~= **3334**