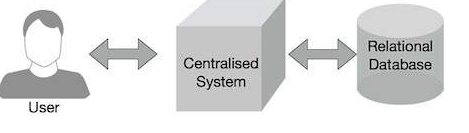
**Map Reduce**

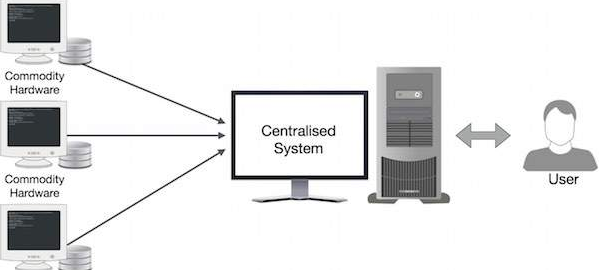
Hadoop **MapReduce** is the heart of the Hadoop system. It provides all the capabilities you need to break **big data** into manageable chunks, process the **data**in parallel on your distributed cluster, and then make the **data** available for user consumption or additional processing.

Why MapReduce?

Traditional Enterprise Systems normally have a centralized server to store and process data



Google solved this bottleneck issue using an algorithm called MapReduce. Map Reduce divides a task into small parts and assigns them to many computers. Later, the results are collected at one place and integrated to form the result dataset.



## How Map Reduce Works?

The Map Reduce algorithm contains two important tasks, namely Map and Reduce.

* The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).
* The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.

The reduce task is always performed after the map job.

**The Mapper basically contain:**

**RecoredReader:**

Record Reader cover a byte oriented view of input into a record oriented view and put it to mapper task.it put record in key and value pair.

**Map**

It work on key value pair generated by Record Reader and generate zero or more key value pair. The Map Reduce decide Key value pair based on context

**Combiner**

it take intermediate key value pair provide by mapper and apply user specific aggregate function to only that mapper it is also known as local reducer.

**Partitioner:**

it take intermediate key value pair produce by mapper split them into shared and send the shared to particular reducer as per user specific code.

REDUCER:

Shuffle and sort:

the phase take the output of all the partitioned and download them into a local where reducer is running machine. The main purpose is grouping similar word.

Reduce: these reduce group data produce by shuffle and sort phase. apply reduce function and process one group at a time. It provide various function aggregate filter group join

output phase: it separate key-value pair to tab(default)and write our output in record writer