# **What is Apache spark?**

**Spark is a parallel execution framework which enables us to perform operations on big data sets in a highly parallel fashion.**

**Spark is 10X times powerful than hadoop**

**Terms**

DAG= Direct Acyclic graph==execution plan

## Comparison with Hadoop

Hadoop is a popular parallel execution framework, which uses the map reduce framework to crunch big data sets.

And it was a revolutionary technology, but it suffers from two problems.

Firstly, it's a very rigid model.

You must do a map and then a reduced process, which is powerful but maybe not suitable for all requirements. Its second weakness is that if you have complex requirements, then you need to chain together, map

reduce jobs, and due to the way that Hadoop map produce is designed, after one map produces finished,

**the results have to be written to disk** and then reloaded into the next map task.

So there's quite a hit on performance there.

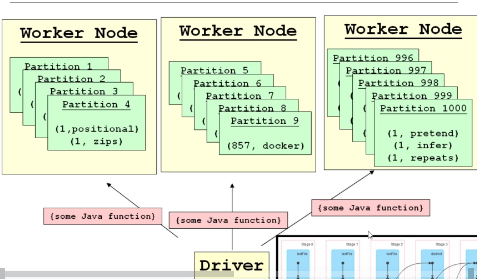
So clearly Spark kind of defines itself in comparison with Hadoop map reduce.

What happened when we upload the code the driver node will create the

## Partitions

A Partition is a smaller block of data, it’s not a node.

A node will have multiple partitions in it.



Worker Nodes are physically separated computers is a just an another physical computer

1. Driver is responsible for dividing and sending the data.
2. Data will be divided and distributed across the worker nodes , among partitions(means huge data will be divided and distributed across many physical computers /worker nodes because 1 single computer cannot process all these huge data) and
3. Once the above java functions reach the worker nodes , these functions will be executed on each partition. Lets say we have the sort function the sorting will be applied on each partition and many parallel threads will be running on each partition and sorting will happen on all partitions simultaneously

### Task

When a java function is applied on a partition –means on a partition level in a worker node then it is called a task.

## **RDD- Resilient distributed Datasets**

Resilient means if any node fails then data on that node can be recovered or recreated



Here only execution plan is created , only at last it will be executed



In above local means , we are running on local machine and \* means use all cores of a processor

If u don’t mention \* by default it will use only single core