Article

Get Details About Distributed Computing Environments

By **Abdul Rasheed Feroz Khan** on **Dec 07 2016**

Introduction

This article will help you to understand the need for a distributed computing environment, understand the role of Hadoop in a distributed computing setup and the core technologies, which works with Hadoop.

Just as a kick start, let's move across the data set that the giants like Facebook, Google and NSA have.

Note

All data specified below is sourced from Google.

Facebook

Data that Facebook holds	300 Petabytes
Data processed by Facebook in a day	600 Terabytes
Users in a month on Facebook	1 Billion
Likers in a day on Facebook	3 Billion
Photos uploaded in a day on Facebook	300 Million

NSA

Data stored in NSA	5 Exabytes
Processing per day	30 Petabytes

It is said that NSA will be touching 1.6% of internet traffic per day as we know, the NSA monitors Web searches, Websites visited, phone calls, financial transactions, health information etc.

Note - This is not an appropriate report as NSA doesn't release its statistical values for the third parties.

Google

Oops!! Google beats the amount of data transaction compared to the other two giants!!

Data that Google has	15 Exabytes
Data processed in a day	100 Petabytes
Number of pages indexed	60 Trillion
Unique searches per month	1 Billion plus
Searches per second	2.3 Million

Thus, having a huge data set like single machine, however powerful it may be, cannot compute all these data at such a huge scale. Hence, we move towards a Big Data System.

Big Data System Requirements

Store - Raw Data

It should be able to store the massive amount of data as the raw data is always huge.

Process

We should be able to extract the useful information alone from it in a timely manner. Finally, we need a scaling infrastructure that can keep up with the data, which keeps growing. This should help us to store the data as volume increases and it can also process the same.

Note - Storage and processing is useless without having a scaling Infrastructure that can accommodate the needs that keep on increasing with the data.

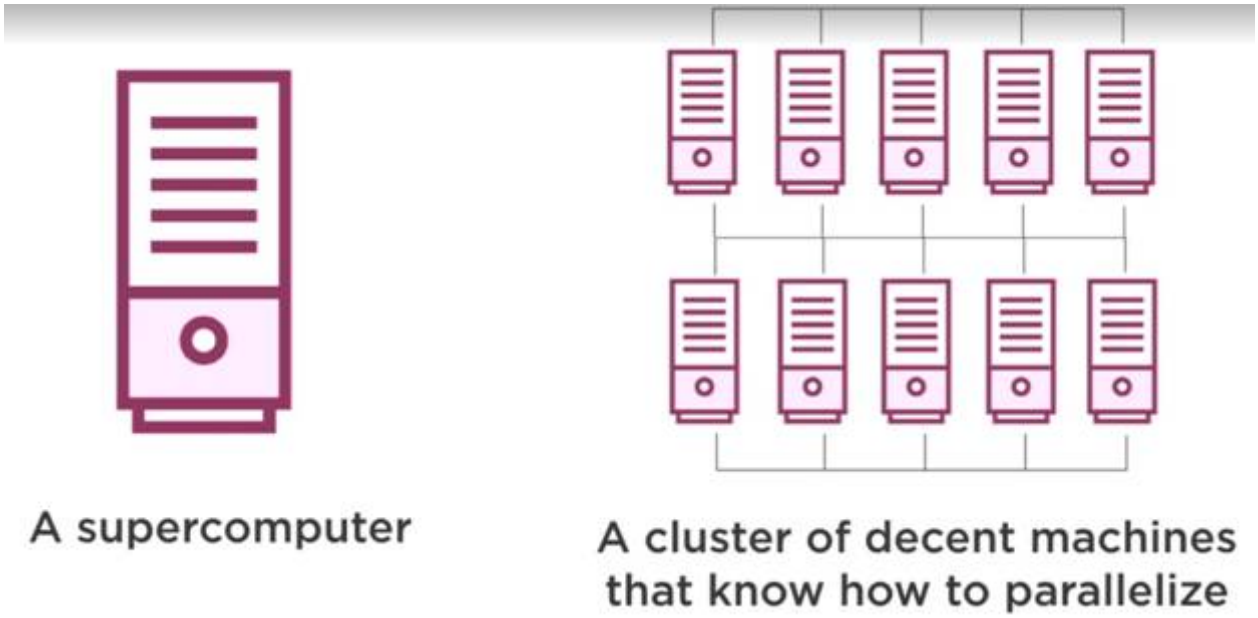
Thus, to work with such an environment, which holds various insights, we need a distributed computing framework and there goes Hadoop.

Build System for Computation

Any system that can make computation for us can be built in two ways.

1. *Monolithic System*
   Everything on a single machine, which is a super computer and processes all the data instructions.

2. *Distributed System*
   There will be multiple machines and multiple processors. It holds a set of nodes called cluster, where no node is a Super computer , which acts as a single entity. This system can scale linearly and if you double the number of nodes in your system, you will be getting double the storage. It also increases the speed of the machine to twice.



This distributed system is where the companies like Facebook, Google, Amazon, etc. survive with the cluster of machines with the vast Server environments and this is the place, where the actual data processing takes place.

Now, all these Distributed Environment machines and not Servers need to be coordinated by a single software. This software will take care of the needs of the distributed system like partitioning the data across the multiple nodes – replicating the data on the node and coordinating the computing tasks, which

will be running on the parallel machines. This software will also handle fault tolerance and recovery like disk failures and node failures. This software should also allocate the process, as managed by the capacity of the node in terms of memory and hard disk space.

My future articles will be all about Big Data – Hadoop. Thus, keep surfing for my articles.

Thank you for using C# Corner