**O QUE É UM CLUSTER DE COMPUTADORES?**

Um servidor é um computador geralmente com alta capacidade computacional, que serve ( fornece) serviços de armazenamento, aplicações ou banco de dados

Um servidor possui escalabilidade vertical, ou seja, há um limite até onde conseguimos incluir mais espaço em disco, mais processadores e mais memória RAM

Um cluster de computadores é um conjunto de servidores com um mesmo propósito visando fornecer um tipo de serviço, como armazenamento ou processamento de dados

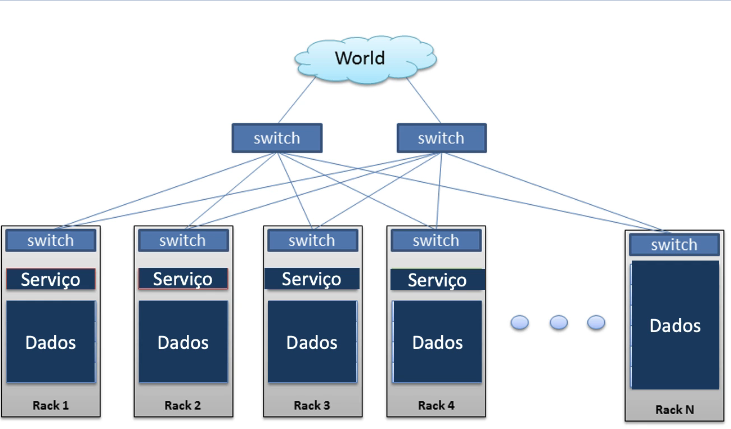
Um cluster possui escalabilidade horizontal, ou seja, se quisermos aumentar a capacidade computacional incluímos mais máquinas no cluster ( além da escalabilidade vertical de cada máquina individual no cluster )

Clusters de computadores são cada vez mais usados em Big data, o que nos permite realizar armazenamento e processamento paralelo através de diversas máquinas ( diversos servidores )

* **O que é armazenamento Paralelo?**

Com clusters de computadores aumentamos de forma considerável a capacidade computacional

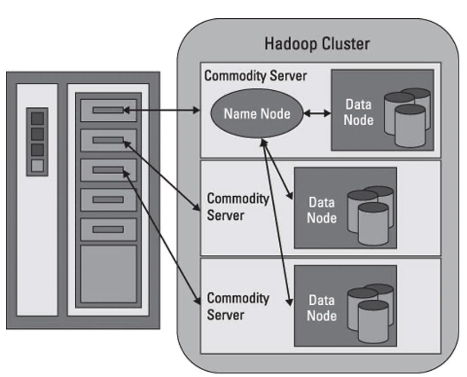
* O armazenamento paralelo consiste em distribuir o armazenamento de dados através de diversos servidores ( computadores ), o que permite aumentar de forma considerável a capacidade de armazenamento usando hardware de baixo custo



* **Software para armazenamento paralelo - APACHE HADOOP**

Precisamos de um sistema de arquivos distribuido. Seu computador pessoal tem um sistema de arquivos ( NTFS, ext3, etc ) mas ele não foi desenvolvido para armazenamento distribuido

Entre algumas opções, o Apache Hadoop HDFS ( hadoop distributed file system) tem se mostrado a solução ideal para gerenciar o armazenamento distribuido em um cluster de computadores



* O HDFS é o software responsável pela gestão do cluster de computadores definido como os arquivos serão distribuídos através do cluster
* Com o HDFS podemos construir um data lake que roda sobre um cluster de computadores e permite o armazenamento de grandes volumes de dados com hardware commodity ( de baixo custo )

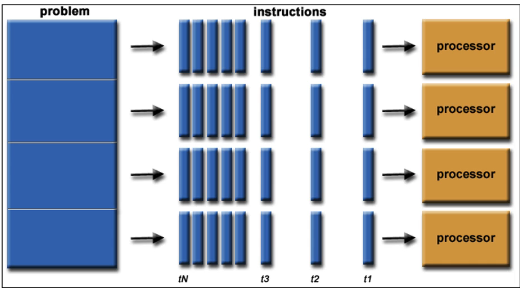
O HDFS permitiu que o big data pudesse ser usado em larga escala !

* **Processamento paralelo de Big Data**

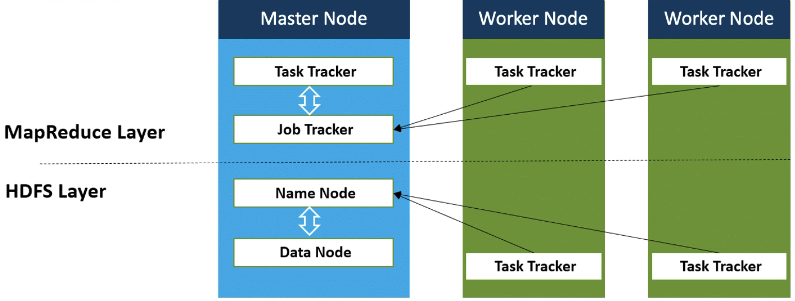
Resolvemos um problema. Podemos agora armazenar grandes quantidades de dados em um cluster computadores através de armazenamento paralelo de dados

mas como vamos processar os dados se eles estão agora distribuidos em diversos computadores?

* No Processamento paralelo o objetivo é dividir uma tarefa em várias sub-tarefas e executá las em paralelo
* O apache hadoop Map reduce e o Apache Spark são dois frameworks para esse propósito
* Ao usar um framework de processamento paralelo, as sub-tarefas são levadas para o processador da máquina do cluster onde os dados estão armazenados, aumentando assim a velocidade de processamento de grandes volumes de dados



* **Arquitetura de armazenamento e processamento paralelo**
* O **HDFS** é um serviço rodando em todas as máquinas do cluster, sendo um NameNode para gerenciar o cluster e os dataNodes que fazem o trabalho de armazenamento propriamente dito
* O **MapReduce** também é um serviço rodando em todas as máquinas do cluster, sendo um Job Tracker para gerenciar o processamento e os Task Trackers que fazem o trabalho de processamento
* O **Job Tracker** consulta o NameNode a fim de saber a localização dos blocos de dados nas máquinas do cluster
* Os **task trackers** se comunicam com os DataNodes para obter os dados do disco, executar o processamento e então retornar o resultado ao Job Tracker
* Essa arquitetura permite armazenar e processar grandes quantidades de dados e assim extrair valor do Big Data atraves da análise de dados



* **Soluções de armazenamento e processamento Paralelo**

Na nuvem:

* Azure HDInsight
* Amazon EMR

Databricks foi a mesma equipe que desenvolveu o SPARK para criar lakehouse plataform

REFERÊNCIAS :

Apache Hadoop

<http://hadoop.apache.org/>

Cloudera

<https://www.cloudera.com/>

Microsoft Azure HDInsight

<https://azure.microsoft.com/pt-br/services/hdinsight/#overview>

AWS EMR

[https://aws.amazon.com/pt/emr/ Databricks https://databricks.com/](https://aws.amazon.com/pt/emr/)

Engenharia de Dados com Hadoop e Spark <https://www.datascienceacademy.com.br/course/big-data-analytics-engineer>