Artigo MapReduce

Este formulário tem o objetivo de fixar o que foi abordado no artigo (data de entrega 24/09/2020)

1. Como funciona a Função Map/Reduce

Na verdade, entendi que são duas funções que dão nome ao modelo: a função map e a função reduce. Em geral a operação de map é usada para encontrar algo, e a operação de reduce é usada para fazer a sumarização do resultado.

1. Descreva resumidamente o seu modelo de programação

O modelo MapReduce é baseado no paradigma de programação funcional, desenhado para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes, em cluster de máquinas em um sistema distribuído. Primeiro divide os dados de entrada em M pedaços. Em seguida inicia várias cópias do programa em um cluster de computadores. Uma das cópias do programa é o *master*. As outras cópias são denominadas *workers*. Um *worker* que possui uma tarefa de *map* le o conteúdo correspondente ao pedaço da entrada, interpreta os pares chave/valor a partir dos dados de entrada e passa como parâmetro para a função de *map* do usuário. Os pares chave/valor intermediários produzidos pela função de map são armazenados em memória. A localização desses pares é informada ao *master*, que irá repassar essa localização para os *workers* com tarefas de *reduce*. Este usa uma chamada de procedimento remota para buscar os dados do disco local dos *workers* de *map*. Quando os dados já foram lidos, ele ordena os dados pelas chaves intermediárias, para que todas as ocorrências de uma mesma chave sejam agrupadas junto. A saída de cada função de *reduce* é adicionada ao final de um arquivo de saída para aquela partição de *reduce.* Quando todas as tarefas de map e reduce foram terminadas, o master retorna o programa do usuário.

1. De um exemplo de programa que se beneficie do Map/Reduce e que não consta no artigo. Escreva seu funcionamento em poucas linhas.

Pig: uma plataforma para grandes conjuntos de dados que possui uma linguagem de programação de alto nível para realizar a análise desses dados. Além disso, possui a [infraestrutura necessária](https://www.devmedia.com.br/infraestrutura-de-data-centers/25679) para avaliar os programas criados, como um compilador especial que transforma as aplicações desenvolvidas nessa linguagem em uma sequência de programas MapReduce.

1. O que é um cluster, cite seus componentes.

Cluster é o nome dado a um sistema que relaciona um ou mais computadores para que trabalhem de forma unificada e entreguem um resultado ou objetivo. Cada computador inserido em um Cluster é chamado de “nó”. E todos os nós são interconectados por uma tecnologia de rede, onde cada um desempenha uma função específica e pré-determinada e dividem entre si as tarefas de processamento e as executam de forma simultânea.

As máquinas são normalmente processadores x86 de processador duplo executando Linux, com 2 a 4 GB de memória por máquina. Hardware de rede de commodities é usado – normalmente 100 megabits / segundo ou 1 gigabit / segundo no nível da máquina, mas com média consideravelmente menor em geral na bissecção de banda larga.

1. O que é um job scheduler e como funciona? Cite o nome de algum utilizado no mercado.

É um aplicativo de computador para controlar a execução automática de tarefas, geralmente fornecem uma interface gráfica com o usuário e um único ponto de controle para definição e monitoramento de execuções em segundo plano em uma rede distribuída de computadores.

## Exemplo: Quartz Enterprise Job Scheduler, Oracle Scheduler

1. Descreva a tolerância a falhas de um nó e do o que acontece se o mestre parar de funcionar.

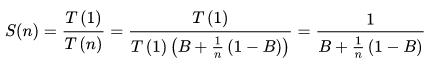
O processo master envia um ping periodicamente para cada nó. Se o master não receber uma resposta em um certo período, assume que aquela máquina falhou. Todas as tarefas de map completadas ali são resetadas para seu estado inicial e são re-escalonadas para outro nó.

Em caso de falha no master, uma nova instância pode ser levantada, recuperando a partir do último estado que foi salvo, mas aqui deve ser bem maior o controle e indesejável a falha, pois o processo master é o elo entre a execução das tarefas de map e reduce*.*

1. O que é e como é calculado o Speedup?

É quanto tempo seu programa leva sendo executado, isto é, vai gastar em paralelo x serial. Assim é possível estudar e reescrever o algoritmo. Executa o programa 10 vezes e faz o cálculo com a medição de tempo. Cálculos: média, mediana e desvio padrão.

Speedup pode ser definido como a relação entre o tempo gasto para executar uma tarefa com um único processador e o tempo gasto com N processadores, ou seja, Speedup é a Medida do ganho em tempo. O speedup teórico {\displaystyle S(n)} que pode ser obtido pela execução de um dado algoritmo, em um sistema capaz da execução de {\displaystyle n}n threads de execução, é:

{\displaystyle S={T(1) \over T(N)}} 

Onde 'S' é o Speedup, n o número de threads de execução, B fração de um algoritmo estritamente serial, e 'T'(N) é o tempo gasto para 'N' processadores

1. Cite alguns frameworks que trabalham com Map/Reduce

Apache Hadoop, é o principal framework utilizado no processamento e armazenamento de grandes conjuntos de dados – Big Data. Inicialmente inspirado no MapReduce e no GoogleFS.

O Spark surgiu para suprir a principal deficiência do modelo MapReduce e da ferramenta Hadoop, ele é um framework para clusterização que executa processamento em memória - sem utilização de escrita e leitura em disco rígido - com o objetivo de ser superior aos motores de busca baseados em disco como o MapReduce.