PROJETO FINAL

Universidade Federal do Piauí
Sistemas de Informação
Sistemas Distribuídos
Márcio Silvano de Sousa, Artur Pereira da Silva, Amanda Gonçalves Bernardes

Resumo. *Atualmente uma questão em que os profissionais da área de TI vem tentando resolver é como fazer para administrar aplicações que trabalham com grandes quantidades de dados, pois cada vez mais se tem a necessidade de processar dados de forma mais rápida e eficaz. Por tal motivo, esse trabalho tem por objetivo implementar um cluster de computadores para contar a quantidade de ocorrências de palavras em um determinado texto, distribuindo partes do texto para máquinas distintas diminuindo o tempo de processamento para realizar tal operação.*

Palavras chave: dados, processamento, cluster

Introdução

De acordo com bacellar (2010) Atualmente pode-se observar um grande numero de aplicações que exigem cada vez mais uma grande quantidade de processamento de dados, aplicações de mapeamento genético, computação gráfica, previsões metrológicas e até mesmo programas que exigem um grande numero de variáveis de entrada.

Segundo cetax (2018) Uma única máquina/servidor não consegue processar todo esse grande volume de dados (Big Data). O volume de dados é imenso e exige mais de uma única máquina para processar. Por isso, criou-se o cluster, uma forma de gerenciar diversas máquinas que funcionam como uma única máquina.

Tendo em vista contar de forma mais rápida e eficás a quantidade de ocorrência de uma determinada palavra em um arquivo de texto, que possui uma grande quantidade de palavras, esse trabalho tem como objetivo principal implementar um cluster de computadores com o intuito de diminuir o tempo de processamento da atividade.

Arquitetura do Sistema

O sistema foi dividido em três partes principais: o cliente, que é a parte do sistema que será executada na máquina do cliente e é a parte onde o cliente irá selecionar o arquivo com o texto e irá informar as palavras a serem buscadas no texto.Por fim a aplicação irá enviar as informações para serem processadas em outras máquinas e irá receber como retorno o resultado da contagem das palavras.

O HeadNode que será executado em uma outra máquina será responsavel por fazer o controle do cluster, onde no mesmo será possível adicionar e remover as máquinas ao cluster, bem como setar o estado da máquina como ativo ou inativo. Essa parte também será responsável por receber as informações que chegam do cliente, dividir o texto de acordo com a quantidade de máquinas inseridas e enviar cada parte do texto juntamente com as palavras mapeadas para cada uma das máquinas.

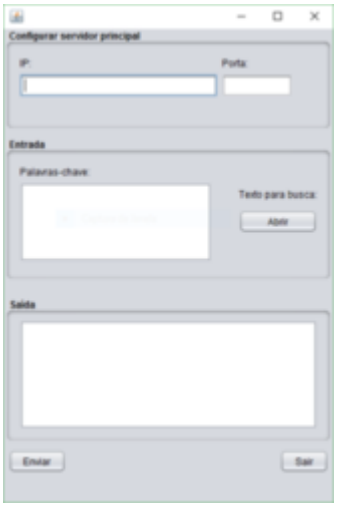

Figura 1: Tela do Cliente

O Node será instalado em cada uma das máquinas que serão uilizadas para realizar o processamento, cada máquina receberá sua fatia de texto enviada pelo HeadNode e o sistema irá realizar o reducer para contar a ocorrência das palavras na fatia de texto correspondente. O resultado do processamento de cada máquina será retornado ao HeadNode que irá calcular o resultado dos valores processados nas máquinas do cluster e retornará o resultado total para a máquina do cliente que será responsável por exebir ao usuário.

Toda a comunicação entre as máquinas para o envio de informações foi feita utilizando sockets e todas as informações trocadas passam por por um processo de criptografia ao saírem para envio e de decriptográfia ao chegarem nas máquinas destinatárias.

Resultados e Discussão

O sistema desenvolvido solucionou bem o problema proposto, todos os reultados apresentados con-

dizem com o que era esperado ao se desenvolver a aplicação, bem como o tempo de processamento diminui ao distribuir tarefas a máquinas diferentes.

Poré, observou-se também uma variação no tempo de processamento, o que pode ter ocorrido devido ao tipo de comunicação utilizada, no caso sockets, como também devido ao fato de que as máquinas utilizadas possuírem arquiteturas diferentes.

Conclusão

O trabalho desenvolvido serviu como grande aprendizado de como funciona na prática todo o conteúdo ministrado na disciplina de Sistemas Distribuídos. Bem como aprender sobre a importância do uso da distribuição de tarefas em sistemas que necessitam processar uma grande quantidade de dados.

O mesmo ajudou a solucionar de certa forma bem o problema proposto, onde os resultados apresentados foram em encontro ao que foi requerido pelo professor da disciplina.

Referências

BACELLAR, Hilário. Cluster: Computação de Alto Desempenho. 2010.

CETAX. Apache Hadoop: Tudo o que você precisa saber. 2018. Disponível em: <https://www.cetax.com.br/apache-hadoop-tudo-o-que-voce-precisa-saber/>. Acesso em: 22 nov. 2018.