Sistema de Computação Paralela e Distribuída Utilizando Raspberry Pi e Apache Hadoop

Hemerson R. P. Pontes[1], Gilvandro C. de Medeiros[2], Joanderson L. L. Borges[2], Helton Maia[2]

[1]Departamento de Engenharia de Computação e Automação – Universidade Federal do Rio Grande do norte (UFRN) Natal – RN – Brasil

[2]Escola de Ciência e Tecnologia – Universidade Federal do Rio Grande do norte (UFRN) Natal – RN – Brasil

{hrpp, gilvandrocesar, joandersonlucas}@ufrn.edu.br, helton.maia@ect.ufrn.br

Abstract*. In the context of Big Data, the large flow and complexity of the data generated requires a high computational cost for processing tasks and extracting information, and it is a challenge to complete such executions in a timely manner for technical or business decision making. However, in computational clusters, it is possible to manage and distribute data packets between different processing units, making it possible and feasible to work with a large volume of data, processing them in parallel and distributed. Therefore, the present work disposes to build a structure and study the operation of a cluster, managed by the Apache Hadoop framework, for distributed data processing.*

Resumo. *No contexto de Big Data, o grande fluxo e a complexidade dos dados gerados exigem elevado custo computacional para tarefas de processamento e extração de informação, sendo um desafio concluir tais execuções em tempo hábil para tomadas de decisões técnicas ou empresariais. No entanto, em clusters computacionais, pode-se gerenciar e distribuir pacotes de dados entre diferentes unidades de processamento, tornando-se possível e viável trabalhar com um grande volume de dados, processando-os de forma paralela e distribuída. Portanto, o presente trabalho se dispõe a construir a infraestrutura de um cluster e estudar seu funcionamento utilizando, para isso, a ferramenta Apache Hadoop para o processamento distribuído de dados.*

1. Introdução

A popularização do acesso à Internet, o desenvolvimento da indústria de eletrônicos e a conectividade alicerçam a realidade de um modelo social em transição. A revolução tecnológica e informacional dos últimos anos, por muitos denominada como Indústria 4.0, tem como bases a inteligência artificial, o poder computacional e a conectividade, se relacionando com o paradigma Big Data (BIGDATABUSINESS, 2018).

O conceito de Big Data destaca-se como uma forma de se analisar grandes conjuntos de dados heterogêneos (MANYIKA et al., 2011). Atualmente, Big Data é identificado a partir de 5 V's: velocidade, volume, variedade, valor e veracidade. O volume representa grandes conjuntos de dados; velocidade é a demanda por respostas em tempo real; variedade é a forma como os dados são recebidos, alternado entre os

estruturados, semi-estruturados e não estruturado; valor é a informação extraída a partir dos dados; e veracidade tem relação com o sentido e à autenticidade (ANSELMO, 2015).

No cenário de Big Data, o desafio computacional de processar rapidamente um grande conjunto de dados muitas vezes inutiliza o uso de microprocessadores comerciais de forma isolada, devido a capacidade de processamento solicitada ser muito superior ao que pode ser oferecido (GOLDMAN et al., 2018). Muitas vezes, também, é inviável a compra de melhores processadores para trabalhar com esses dados, seja pelo elevado custo ou por limitações técnicas. Para contornar esta problemática, pode-se, ao invés de trocar os processadores por outros mais potentes, aplicar princípios da computação paralela e distribuída, subdividindo o conjunto de dados entre diferentes unidades de processamento em um cluster, aumentando, assim, a capacidade de processamento.

No referido contexto, o presente trabalho se propõe a estudar e desenvolver aplicações de computação paralela e distribuída em um cluster, utilizando placas Raspberry Pi e o software de gerenciamento Apache Hadoop para processar dados de forma distribuída. Além disso, também foram feitas medições de temperatura da CPU para cada placa e de corrente elétrica consumida pelo cluster, avaliando o desempenho e assim validando o funcionamento do sistema proposto.

2. Metodologia

Em um cluster, pode-se diferenciar as unidades de processamento em dois tipos, os *nodes* (nós) e o *master* (mestre). Em linhas gerais, os nós e o mestre se relacionam de forma gerencial centralizada, sendo o mestre responsável por distribuir e controlar os dados entre os nós, que lá são processados e retornando para o mestre o resultado após o devido tempo de execução. Um dos *softwares* mais importantes para gerenciamento de grande volumes de dados é o Apache Hadoop. De código aberto, o Hadoop fornece uma estrutura que permite o processamento distribuído de grandes conjuntos de dados e gerenciamento de clusters (APACHE HADOOP, 2018).

No *framework* do Hadoop, podem-se destacar e definir os seguintes módulos:

(a) *Common*: Ferramentas de suporte para os demais módulos do Hadoop;
(b) *Hadoop Distributed File System* (HDFS): Um sistema de arquivos distribuído que possibilita o acesso aos dados da aplicação;
(c) *Yet Another Resource Negotiator* (YARN): Uma estrutura para o planejamento de tarefas e gerenciamento de recursos como Unidade Central de Processamento (Central Processing Unit - CPU) e Memória de acesso aleatório (Random-access memory - RAM) do cluster;
(d) *MapReduce*: Um paradigma de programação baseado no YARN para processamento paralelo de grandes conjuntos de dados. Abstraindo toda a complexidade das implementações paralelizadas em apenas duas funções: Mapeamento e Redução.

Para este trabalho, utilizou-se de mini-computadores Raspberry Pi 3 Model B para os nós escravos e um computador modelo HP 6005 *Desktop* PC para servir de nó mestre. A partir da definição dos componentes, oss Sistema Operacionais (SO) escolhidos para operar nas referidas unidades do cluster foram as distribuições Linux Raspbian Jessie e Ubuntu 16.04 LTS compatíveis com os *hardwares* selecionados e com o Hadoop na versão 2.7.5, devido a natureza *open source* desses SOs e a consequente facilidade para implementar aplicações distribuídas.

3. Resultados Preliminares

Construiu-se um pequeno cluster escalável, compacto, funcional e gerenciado pelo software Apache Hadoop, capaz de realizar processamento distribuído de dados. O cluster é composto por 4 nós, 3GB RAM, 16 núcleos de processamento e armazenamento com uma capacidade total de 116,63 GB. Sua estrutura pode ser conferida na Figura 1, onde todos os componentes estão identificados: A) Quinas para junções de placas de acrílico impressas em ABS; B) Switch 8-port 10/100; C) 4 placas Raspberry PI model B; D) Fonte de alimentação 5v 10A; E) Hub USB de 4 portas e F) Painel de chaveamento das placas.

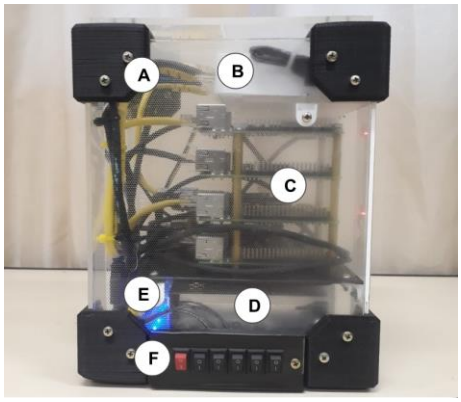

Figura 1. Estrutura construída para acomodar os nós escravos do cluster.

Para validar o funcionamento do cluster, gerou-se, aleatoriamente, um arquivo de texto (*.txt) com aproximadamente 1 MB de caracteres, que, a partir de replicação, deu origem também a outro arquivo do mesmo tipo, mas com 10 MB. Em seguida, aplicou-se o teste contador de palavras (*Word Count*), fazendo com que as placas trabalhassem em conjunto para contar o número de palavras existentes no arquivo. Na Figura 2, seguem gráficos referentes às medições de Temperatura da CPU (ºC) para cada um dos 4 *nodes*, diferenciados por cores, e Corrente Elétrica (A) para o cluster, ao aplicar o *Word Count* para cada um dos arquivos:

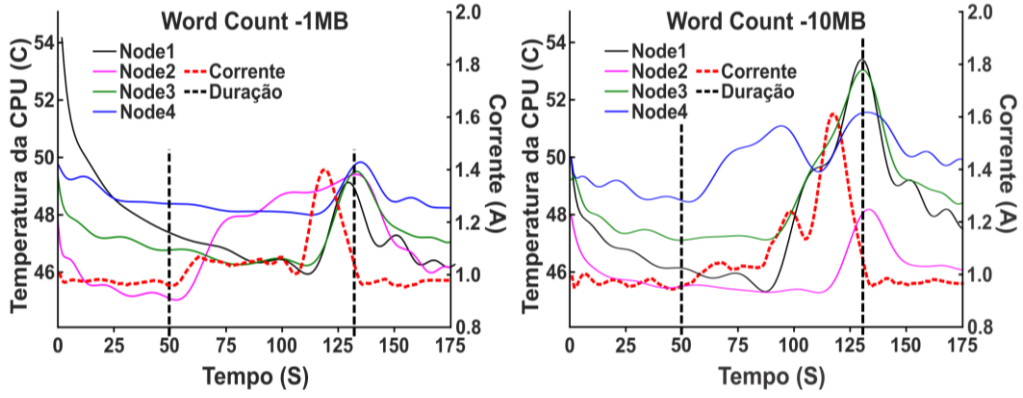

Figura 2 - Consumo de corrente elétrica e temperatura da CPU em função do tempo durante o processo de execução do teste Word Count para os arquivos de 1MB e 10MB.

Verifica-se que na linha tracejada na cor vermelha, referente ao valor de corrente elétrica mensurada, o início da execução do código é marcado por um aumento gradativo no consumo, que em seguida se mantém estável durante a etapa de mapeamento, e, por fim, durante a redução, verifica-se um aumento relativamente alto deste consumo de corrente. Já em relação a temperatura da CPU, apesar de não parecer existir um padrão

entre os nós, provavelmente devido a própria arquitetura e distribuição das placas no cluster, em alguns instantes após o início da etapa de redução a temperatura passa a aumentar como resposta ao aumento do consumo de corrente, sendo um comportamento esperado, devido ao tempo necessário para propagação do calor.

Observa-se, ainda, que o comportamento do cluster para ambos os testes se diferencia levemente quanto ao tempo de duração, sendo praticamente iguais, e quanto ao consumo de corrente, mostrando que o teste de 10 MB exigiu maior consumo energético que o de 1 MB. Conclui-se, portanto, que o sistema de processamento distribuído pode realizar um maior número de tarefas em um mesmo tempo de referência, exigindo um consumo energético maior.

4. Conclusão

Foi realizado o desenvolvimento de um pequeno cluster de baixo custo aquisitivo baseado em sistemas inteligentes. Conclui-se que o cluster, com a metodologia adotada, funciona e atende ao seu propósito, mas sabe-se que o tamanho dos arquivos utilizados ainda está bem abaixo do que pode ser considerado como base de dados no paradigma de Big Data. Porém, sabe-se também que o cluster, para os testes realizados, operou abaixo da sua capacidade máxima de trabalho, demandando, portanto, testes com arquivos bem maiores.

Foram realizados testes fundamentais para análise do funcionamento do cluster, que incluem medições de temperatura e consumo de corrente elétrica, viabilizando possíveis cálculos de consumo energético durante a execução de tarefas que exigem alto poder de processamento.

Como sugestões de trabalhos futuros, pretende-se utilizar a proposta do trabalho como base para criação de materiais didáticos para aulas práticas de sistemas de processamento distribuídos, aplicar a computação em cluster em cenários que exigem processamento em tempo real, utilizar o cluster para o processamento de sinais e imagens. Este protótipo desenvolvido também pode se tornar uma alternativa comercial para construção de clusters de baixo custo, principalmente para utilização em pequenas empresas.

Referências

Anselmo, F. (2015) “Big data: uma análise conceitual e funcional”. 2015. 99 f. Trabalho de conclusão de curso – Curso Sistemas de Informação. Universidade do Planalto Catarinense, Lages.

Apache Hadoop. (2018) “Homepage”. http://hadoop.apache.org, Novembro.

Big Data Business. (2018) “A importância de Big Data para a Indústria 4.0”, http://www.bigdatabusiness.com.br/big-data-na-industria-4-0/, Novembro.

Goldman, Alfredo & Kon, Fabio & Pereira Junior, Francisco & Polato, Ivanilton & De, Rosangela & Pereira, Fátima. (2018) “Capítulo 3 Apache Hadoop: conceitos teóricos e práticos, evolução e novas possibilidades”.

Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxbugh, C.; Byers, A. H. (2011) “Big data: The next frontier for innovation, competition, and productivity”. The McKinsey Global Institute.