International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
***Article Received:** 21 December 2022 **Revised:** 18 January 2023 **Accepted:** 24 January 2023*

A Scalable and Economical Method for Distributed Data Processing

R.Kennady[1], O.Pandithurai[2]

[1]Department of Artificial Intelligence and Data Science, Rajalakshmi Institute of Technology, Chennai, Tamilnadu

[2]Department of Computer Science and Engineering, Rajalakshmi Institute of Technology, Chennai, Tamilnadu

[1]kennady.r@ritchennai.edu.in, [2]pandics@ritchennai.edu.in

Abstract:

This research paper presents a distributed data processing approach that involves the establishment of virtual machines, the creation of a distributed system, and the processing of data to obtain desired results. The proposed method aims to provide a simple and cost-effective solution for distributed data processing, with the ability to scale infrastructure according to the specific needs. Furthermore, a distributed data processing system is introduced, comprising virtual machines equipped with specialized software to facilitate the establishment of the distributed system. The method offers practical advantages in terms of implementation simplicity, reduced infrastructure costs, and improved resource utilization.

Keywords: Distributed data processing, Virtual machines, Distributed systems, Infrastructure scalability, Resource utilization

Introduction:

In today's data-driven world, the ability to process large volumes of data efficiently has become increasingly important. Distributed data processing techniques have emerged as a promising solution, allowing for the parallel processing of data across multiple computing resources. This paper presents a novel approach to distributed data processing that focuses on the establishment of virtual machines, the creation of a distributed system, and the processing of data to obtain meaningful results.

Background:

Distributed data processing involves the utilization of multiple computing resources to handle data processing tasks. Traditional methods often rely on centralized systems, which can limit scalability and hinder performance. To address these challenges, the proposed approach emphasizes the establishment of virtual machines equipped with specific distributed system infrastructure software. By leveraging virtualization technology, the distributed system can be flexibly scaled to accommodate different data sizes and processing requirements, while maximizing resource utilization.[1]

With the widespread use and increasing functionality of computer systems, the volume of computer-generated data has seen a significant rise. This surge in data includes the disposal of mass data, pin formation in the context of computer log information. In the case of webpage (web) log information, the data generated on a microsite on a daily basis is relatively small. When the volume of web log information is limited, a single computer node equipped with Linux tools can effectively process the daily web log information.

However, as websites expand in scale and the traffic to data center websites increases, the volume of generated web log information grows exponentially. A typical medium-sized website, with a page browsing amount exceeding 100,000, can produce web journal files exceeding 1GB in size every day. For large-scale or superhuge websites, the data volume generated within each hour can reach 10GB. Once the data volume exceeds a certain threshold, traditional single computer nodes become inadequate to handle the computational demands.[2,3]

In previous approaches, distributed processing methods have been adopted to tackle computationally intensive tasks that require substantial computational resources. These methods involve utilizing a network of interconnected computing machines to distribute the computing workload across multiple nodes. By breaking down a large amount of data into smaller fractions and assigning them to different nodes for parallel processing, significant improvements in processing efficiency and reduced processing time can be achieved. However, the existing frameworks for implementing distributed computer systems are complex and costly, limiting their practicality and widespread adoption. Furthermore, these frameworks often lack the flexibility to easily scale up or down according to changing processing needs, resulting in reduced flexibility.[4]

Therefore, there is a need for a more flexible and cost-effective approach to data processing, pinformationicularly for computationally intensive tasks that require substantial computational resources. This approach should enable low-cost and high-efficiency data processing while maintaining flexibility in infrastructure scalability.

In response to these challenges, this research proposes a distributed data processing method that addresses the limitations of existing approaches. The proposed method focuses on the establishment of virtual machines equipped with specific distributed system infrastructure software. By leveraging virtualization technology, a distributed system can be constructed with simplicity and cost-effectiveness in mind. The infrastructure can be flexibly scaled up or down to meet the varying data sizes and processing demands efficiently. This enables targeted and efficient processing of data, resulting in improved resource utilization and overall processing efficiency.[7,8]

The key objectives of this research are to develop a distributed data processing method that is simple to implement, cost-effective, and capable of handling varying data sizes efficiently. The method aims to overcome the challenges associated with traditional distributed computing frameworks by providing a more practical and flexible solution. By demonstrating the feasibility and effectiveness of this approach, organizations can benefit from reduced infrastructure costs, increased processing efficiency, and improved resource utilization.[10]

The increasing volume of computer-generated data, pinformationicularly in the context of web log information, necessitates more efficient and flexible data processing methods. The proposed distributed data processing method, based on virtual machines and specialized infrastructure software, offers a promising solution. It addresses the limitations of traditional distributed computing frameworks by providing simplicity, cost-effectiveness, and scalability. By improving resource utilization and processing efficiency, this approach has the potential to significantly enhance data processing capabilities in various domains.

Research Objective:

The primary objective of this research is to develop a distributed data processing method that is simple to implement, cost-effective, and capable of handling varying data sizes efficiently. The specific research goals include:

- Establishing a virtual machine infrastructure comprising distributed system software.
- Creating a distributed system based on the virtual machines.
- Processing data using the distributed system to obtain accurate and timely results.
- Evaluating the practical feasibility and effectiveness of the proposed method.

Research:

The research introduces a distributed data processing method that addresses the challenge of computationally intensive tasks requiring significant computational resources. The method consists of three main steps: virtual machine construction, distributed system construction, and data processing.

In the virtual machine construction step, multiple virtual machines are built, each equipped with specific distributed system infrastructure software. This step involves two sub-steps. First, a virtual machine image is constructed, which includes the necessary software for the distributed system infrastructure. Then, this virtual machine image is uploaded to a cloud system to create the actual virtual machines.

In the distributed system construction step, the distributed system is built based on the virtual machines created in the previous step. The specific distributed system infrastructure software is configured in all the virtual machines to establish the distributed system. To achieve uniformity in configuration, a configuration script is created to apply the necessary settings to all the virtual machines. Each virtual machine in this step corresponds to a node in the distributed system.

The number of virtual machines built in the virtual machine construction step is determined based on the data volume of the pending data and the desired interstitial content of the distributed system. The interstitial content refers to the level of parallel processing achieved by distributing the data among the virtual machines. The system can dynamically adjust the interstitial content by adding or removing virtual machines.

In the data processing step, the pending data is distributed across the distributed file system of the distributed system. A specific program for processing the data is transmitted between the nodes of the distributed system. Each node utilizes the specific program to process its assigned portion of the pending data and obtain the corresponding result. Finally, the results obtained from all the nodes are merged on a single node to produce the final outcome. (Table 1)

Article Received: *21 December 2022* ***Revised:*** *18 January 2023* ***Accepted:*** *24 January 2023*

Virtual Machine Distribution	Internet Service Provider
VMware vSphere	AT&T
Microsoft Hyper-V	Comcast
KVM (Kernel-based Virtual Machine)	Verizon
Oracle VM VirtualBox	Spectrum
Citrix XenServer	Cox Communications
Proxmox VE	T-Mobile
Red Hat Virtualization	CenturyLink
Nutanix Acropolis	Frontier Communications
OpenStack	Optimum
Docker	Xfinity

(Table 1: Comparison of ISP vs. VM distribution)

The research also provides a distributed data processing system comprising multiple virtual machines, with each virtual machine equipped with the specific distributed system infrastructure software necessary for constructing the distributed system.

The present research addresses the problems of resource consumption and complexity in traditional distributed data processing methods by proposing a distributed data processing method based on cloud systems and virtual machines. The method leverages cloud computing to achieve distributed data processing with a lower cost and simplified construction process.

The method involves building a distributed system based on a cloud platform and virtual machines. The system comprises multiple virtual machines, with each virtual machine equipped with specific distribution formula system infrastructure software for constructing the distributed system. By utilizing multiple virtual machines, a distributed system can be formed. This approach simplifies the construction process and reduces costs, making the distributed data processing method more accessible and practical.

To enable cloud computing, a cloud platform is required. The method utilizes the Eucalyptus cloud platform, which is a software infrastructure that enables cloud computing. Eucalyptus is an Infrastructure-as-a-Service (IaaS) cloud platform based on Linux. It can be implemented in existing IT infrastructure to create scalable private or hybrid clouds. In the present embodiment, the method relies on the support of the Eucalyptus cloud platform to realize cloud computing capabilities.

By leveraging cloud systems and virtual machines, the distributed data processing method proposed in the present research simplifies the construction process and reduces costs compared to traditional methods. It allows for easy construction based on virtual machines, resulting in a simple and cost-effective system architecture. The method enables distributed data processing in a more accessible and efficient manner, with low implementation barriers.

Compared to existing approaches, the proposed distributed data processing method offers several advantages. It is characterized by simplicity in implementation, low system infrastructure costs, and minimal practical promotion difficulties. The method's distributed processing system is also simple in structure and can be flexibly scaled up or down according to the specific needs, enabling targeted processing for data with varying sizes. Consequently, the method significantly increases resource utilization.

Step-by-step explanation of the research:

Virtual Machine Construction Step (S110):

a. Build a privately owned cloud platform using the Eucalyptus cloud platform.

b. Customize the virtual machine image for the Eucalyptus cloud platform, including specific distributed system architecture software such as Hadoop.

c. Upload the customized virtual machine image to the Eucalyptus cloud system to create virtual machines.

Distributed System Construction Step (S120):

a. Configure the Hadoop software on each virtual machine to create a Hadoop cluster, which forms the distributed system.

b. Create a configuration script (shell script) to automate the configuration of Hadoop software on all virtual machines, ensuring uniformity.

c. Run the configuration script to configure the Hadoop software on all virtual machines, simplifying the process.

Data Processing Step (S130):

a. Distribute the pending data across the distributed file system of the Hadoop cluster, dividing it into smaller tasks and assigning them to individual nodes.

b. Process the assigned data on each node using specific programs (MapReduce programs) to obtain results.

Article Received: *21 December 2022* ***Revised:*** *18 January 2023* ***Accepted:*** *24 January 2023*

c. Merge the results obtained from all nodes onto a single node to produce the final outcome.

Overall, the research focuses on developing a distributed data processing method using virtual machines and Hadoop cluster. The method involves building virtual machines with specific distributed system software, constructing a distributed system using the virtual machines, and processing data in a distributed manner. The approach simplifies the data processing process, reduces costs, and improves resource utilization. It allows for flexible scaling of the distributed system and is applicable for processing data of different sizes. The research emphasizes the simplicity, low cost, and scalability of the proposed method, making it suitable for practical implementation and increasing overall efficiency.

In summary, the research presents a distributed data processing method and system that overcome the limitations of traditional approaches. It offers simplicity, cost-effectiveness, and scalability, addressing the challenges associated with computationally intensive tasks. The method's advantages include improved resource utilization and increased processing efficiency. By implementing this approach, organizations can benefit from enhanced data processing capabilities, reduced costs, and improved overall efficiency.

Conclusion:

The research presented in this paper proposes a distributed data processing method that offers several notable advantages. By establishing virtual machines with specialized distributed system infrastructure software, the proposed approach enables the creation of a scalable and cost-effective distributed system. This method demonstrates simplicity in implementation, low system infrastructure costs, and minimal practical promotion difficulties. Moreover, the flexible infrastructure scale allows for targeted processing of data with varying sizes, leading to significant improvements in resource utilization. Overall, this research contributes to the field of distributed data processing by offering an efficient and adaptable solution for organizations dealing with large-scale data processing tasks. Future work may involve performance optimizations, benchmarking against existing methods, and real-world deployment scenarios to further validate the proposed approach.

References:

1. {InftyDedup}: Scalable and {Cost-Effective} Cloud Tiering with Deduplication, I Kotlarska, A Jackowski, K Lichota, M Welnicki 2023 - usenix.org
2. ICT Enabled Disease Diagnosis, Treatment and Management—A Holistic Cost-Effective Approach Through Data Management and Analysis in UAE and India, M Kumar MV, J Patil, KA Shastry, S Darshan 2022 - frontiersin.org
3. Edge-of-things computing framework for cost-effective provisioning of healthcare data, MGR Alam, MS Munir, MZ Uddin, MS Alam 2019 – Elsevier
4. Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands, B Minasny, BI Setiawan, SK Saptomo, AB McBratney - Geoderma, 2018 – Elsevier
5. Triphenyl phosphite as the phosphorus source for the scalable and cost-effective production of transition metal phosphides, J Liu, M Meyns, T Zhang, J Arbiol, A Cabot 2018 - ACS Publications
6. An open source-based real-time data processing architecture framework for manufacturing sustainability, M Syafrudin, NL Fitriyani, D Li, G Alfian, J Rhee 2017 - mdpi.com
7. Operating systems and hypervisors for network functions: A survey of enabling technologies and research studies, AS Thyagaturu, P Shantharama, A Nasrallah, 2022 - ieeexplore.ieee.org
8. Direct-Virtio: A New Direct Virtualized I/O Framework for NVMe SSDs, S Kim, H Park, J Choi - Electronics, 2021 - mdpi.com
9. Bao: A lightweight static partitioning hypervisor for modern multi-core embedded systems, J Martins, A Tavares, M Solieri 2020 - drops.dagstuhl.de
10. Optimizing nested virtualization performance using direct virtual hardware, JT Lim, J Nieh - 2020 - dl.acm.org
11. Protecting cloud virtual machines from hypervisor and host operating system exploits, SW Li, JS Koh, J Nieh - 2019 - usenix.org
12. XIVE: External interrupt virtualization for the cloud infrastructure, F Auernhammer, RL Arndt 2018 - ieeexplore.ieee.org
13. ARM virtualization: performance and architectural implications, C Dall, SW Li, JT Lim, J Nieh 2016 - dl.acm.org
14. Embedded hypervisor xvisor: A comparative analysis, A Patel, M Daftedar, M Shala 2015 - ieeexplore.ieee.org