Data Processing Model to Perform Big Data Analytics in Hybrid Infrastructures

*

Team Members:
1.Anil Varikuppala
2.Vivek Reddy Suresh Puttireddy
3.Pavan Sai Kumar Jalluri
4.Om Sai Kumar Vaddi

***Abstract*—This project explores the implementation of a hybrid data processing model integrating cloud computing and volunteer computing infrastructures for efficient big data analytics. The approach aims to optimize resource allocation and performance while minimizing costs, based on the methodology proposed by Dos Anjos et al. (2020) and principles outlined by Marz and Warren (2015). Additionally, it leverages insights from Zaharia et al. (2010) on Spark.**

***Index Terms*—Hybrid Data Processing Model, Cloud Computing, Volunteer Computing, Big Data Analytics, Resource Allocation, Cost Efficiency, Performance Optimization, Scalability, HR Alloc Algorithm, Dynamic Data Placement.**

I. Research Statement and Conjecture:

Our approach will work by leveraging the benefits of both cloud and volunteer computing infrastructures to address resource management challenges and cost efficiency in big data analytics. By integrating these environments, we hypothesize that our model will achieve significant improvements in performance and cost savings compared to traditional cloud-based solutions.

Cloud computing offers robust resources and seamless scalability, ensuring consistent performance for demanding big data tasks. However, it can be costly, especially for extensive, continuous operations. Volunteer computing, on the other hand, utilizes the idle processing power of personal computers volunteered by individuals, offering a virtually free and abundant resource pool.

By integrating these two environments, our model can dynamically allocate tasks based on real-time resource availability and cost considerations. High-priority and resource-intensive tasks can be assigned to the cloud for reliability, while less critical or parallelizable tasks can be offloaded to volunteer networks, thus reducing reliance on expensive cloud resources.

We hypothesize that this hybrid approach will not only cut costs but also improve overall system performance by efficiently distributing workloads. This integration can potentially transform big data analytics, making it more accessible and affordable without compromising on performance.

II. Related Work (Literature Survey):

We will conduct a comprehensive literature review focusing on big data analytics, cloud computing, and volunteer computing, based on the paper by Dos Anjos et al. (2020), as well as the works of Marz and Warren (2015), and Zaharia et al. (2010). Our methodology will include implementing and evaluating the HR Alloc algorithm for data placement and resource allocation in a hybrid infrastructure setup. This review will explore the latest advancements and best practices in managing and processing large-scale data, leveraging cloud resources for scalability, and utilizing volunteer computing for distributed tasks. The HR Alloc algorithm's effectiveness will be assessed in performance, efficiency, and adaptability to varying workloads.

III. Methodology:

Our methodology involves the following steps:

Development of HR Alloc Algorithm: The HR Alloc algorithm will be designed to facilitate dynamic data placement and resource allocation across cloud and volunteer computing platforms. This involves creating mechanisms for monitoring resource availability and workload distribution in real time. Prototype Implementation: We will develop a prototype of the HR Alloc algorithm to test its functionality and performance in a controlled environment. This prototype will simulate a hybrid infrastructure, allowing us to assess the algorithm's ability to manage resources efficiently.

Performance Testing: A series of performance tests will be conducted to evaluate the HR Alloc algorithm's effectiveness. These tests will compare the hybrid model's performance with traditional cloud-only models, focusing on metrics such as processing speed, resource utilization, and cost efficiency.

Data Analysis: The results from the performance tests will be analyzed to quantify the improvements achieved by the hybrid model. This analysis will include statistical comparisons and visualizations to highlight the benefits and potential drawbacks of the proposed approach.

IV. Progress:

- Literature Review: Completed by June 26.
- Prototype Development: Completed by July 2. The prototype has been developed using a combination of cloud-based services and volunteer computing resources.
- Performance Testing: Initial tests completed by July 10. These tests involved running various data processing tasks to evaluate the prototype's performance.
- Abstract Submission: A draft is currently in progress and will be completed by July 17.
- July 24: Finalize project report and submit for evaluation.

V. Preliminary Results and Analyses:

Preliminary tests indicate that the hybrid model shows promise in reducing processing times and costs compared to traditional cloud-based models. Detailed performance metrics and cost analyses are currently being compiled and will be included in the final report.

VI. Conclusion and Future Work:

The initial results support our conjecture that integrating cloud and volunteer computing infrastructures can enhance performance and cost efficiency in big data analytics. Future work will focus on refining the algorithm, conducting more extensive testing, and preparing the final project report.

References

[1] 1. Julio C. S. Dos Anjos et al., ”Data Processing Model to Perform Big Data Analytics in Hybrid Infrastructures,” IEEE Access, DOI: 10.1109/ACCESS.2020.3023344, September 2020.

[2] 2. Marz, N., Warren, J. (2015). Big Data: Principles and best practices of scalable realtime data systems. Manning Publications

[3] 3. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., Stoica, I. (2010). Spark: Cluster computing with working sets. HotCloud, 10(10-10), 95.